More Mechanisms for Generating Power-Law Size Distributions II

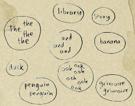
Principles of Complex Systems
CSYS/MATH 300, Spring, 2013 | #SpringPoCS2013

Prof. Peter Dodds
@peterdodds

Department of Mathematics & Statistics | Center for Complex Systems | Vermont Advanced Computing Center | University of Vermont

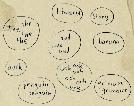
Outline

Growth Mechanisms
Random Copying
Words, Cities, and the Web
Optimization
Minimal Cost
Mandelbrot vs. Simon
Assumptions
Model
Analysis
Extra
And the winner is...?
References



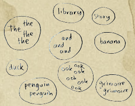
Aggregation:

- Random walks represent additive aggregation
- Mechanism: Random addition and subtraction
- Compare across realizations, no competition.
- Next: Random Additive/Copying Processes involving Competition.
- Widespread: Words, Cities, the Web, Wealth, Productivity (Lotka), Popularity (Books, People, ...)
- Competing mechanisms (trickiness)



Work of Yore:

- 1924: G. Udny Yule [24]:
  # Species per Genus
- 1926: Lotka [10]:
  # Scientific papers per author (Lotka's law)
- 1953: Mandelbrot [12]:
  Optimality argument for Zipf's law; focus on language.
- 1955: Herbert Simon [20, 26]:
  Zipf's law for word frequency, city size, income, publications, and species per genus.
- 1965/1976: Derek de Solla Price [18, 19]:
  Network of Scientific Citations.
- 1999: Barabasi and Albert [1]:
  The World Wide Web, networks-at-large.



Examples:

Recent evidence for Zipf's law...

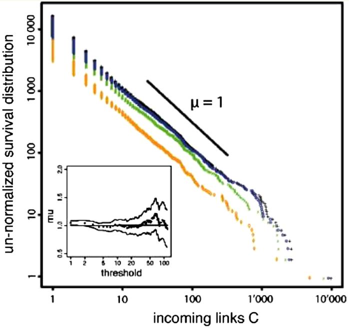
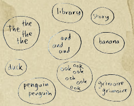


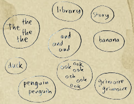
FIG. 1 (color online). (Color Online) Log-log plot of the number of packages in four Debian Linux Distributions with more than C in-directed links. The four Debian Linux Distributions are Woody (19.07.2002) (orange diamonds), Sarge (06.06.2005) (green crosses), Etch (15.08.2007) (blue circles), Lenny (15.12.2007) (black+'s). The inset shows the maximum likelihood estimate (MLE) of the exponent μ together with two boundaries defining its 95% confidence interval (approximately given by $1 \pm 2/\sqrt{n}$, where n is the number of data points using in the MLE), as a function of the lower threshold. The MLE has been modified from the standard Hill estimator to take into account the discreteness of C.

Maillart et al., PRL, 2008:
"Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution" [11]



Herbert Simon (⊞) (1916–2001):

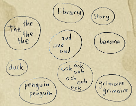
- Political scientist
- Involved in Cognitive Psychology, Computer Science, Public Administration, Economics, Management, Sociology
- Coined ‘bounded rationality’ and ‘satisficing’
- Nearly 1000 publications
- An early leader in Artificial Intelligence, Information Processing, Decision-Making, Problem-Solving, Attention Economics, Organization Theory, Complex Systems, And Computer Simulation Of Scientific Discovery.
- Nobel Laureate in Economics



Essential Extract of a Growth Model:

Random Competitive Replication (RCR):

1. Start with 1 elephant (or element) of a particular flavor at $t = 1$
2. At time $t = 2, 3, 4, \ldots$, add a new elephant in one of two ways:
   - With probability ρ, create a new elephant with a new flavor
     = Mutation/Innovation
   - With probability $1 - \rho$, randomly choose from all existing elephants, and make a copy.
     = Replication/Imitation
   - Elephants of the same flavor form a group

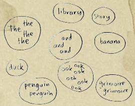


Random Competitive Replication:

Example: Words appearing in a language

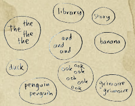
- Consider words as they appear sequentially.
- With probability ρ, the next word has not previously appeared
  = Mutation/Innovation
- With probability $1 - \rho$, randomly choose one word from all words that have come before, and reuse this word
  = Replication/Imitation

Note: This is a terrible way to write a novel.



For example:

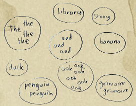

library
Story
The the the the
and and and
banana
duck
ook ook ook ook ook ook
penguin penguin
grimoire grimoire
• 21 words used
• next word is new with prob ρ
• next word is a copy with prob 1−ρ
prob: next word:
6/21 ook
4/21 the
3/21 and
2/21 penguin
1/21 library



Random Competitive Replication:

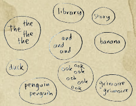
Some observations:

- Fundamental Rich-get-Richer story;
- Competition for replication between individual elephants is random;
- Competition for growth between groups of matching elephants is not random;
- Selection on groups is biased by size;
- Random selection sounds easy;
- Possible that no great knowledge of system needed (but more later ...).



Random Competitive Replication:

- Steady growth of system: +1 elephant per unit time.
- Steady growth of distinct flavors at rate ρ
- We can incorporate
  1. Elephant elimination
  2. Elephants moving between groups
  3. Variable innovation rate ρ
  4. Different selection based on group size
     (But mechanism for selection is not as simple...)



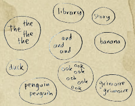
Random Competitive Replication:

Definitions:

- k_i = size of a group i
- $N_k(t)$ = # groups containing k elephants at time t.

Basic question: How does $N_k(t)$ evolve with time?

First: $\sum_k kN_k(t) = t =$ number of elephants at time t

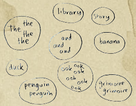


Random Competitive Replication:

$P_k(t)$ = Probability of choosing an elephant that belongs to a group of size k:

- $N_k(t)$ size k groups
- $\Rightarrow kN_k(t)$ elephants in size k groups
- t elephants overall

$$P_k(t) = \frac{kN_k(t)}{t}$$

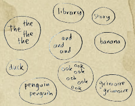


Random Competitive Replication:

$N_k(t)$, the number of groups with k elephants, changes at time t if

1. An elephant belonging to a group with k elephants is replicated
   $N_k(t+1) = N_k(t) - 1$
   Happens with probability $(1-\rho)kN_k(t)/t$

2. An elephant belonging to a group with $k-1$ elephants is replicated
   $N_k(t+1) = N_k(t) + 1$
   Happens with probability $(1-\rho)(k-1)N_{k-1}(t)/t$

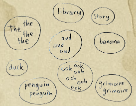


Random Competitive Replication:

Special case for $N_1(t)$:

1. The new elephant is a new flavor:
   $N_1(t+1) = N_1(t) + 1$
   Happens with probability ρ

2. A unique elephant is replicated.
   $N_1(t+1) = N_1(t) - 1$
   Happens with probability $(1-\rho)N_1/t$



Random Competitive Replication:

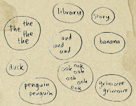
Put everything together:

For $k > 1$:

$$\langle N_k(t+1) - N_k(t) \rangle = (1-\rho)\left((k-1)\frac{N_{k-1}(t)}{t} - k\frac{N_k(t)}{t}\right)$$

For $k = 1$:

$$\langle N_1(t+1) - N_1(t) \rangle = \rho - (1-\rho) 1 \cdot \frac{N_1(t)}{t}$$

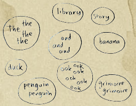


Random Competitive Replication:

Assume distribution stabilizes: $N_k(t) = n_k t$

(Reasonable for t large)

- Drop expectations
- Numbers of elephants now fractional
- Okay over large time scales
- n_k / ρ = the fraction of groups that have size k.



Random Competitive Replication:

Stochastic difference equation:

$$\langle N_k(t+1) - N_k(t) \rangle = (1-\rho)\left((k-1)\frac{N_{k-1}(t)}{t} - k\frac{N_k(t)}{t}\right)$$

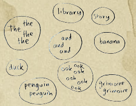
becomes

$$n_k(t+1) - n_k t = (1-\rho)\left((k-1)\frac{n_{k-1}t}{t} - k\frac{n_k t}{t}\right)$$

$$n_k(\cancel{t}+1-\cancel{t}) = (1-\rho)\left((k-1)\frac{n_{k-1}\cancel{t}}{\cancel{t}} - k\frac{n_k \cancel{t}}{\cancel{t}}\right)$$

$$\Rightarrow n_k = (1-\rho)\left((k-1)n_{k-1} - kn_k\right)$$

$$\Rightarrow n_k\left(1 + (1-\rho)k\right) = (1-\rho)(k-1)n_{k-1}$$



Random Competitive Replication:

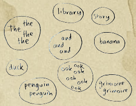
We have a simple recursion:

$$\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1+(1-\rho)k}$$

- Interested in k large (the tail of the distribution)
- Can be solved exactly.

  Insert question from assignment 3 (⊞)
- To get at tail: Expand as a series of powers of $1/k$

  Insert question from assignment 3 (⊞)



Random Competitive Replication:

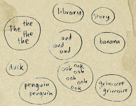
- We (okay, you) find

$$\frac{n_k}{n_{k-1}} \simeq (1 - \frac{1}{k})^{\frac{(2-\rho)}{(1-\rho)}}$$

- $$\frac{n_k}{n_{k-1}} \simeq \left(\frac{k-1}{k}\right)^{\frac{(2-\rho)}{(1-\rho)}}$$

- $$n_k \propto k^{-\frac{(2-\rho)}{(1-\rho)}} = k^{-\gamma}$$

$$\boxed{\gamma = \frac{(2-\rho)}{(1-\rho)} = 1 + \frac{1}{(1-\rho)}}$$



- Micro-to-Macro story with ρ and γ measurable.

$$\boxed{\gamma = \frac{(2-\rho)}{(1-\rho)} = 1 + \frac{1}{(1-\rho)}}$$

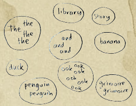
- Observe $2 < \gamma < \infty$ for $0 < \rho < 1$.
- For $\rho \simeq 0$ (low innovation rate):

$$\gamma \simeq 2$$

- 'Wild' power-law size distribution of group sizes, bordering on 'infinite' mean.
- For $\rho \simeq 1$ (high innovation rate):

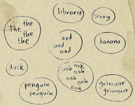
$$\gamma \simeq \infty$$

- All elephants have different flavors.
- Upshot: Tunable mechanism producing a family of universality classes.



Random Competitive Replication:

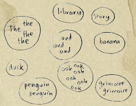
- Recall Zipf's law: $s_r \sim r^{-\alpha}$
  (s_r = size of the rth largest elephant)
- We found $\alpha = 1/(\gamma - 1)$
- $\gamma = 2$ corresponds to $\alpha = 1$
- We (roughly) see Zipfian exponent [26] of $\alpha = 1$ for many real systems: city sizes, word distributions, ...
- Corresponds to $\rho \to 0$, low innovation.
- Krugman doesn't like it) [9] but it's all good.
- Still, other quite different mechanisms are possible...
- Must look at the details to see if mechanism makes sense... more later.



What about small k?:

We had one other equation:

- $$\langle N_1(t+1) - N_1(t) \rangle = \rho - (1-\rho) 1 \cdot \frac{N_1(t)}{t}$$
- As before, set $N_1(t) = n_1 t$ and drop expectations
- $$n_1(t+1) - n_1 t = \rho - (1-\rho) 1 \cdot \frac{n_1 t}{t}$$
- $$n_1 = \rho - (1-\rho) n_1$$
- Rearrange:
$$n_1 + (1-\rho) n_1 = \rho$$
- $$\boxed{n_1 = \frac{\rho}{2-\rho}}$$

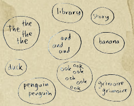


So... $$N_1(t) = n_1 t = \frac{\rho t}{2 - \rho}$$

- Recall number of distinct elephants = ρt.
- Fraction of distinct elephants that are unique (belong to groups of size 1):

$$\frac{N_1(t)}{\rho t} = \frac{1}{2 - \rho}$$

  (also = fraction of groups of size 1)
- For ρ small, fraction of unique elephants $\sim 1/2$
- Roughly observed for real distributions
- ρ increases, fraction increases
- Can show fraction of groups with two elephants $\sim 1/6$
- Model does well at both ends of the distribution



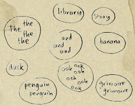
Words:

From Simon [20]:

Estimate $\rho_{\text{est}} = \#$ unique words$/\#$ all words

For Joyce's Ulysses: $\rho_{\text{est}} \simeq 0.115$

N_1 (real)	N_1 (est)	N_2 (real)	N_2 (est)
16,432	15,850	4,776	4,870



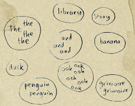
Evolution of catch phrases:

- Yule's paper (1924) [24]:
  "A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S."
- Simon's paper (1955) [20]:
  "On a class of skew distribution functions" (snore)

From Simon's introduction:

It is the purpose of this paper to analyse a class of distribution functions that appear in a wide range of empirical data—particularly data describing sociological, biological and economic phenomena.

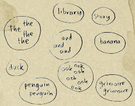
Its appearance is so frequent, and the phenomena so diverse, that one is led to conjecture that if these phenomena have any property in common it can only be a similarity in the structure of the underlying probability mechanisms.



Evolution of catch phrases:

Derek de Solla Price:

- First to study network evolution with these kinds of models.
- Citation network of scientific papers
- Price's term: Cumulative Advantage
- Idea: papers receive new citations with probability proportional to their existing # of citations
- Directed network
- Two (surmountable) problems:
  1. New papers have no citations
  2. Selection mechanism is more complicated

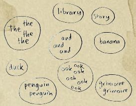


Evolution of catch phrases:

Robert K. Merton: the Matthew Effect (⊞)

- Studied careers of scientists and found credit flowed disproportionately to the already famous

  From the Gospel of Matthew:
  "For to every one that hath shall be given...
  (Wait! There's more....)
  but from him that hath not, that also which he seemeth to have shall be taken away.
  And cast the worthless servant into the outer darkness; there men will weep and gnash their teeth."
- (Hath = suggested unit of purchasing power.)
- Matilda effect: (⊞) women's scientific achievements are often overlooked



Evolution of catch phrases:

Merton was a catchphrase machine:

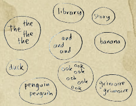
1. Self-fulfilling prophecy
2. Role model
3. Unintended (or unanticipated) consequences
4. Focused interview → focus group

And just to be clear...

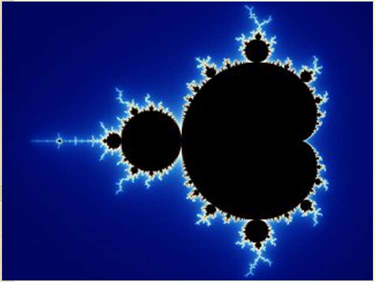
Merton's son, Robert C. Merton, won the Nobel Prize for Economics in 1997.

Evolution of catch phrases:

- Barabasi and Albert [1]—thinking about the Web
- Independent reinvention of a version of Simon and Price's theory for networks
- Another term: "Preferential Attachment"
- Considered undirected networks (not realistic but avoids 0 citation problem)
- Still have selection problem based on size (non-random)
- Solution: Randomly connect to a node (easy) . . .
- . . . and then randomly connect to the node's friends (also easy)
- Scale-free networks = food on the table for physicists



Benoît Mandelbrot (⊞)



Nassim Taleb's tribute:

Benoit Mandelbrot, 1924-2010

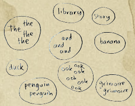
A Greek among Romans

- Mandelbrot = father of fractals
- Mandelbrot = almond bread
- Bonus Mandelbrot set action: here (⊞).

Another approach:

Benoît Mandelbrot

- Derived Zipf's law through optimization [12]
- Idea: Language is efficient
- Communicate as much information as possible for as little cost
- Need measures of information (H) and average cost (C)...
- Language evolves to maximize H/C, the amount of information per average cost.
- Equivalently: minimize C/H.
- Recurring theme: what role does optimization play in complex systems?



The Quickening (⊞)—Mandelbrot v. Simon:

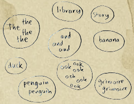
There Can Be Only One: (⊞)

- Things there should be only one of: Theory, Highlander Films.
- Feel free to play Queen's It's a Kind of Magic (⊞) in your head (funding remains tight).

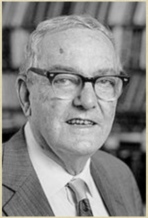
We were born to be Princes of the Universe

Mandelbrot vs. Simon:

- Mandelbrot (1953): “An Informational Theory of the Statistical Structure of Languages” [12]
- Simon (1955): “On a class of skew distribution functions” [20]
- Mandelbrot (1959): “A note on a class of skew distribution functions: analysis and critique of a paper by H.A. Simon” [13]
- Simon (1960): “Some further notes on a class of skew distribution functions” [21]



I have no rival, No man can be my equal



Mandelbrot vs. Simon:

- Mandelbrot (1961): "Final note on a class of skew distribution functions: analysis and critique of a model due to H.A. Simon" [15]
- Simon (1961): "Reply to 'final note' by Benoit Mandelbrot" [23]
- Mandelbrot (1961): "Post scriptum to 'final note'" [15]
- Simon (1961): "Reply to Dr. Mandelbrot's post scriptum" [22]

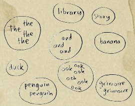
I am immortal, I have inside me blood of kings

Mandelbrot:

"We shall restate in detail our 1959 objections to Simon's 1955 model for the Pareto-Yule-Zipf distribution. Our objections are valid quite irrespectively of the sign of p-1, so that most of Simon's (1960) reply was irrelevant." [14]

Simon:

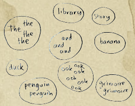
"Dr. Mandelbrot has proposed a new set of objections to my 1955 models of the Yule distribution. Like his earlier objections, these are invalid." [23]



Zipfarama via Optimization:

Mandelbrot's Assumptions:

- Language contains n words: $w_1, w_2, \ldots, w_n$.
- ith word appears with probability p_i
- Words appear randomly according to this distribution (obviously not true...)
- Words = composition of letters is important
- Alphabet contains m letters
- Words are ordered by length (shortest first)



Zipfarama via Optimization:

Word Cost

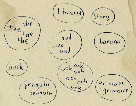
- Length of word (plus a space)
- Word length was irrelevant for Simon's method

Objection

- Real words don't use all letter sequences

Objections to Objection

- Maybe real words roughly follow this pattern (?)
- Words can be encoded this way
- Na na na-na naaaaa...

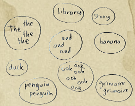


Zipfarama via Optimization:

Binary alphabet plus a space symbol

i	1	2	3	4	5	6	7	8
word	1	10	11	100	101	110	111	1000
length	1	2	2	3	3	3	3	4
$1 + \ln_2 i$	1	2	2.58	3	3.32	3.58	3.81	4

- Word length of 2^kth word: $= k + 1 = 1 + \log_2 2^k$
- Word length of ith word $\simeq 1 + \log_2 i$
- For an alphabet with m letters,
  word length of ith word $\simeq 1 + \log_m i$.



Zipfarama via Optimization:

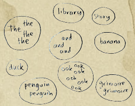
Total Cost C

- Cost of the ith word: $C_i \simeq 1 + \log_m i$
- Cost of the ith word plus space: $C_i \simeq 1 + \log_m(i+1)$
- Subtract fixed cost: $C_i' = C_i - 1 \simeq \log_m(i+1)$
- Simplify base of logarithm:

$$C_i' \simeq \log_m(i+1) = \frac{\log_e(i+1)}{\log_e m} \propto \ln(i+1)$$

- Total Cost:

$$C \sim \sum_{i=1}^{n} p_i C_i' \propto \sum_{i=1}^{n} p_i \ln(i+1)$$



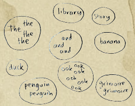
Zipfarama via Optimization:

Information Measure

- Use Shannon's Entropy (or Uncertainty):

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- (allegedly) von Neumann suggested 'entropy'...
- Proportional to average number of bits needed to encode each 'word' based on frequency of occurrence
- $-\log_2 p_i = \log_2 1/p_i$ = minimum number of bits needed to distinguish event i from all others
- If $p_i = 1/2$, need only 1 bit ($log_2 1/p_i = 1$)
- If $p_i = 1/64$, need 6 bits ($log_2 1/p_i = 6$)



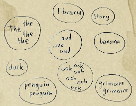
Zipfarama via Optimization:

Information Measure

- Use a slightly simpler form:

$$H = -\sum_{i=1}^{n} p_i \log_e p_i / \log_e 2 = -g \sum_{i=1}^{n} p_i \ln p_i$$

where $g = 1/\ln 2$



Zipfarama via Optimization:

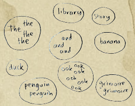
- Minimize

$$F(p_1, p_2, \ldots, p_n) = C/H$$

  subject to constraint

$$\sum_{i=1}^{n} p_i = 1$$

- Tension:
  (1) Shorter words are cheaper
  (2) Longer words are more informative (rarer)



Zipfarama via Optimization:

Time for Lagrange Multipliers:

- Minimize

$$\Psi(p_1, p_2, \ldots, p_n) =$$

$$F(p_1, p_2, \ldots, p_n) + \lambda G(p_1, p_2, \ldots, p_n)$$

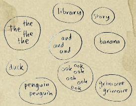
where

$$F(p_1, p_2, \ldots, p_n) = \frac{C}{H} = \frac{\sum_{i=1}^{n} p_i \ln(i+1)}{-g \sum_{i=1}^{n} p_i \ln p_i}$$

and the constraint function is

$$G(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i - 1 = 0$$

Insert question from assignment 4 (⊞)

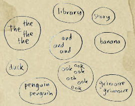


Zipfarama via Optimization:

Some mild suffering leads to:

- $$p_j = e^{-1-\lambda H^2/gC}(j+1)^{-H/gC} \propto (j+1)^{-H/gC}$$
- A power law appears [applause]: $\boxed{\alpha = H/gC}$
- Next: sneakily deduce λ in terms of g, C, and H.
- Find

$$p_j = (j+1)^{-H/gC}$$



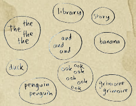
Zipfarama via Optimization:

Finding the exponent

- Now use the normalization constraint:

$$1 = \sum_{j=1}^{n} p_j = \sum_{j=1}^{n} (j+1)^{-H/gC} = \sum_{j=1}^{n} (j+1)^{-\alpha}$$

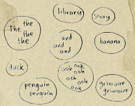
- As $n \to \infty$, we end up with $\zeta(H/gC) = 2$ where ζ is the Riemann Zeta Function
- Gives $\alpha \simeq 1.73$ (> 1, too high)
- If cost function changes ($j + 1 \to j + a$) then exponent is tunable
- Increase a, decrease α



Zipfarama via Optimization:

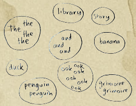
All told:

- Reasonable approach: Optimization is at work in evolutionary processes
- But optimization can involve many incommensurate elephants: monetary cost, robustness, happiness,...
- Mandelbrot's argument is not super convincing
- Exponent depends too much on a loose definition of cost



From the discussion at the end of Mandelbrot's paper:

- A. S. C. Ross: "M. Mandelbrot states that 'the actual direction of evolution (sc. of language) is, in fact, towards fuller and fuller utilization of places'. We are, in fact, completely without evidence as to the existence of any 'direction of evolution' in language, and it is axiomatic that we shall remain so. Many philologists would deny that a 'direction of evolution' could be theoretically possible; thus I myself take the view that a language develops in what is essentially a purely random manner."
- Mandelbrot: "As to the 'fundamental linguistic units being the least possible differences between pairs of utterances' this is a logical consequence of the fact that two is the least integer greater than one."



More:

Reconciling Mandelbrot and Simon

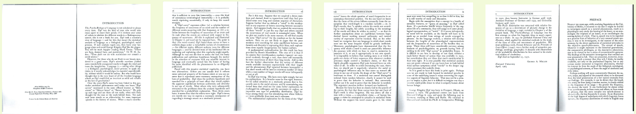
- Mixture of local optimization and randomness
- Numerous efforts...

1. Carlson and Doyle, 1999:
   Highly Optimized Tolerance
   (HOT)—Evolved/Engineered Robustness [4, 5]
2. Ferrer i Cancho and Solé, 2002:
   Zipf's Principle of Least Effort [8]
3. D'Souza et al., 2007:
   Scale-free networks [6]

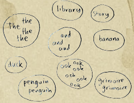
More

Other mechanisms:

- Much argument about whether or not monkeys typing could produce Zipf's law... (Miller, 1957) [16]
- Miller gets to slap Zipf rather rudely in an introduction to a 1965 reprint of Zipf's "Psycho-biology of Language" [17, 25]
- Let us now slap Miller around by simply reading his words out:



- Side note: Miller mentions "Genes of Language."
- Still fighting: "Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution" [7] by Ferrer-i-Cancho and Elvevåg, 2010.



Others are also not happy:

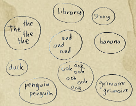
Krugman and Simon

- "The Self-Organizing Economy" (Paul Krugman, 1995) [9]
- Krugman touts Zipf's law for cities, Simon's model
- "Déjà vu, Mr. Krugman" (Berry, 1999)
- Substantial work done by Urban Geographers

Who needs a hug?

From Berry [2]

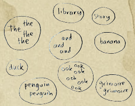
- Déjà vu, Mr. Krugman. Been there, done that. The Simon-Ijiri model was introduced to geographers in 1958 as an explanation of city size distributions, the first of many such contributions dealing with the steady states of random growth processes, ...
- But then, I suppose, even if Krugman had known about these studies, they would have been discounted because they were not written by professional economists or published in one of the top five journals in economics!



Who needs a hug?

From Berry [2]

- ... [Krugman] needs to exercise some humility, for his world view is circumscribed by folkways that militate against recognition and acknowledgment of scholarship beyond his disciplinary frontier.
- Urban geographers, thank heavens, are not so afflicted.



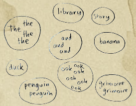
So who's right?

Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution

T. Maillart,[1] D. Sornette,[1] S. Spaeth,[2] and G. von Krogh[2]

[1]*Chair of Entrepreneurial Risks, Department of Management, Technology and Economics, ETH Zurich, CH-8001 Zurich, Switzerland*
[2]*Chair of Strategic Management and Innovation, Department of Management, Technology and Economics, ETH Zurich, CH-8001 Zurich, Switzerland*
(Received 30 June 2008; published 19 November 2008)

Zipf's power law is a ubiquitous empirical regularity found in many systems, thought to result from proportional growth. Here, we establish empirically the usually assumed ingredients of stochastic growth models that have been previously conjectured to be at the origin of Zipf's law. We use exceptionally detailed data on the evolution of open source software projects in Linux distributions, which offer a remarkable example of a growing complex self-organizing adaptive system, exhibiting Zipf's law over four full decades.



So who's right?

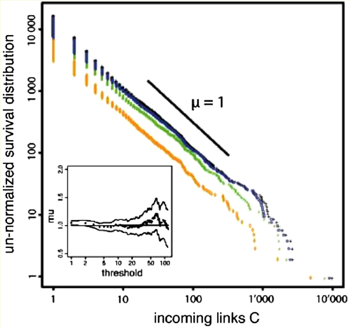


FIG. 1 (color online). (Color Online) Log-log plot of the number of packages in four Debian Linux Distributions with more than C in-directed links. The four Debian Linux Distributions are Woody (19.07.2002) (orange diamonds), Sarge (06.06.2005) (green crosses), Etch (15.08.2007) (blue circles), Lenny (15.12.2007) (black+'s). The inset shows the maximum likelihood estimate (MLE) of the exponent μ together with two boundaries defining its 95% confidence interval (approximately given by $1 \pm 2/\sqrt{n}$, where n is the number of data points using in the MLE), as a function of the lower threshold. The MLE has been modified from the standard Hill estimator to take into account the discreteness of C.

Maillart et al., PRL, 2008:
"Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution" [11]

So who's right?

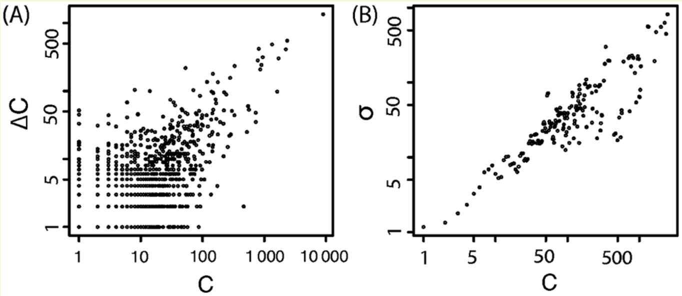


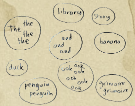
FIG. 2. Left panel: Plots of ΔC versus C from the Etch release (15.08.2007) to the latest Lenny version (05.05.2008) in double logarithmic scale. Only positive values are displayed. The linear regression $\Delta C = R \times C + C_0$ is significant at the 95% confidence level, with a small value $C_0 = 0.3$ at the origin and $R = 0.09$. Right panel: same as left panel for the standard deviation of ΔC.

- Rough, approximately linear relationship between C

So who's right?

Bornholdt and Ebel (PRE), 2001:
"World Wide Web scaling exponent from Simon's 1955 model" [3].

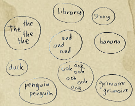
- Show Simon's model fares well.
- Recall ρ = probability new flavor appears.
- Alta Vista (⊞) crawls in approximately 6 month period in 1999 give $\rho \simeq 0.10$
- Leads to $\gamma = 1 + \frac{1}{1-\rho} \simeq 2.1$ for in-link distribution.
- Cite direct measurement of γ at the time: 2.1 ± 0.1 and 2.09 in two studies.



So who's right?

Nutshell:

- Simonish random 'rich-get-richer' models agree in detail with empirical observations.
- Power-lawfulness: Mandelbrot's optimality is still apparent.
- Optimality arises for free in Random Competitive Replication models.



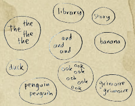
References I

[1] A.-L. Barabási and R. Albert.
Emergence of scaling in random networks.
Science, 286:509–511, 1999. pdf (⊞)

[2] B. J. L. Berry.
Déjà vu, Mr. Krugman.
Urban Geography, 20:1–2, 1999. pdf (⊞)

[3] S. Bornholdt and H. Ebel.
World Wide Web scaling exponent from Simon's 1955 model.
Phys. Rev. E, 64:035104(R), 2001. pdf (⊞)

[4] J. M. Carlson and J. Doyle.
Highly optimized tolerance: A mechanism for power laws in designed systems.
Phys. Rev. E, 60(2):1412–1427, 1999. pdf (⊞)



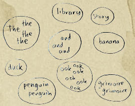
References II

[5] J. M. Carlson and J. Doyle.
Complexity and robustness.
Proc. Natl. Acad. Sci., 99:2538–2545, 2002. pdf (⊞)

[6] R. M. D'Souza, C. Borgs, J. T. Chayes, N. Berger, and R. D. Kleinberg.
Emergence of tempered preferential attachment from optimization.
Proc. Natl. Acad. Sci., 104:6112–6117, 2007. pdf (⊞)

[7] R. Ferrer-i Cancho and B. Elvevåg.
Random texts do not exhibit the real Zipf's law-like rank distribution.
PLoS ONE, 5:e9411, 03 2010.

[8] R. Ferrer i Cancho and R. V. Solé.
Zipf's law and random texts.
Advances in Complex Systems, 5(1):1–6, 2002.

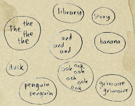


References III

[9] P. Krugman.
The self-organizing economy.
Blackwell Publishers, Cambridge, Massachusetts, 1995.

[10] A. J. Lotka.
The frequency distribution of scientific productivity.
Journal of the Washington Academy of Science, 16:317–323, 1926.

[11] T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh.
Empirical tests of Zipf's law mechanism in open source Linux distribution.
Phys. Rev. Lett., 101(21):218701, 2008. pdf (⊞)

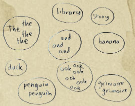


References IV

[12] B. B. Mandelbrot.
An informational theory of the statistical structure of languages.
In W. Jackson, editor, Communication Theory, pages 486–502. Butterworth, Woburn, MA, 1953. pdf (⊞)

[13] B. B. Mandelbrot.
A note on a class of skew distribution function. Analysis and critique of a paper by H. A. Simon.
Information and Control, 2:90–99, 1959.

[14] B. B. Mandelbrot.
Final note on a class of skew distribution functions: analysis and critique of a model due to H. A. Simon.
Information and Control, 4:198–216, 1961.



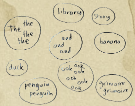
References V

[15] B. B. Mandelbrot.
Post scriptum to 'final note'.
Information and Control, 4:300–304, 1961.

[16] G. A. Miller.
Some effects of intermittent silence.
American Journal of Psychology, 70:311–314, 1957.
pdf (⊞)

[17] G. A. Miller.
Introduction to reprint of G. K. Zipf's "The Psycho-Biology of Language." MIT Press, Cambridge MA, 1965. pdf (⊞)

[18] D. J. d. S. Price.
Networks of scientific papers.
Science, 149:510–515, 1965. pdf (⊞)



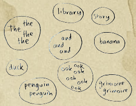
References VI

[19] D. J. d. S. Price.
A general theory of bibliometric and other cumulative advantage processes.
J. Amer. Soc. Inform. Sci., 27:292–306, 1976.

[20] H. A. Simon.
On a class of skew distribution functions.
Biometrika, 42:425–440, 1955. pdf (⊞)

[21] H. A. Simon.
Some further notes on a class of skew distribution functions.
Information and Control, 3:80–88, 1960.

[22] H. A. Simon.
Reply to Dr. Mandelbrot's post scriptum.
Information and Control, 4:305–308, 1961.



References VII

[23] H. A. Simon.
Reply to 'final note' by Benoît Mandelbrot.
Information and Control, 4:217–223, 1961.

[24] G. U. Yule.
A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S.
Phil. Trans. B, 213:21–, 1924.

[25] G. K. Zipf.
The Psycho-Biology of Language.
Houghton-Mifflin, New York, NY, 1935.

[26] G. K. Zipf.
Human Behaviour and the Principle of Least-Effort.
Addison-Wesley, Cambridge, MA, 1949.