Power Laws and Rich-Get-Richer Phenomena

Objectives

- Examine phenomena related to popularity
- Specific instance: popularity of Web pages in terms of number of in-links
- Power-law distribution of number of in-links
- Simple model to explain why power-laws emerge
- Approximate mathematical analysis of the model

Popularity as a Network Phenomenon

- Popularity is characterized by extreme imbalances
  - almost everyone goes through life known only to people in their immediate social circles,
  - a few people achieve wider visibility, and
  - a very, very few attain global name recognition
- Analogous things could be said of books, movies, or almost anything that commands an audience

Popularity as a Network Phenomenon

- How can we quantify these imbalances?
- Why do they arise?
- Are they somehow intrinsic to the whole idea of popularity?

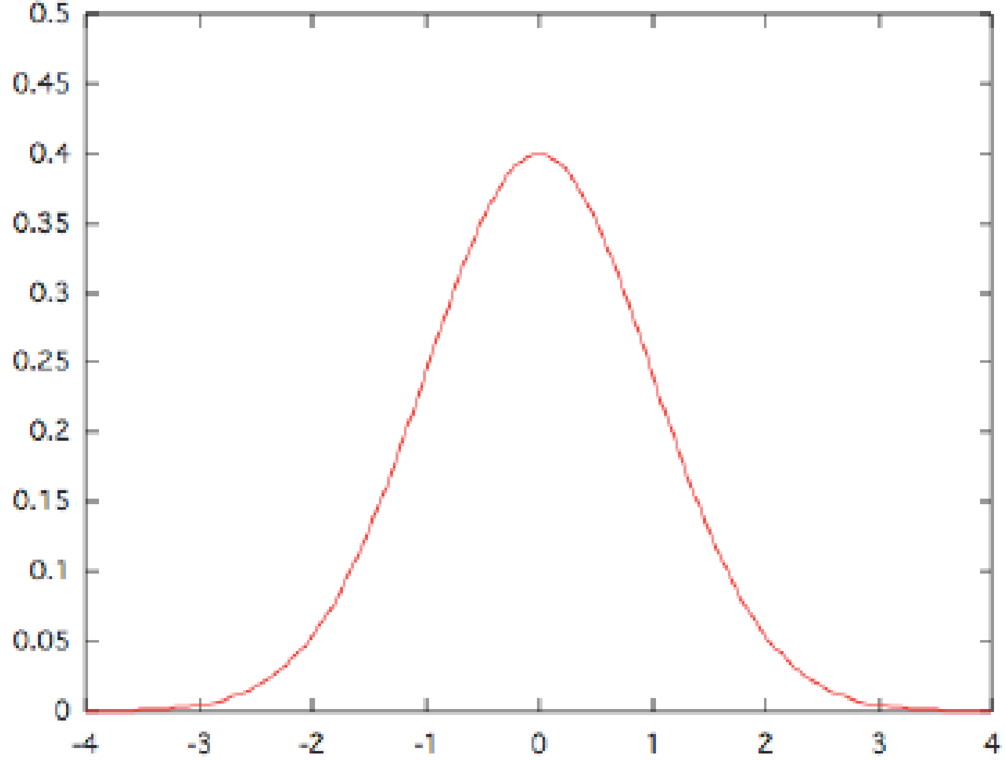
Popularity in Web

- Focus on the Web as a concrete domain in which it is possible to measure popularity very accurately
  - difficult to estimate the number of people who have heard of Barack Obama or Bill Gates
  - easy to count the number of links to high-profile Web sites such as Google, Amazon, or Wikipedia
- Number of in-links to a Web page as a measure of the page's popularity
  - but as just an example of a much broader phenomenon

Basic question

As a function of k, what fraction of pages on the Web have k in-links?

A Simple Hypothesis: The Normal Distribution



$\mu = 0$
$\sigma = 1$

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

- Why is normal distribution ubiquitous across the natural sciences?
- Central Limit Theorem: if we take any sequence of small independent random quantities, then in the limit their sum (or average) will be distributed according to the normal distribution
- Ex:
  - perform repeated measurements of a physical quantity, the the variations are the cumulative result of many independent sources of error in each trial, then the distribution of measured values is normal

The Normal Distribution in the Web

- How would this apply in the Web?
  - Assume that each page decides independently at random whether to link to any other given page
  - The number of in-links to a given page is the sum of many independent random quantities (i.e. the presence or absence of a link from each other page),
  - We'd expect it to be normally distributed
  - The number of pages with k in-links should decrease exponentially in k, as k grows large

The conclusion is not verified by reality, because the assumption is not valid

Power Laws

- What does reality say?
  - The fraction of Web pages that have k in-links is approximately proportional to $1/k^2$
  - recurring finding in studies over many different Web snapshots, taken at different points in the Web's history
- Why is this so different from the normal distribution?
  - exponential decrease: e^{-k^2} or e^{-k} or 2^{-k}
  - power law: k^{-2}
  - Ex: k = 1000
    - exp -> 0
    - power law -> 10^{-6}

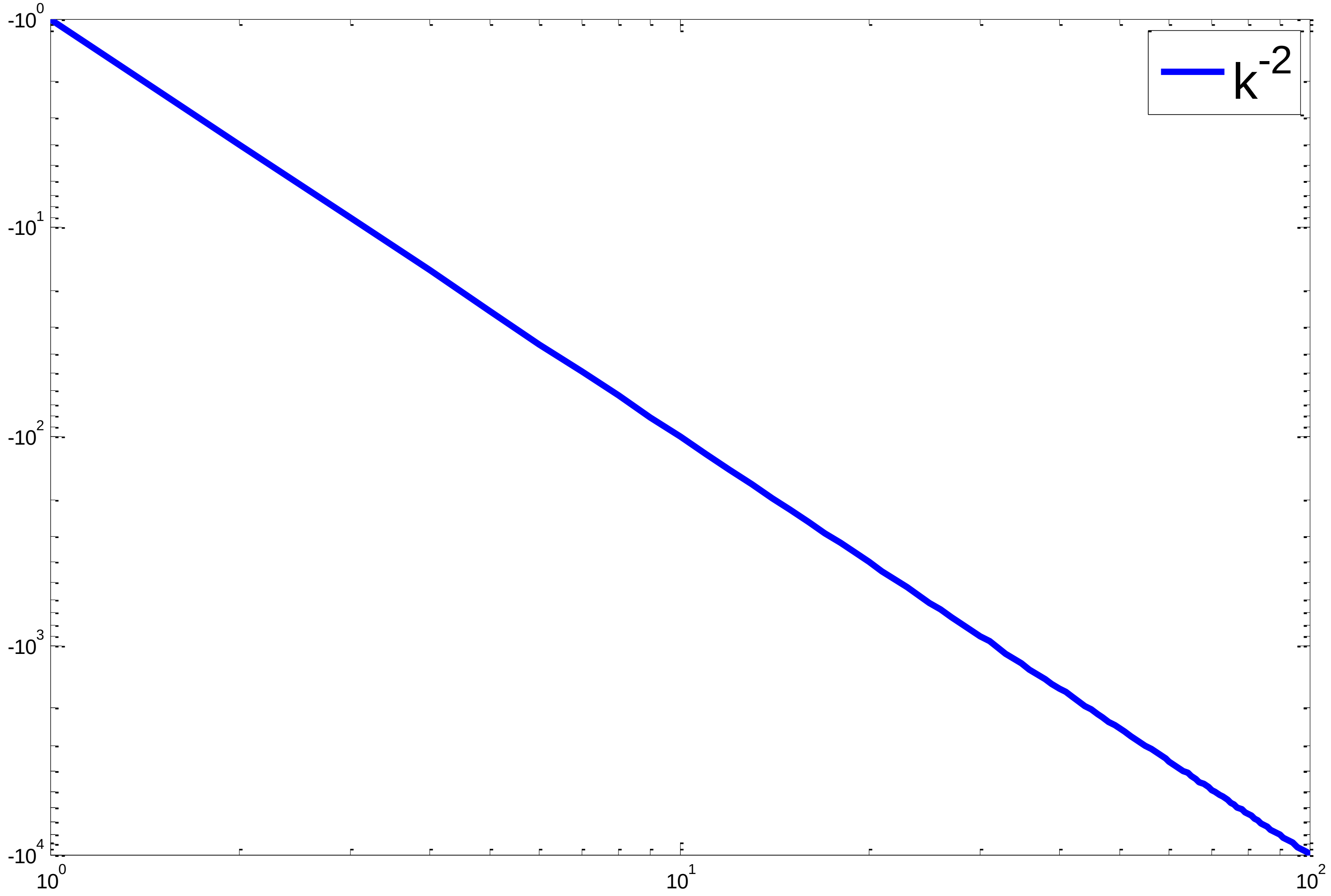
Power Laws

- What does this mean?
  - With power laws it’s possible to see very large values of *k*
  - Remember: large values of popularity are likely to arise
- Where else do we observe power laws?
  Fractions of:
  - telephone numbers receiving *k* calls per day ($1/k^2$)
  - Books bought by *k* people ($1/k^3$)
  - scientific papers receiving *k* citations ($1/k^3$)
- As normal distribution is widespread in natural sciences, power laws dominate when we measure a type of popularity

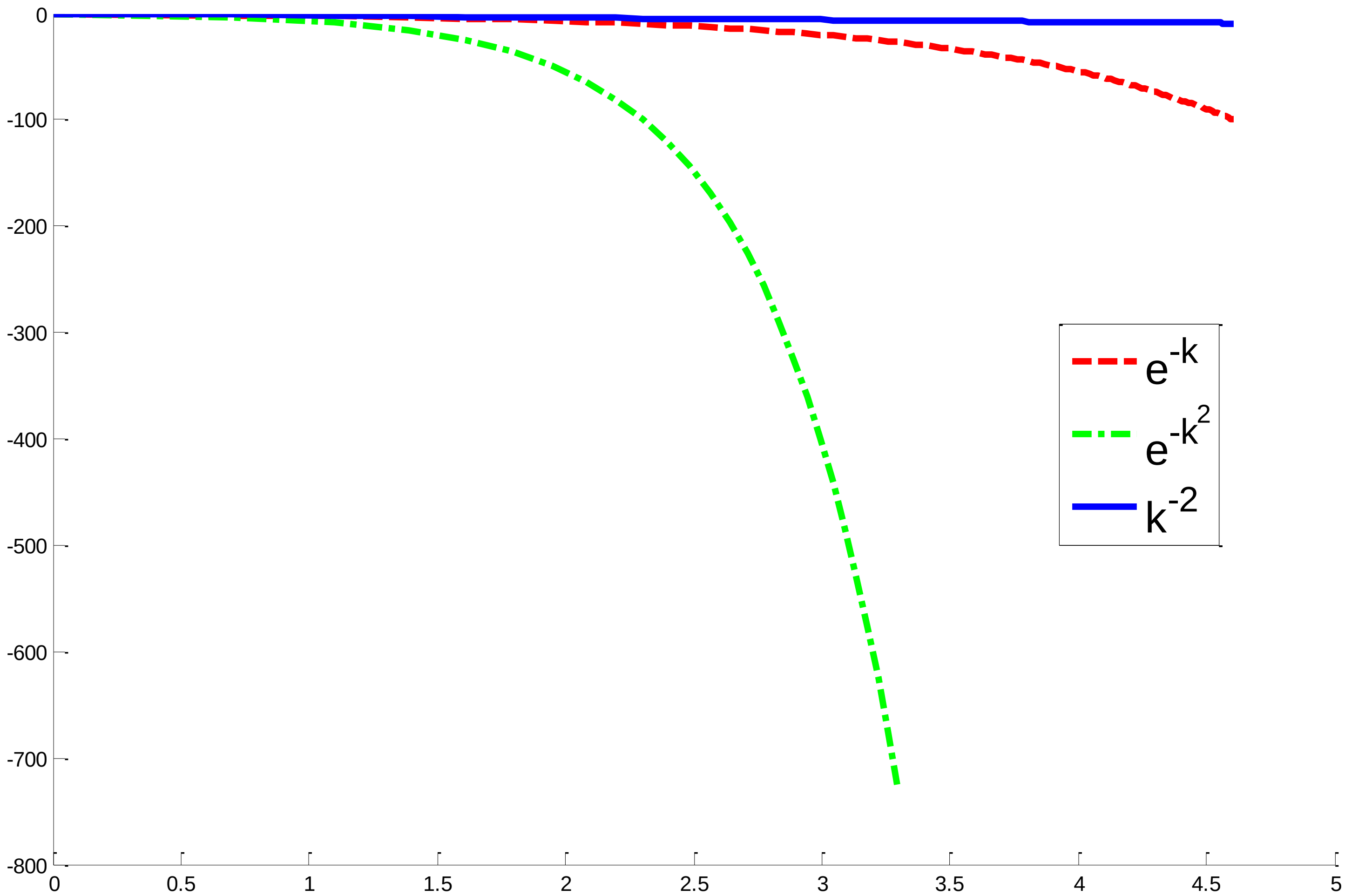
Power Laws

- Quick test for whether a dataset exhibits a power-law distribution
  - $f(k)$ be the fraction of items that have value k
  - test if $f(k) = a/k^c$ for some a and c
  - $f(k) = ak^{-c}$
  - $\log f(k) = \log(ak^{-c}) = \log a - c \log k$
- What does this mean?
  - loglog plot: plot log $f(k)$ as a function of $\log k$,
  - then we should see a straight line: $-c$ will be the slope, and $\log a$ will be the y-intercept

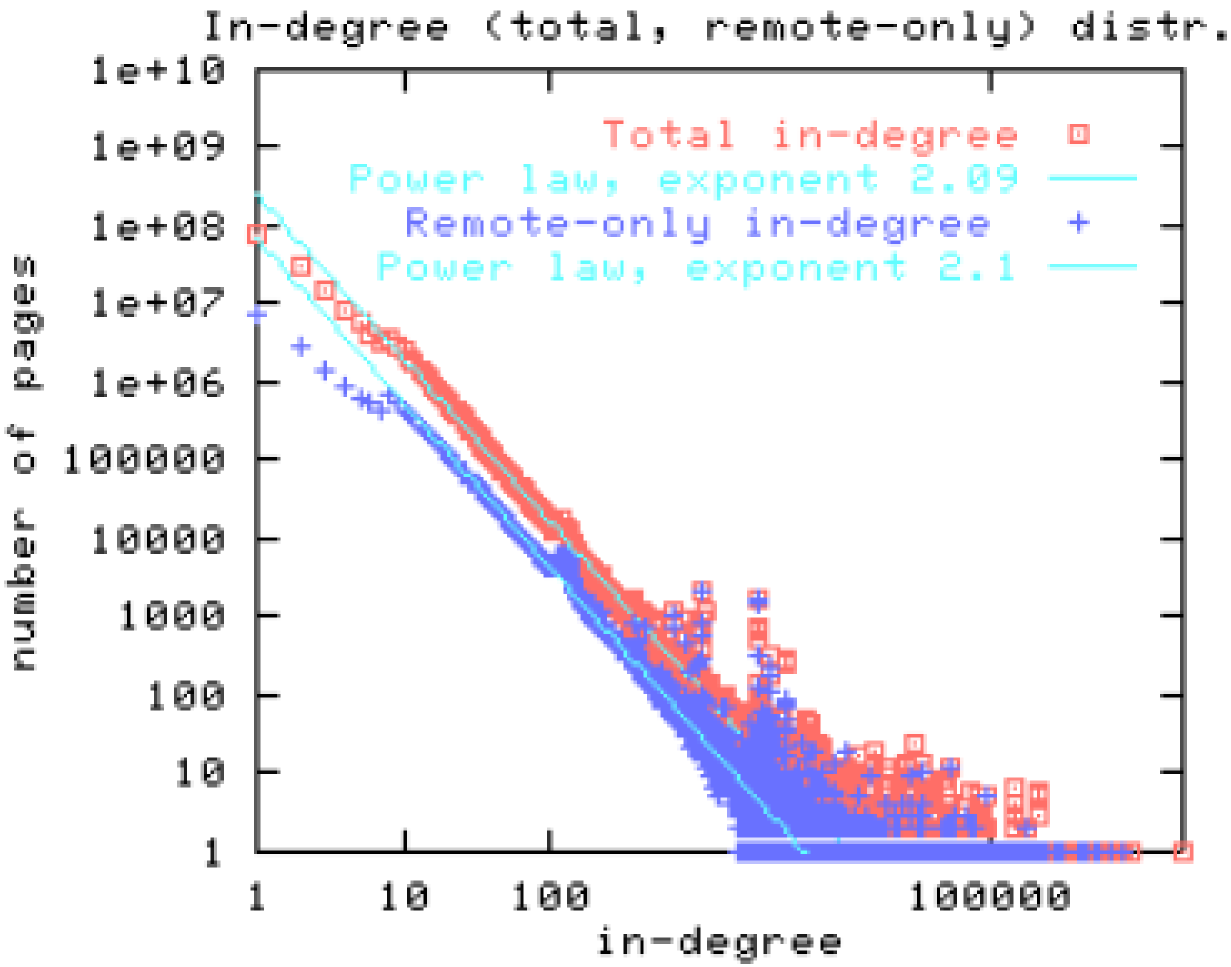
loglog plot



loglog plot



loglog plot of in-links in Web



What causes power laws?

- We need a simple explanation for what is causing power laws?
- loglog of in-links in Web: a straight line for much of the distribution
  - even when many utterly uncontrollable factors come into play in the formation of Web links
- What underlying process is keeping the line so straight?

Rich-Get-Richer Models

(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.)

Rich-Get-Richer Models

- Copying mechanism in (2b) implements "rich-get-richer" dynamics
  - when you copy the decision of a random earlier page, the probability that you end up linking to some page ξ is directly proportional to the total number of pages that currently link to ξ
- We can equivalently write:

(2) ...

(b) With probability $1 - p$, page j chooses a page ξ *with probability proportional to* ξ*'s current number of in-links, and creates a link to* ξ.

- Why "Rich-get-richer" (a.k.a. *preferential attachment*)? the probability that page ξ experiences an increase in popularity is directly proportional to ξ*'s* current popularity

Analysis of Rich-Get-Richer

- Probabilistic model:
  - we have specified a randomized process that runs for *N* steps (as the *N pages* are created one at a time)
  - we determine the expected number of pages with *k* in-links at the end of the process
- Random variable $X_j(t)$ is the number of in-links to a node *j* at a time step $t \geq j$
  - Initial condition: $X_j(j) = 0$, because n*ode j* starts with no in-links when it is first created at time *j*
  - probability that node *t + 1* links to node *j* is:

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t}$$

Analysis of Rich-Get-Richer

- Approximate the probabilistic model with a deterministic model (no clear proof for the probabilistic)
  - time runs not in discrete steps but continuously in [0,*N*]
  - $X_j(t)$ is approximated by a continuous function of time $x_j(t)$

$$\frac{p}{t} + \frac{(1-p)X_j(t)}{t} \longrightarrow \frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}$$

Analysis of Rich-Get-Richer

Set $q = 1\text{-}p$

$$\frac{dx_j}{dt} = \frac{p + qx_j}{t} \quad \Rightarrow \quad \frac{1}{p + qx_j}\frac{dx_j}{dt} = \frac{1}{t} \quad \Rightarrow$$

$$\int \frac{1}{p + qx_j}\frac{dx_j}{dt}dt = \int \frac{1}{t}dt \quad \Rightarrow \quad \ln(p + qx_j) = q\ln t + c$$

$$p + qx_j = At^q \ \text{ writing } A = e^c \quad \Rightarrow \quad x_j(t) = \frac{1}{q}\left(At^q - p\right)$$

Analysis of Rich-Get-Richer

$$0 = x_j(j) = \frac{1}{q}\left(Aj^q - p\right) \quad \Rightarrow \quad A = p/j^q \quad \Rightarrow$$

$$x_j(t) = \frac{1}{q}\left(\frac{p}{j^q}\cdot t^q - p\right) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right]$$

- With the probabilistic model: For a given value of *k*, and a time *t*, what fraction of all nodes have at least *k* in-links at time *t*?
- With the approximate model: For a given value of *k*, and a time *t*, what fraction of all functions $x_j(t)$ satisfy $x_j(t) \geq k$?

Analysis of Rich-Get-Richer

$$x_j(t) = \frac{p}{q}\left[\left(\frac{t}{j}\right)^q - 1\right] \geq k \quad \Rightarrow \quad j \leq t\left[\frac{q}{p} \cdot k + 1\right]^{-1/q}$$

The fraction of values *j*, out of total *t* values, that that satisfy this is:

$$\frac{1}{t} \cdot t\left[\frac{q}{p} \cdot k + 1\right]^{-1/q} = \left[\frac{q}{p} \cdot k + 1\right]^{-1/q}$$

This fraction approximates the fraction of nodes *F(k)* with at least *k* in-links. We want to approximate the fraction of nodes *f(k)* with exactly *k* in-links:

f(k) = -d*F(k)*/d*k*

Analysis of Rich-Get-Richer

Differentiating $\left[\frac{q}{p}\cdot k+1\right]^{-1/q}$ we get $\frac{1}{q}\frac{q}{p}\left[\frac{q}{p}\cdot k+1\right]^{-1-1/q}$

The fraction of nodes *f(k)* with *k* in-links is proportional to $k^{-(1+1/q)}$

This a power law with exponent $1+\frac{1}{q}=1+\frac{1}{1-p}$

$0 \leq p \leq 1$

- If *p* close to 1 -> no copying -> exponent tends to infinity -> nodes with very large numbers of in-links become increasingly rare
- If *p* close to 0 -> exponent becomes 2 -> allowing for many nodes with very large numbers of in-links
- We see why exponent close to 2 has been observed in real measurements in Web

The Unpredictability of Rich-Get-Richer Effects

- Once an item becomes well established, the rich-get-riche push it even higher
- But the initial stages of its rise to popularity is a relatively fragile thing
  - random effects early in the process
  - Ex: if we could roll time back 15 years, and then run history forward again, would the Harry Potter books again sell hundreds of millions of copies?
- If history were to be replayed multiple times, a power-law distribution of popularity emerges each of these times, but it’s far from clear that the most popular items would always be the same

The Unpredictability of Rich-Get-Richer Effects

- Salgankik, Dodds, and Watts study:
  - They created a music download site with 48 obscure songs of varying quality
  - Visitors 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
  - At the end of a session, visitors were given the opportunity to download copies of the songs that they liked
- Simulate the “history replayed multiple times”:
  - upon arrival they were actually being assigned (without knowing) 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
  - Each parallel copy then evolved differently as users arrived

- Goal: observe what happens to the popularities of 48 songs when history runs forward eight different times

The Unpredictability of Rich-Get-Richer Effects

- Results of the study:
  - The "market share" (popularity measured through downloads) 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
- Second goal: is feedback producing greater inequality in outcomes (copying -> power-laws)
  - Assigned some users to a ninth version of the site with no feedback about download counts
  - Result: significantly less variation in the market share of different songs

Conclusion

- This was a simple model
- Goal is not to capture all reasons why people create links on the Web, but to show that a simple and natural principle behind link creation leads directly to power laws
- Rich-get-richer models can suggest a basis for power laws in a wide array of settings
  - Ex: populations of cities have been observed to follow a power law distribution
  - Why? once formed, a city grows in proportion to its current size simply as a result of people having children -> a rich-get-richer model
- Finding similar laws governing Web page popularity, city populations, gene copies, river sizes, etc. is quite mysterious
- If one views all these as outcomes of processes exhibiting rich-get-richer effects, then the picture starts to become clearer