Mechanisms for Generating Power-Law Size Distributions, Part 4

Last updated: 2022/08/27, 23:54:10 EDT

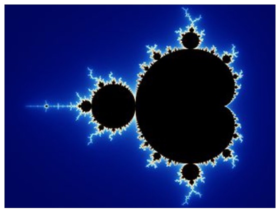
Principles of Complex Systems, Vols. 1, 2, & 3D
CSYS/MATH 300, 303, & 394, 2022–2023 | @pocsvox

Prof. Peter Sheridan Dodds | @peterdodds

Computational Story Lab | Vermont Complex Systems Center
Santa Fe Institute | University of Vermont

Outline

Benoît Mandelbrot



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

Another approach:

Benoît Mandelbrot

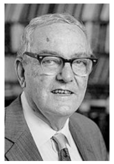
- Derived Zipf's law through optimization [8]
- Idea: Language is efficient
- Communicate as much information as possible for as little cost
- Need measures of information (H) and average cost (C)...
- Language evolves to maximize H/C, the amount of information per average cost.
- Equivalently: minimize 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

 vs. 

Mandelbrot vs. Simon:

- Mandelbrot (1953): "An Informational Theory of the Statistical Structure of Languages" [8]
- Simon (1955): "On a class of skew distribution functions" [14]
- Mandelbrot (1959): "A note on a class of skew distribution functions: analysis and critique of a paper by H.A. Simon" [9]
- Simon (1960): "Some further notes on a class of skew distribution functions" [15]

I have no rival, No man can be my equal

 vs. 

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" [10]
- Simon (1961): "Reply to 'final note' by Benoit Mandelbrot" [17]
- Mandelbrot (1961): "Post scriptum to 'final note'" [11]
- Simon (1961): "Reply to Dr. Mandelbrot's post scriptum" [16]

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." [10]

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." [17]

Two theories enter, one theory leaves:

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
$1 + \log_2 i$	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:

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

- Total Cost:

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

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:

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

where $g = 1/\log_e 2$

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

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

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 5

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]: $\alpha = H/gC$
- 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:

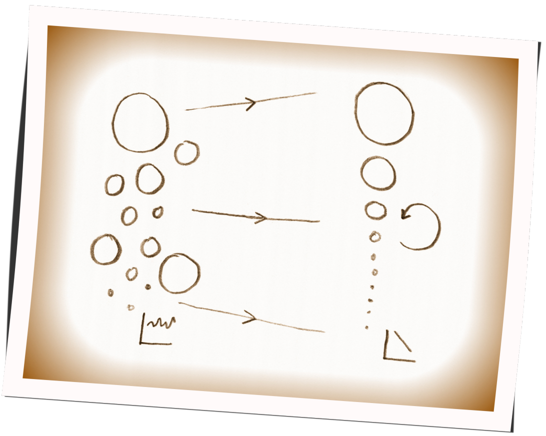
$$1 = \sum_{j=1}^{n} p_j = \sum_{j=1}^{n} (j+1)^{-H/gC} = \sum_{j=1}^{n} (j+1)^{-\alpha}$$

- 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) or $\gamma = 1 + \frac{1}{\alpha} \simeq 1.58$ (very wild)
- 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 [2, 3]
2. Ferrer i Cancho and Solé, 2002: Zipf's Principle of Least Effort [6]
3. D'Souza et al., 2007: Scale-free networks [4]

More

Other mechanisms:

- Much argument about whether or not monkeys typing could produce Zipf's law... (Miller, 1957) [12]
- Miller gets to slap Zipf rather rudely in an introduction to a 1965 reprint of Zipf's "Psycho-biology of Language" [13, 18]
- Let us now slap Miller around by simply reading his words out (see next slides):

- Side note: Miller mentions "Genes of Language."
- Still fighting: "Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution" [5] by Ferrer-i-Cancho and Elvevåg, 2010.

INTRODUCTION

The Psycho-Biology of Language is not calculated to please every taste. Zipf was the kind of man who would take roses apart to count their petals; if it violates your sense of values to tabulate the different words in a Shakespearean sonnet, this is not a book for you. Zipf took a scientist's view of language — and for him that meant the statistical analysis of language as a biological, psychological, social process. If such analysis repels you, then leave your language alone and avoid George Kingsley Zipf like the plague. You will be much happier reading Mark Twain: "There are liars, damned liars, and statisticians." Or W. H. Auden: "Thou shalt not sit with statisticians nor commit a social science."

However, for those who do not flinch to see beauty murdered in a good cause, Zipf's scientific exertions yielded some wonderfully unexpected results to boggle the mind and tease the imagination. Language *is* — among other things — a biological, psychological, social process; to apply statistics to it merely acknowledges its essential unpredictability, without which it would be useless. But who would have thought that in the very heart of all the freedom language allows us Zipf would find an invariant as solid and reliable as the law of gravitation?

Put it this way. Suppose that we acquired a dozen monkeys and chained them to typewriters until they had produced some very long and random sequence of characters. Suppose further that we defined a "word" in this monkey-text as any sequence of letters occurring between successive spaces. And suppose finally that we counted the occurrences of these "words" in just the way Zipf and others counted the occurrences of real words in meaningful texts. When we plot our results in the same manner, we will find exactly the same "Zipf curves" for the monkeys as for the human authors. Since we are not likely to argue that the poor monkeys were searching for some equilibrium between uniformity and diversity in expressing their ideas, such explanations seem equally inappropriate for human authors.

A mathematical rationalization for this result has been provided by Benoit Mandelbrot. The crux of it is that if we assume that word-boundary markers (spaces) are scattered randomly through a text, then there will necessarily be more occurrences of short than long words. Add to this fact the further observation that the variety of different words available increases exponentially with their length and the phenomenon Zipf reported becomes inescapable: a few short words will be used an enormous number of times while a vast number of longer words will occur infrequently or not at all.

So Zipf was wrong. His facts were right enough, but not his explanations. In a broader sense he was right, however, for he called attention to a stochastic process that is frequently seen in the social sciences, and by accumulating statistical data that cried out for some better explanation he challenged his colleagues and his successors to explore an important new type of probability distribution. Zipf belongs among those rare but stimulating men whose failures are more profitable than most men's successes.

So who's right?

Bornholdt and Ebel (PRE), 2001: "World Wide Web scaling exponent from Simon's 1955 model" [1].

- 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?

Recent evidence for Zipf's law...

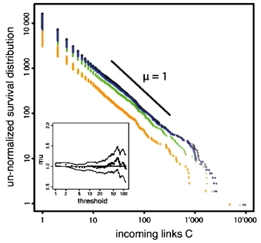


FIG. 1 (color online). (Color Online) Log-log plot of the number of packages in four Debian Linux Distributions with more than C in-directed links. The four Debian Linux Distributions are Woody (19.07.2002) (orange diamonds), Sarge (06.06.2005) (green crosses), Etch (15.08.2007) (blue circles), Lenny (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" [7]

So who's right?

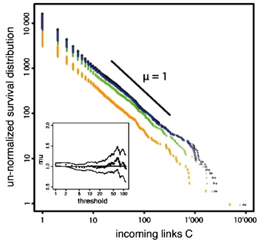


FIG. 1 (color online). (Color Online) Log-log plot of the number of packages in four Debian Linux Distributions with more than C in-directed links. The four Debian Linux Distributions are Woody (19.07.2002) (orange diamonds), Sarge (06.06.2005) (green crosses), Etch (15.08.2007) (blue circles), Lenny (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" [7]

So who's right?

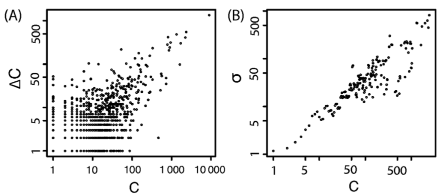


FIG. 2. Left panel: Plots of ΔC versus C from the Etch release (15.08.2007) to the latest Lenny version (05.05.2008) in double logarithmic scale. Only positive values are displayed. The linear regression $\Delta C = R \times C + C_0$ is significant at the 95% confidence level, with a small value $C_0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

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

[2] 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

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

References II

[4] 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

[5] 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.

[6] 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

[7] 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

[8] 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

[9] 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.

References IV

[10] 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.

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

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

References V

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

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

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

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

References VI

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

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