

Power Laws & Rich Get Richer

Advanced Social Computing

Department of Computer Science
University of Massachusetts, Lowell
Spring 2020

Hadi Amiri
hadi@cs.uml.edu



Lecture Topics

- Popularity
- Power Laws
- Rich Get Richer model

Popularity

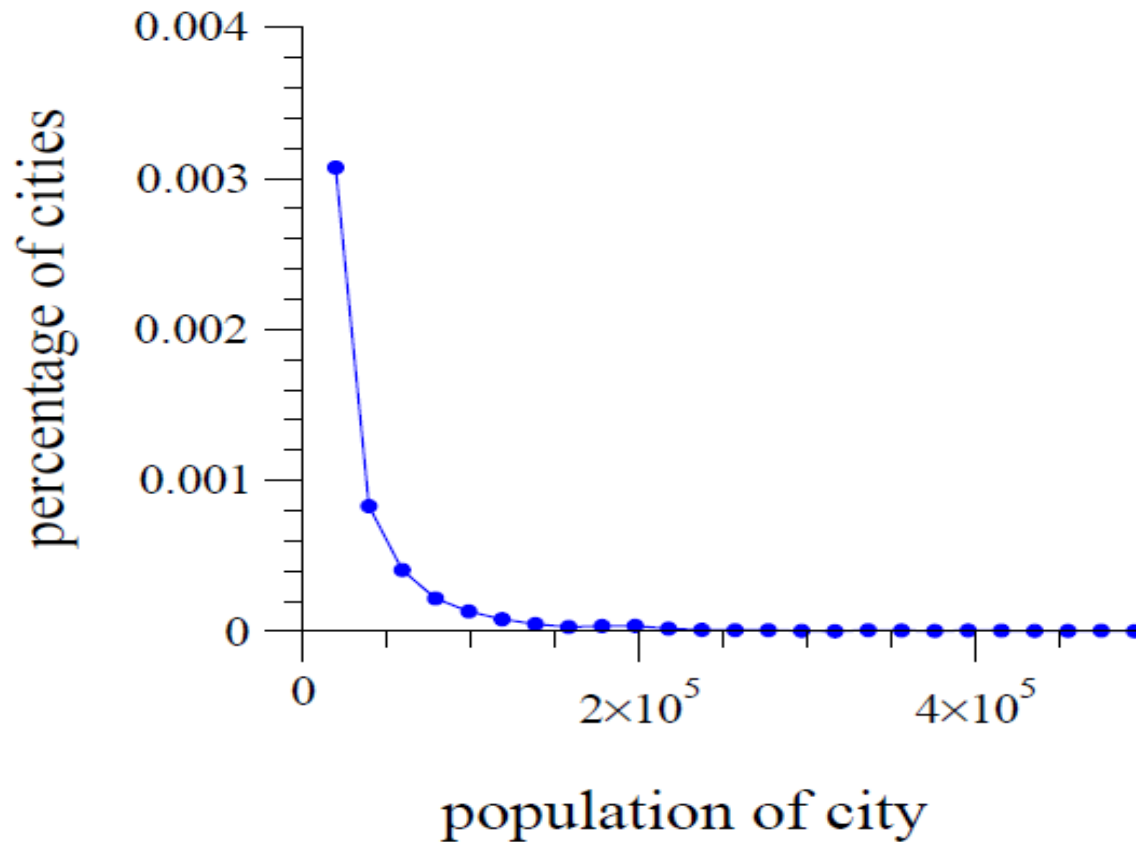
- Popularity can be characterized by **extreme imbalances!**
 - People are known to their immediate social circle!
 - Few people achieve wider visibility!
 - Very few achieve global name recognition.
- Learning objectives:
 - How can we quantify these imbalances?
 - Why do they arise?

Power Law

- A function that decreases as k to some fixed power, e.g. $1/k^2$, is called a **power law**!
 - It allows to see very large values of k in data!
- Extreme imbalances are likely to arise!

Power Law- Cnt.

- Histogram of the populations of all US cities with population of 10,000 or more.



Power Law- Cnt.

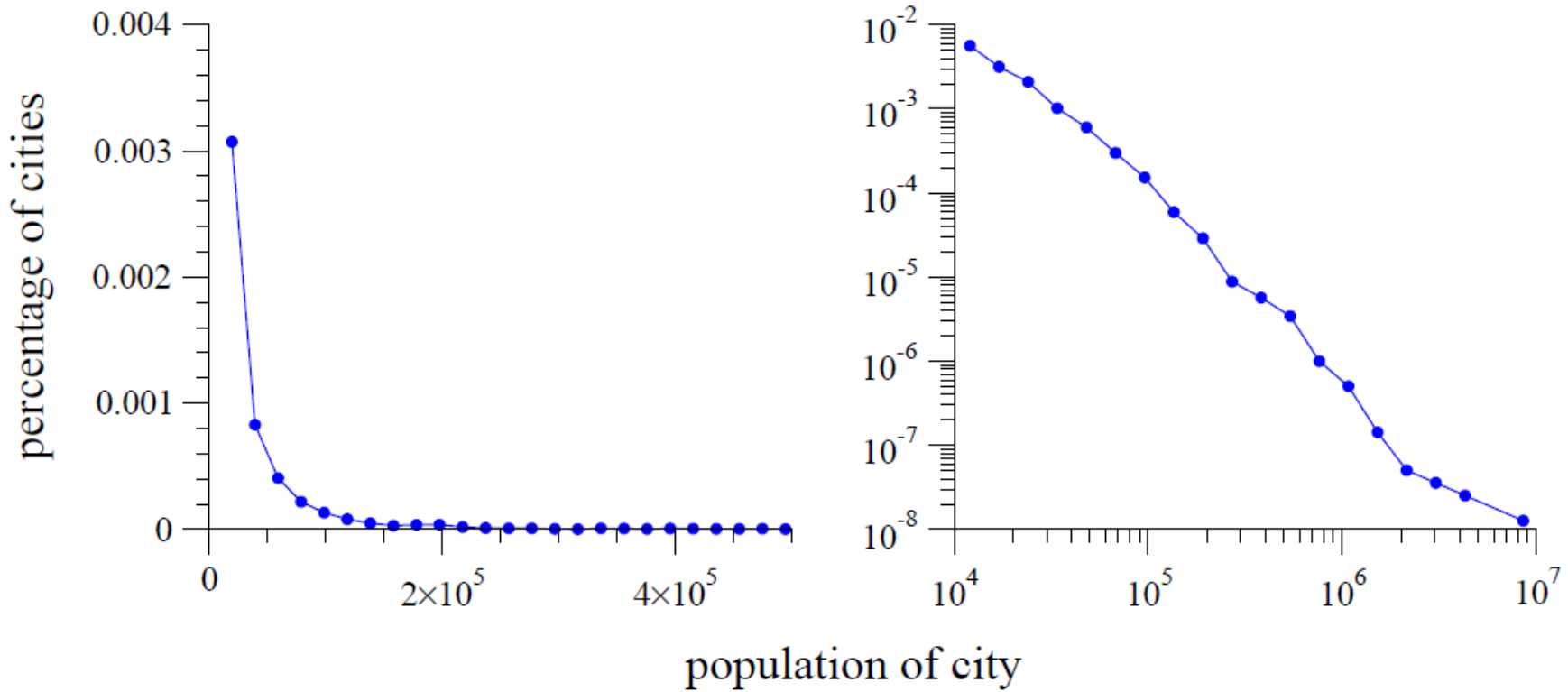
- **Power law Test:** Given a dataset, test if it exhibits a power law distribution?
 1. Compute histogram of values wrt a popularity measure (e.g. *#in-links, #downloads, population of cities, etc.*)
 2. Test if the result approximately estimates a power law $1/k^c$ for some c , and if so, estimate the exponent c .

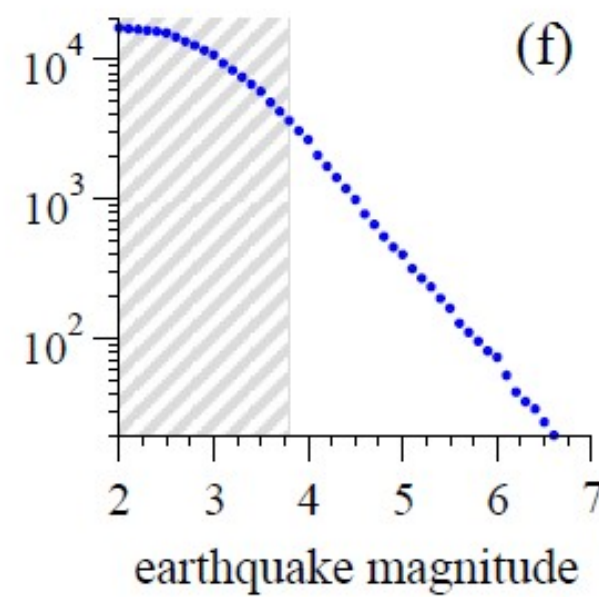
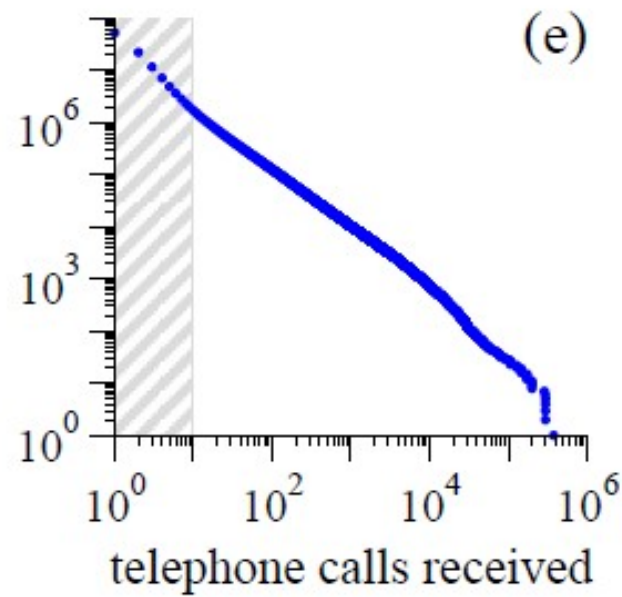
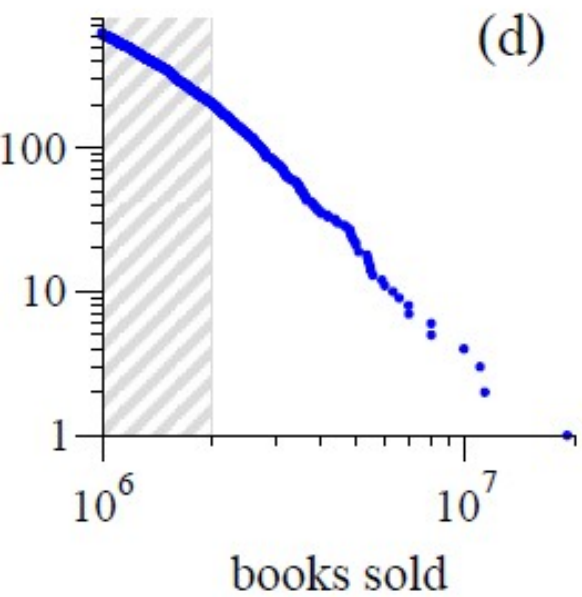
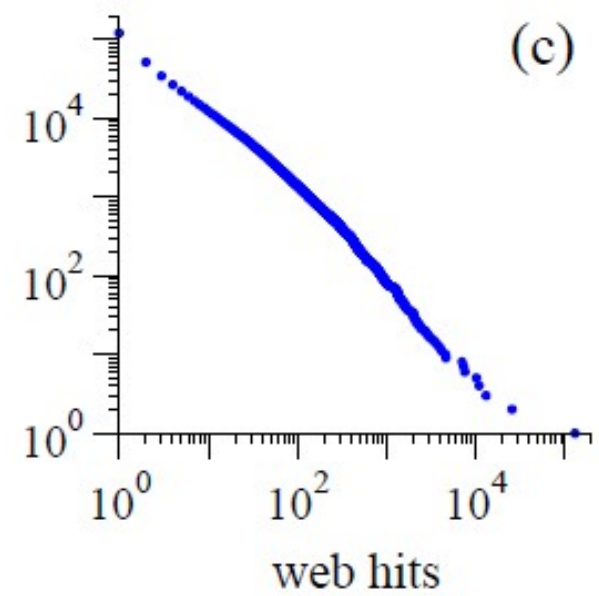
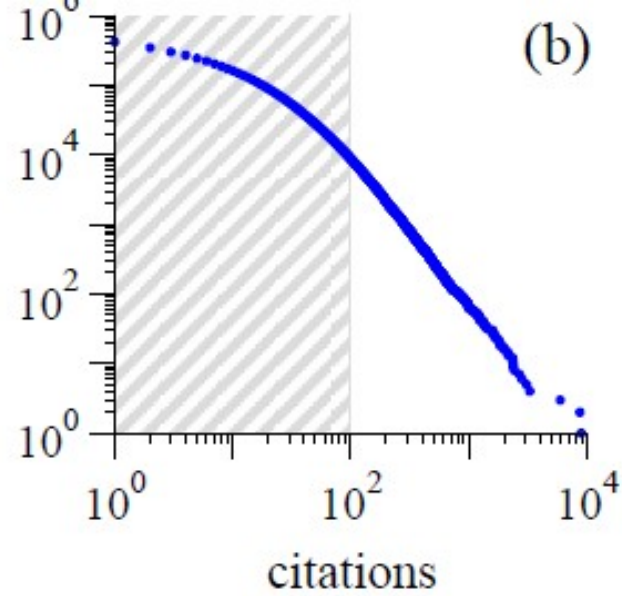
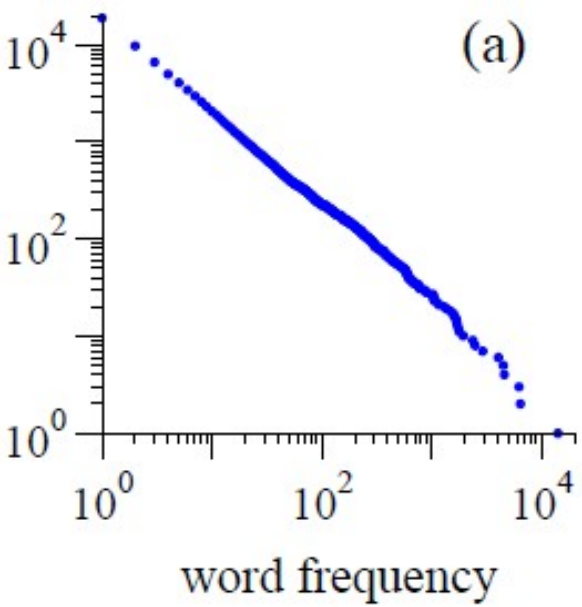
Power Law- Cnt.

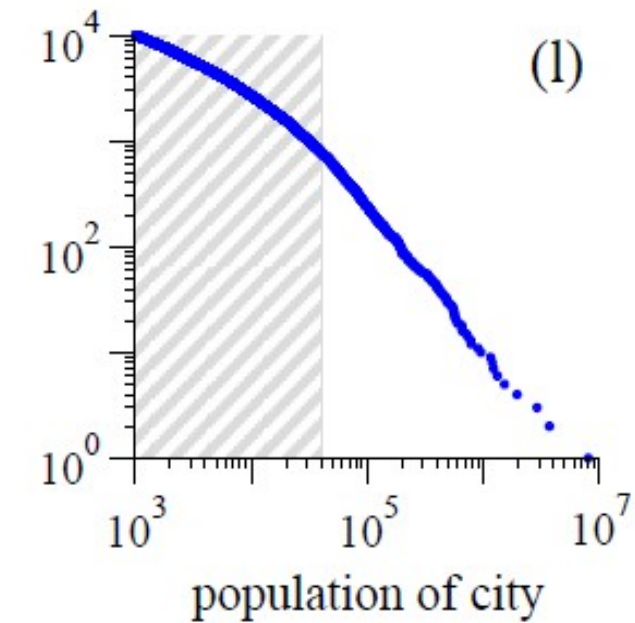
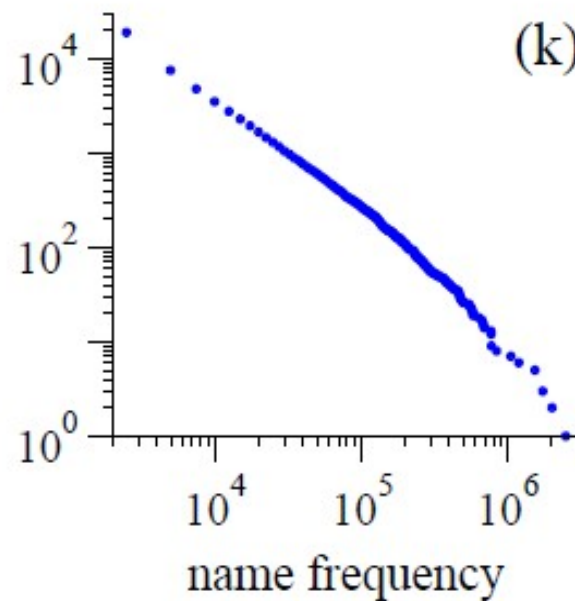
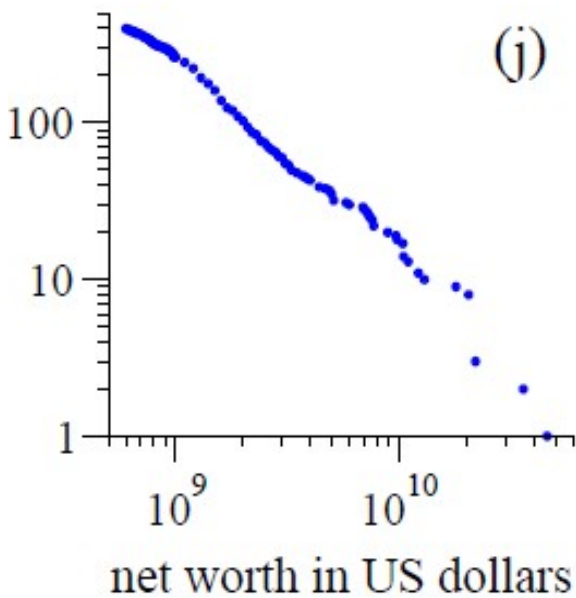
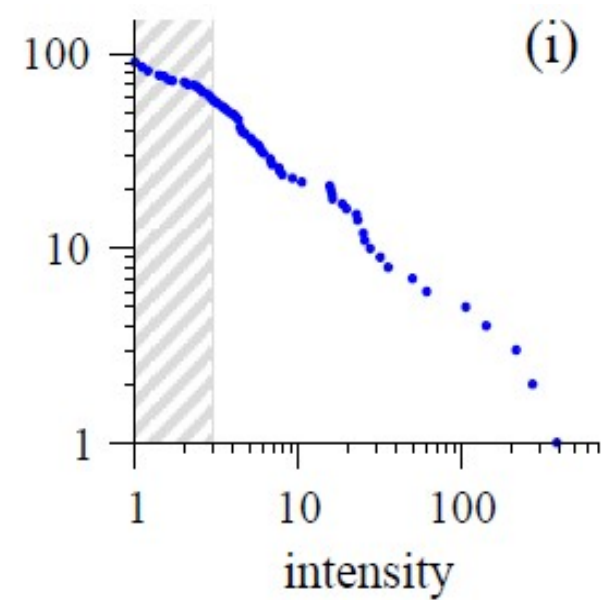
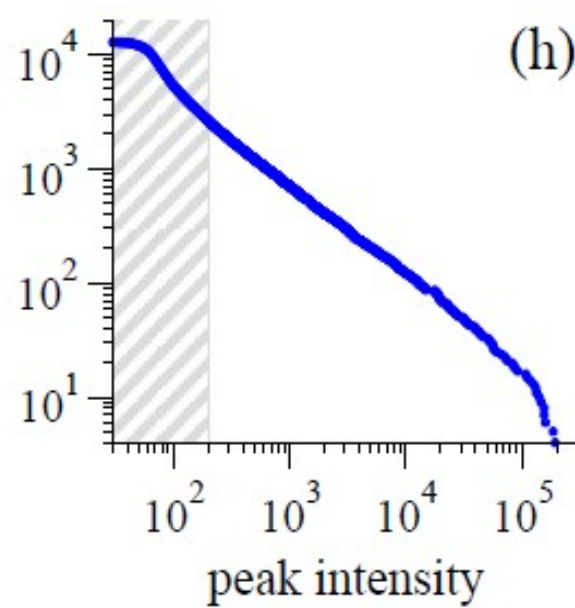
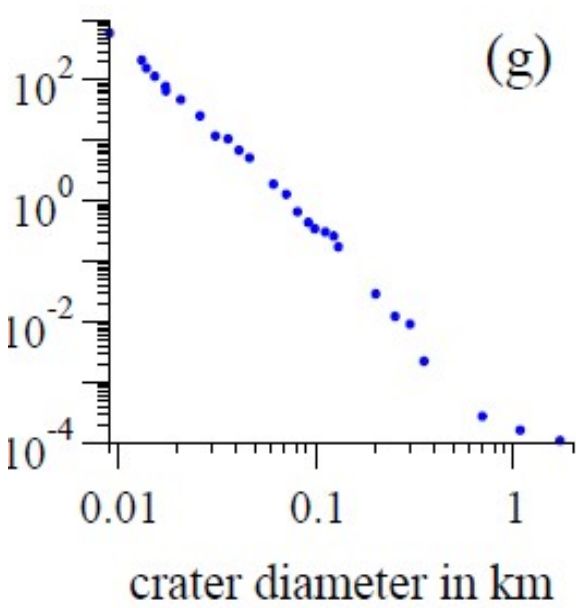
- What should a power law plot look like?
 - $f(k)$: the fraction of items that have value k
 - If power law holds, $f(k) = a/k^c$?
 - for some constant c and a .
 - $f(k) = a/k^c = ak^{-c}$
 - $\log f(k) = \log a - c \log k$
 - **straight line!** “ $\log f(k)$ ” as a function of “ $\log k$ ”
 - “ c ”: slope, and
 - “ $\log a$ ”: y-intercept.
 - log-log plot!

Power Law- Cnt.

- If power-law holds, the “log -log” plot should be a **straight line**.







Popularity

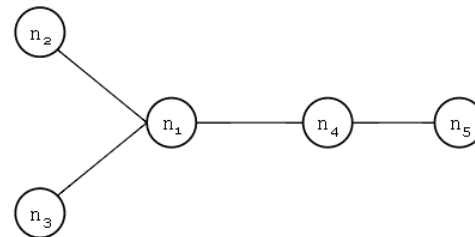
- Let's focus on the Web in which we can measure popularity accurately!
 - Popularity of a page

Popularity- Cnt.

- Let's focus on the Web in which we can measure popularity accurately!
 - Popularity of a page ~ number of its **in-links**
 - Easy to count!

Degree Centrality- Cnt.

- A node is central if it has ties to many other nodes
 - Look at the node degree



$$C(n_1) = \sum_{j=1}^n A_{1j} = \sum_{i=1}^n A_{i1} = 3$$

	n1	n2	n3	n4	n5	$\sum_{j=1}^n A_{ij}$
n1	0	1	1	1	0	3
n2	1	0	0	0	0	1
n3	1	0	0	0	0	1
n4	1	0	0	0	1	2
n5	0	0	0	1	0	1
$\sum_{i=1}^n A_{ij}$	3	1	1	2	1	

Adjacency Matrix (A)

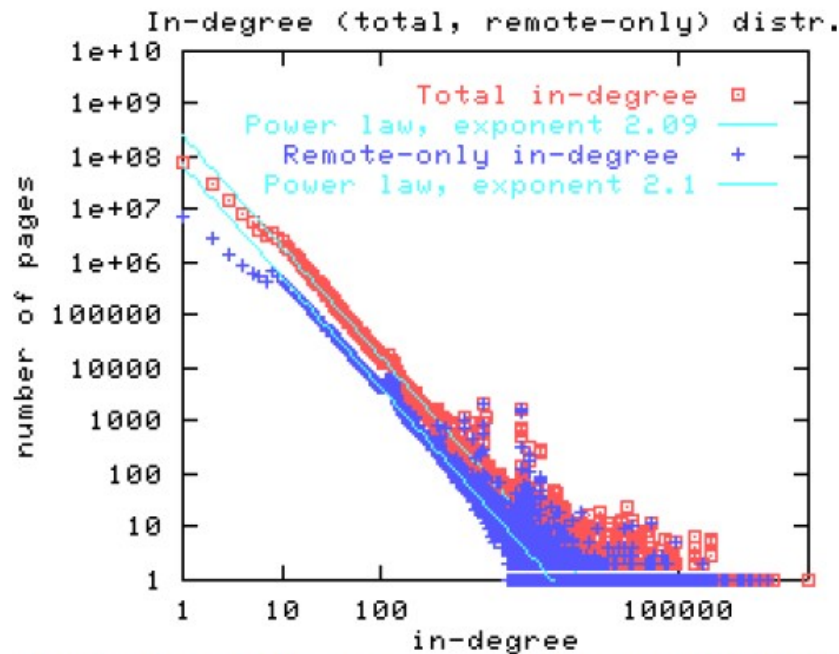
Popularity- Cnt.

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

Popularity- Cnt.

- Question:

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



Remote-only: older crawl

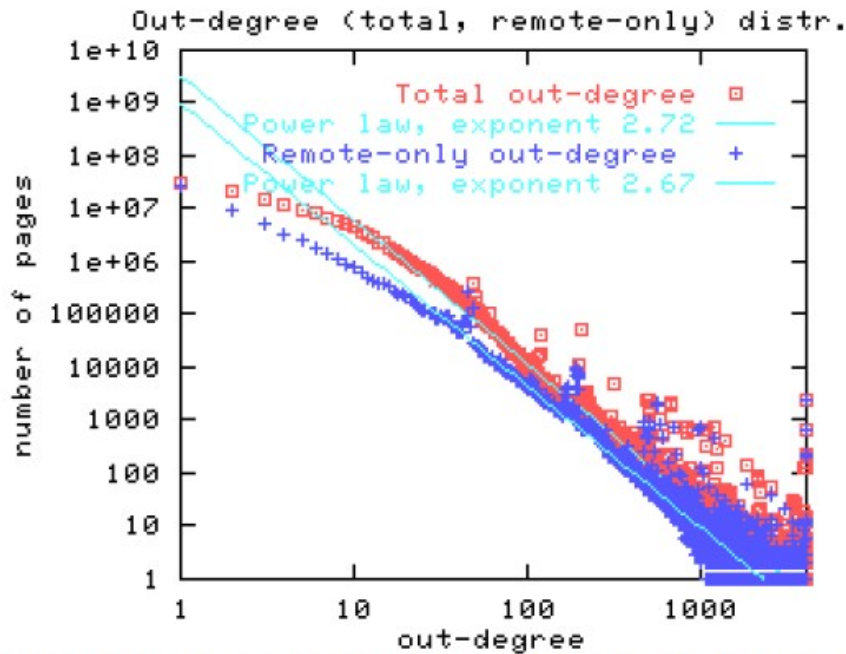
- $c \approx 2.1$
- Straight lines are linear regressions for the best power law fit.
- The anomalous bump at 120 on the x-axis is due to a large *clique** formed by a single spammer.

* Subset of nodes such that every two distinct nodes are adjacent.

Popularity- Cnt.

- Question:

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

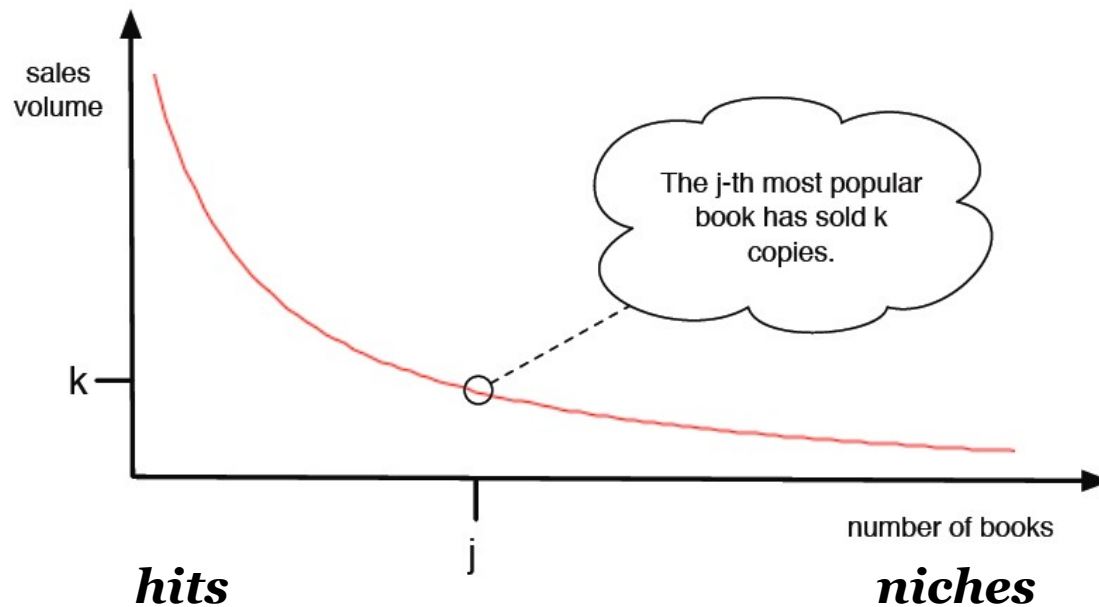


Remote-only: older crawl

- $c \approx 2.7$
- Initial segment of the out-degree distribution deviates significantly from the power law:
 - pages with low out-degree follow a different distribution.

Popularity- The Long Tail

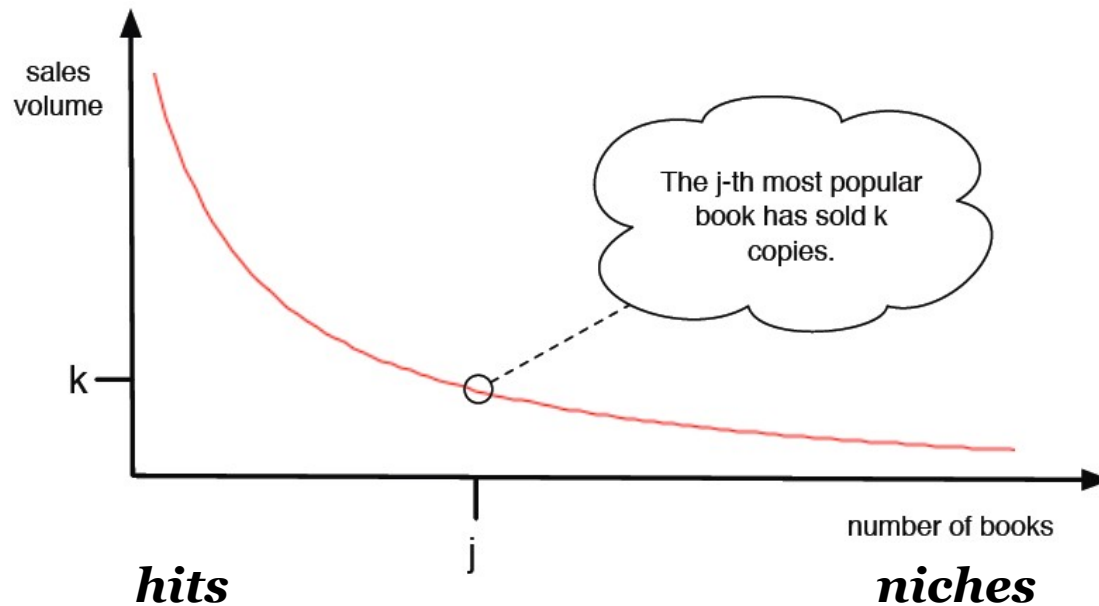
- **Question:** Are most sales generated by a
 - **small set** of **popular items** (*hits*), or
 - **large set** of **less popular items** (*niches*)?



Check if this curve is changing shape over time, adding more area under the right at the expense of the left!

Popularity- The Long Tail

- **Question:** Would personalization be useful?
 - E.g. through exposing people to items that (may not be popular but) match with their interests!



Popularity- Cause

- What is causing Power laws / Popularity?

Rich Get Richer (RGR)

Rich-Get-Richer: A simple model for the creation of links as a basis for power laws!

1. Pages are created in order and named 1, 2, ..., N.
2. When page j is created, it produces a link to an **earlier page $i < j$** according to the following rules:
 - a) With probability p , page j chooses page i uniformly at random, and creates **a link to i** .
 - b) With probability $(1 - p)$, page j chooses page i uniformly at random and creates **a link to the page that i points to** (copies decision made by i).
- Let's assume that each page creates just 1 link
 - We can extend this model to multiple links as well.

RGR - Power Law

- We observe power law, if we run this model for many pages
 - the fraction of pages with k in-links will be distributed according to a power law $1/k^c$!
 - Value of the exponent c depends on the choice of p .
- Correlation between c and p ?

RGR - Power Law

- We observe power law, if we run this model for many pages
 - the fraction of pages with k in-links will be distributed according to a power law $1/k^c$!
 - Value of the exponent c depends on the choice of p .
- Correlation between c and p ?
 - Smaller p
 - Copying becomes more frequent -> more likely to see extremely popular pages ->
 - c gets smaller as well

RGR - Preferential Attachment

- Due to copying mechanism: the probability of linking to a page is proportional to the total number of pages that currently link to that page!
- Preferential Attachment: restating rule 2 (b):
 - **b)** With probability $(1-p)$, page j chooses page i with probability **proportional to i 's current number of in-links** and creates a link to i .
 - links are formed “preferentially” to pages that already have high popularity.

RGR - Preferential Attachment

Rich-Get-Richer:

1. Pages are created in order and named $1, 2, \dots, N$.
2. When page j is created, it produces a link to an **earlier page $i < j$** according to the following rules:
 - a) With probability p , page j chooses page i uniformly at random and creates **a link to i** .
 - b) With probability $(1 - p)$, page j chooses page i with probability **proportional to i 's current number of in-links** and creates a link to i .

RGR - Probabilistic Model

- Probabilistic model
 - $X_j(t)$: number of in-links to node j at a time t
- Two points about $X_j(t)$
 1. Value of $X_j(t)$ at time $t=j$
 - $X_j(j) = 0$
 - node j starts with 0 in-link when it's first created at time j !
 2. Expected Change to $X_j(.)$ over time

Compute the probability that node j gains an in-link in step $t+1$?

RGR - Probabilistic Model

- Expected Change to $X_j(\cdot)$ over time
 - Probability that node j gains an in-link in step $t+1$?

RGR - Probabilistic Model

2. Expected Change to $X_j(\cdot)$ over time

- Probability that node j gains an in-link in step $t+1$?
 - Happens if the newly created node $t+1$ points to node j .
 - Two cases:
 1. With probability p , node $t+1$ links to an earlier node chosen uniformly at random:
 - Thus, node $t + 1$ links to node j with probability $1/t$
 2. With probability $1 - p$, node $t+1$ links to an earlier node with probability proportional to the node's current number of in-links.
 - At time $t+1$:
 - total number of links in the network?
 - t (one out of each prior node)
 - How many of them point to node j ?
 - $X_j(t)$ (based on the definition)
 - Thus, node $t + 1$ links to node j with probability $X_j(t)/t$.

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

RGR - Probabilistic Model

- Deterministic approximation
 - Approximate $X_j(t)$ —the # of in-links of node j —by a continuous function of time $x_j(t)$.
 - Model for rate of growth:

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

$$\frac{dx_j}{dt} = \frac{p}{t} + \frac{(1-p)x_j}{t}. \quad \longrightarrow \quad x_j(t) = \frac{p}{q} \left[\left(\frac{t}{j} \right)^q - 1 \right].$$

RGR - Probabilistic Model

- Identifying power law in DA $x_j(t) = \frac{p}{q} \left[\left(\frac{t}{j} \right)^q - 1 \right]$.
 - For a given value of k and time t , what fraction of nodes have at least k in-links at t , OR
 - For a given value of k and time t , what fraction of all j s satisfy $x_j(t) \geq k$?

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

Power law:

The fraction of x_j that are at least k is proportional to $k^{-1/q}$.

RGR - Probabilistic Model

- Explain power laws using the Rich-Get-Richer model:
 - Fraction of numbers receiving k calls per day: $1/k^2$
 - Fraction of books bought by k people: $1/k^3$
 - Fraction of papers with k citations: $1/k^3$
 - Fraction of cities with population k : $1/k^c$
 - Cities grow in proportion to their size, simply as a result of people having children!
- Once an item becomes popular, the rich-get-richer dynamics are likely to push it even higher!

RGR - Unpredictability

- If we replay the history:
 - Do you think the most popular items will remain the same as they are now?

- Do we observe power law?

RGR - Unpredictability

- If we replay the history:
 - Do you think the most popular items will remain the same as they are now?
 - Less likely
 - Random effects early in the process play a role in the future popularity.
 - Do we observe power law?
 - Power-law distribution of popularity would probably exist in each replay!

How to properly investigate unpredictability in the contents of RGR?

RGR - Unpredictability

- Music download site
 - 48 obscure songs/bands.
 - >14K visitors
 - can participate only once and can't share opinions.
 - Visitors/subjects could listen and download songs
 - “download count” for each song is shown to visitors.
 - the number of times it had been downloaded thus far.
 - Parallel World - two settings:
 1. Visitors upon arrival were being assigned at random to one of 8 “parallel” copies of the site.
 2. Visitors upon arrival were being assigned to a copy of the site in which “download counts” info was removed.

RGR - Unpredictability

- Music download site

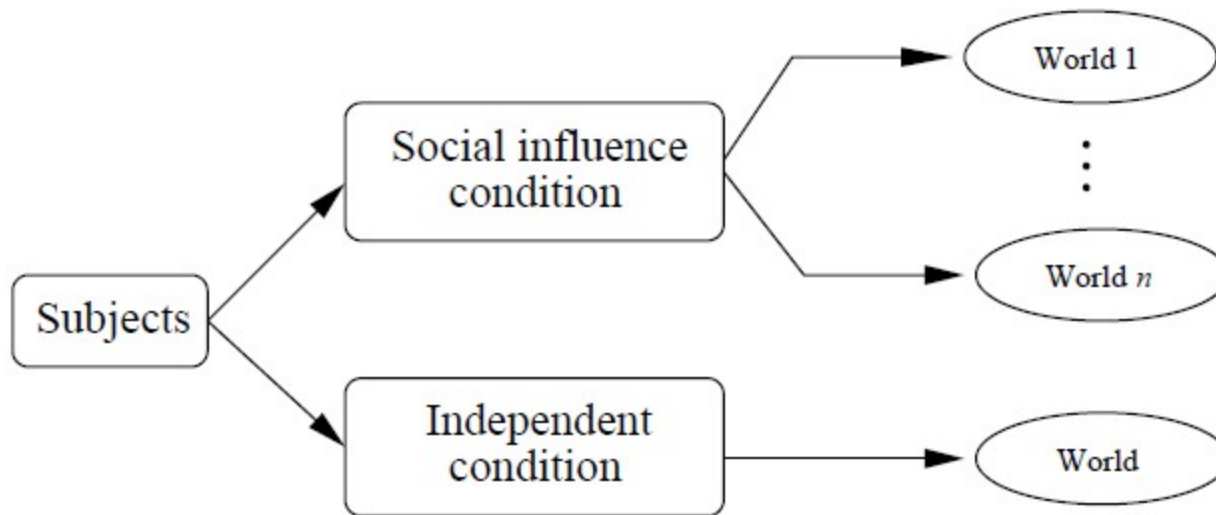


Figure S1: Schematic of the experimental design.

RGR - Unpredictability

Experiment 1

- Social Influence:
 - Each visitor was given information only about the behavior of others in its copy of the site!
 - Opportunity to contribute to RGR dynamics!
 - **Songs presented in grid & were not ordered by download counts!**
 - The parallel copies started out identically
 - same songs, download counts for all songs set to zero.
- Independent:
 - No direct contribution to RGR dynamics!
 - **Songs presented in grid & in random order.**

RGR - Unpredictability

Music Lab – Song Selection - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.musiclab.columbia.edu/me/songs

	# of down loads	[Help] [Log off]	# of down loads	# of down loads	
HARTSFIELD: "enough is enough"	20	GO MORECAI: "it does what its told"	12	UNDO: "while the world passes"	24
DEEP ENOUGH TO DIE: "for the sky"	17	PARKER THEORY: "she said"	47	UP FOR NOTHING: "in sight of"	13
THE THRIFT SYNDICATE: "2003 a tragedy"	20	MISS OCTOBER: "pink aggression"	27	SILVERFOX: "gnaw"	17
THE BROKEN PROMISE: "the end in friend"	19	POST BREAK TRAGEDY: "florence"	14	STRANGER: "one drop"	10
THIS NEW DAWN: "the belief above the answer"	12	FORTHFADING: "fear"	24	FAR FROM KNOWN: "route 9"	18
NOONER AT NINE: "walk away"	6	THE CALEFACTION: "trapped in an orange peel"	20	STUNT MONKEY: "inside out"	46
MORAL HAZARD: "waste of my life"	8	S2METRO: "lockdown"	17	DANTE: "lifes mystery"	14
NOT FOR SCHOLARS: "as seasons change"	27	SIMPLY WAITING: "went with the count"	16	FADING THROUGH: "wish me luck"	10
SECRETARY: "keep your eyes on the ballistics"	5	STAR CLIMBER: "tell me"	38	UNKNOWN CITIZENS: "falling over"	34
ART OF KANLY: "seductive intro, melodic breakdown"	10	THE FASTLANE: "til death do us part (i dont)"	31	BY NOVEMBER: "if i could take you"	20
HYDRAULIC SANDWICH: "separation anxiety"	20	A BLINDING SILENCE: "miseries and miracles"	17	DRAWN IN THE SKY: "tap the ride"	12
EMBER SKY: "this upcoming winter"	25	SUM RANA: "the bolshevik boogie"	15	SELSIUS: "stars of the city"	22
SALUTE THE DAWN: "i am error"	13	CAPE RENEWAL: "baseball warlock v1"	12	SIBRIAN: "eye patch"	14
RYAN ESSMAKER: "detour_(be still)"	14	UP FALLS DOWN: "a brighter burning star"	11	EVAN GOLD: "robert downey jr"	10
BEERBONG: "father to son"	12	SUMMERSWASTED: "a plan behind destruction"	17	BENEFIT OF A DOUBT: "run away"	38
HALL OF FAME: "best mistakes"	19	SILENT FILM: "all i have to say"	61	SHIPWRECK UNION: "out of the woods"	16

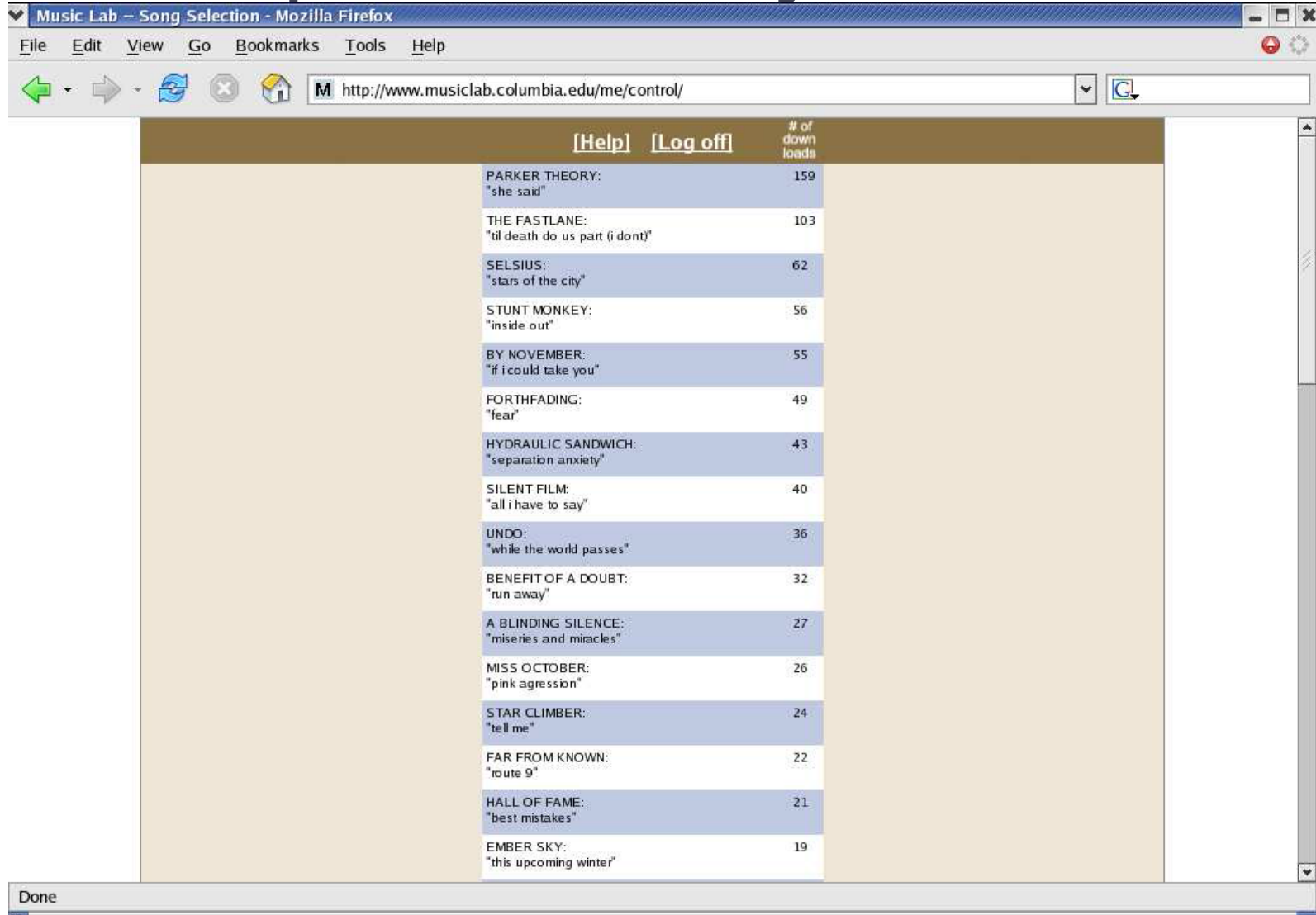
Subjects could participate only once and could not share opinions.

RGR - Unpredictability

Experiment 2

- Social Influence:
 - Each visitor was given information only about the behavior of others in its copy of the site!
 - Opportunity to contribute to RGR dynamics!
 - **Songs presented in one column & in descending order of download counts!**
 - The parallel copies started out identically
 - same songs, download counts for all songs set to zero.
- Independent:
 - No direct contribution to RGR dynamics!
 - **Songs presented in one column & random order.**

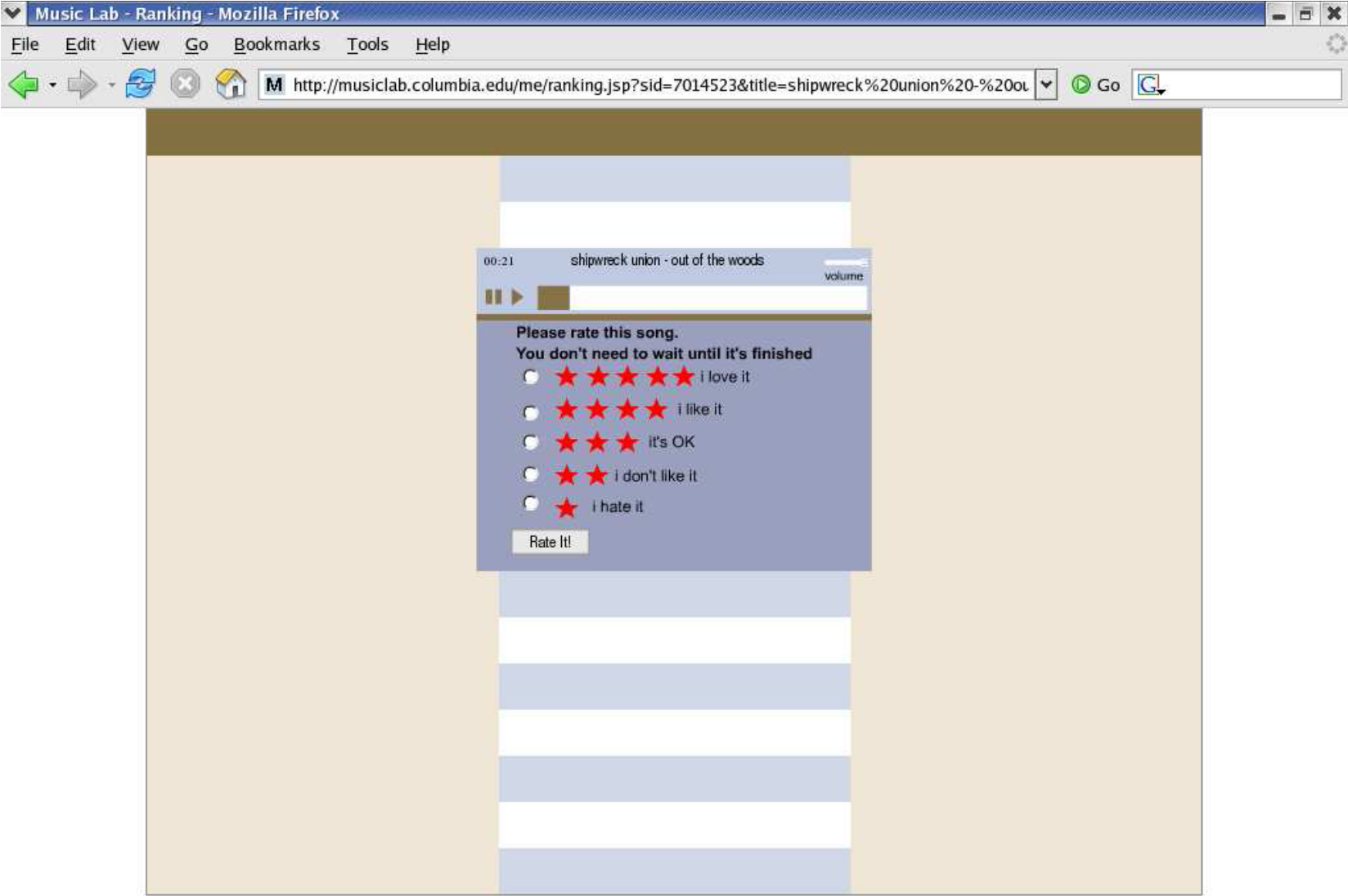
RGR - Unpredictability



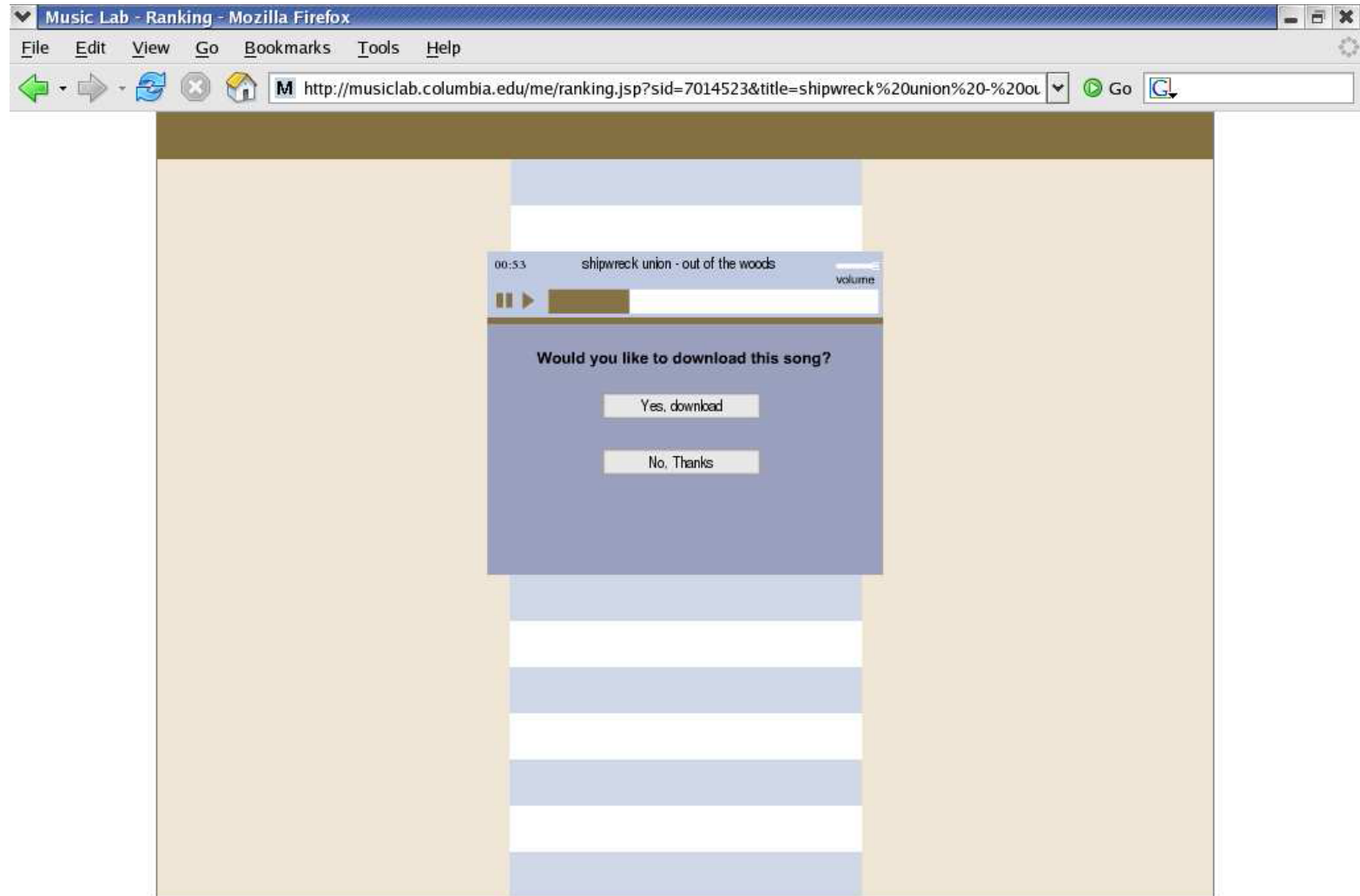
	[Help]	[Log off]	# of down loads
PARKER THEORY: "she said"			159
THE FASTLANE: "til death do us part (i dont)"			103
SELSIUS: "stars of the city"			62
STUNT MONKEY: "inside out"			56
BY NOVEMBER: "if i could take you"			55
FORTHFADING: "fear"			49
HYDRAULIC SANDWICH: "separation anxiety"			43
SILENT FILM: "all i have to say"			40
UNDO: "while the world passes"			36
BENEFIT OF A DOUBT: "run away"			32
A BLINDING SILENCE: "miseres and miracles"			27
MISS OCTOBER: "pink agression"			26
STAR CLIMBER: "tell me"			24
FAR FROM KNOWN: "route 9"			22
HALL OF FAME: "best mistakes"			21
EMBER SKY: "this upcoming winter"			19

Subjects could participate only once and could not share opinions.

RGR - Unpredictability



RGR - Unpredictability



RGR - Unpredictability

- Music Unknownness!

	How familiar are you with the following bands?		
	Don't know it at all (% of subjects)	Heard of it (% of subjects)	Know it pretty well (% of subjects)
Real Bands			
GUYS ON COUCH	87.9	11.0	1.1
GROVER DILL	88.4	10.5	1.1
REMNANT SOLDIER	77.2	19.9	2.9
Fake Band			
PETER ON FIRE	84.5	13.7	1.8

Table S4: Comparing the popularity of the potential bands from our sample to a fake band. Subjects reported being about as familiar with an fake band (Peter on Fire) as three potential bands from our sample. The high recognition rate for Remnant Soldier is likely a question ordering effect — it was asked immediately after the well known band U2.

These results, along with screening, led authors believe that the music used in the experiment was essentially unknown.

RGR - Unpredictability

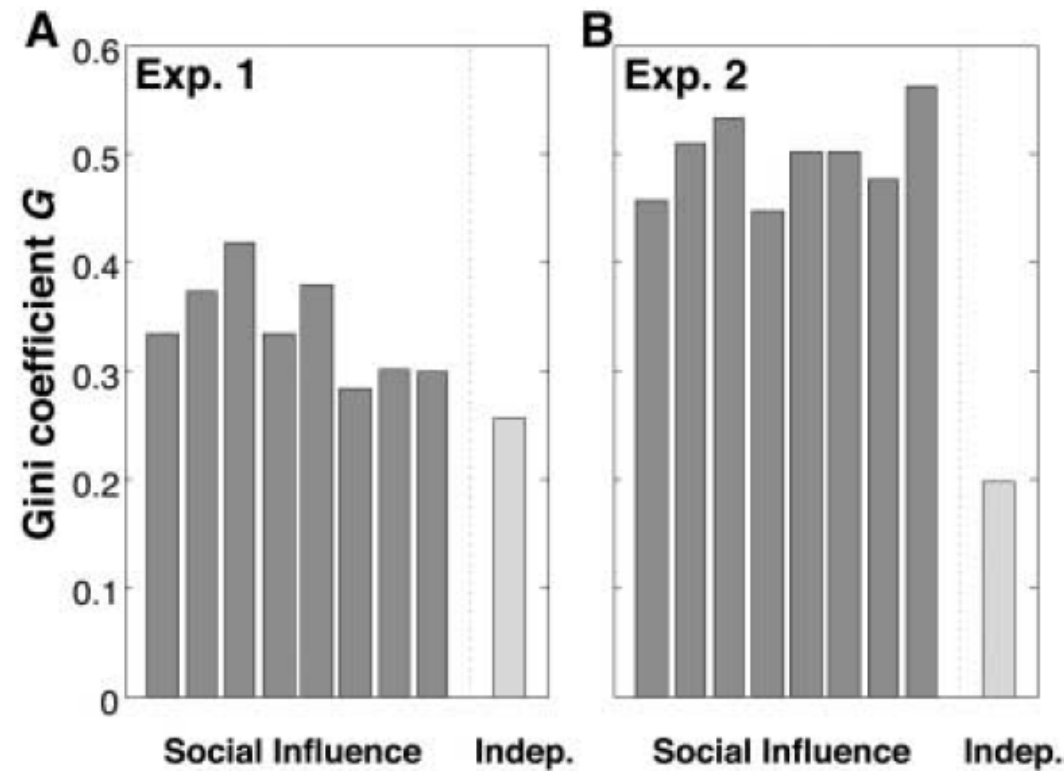


Fig. 1. Inequality of success for social influence (dark bars) and independent (light bars) worlds for **(A)** experiment 1 and **(B)** experiment 2. The success of a song is defined by m_i , its market share of downloads ($m_i = d_i / \sum_{k=1}^S d_k$, where d_i is song i 's download count and S is the number of songs). Success inequality is defined by the Gini coefficient $G = \frac{\sum_{i=1}^S \sum_{j=1}^S |m_i - m_j|}{2S \sum_{k=1}^S m_k}$, which represents the average difference in market share for two songs normalized to fall between 0 (complete equality) and 1 (maximum inequality). Differences between independent and social influence conditions are significant ($P < 0.01$) (18).

1. The social influence worlds exhibit greater inequality—popular songs are more popular and unpopular songs are less popular—than the independent world.
2. Inequality increased from experiment 1 to experiment 2: not only that social influence contributes to inequality, but as individuals are subject to stronger forms of social influence, the collective outcomes will become increasingly unequal.

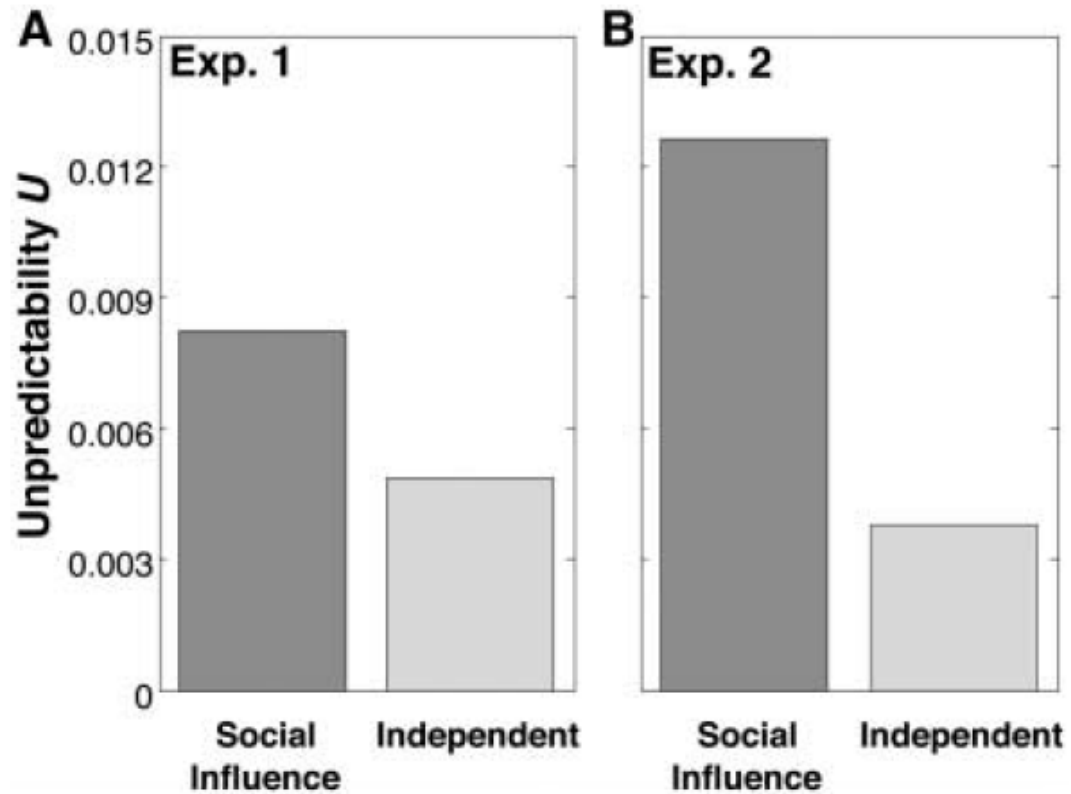
RGR - Unpredictability

Fig. 2. Unpredictability of success for (A) experiment 1 and (B) experiment 2. In both experiments, success in the social influence condition was more unpredictable than in the independent condition. Moreover, the stronger social signal in experiment 2 leads to increased unpredictability. The measure of unpredictability u_i for a single song i is defined as the average difference in market share for that song between all pairs of realizations; i.e.,

$$u_i = \frac{1}{\binom{W}{2}} \sum_{j=1}^W \sum_{k=j+1}^W |m_{i,j} - m_{i,k}|$$

where $m_{i,j}$ is song i 's market share in world j and W is the number of worlds. The overall unpredictability measure $U = \frac{1}{S} \sum_{i=1}^S u_i$ is then the

average of this measure over all S songs. For the independent condition, we randomly split the single world into two subpopulations to obtain differences in market shares, and we then averaged the results over 1000 of these splits. All differences are significant ($P < 0.01$) (18).



- the average difference in market share for a song between distinct social influence worlds is higher than it is between different subpopulations of individuals making independent decision

RGR - Unpredictability

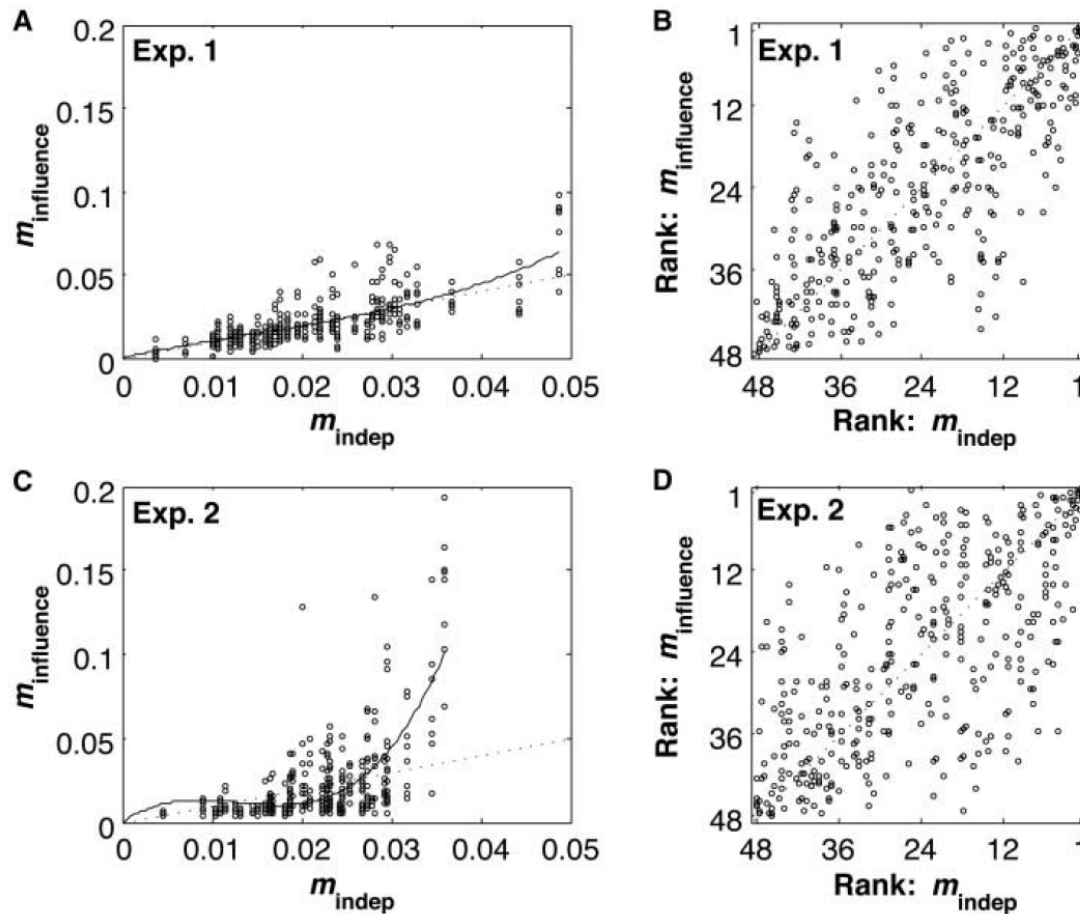


Fig. 3. Relationship between quality and success. (A) and (C) show the relationship between m_{indep} , the market share in the one independent world (i.e., quality), and $m_{influence}$, the market share in the eight social influence worlds (i.e., success). The dotted lines correspond to quality equaling success. The solid lines are third-degree polynomial fits to the data, which suggest that the relationship between quality and success has greater convexity in experiment 2 than in experiment 1. (B) and (D) present the corresponding market rank data.

- On average, quality is positively related to success.
- Songs of any given quality can experience a wide range of success.
- The best songs never do very badly, and the worst songs never do extremely well, but almost any other result is possible.
- Unpredictability also varies with quality, the best songs are the most unpredictable, whereas when measured in terms of rank, intermediate songs are the most unpredictable.

Reading

- Ch.18 Power Laws and Rich-Get-Richer Phenomena [NCM]
- Experimental study of inequality and unpredictability in an artificial cultural market. Salganik et. al. Science'06.