Rich-Get-Richer and The Long Tail

CSC200 Lecture 32

February 22, 2016

Allan Borodin and Amirali Salehi-Abari

CSC200: Lecture 32

- Today:
  - Review of power laws and their natural occurrence
  - Rich-get-richer, The Long Tail Ch. 18.3-18.5
- Announcements
  - Assignment 4: March 30 (Initial questions will be posted this week)
  - Office hours tomorrow: 10-11 and 1-2 instead of 2-4
  - Quiz 7: March 11
  - Final exam schedule has been posted
- Acknowledgement:
  - Some slides' materials are borrowed form the slides of the last offering of this course. Thanks to Professor Boutilier!

Back to our Web Graph Case Study

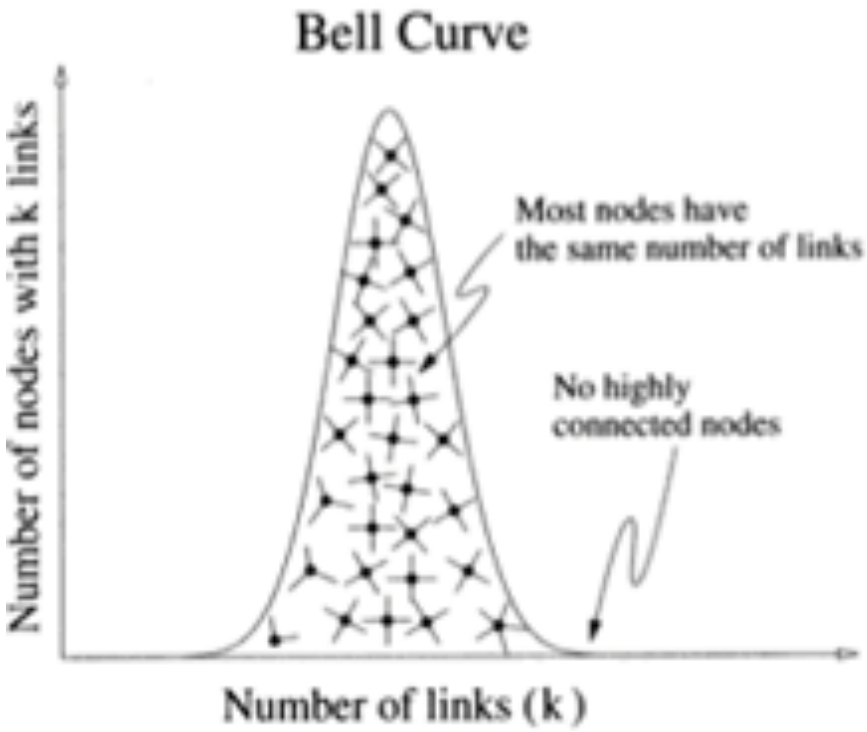
- What is the implication of Central Limit Theorem for webpages (or social networks)?
- If web pages (or people) decide independently at random on whether or not to link to any other web page, then what can you say about degree distribution?
- Based on Central Limit Theorem, as the number of in-links is the sum of many independent random quantities, the in-degree should be approximately normally distributed.
  - So, the number of pages with k in-links should decrease exponentially as k grows large.
  - Also, very large or small numbers of links are extremely unlikely.
- Does this happen in real-world?

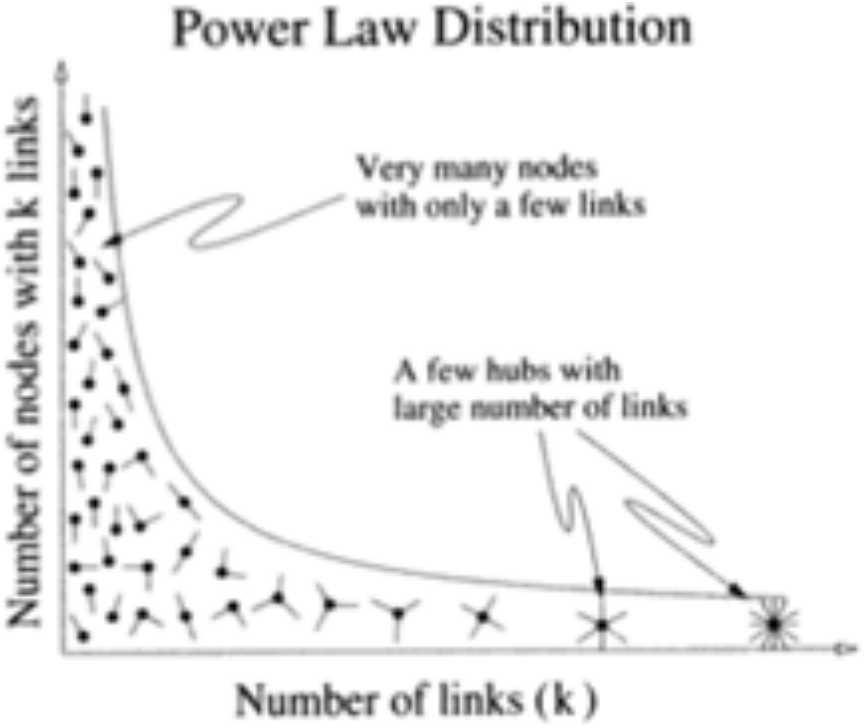
  No.

Power Law Distribution.

- By crawling large sets of web pages and measuring the in-degree distributions, one (somewhat disputed) finding is that the fraction of web pages with *k* in-links is approximately proportional to k^{-2} (not to normal distribution).
  - k^{-2} decreases (with k) much slower than normal dist. does. ``Heavy Tail''
  - Small or large in-links values are more likely to occur compared to normal distribution. (We mentioned "heavy-tailed distributions" in Lecture 23.)
- A function that decreases proportionately with *k* to some fixed power is called a *power law*
  - e.g., fraction of web pages with *k* in-links: $f(k) = \alpha k^{-2} = \alpha\ 1/k^2$
  - α is a normalizing constant (varies with total numbrt of pages, links)
- Power laws occur in:
  - Distribution of wealth follows a power law (Pareto distribution)
  - The fraction of books bought by *k* people $\propto k^{-3}$
  - The fraction of phones receiving *k* calls per day $\propto k^{-2}$
  - Citations to scientific articles, roughly $f(k) = \alpha/k^3$
  - Zipf observed that English word usage (in say a novel) follows a power law.
  - City sizes follow a power law.

Power Law vs. Normal Distribution



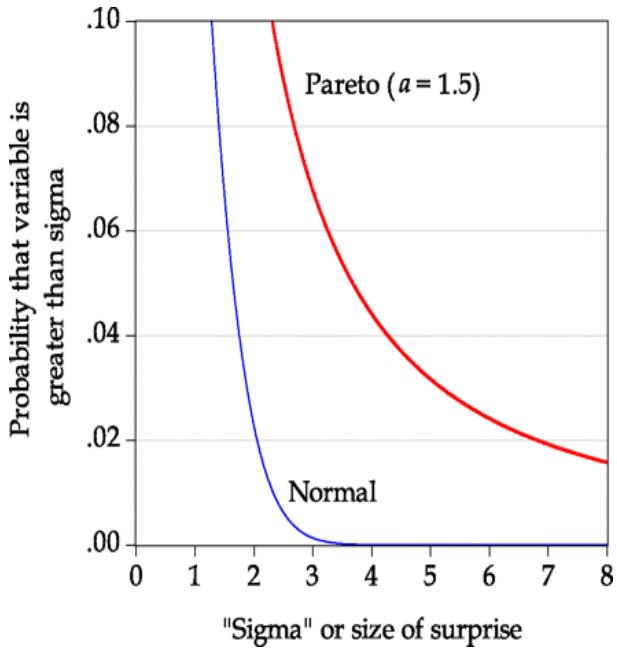


From "Linked: The New Science of Networks"

The "hub "terminology" here is inconsistent with E&K definition.

Power Laws are Scale-free

- The ratio of *f(k)* to *f(k')* depends only on the ratio *k/k'* not on their magnitude or "scale"
  - $f(k)/f(k') = \alpha k^{-c}/\alpha k'^{-c} = (k/k')^{-c}$
  - $f(2k)/f(2k') = \alpha 2k^{-c}/\alpha 2k'^{-c} = (k/k')^{-c}$
  - If you "slide along" on distribution, "relative picture" stays the same
  - Scale-free: the unit of measurement does not matter.
- They also have *long (or fat) tails*
  - significant numbers of events occur with large values of *k*
  - due to scale-free property: the relative reduction from *f(5)* to *f(10)* is same as from *f(50)* to *f(100)*, *f(1000)* to *f(2000)*, etc.: very stretched out!
  - compare Pareto distribution to normal distribution as *k* grows



Example 1: Links to Web Pages

- f(k): fraction of web pages with *k* in-links.
- f(k) = $\alpha\, k^{-2}$
  - So $\frac{f(1)}{f(2)} = 2^2 = 4$ times as many pages with 1 in-link as 2.
  - So $\frac{f(2)}{f(3)} = \frac{2^{-2}}{3^{-2}} = \frac{9}{4}$ times as many pages with 2 links as 3.
  - So $\frac{f(3)}{f(4)} = \frac{3^{-2}}{4^{-2}} = \frac{16}{9}$ times as many pages with 3 links as 4.
  - … 1.21 times as many pages with 10 links as 11 (ratio is 121/100)
  - … 1.02 times as many pages with 100 links as 101 (ratio is $101^2/100^2$)
- Notice that the *relative decrease* in number of pages with 1 additional links slows down very quickly, leaving a reasonable proportion of pages with large numbers of links
- BTW, can you quickly determine $\frac{f(6666666666)}{f(8888888888)}$ = ?

Example 2: What does $1/k^2$ Look Like?

- Suppose 1000 pages link to each other.
- Limit ourselves to maximum of *10 in-links* per page (for simplicity).
  - In-link distribution: $f(k) = \alpha\, k^{-2}$.
  - Table shows approx. number of pages with 1, 2, … 10 in-links
- Math is simple:
  - We know $\sum_{k=1}^{10} f(k) = 1$, so $\alpha \approx 0.645$
  - So, number of pages with
    - 1 in-link = $f(1) * 1000 \approx 645$
    - 2 in-link = $f(2) * 1000 = 0.645 * 2^{-2} * 1000 \approx 161$
    - …
- What is the number of edges?
  - 645*1+161*2+…+10* 6 = 1884
  - Edge density p = 1884/(1000*999) ≈ 0.0019

k (# of in links)	$1/k^2$	how many pages
1	1	645
2	1/4	161
3	1/9	72
4	1/16	40
5	1/25	26
6	1/36	18
7	1/49	13
8	1/64	10
9	1/81	8
10	1/100	6

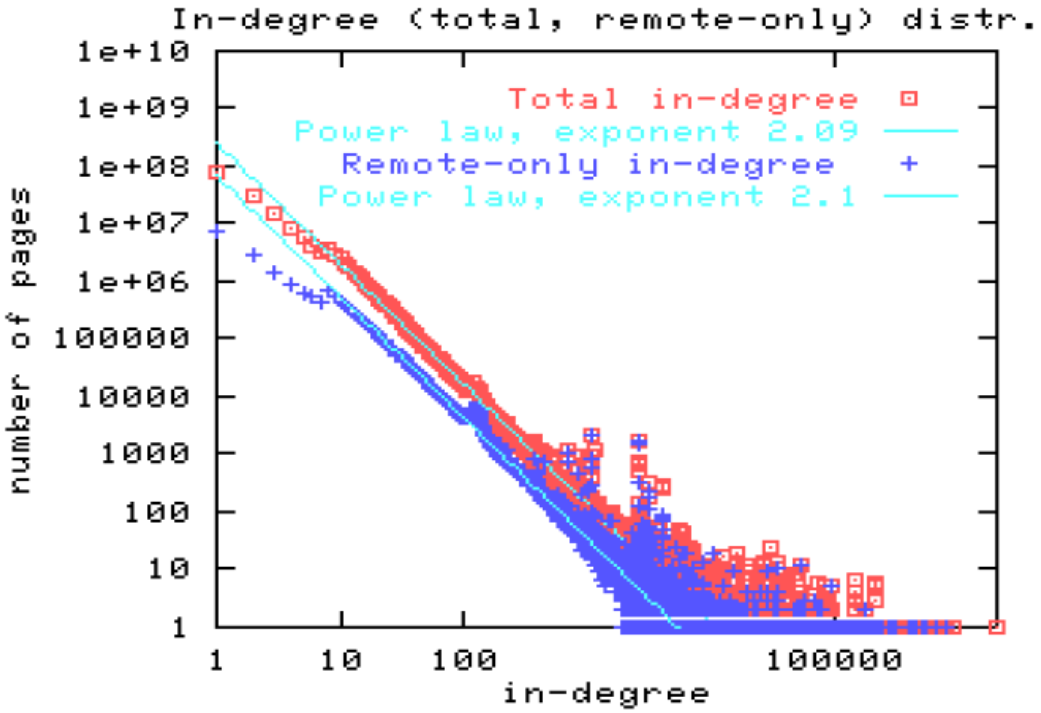
Example 3: uniformly at random

- Table shows approx. number of pages with 1, 2, … 10 in-links for In-link distribution f(k) = $\alpha\, k^{-2}$
  - About 6 of 1000 pages have 10 in-links
- **Contrast:** suppose each page selects its out-link uniformly at random with p = 0.0019 (the same edge density in Example 2).
  - Each target page has p=19/10000 chance of being selected by a specific source page.
  - Chance of a specific page having 10 in-links is $= \binom{999}{10} p^{10} (1-p)^{989} \approx 2.44 * 10^{-5}$
  - Expected number of webpages with 10 in-links ≈ 0.0244.
  - *Comparing this with 6 in power laws model, we see the power laws model have $\approx 6/.0244 \approx 245$ more times as many webpages with 10 in-links!!*

k (# of in links)	$1/k^2$	how many pages
1	1	645
2	1/4	161
3	1/9	72
4	1/16	40
5	1/25	26
6	1/36	18
7	1/49	13
8	1/64	10
9	1/81	8
10	1/100	6

Testing a Power Law: Log-log plot

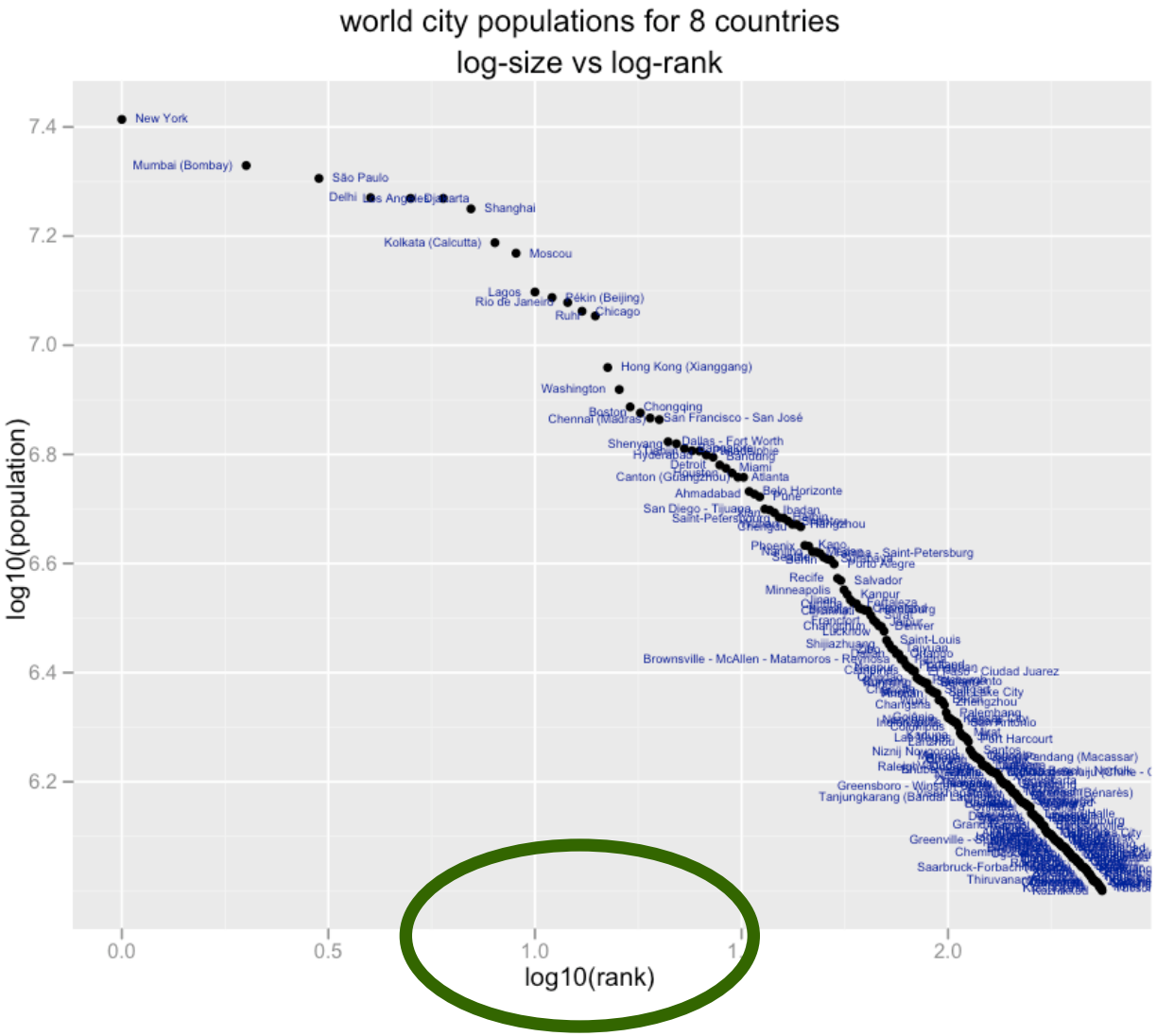
- Given some data set, does it follow a power law?
  - Start with $f(k) = \alpha k^{-c}$
  - Take logarithm of both sides: $\log f(k) = \log(\alpha k^{-c}) = \log \alpha - c \log k$
  - If we plot $\log f(k)$ vs. $\log k$, we should see a *linear relationship* with slope of $-c$ and y-intercept of $\log \alpha$ (recall α is constant)
    - i.e., **log-log plot should be linear**



Power law: in-links to web pages.
E&K Ch.18, Fig.18.2; originally from Broder et al. (2000)

Same article as Bowtie structure of the web!

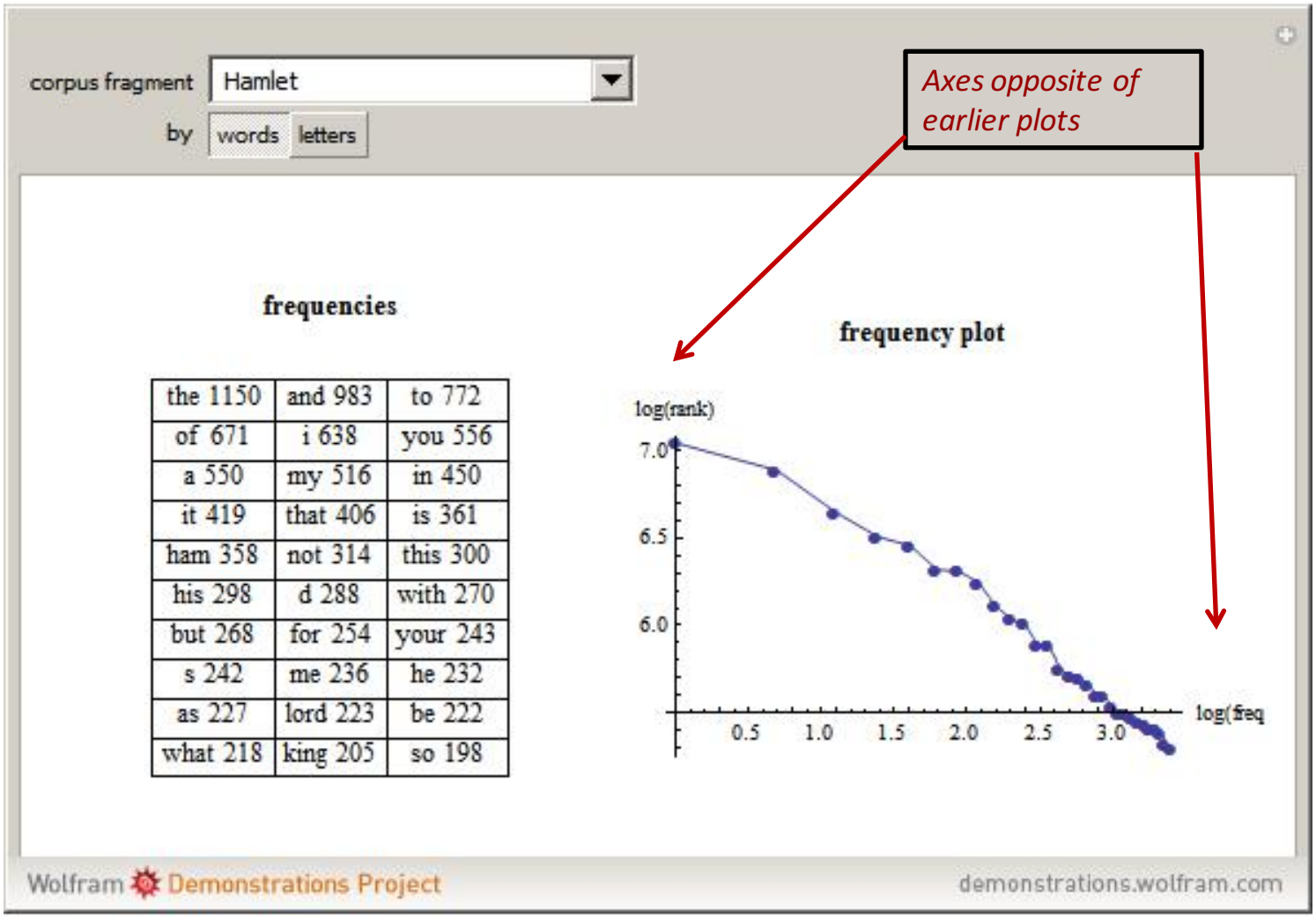
Log-log Plot: City Populations vs. Rank



- Data from 8 countries
  - Brasil
  - China
  - Germany
  - India
  - Indonesia
  - Nigeria
  - Russia
  - US

http://brenocon.com/blog/2009/05/zipfs-law-and-world-city-populations/

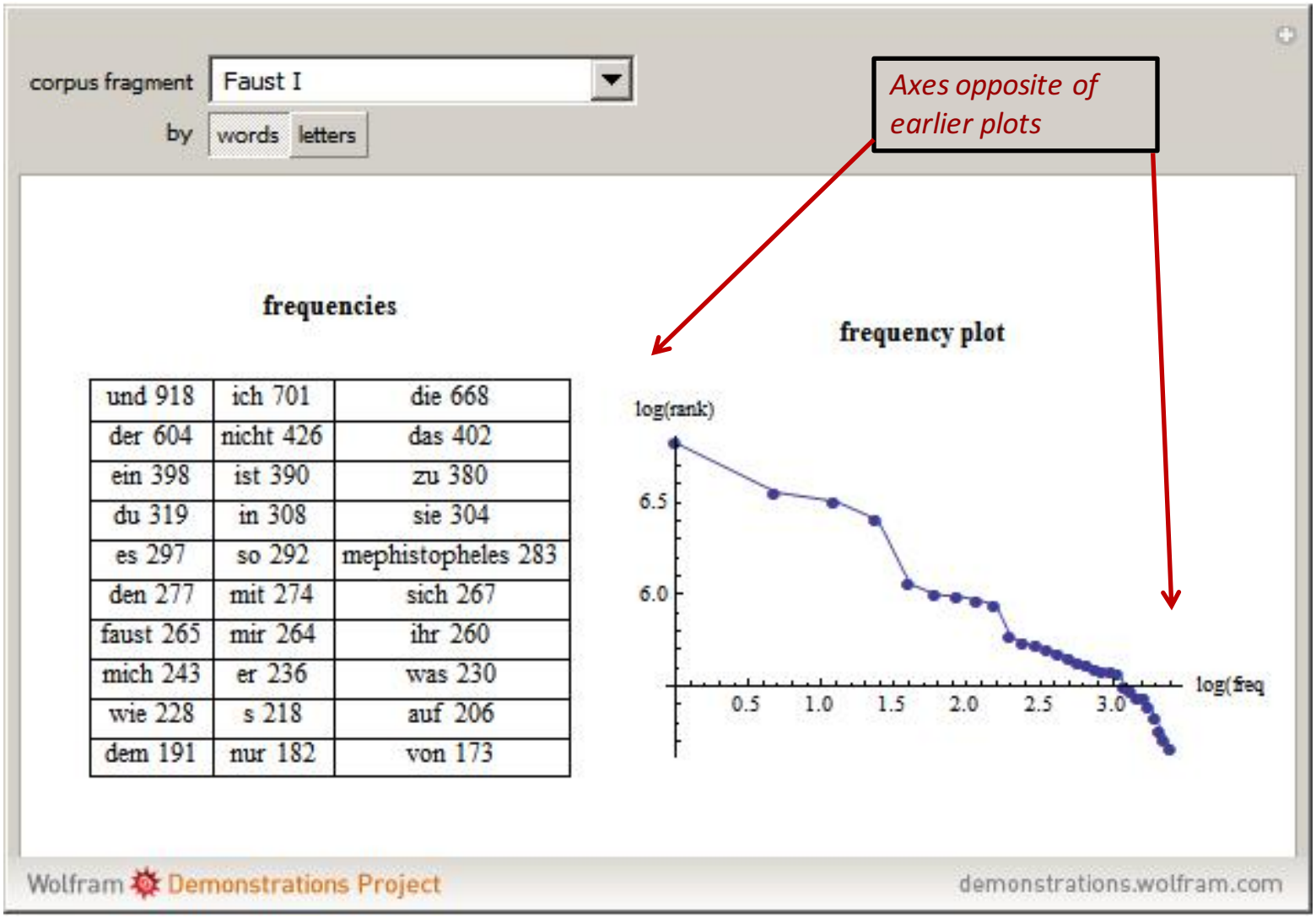
Zipf's Law for Words: Hamlet



the 1150	and 983	to 772
of 671	i 638	you 556
a 550	my 516	in 450
it 419	that 406	is 361
ham 358	not 314	this 300
his 298	d 288	with 270
but 268	for 254	your 243
s 242	me 236	he 232
as 227	lord 223	be 222
what 218	king 205	so 198

http://demonstrations.wolfram.com/ZipfsLawAppliedToWordAndLetterFrequencies/

Zipf's Law for Words: Faust



und 918	ich 701	die 668
der 604	nicht 426	das 402
ein 398	ist 390	zu 380
du 319	in 308	sie 304
es 297	so 292	mephistopheles 283
den 277	mit 274	sich 267
faust 265	mir 264	ihr 260
mich 243	er 236	was 230
wie 228	s 218	auf 206
dem 191	nur 182	von 173

http://demonstrations.wolfram.com/ZipfsLawAppliedToWordAndLetterFrequencies/

What causes a power law distribution?

- By central limit theorems, we learned that if individuals or entities make their decisions independently uniformly at random, the distribution will be normal.
- So what causes a power law?
  - Feedback introduced by correlated decisions across a population.
- Is there any simple mechanism that explains how a power-law distribution arises?
  - Rich-get-richer generative models.

Rich-Get-Richer Models

- Rather than focusing on internal processes of decision making, we assume individuals have a tendency to copy the action of others before them.
- *N* web pages created in order: *1, 2, 3, …, N*
- When page *j* created, it creates a single out-link as follows:
  a) *With probability p:* it chooses an existing page $i < j$ uniformly at random and links to it.
  b) *With probability 1-p:* it chooses an existing page $i < j$ uniformly at random and links to the page that *i* points to.
- We'll assume $m = 1$ out-link (principles are the same for $m>1$)
- What's going on here?
  - (b) is a simple copying process: creator of new web page *j* is "lazy," and just copies the behavior of some random other page.
  - *j* isn't linking to *i*, it is taking *i*'s decision to be "useful" and copying it

Rich-Get-Richer Models

- Process (b) is equivalent to saying:
  - *With probability 1-p:* link to an existing page *k* with probability proportional to *in-degree(k)*
- This is why we call this a *"rich get richer" model*
  - The more links a node has, the more likely it is to attract even more
  - Popularity is self-reinforcing
- Also known as *preferential attachment*
  - a popular model for explaining formation of networks.
- It turns out this process will give rise to a distribution over in-degrees that satisfy a power law: the fraction of pages with in-degree *k* will be (approximately) proportional to $1/k^c$, *i.e.*, $f(k) = \alpha k^{-c}$
  - constant *c* will depend on the probabilities *p*
  - as *1-p* (copying) probability gets larger, *c* gets smaller, making extremely popular pages very likely (more prevalent)

Preferential Attachment Graph

- You can look at a NetLogo simulation to get a sense of the distribution of node popularity
  - File -> Model Libraries -> Networks -> Preferential attachment
- Model is simple (undirected) preferential attachment model in which new nodes connect to old ones with probability proportional to their degree (not in-degree, since it's an undirected graph)
- You will notice that the empirical degree distribution has a long tail and the log-log plot is roughly linear, as expected

Rich-get-richer vs Cascade

- What are the differences between rich-get-richer copying model and cascade model (in Ch. 16)?

	Copying model	Cascade model
Number of Choices	Many	Two
Observability	Limited	Global
Rational Decision Making	No (imitation and copying)	Yes

Ingredients of Rich get Richer?

- Process in which elements (links, new buyers, employees) are added one at a time these elements are attracted to points in proportion to their popularity/size
  - new web pages, population growth in a city, sizes of firms, …
- Is popularity something intrinsic? How much does randomness play a role in popularity of items?
  - Is Harry Potter intrinsically better than other books/series in the same genre? or is it's mega-popularity entirely random?
  - Are Lady Gaga, *Taylor Swift, Bruno Mars, Justin Bieber, Swedish House Mafia, Pitbull, PSY, Maroon5* intrinsically better than the other 10,000 artists vying for your attention?
- If the history replayed multiple times, we know there should be a power-law distribution but would the most popular items be the same?

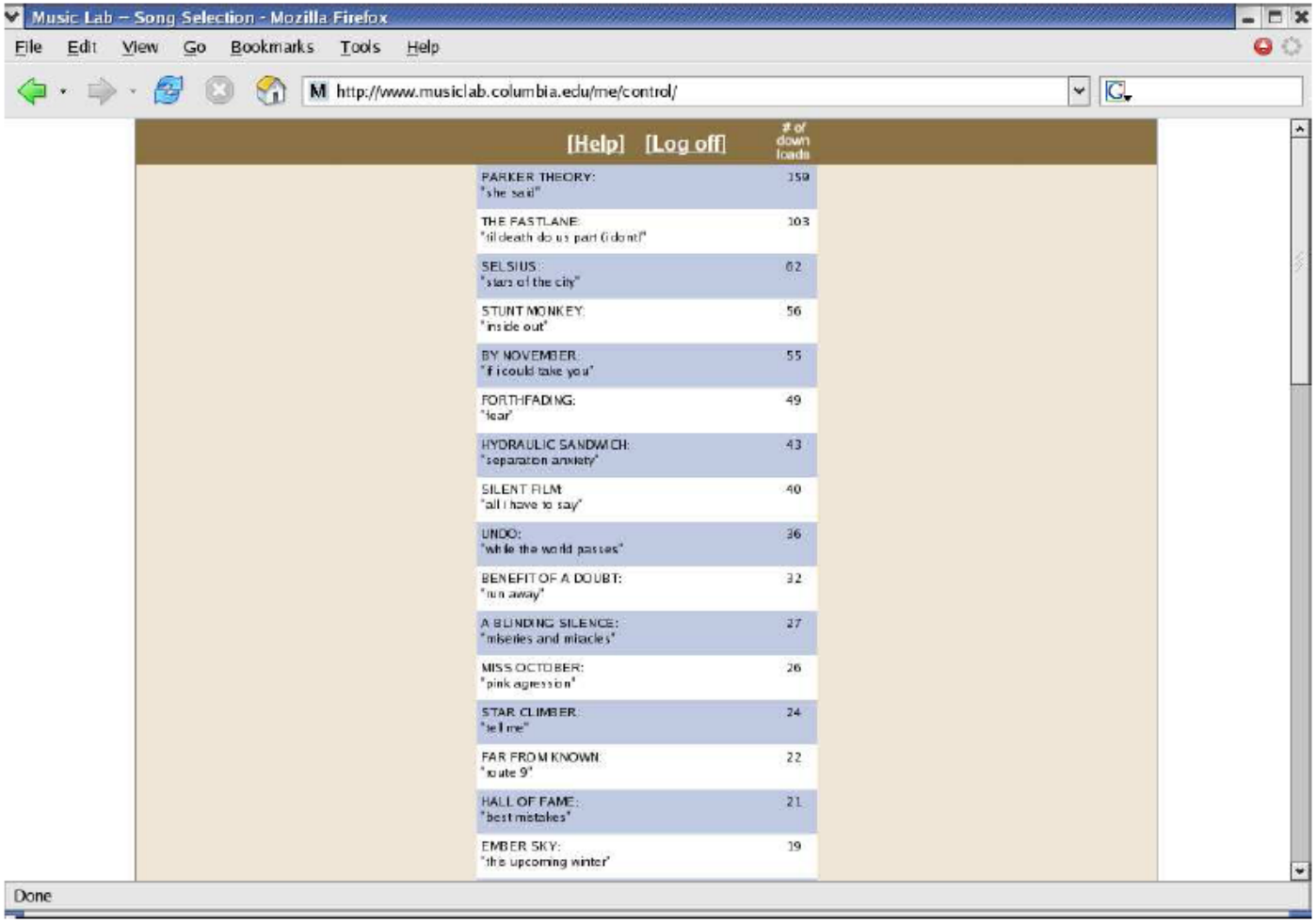
Is Popularity a Chance Happening?

- Salgankik, Dodds, Watts (2006):
  - Created web site with 48 songs varying quality (by real artists).
  - Could listen and then decide to download one's you liked.
  - Also a table is shown listing number of prior downloads of each track.
  - They ran *nine different sites ("worlds") in parallel* all with same tracks
    - Eight sites with social feedback of download counts (each with an initial download count of zero)
    - One site as a control group with no social feedback, referred as independent users.
  - Parallel websites = parallel worlds. Is it a nice way to simulate repeating history?
  - Each participant was slotted to one of the sites at random.
  - In Experiment 1, tracks in those eight websites ordered randomly (not in download count order).
  - In Experiment 2, tracks in those eight websites ordered in download count order (a more salient social signal).

Experiment 1: Interface

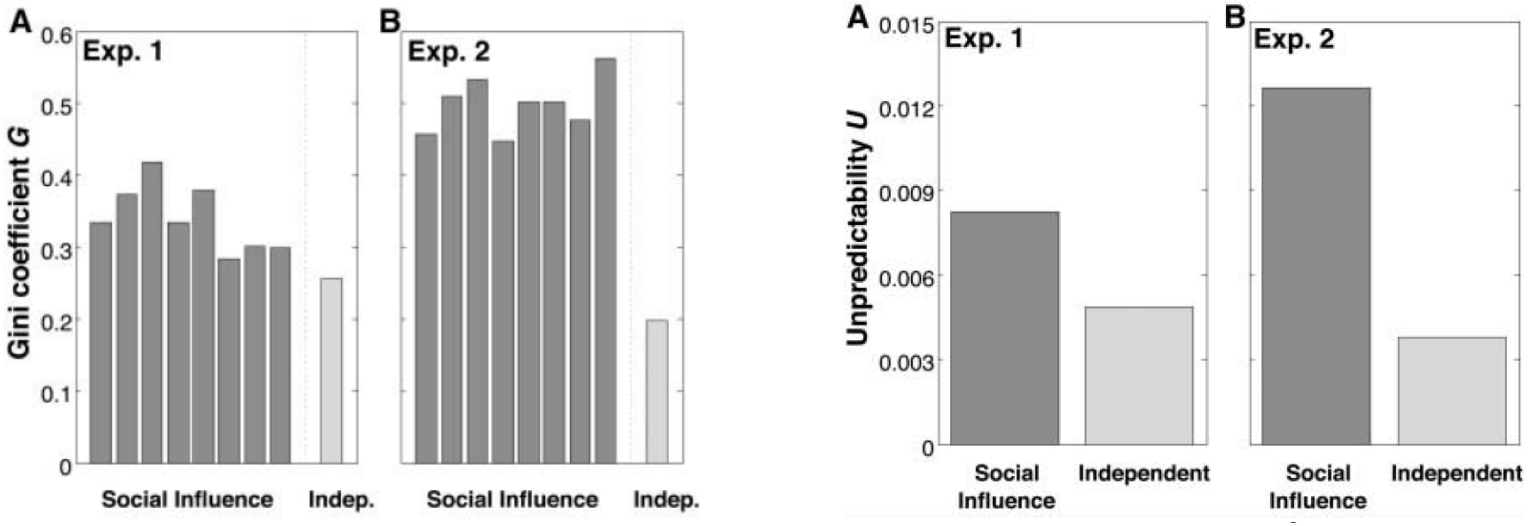
Salgankik, Dodds, Watts (2006)

Experiment 2: Interface



Salgankik, Dodds, Watts (2006)

Results (from Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market, *Science* 311 (2006), Salgankik, Dodds, Watts)



- *Inequality* (Gini index: avg. difference in market share between pairs of songs)
  - more inequality in social conditions, esp. with stronger signal (Exp.2)!
- *Unpredictability = randomness* (difference in a song's "market share", averaged over all comparisons in any two different worlds, then averaged over all songs)
  - a song's market share is much more variable in the social conditions, esp. with the stronger signal (Exp.2)!
- Strongly suggests a real random element to popularity
- *But not completely:* higher quality songs tended to fare better, lower quality worse; so while there's randomness, behavior isn't just pure "copying".
  - The best song never ends at bottom and the worst song never ends at top.

The Long Tail

- The influence of popularity on choices of individuals, when viewed as preferential attachment/rich-get-richer, manifests itself in very non-normal distributions: power laws
  - Very popular items (which won't occur with normal distribution)
  - Power laws also give rise to *a huge number of not very popular items*: this also won't occur with a normal distribution.
- What is the significance of this?
- In the age of e-commerce and digital goods, it can dramatically change the nature of commerce and business.
- Question: Are most sales generated by small set of very popular items ("hits") or very large set of less popular items ("niche products")?
- Chris Anderson (editor of Wired) discussed about this in his book *"The Long Tail: Why the Future of Business is Selling Less of More"*

Why the Long Tail Matters?

- Amazon was among the first to take advantage of the "long tail" so let's discuss books (one might argue Sears-Roebuck did in 1880s)
- In Bricks-and-mortar bookstores, which types of books you find the most? "Hits" or "Niche books".
- Bricks-and-mortar bookstores have limited capacity to hold inventory
  - If you can only stock 5,000-10,000 items, you are going to be sure they appeal to a wide audience.
  - Limits availability to readers, curtails production (authors have no outlets)
- What does an Amazon bring to the table?
  - Can store huge inventories, because of worldwide client base
  - Facilitated by internet, search tools, low-cost shipping, print-on-demand technologies, etc.
  - So does that change the nature of what we buy?