Zipf Law Revisited: A Model of Emergence and Manifestation

Lev B. Levitin
Department of Electrical and
Computer Engineering
Boston University
levitin@bu.edu

1. Introduction

Zipf's law is a famous empirical law that is observed in the behavior of many complex systems of surprisingly different nature. This remarkable frequency-rank relationship known in linguistics as Zipf's law [Zipf 1935], which was first found by Pareto [Pareto 1897] in economics and appears with astonishing invariability in demography [Auerbach 1913], biology [Willis 1922], physics [Nicolis and Tsuda 1989], social sciences, Internet flows, etc. (see also [Guiter and Arapov 1982]). In the realm of linguistics, Zipf's law can be formulated as follows. If we consider a long text and assign ranks to all words that occur in the text in the order of decreasing frequencies, then the frequency f_r of a word of rank r satisfies the empirical law

$$f_r = \frac{B}{r^{\gamma}} \tag{1}$$

where B and γ are constants and $\gamma \approx 1$.

Most theoretical explanations of Zipf's law are based on variational principles similar to those in physics, such as "least effort" [Zipf 1935], "minimum cost" [Mandelbrot 1953], "minimum energy" [Shreider 1967], "equilibrium" [Orlov, 1982], etc. But, in contrast with theoretical physics, where variational principles always rest on the underlying dynamics of the system, here the explanations have somewhat a teleologic flavor, leaving the mechanism of the process concealed. A careful analysis of the assumptions made in the variational derivations of Zipf's law shows that they are all based on a model of noninteracting particles (interpreted as symbols, words, etc.), i.e., on the "ideal gas" model. This approach is expressed in the most explicit form by Shreider [Shreider 1967], who uses a straightforward thermodynamic analogy. Namely, he assigns an "energy" to each "sign" and considers a statistical ensemble of texts formed from these "signs" comprising the text. This is nothing else but an ideal gas of "signs." The same idea in a different form is used in a more recent paper by Li

[Li 1992]. He assumes that symbols (including the "blank space") are generated independently with equal probabilities and shows that this results in (approximately) Zipf's law for the frequencies of words in a long text. Independence of symbols means, of course, absence of interaction, which brings us again to the ideal gas model. Consequently, the author comes to the conclusion that "Zipf's law is not a deep law in natural language as one might first have thought."

We believe the situation is not so trivial. The fact that simple structureless systems can display Zipf-law-like distributions does not preclude Zipf's law--together with more subtle characteristic features--from reflecting mechanisms that govern the behavior of complex systems. Models of such behaviors should be essentially based on the interaction and interdependence of the components of the system and lead to empirically verifiable conclusions different from those provided by "ideal gas" models. A model of the development of an evolutionary system in the form of a nonstationary branching Markov process has been suggested in [Levitin and Schapiro 1993], and [Günther et al. 1996]. Under very simple and general assumptions, this model leads to Zipf's law for the expected values of species populations in an ecosystem. Apparently this is the first model that provides a theoretical explanation of Zipf's law based on a nontrivial interdependence of the system components.

Another result which may prove to be important is the stepwise behavior of the *ranked* expected values in the evolutionary model, in contrast to the smooth Zipf-law behavior of the expected values for the unranked (original) species population numbers. This result is due to the broadness of the distributions of the population numbers in our model, as opposed to narrow (binomial type) distributions in the "ideal gas" models. Thus, there exists an opportunity, by comparison with empirical data, to obtain crucial evidence as to which model relates to reality.

2. An Evolutionary Model

Here we present a model of the development of an evolutionary system suggested in [Levitin and Schapiro 1993] in the form of a nonstationary branching Markov process. We will formulate the model in the language of ecological dynamics, though it can be easily reformulated in terms of demography, linguistics, Internet statistics, etc. Henceforth we will denote random variables by capital letters and their values by lowercase letters.

Consider an ecosystem which consists of populations $N_k(N)$ [$k = 1, 2, \ldots A(N)$] of species s_k, where N is the number of steps of the process interpreted as time (time is discrete in this model), $N_k(N)$ is a random variable which is the population of species s_k at time N, and $A(N)$ is the (random) number of different species at the N^{th} step of the process. The system is assumed to evolve according to the following rules.

1. At the $(N + 1)$th step of the process exactly one individual is created. The probability that the newly created individual belongs to the species s_k is proportional to the population of that species at time N.

$$\Pr\{N_k(N+1) = n_k + 1 \mid N_k(N) = n_k\} = P_{k,N+1}(n_k + 1 \mid n_k) = (1 - c(N))\frac{n_k}{N} \qquad (2)$$

2. The probability that an individual of a new species $s_{A(N)+1}$ will be created at the $(N + 1)$th step of the process (probability of a successful mutation) is

$$\Pr\{N_{A(N)+1}(N+1) = 1 \mid N_{A(N)+1}(N) = 0\} = P_{A+1,N+1} = c(N) \qquad (3)$$

It is mathematically convenient to introduce a "fictitious species" s_0 that "preexisted" at time $N = 1$. The birth of an individual of s_0 can be interpreted as the "noncreation" of an individual of any "real species" s_k. (The linguistic interpretation would be generation of the "empty word.") Then the initial conditions can be expressed as

$$N_0(1) = 1, \quad A(1) = 0 \tag{4}$$

and for any N

$$\sum_{k=0}^{A(N)} N_k(N) = N, \quad \sum_{k=0}^{A(N)+1} P_{k,N+1} = 1$$

Formulas (2)-(4) define a branching Markov process.

We will analyze the behavior of the expected values $E(N_k)$ and the average frequencies $f_k(N) = E(N_k(N))/N$. Consider two special cases corresponding to two different assumptions about the mutation rate.

1. $$c(N) = c = const, \quad c << 1 \tag{5}$$

   Then, the expected number of species at step N is

   $$E(A(N)) = 1 + (N-1)c \tag{6}$$

   Calculation of the explicit expression of $E(N_k(N))$ is complicated by the fact that the step $N^{(k)}$ when species s_k appears is a random variable. After an intricate derivation, we obtain:

   $$E(N_k(N)) = \left[N\left(\frac{c}{1-c}\right)^k \left(\sum_{j=0}^{k-1} \left(\frac{1-c}{c}\right)^j \frac{(-1)^{k-j+1}}{j+1} + (-1)^{k+1} \frac{c \ln c}{1-c} \right) \right]^{1-c} \tag{7}$$

   Hence, for $N>>1$, $c<<1$, the expected values and frequencies of species, numbered by the order of their appearance, are asymptotically equal to

   $$E(N_k(N)) \approx \left(\frac{cN}{k}\right)^{1-c}, \qquad f_k(N) \approx \frac{c^{1-c} N^{-c}}{k^{1-c}} \tag{8}$$

   which is Zipf's law (with the exponent slightly smaller than 1).

2. Now assume that the probability of mutation leading to the emergence of a new species decreases with time:

   $$c(N) = bN^{-q}, \text{ where } q<<1. \tag{9}$$

   Then the expected number of species grows slower then N:

   $$E(A(N)) = \frac{b}{1-q} N^{1-q} \tag{10}$$

   For large ranks $k>>1$ the frequencies are

   $$f_k(N) = \left[\frac{b}{(1-q)k}\right]^{\frac{1}{1-q}} \exp\left[-\frac{b}{q}\left(\frac{b}{(1-q)k}\right)^{\frac{q}{1-q}} + \frac{b}{qN^q} \right] \tag{11}$$

This is also Zipf's law, since the exponential factor is almost constant for $b<<1$, $q<<1$. For example, if b=0.1, q=0.1, the factor changes from 0.45 to 1 when i changes from 1 to infinity. In contrast with case 1 (q=0), now the exponent in Zipf's law is larger then 1, and there exists a counterpart of thermodynamic limit ($N\to\infty$) for the average frequencies:

$$f_k = \lim_{N\to\infty} f_k(N) = \left[\frac{b}{(1-q)k}\right]^{\frac{1}{1-q}} \exp\left[-\frac{b}{q}\left(\frac{b}{(1-q)k}\right)^{\frac{q}{1-q}}\right] \tag{12}$$

3. Information Complexity

Let us address now the question of the complexity of the system described by our model. We expect intuitively that a "good measure" of complexity should reflect both "unpredictability" and "organization" (which implies memory) in the behavior of a complex system. We suggest, as a measure of complexity at time N, the mutual information between two successive states of the system s_N and s_{N-1}.

$$C_N = I(s_n; s_{n-1}) = H(s_n) - H(s_n \mid s_{n-1}) \tag{13}$$

This measure agrees with our intuition since it is nonnegative and vanishes for both extreme cases of chaotic (i.e. memoryless) systems and, on the other hand, strictly deterministic systems.

In our model the state s_N is a random vector with a random number $A(N)$ of components

$$s_N = (N_1(N), N_2(N)..., N_{A(N)}(N)) \tag{14}$$

For large N, we can consider random variables $N_i(N)$ as almost independent. Then in case 1, approximately,

$$C_N \approx \frac{\pi^2}{6} cN - (1-c)\ln N \tag{15}$$

Thus, the limit complexity per one component of the system (one species) is

$$\tilde{C}_{\lim} = \lim_{N\to\infty} \frac{C_N}{E(A(N))} = \frac{\pi^2}{6} \text{ nats,} \tag{16}$$

or 2.37 bits per species.

Similar analysis in case 2 gives the same limit complexity per species (specific complexity) for $q<<1$. Apparently, this complexity is characteristic for all systems which obey Zipf's law with the exponent close to one.

4. The Effect of Ranking

The behavior of the expected values $E(N_k)$ is not sufficient to explain the empirically observed Zipf-law-like distributions. As shown in [Günther *et al.* 1992], the probability distribution for a single species has an asymptotically exponential (geometric) form ($N >> 1$):

$$p_{k,N}(n_k) = \Pr\{N_k(N) = n_k\} = a_k(1-a_k)^{n_k-1}, \tag{17}$$

where

$$a_k = \left(\frac{k}{cN} \right)^{1-c} \tag{18}$$

This is a very broad probability distribution with the standard deviation of the same order of magnitude (in terms of N) as the expected value. Since any empirically observed set of population values is just one random sampling (realization) of the set of random variables $\{N_k\}$, these values listed in the order of species would exhibit a chaotic nonmonotonic behavior, and one would not be able to observe Zipf's law at all! Indeed, looking at Fig. 1 (gray line), it is impossible to recognize Zipf's law in the chaotically fluctuating population values. However, after ranking the same population values in decreasing order, we obtain a much smoother monotonic curve (Fig. 1, solid line) from which Zipf's law can be easily discerned. This phenomenon is explained by the fact that the probability distributions for the new random variables $\widetilde{N}_r$, which are populations of a given rank r, are much narrower than those for N_k -- the populations of the species. Numerical results demonstrating this effect have been presented in [Gtinther *et al.* 1996]. Consequently, *a single realization can serve as a typical representative of the entire statistical ensemble*. (Note that different species may occupy the same rank in different realizations.)

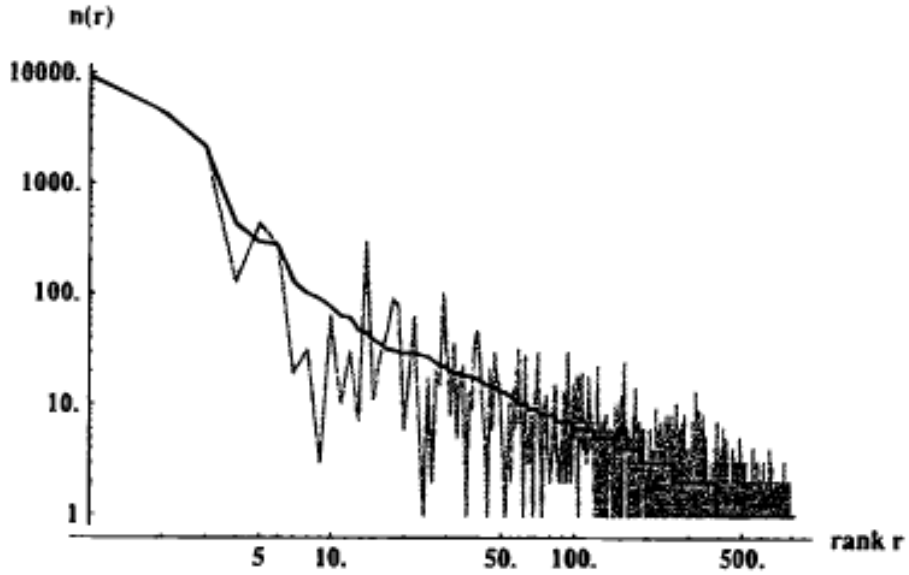

Figure. 1 The gray line shows one realization of the process defined by equations (2)-(4), with the parameters $c = 0.02$ and $N = 40{,}000$. The x axis is the species number as determined by the creation time of this species. The black curve shows the same realization, but now ranked according to the values n_k and plotted against rank. Note the smoothness of this curve.

Obviously, the curve for the expected values of the ranked variables is, in general, steeper than that for the unranked ones. However, it turns out that, in the case of Zipf's law with exponent γ close to 1 ($c \to 0$), the expected values of ranked populations $E(\widetilde{N}_r)$ follow the same law for $r \gg 1$ as $E(N_k)$. Namely, with good approximation,

$$E(\widetilde{N}_r) \approx \left\lfloor \frac{A-1}{r} \exp\left(-\frac{1}{r}\right) \right\rfloor \approx \left\lfloor \frac{A-1}{r} \right\rfloor, \tag{19}$$

which is close to $E(N_k) = \frac{A}{k}$. However, in contrast with expected values for unranked variables, $E(\widetilde{N}_r)$ demonstrate a characteristic "staircase" behavior, which is due to the floor function. More exactly, the expected values $E(\widetilde{N}_r)$ are still changing continuously, but with alternating "steep" and "flat" intervals. This "staircase" shape of the curve for $E(\widetilde{N}_r)$ should not be confused with the steps observed in any empirical realization due to the discreteness of the population numbers. The steps in the expected values curve appear as a result of the ranking procedure, i.e., the transition from $\{N_k\}$ to $\{\widetilde{N}_r\}$, due to the fact that the distribution (17) is very broad. More careful analysis leads to the conclusion that, in fact, the curve remains smooth up to larger ranks. The "steps" become visible if $r \geq 3A^{2/3}$. The length of the i^{th} step, i.e. the interval (r_{i+1}, r_i) for which $E(\widetilde{N}_r) \approx i$ is approximately equal to

$$(\Delta r)_i = r_i - r_{i+1} \approx \frac{A-1}{i(i+1)} = \frac{r_{i+1} r_i}{A-1} \tag{20}$$

These results remain valid for $\gamma \neq 1$, provided that $c = 1 - \gamma \ll 1$. However, the exponent of Zipf's law for $\widetilde{N}_r$ becomes slightly larger then that for N_k. A numerical example given in Fig. 2 demonstrates excellent correspondents between theoretical and simulated data.

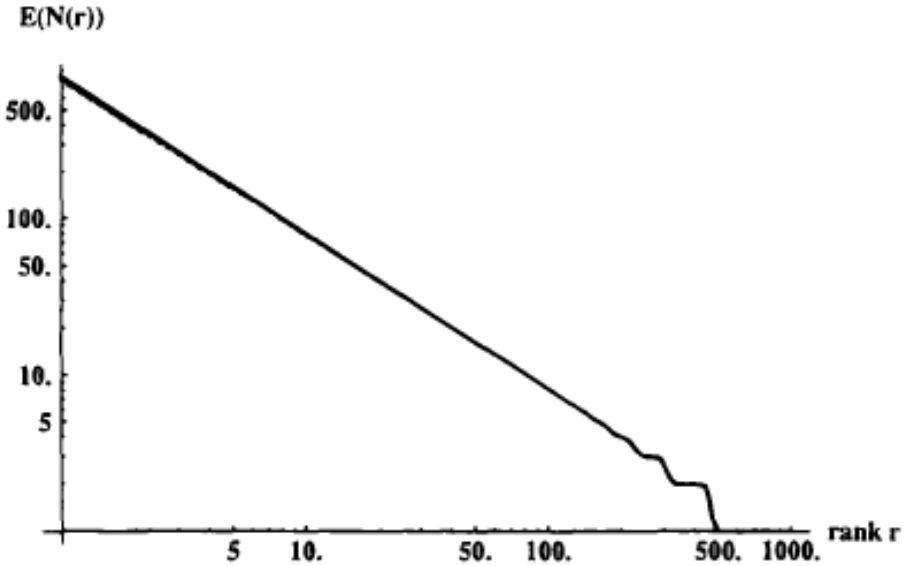

Figure. 2 Comparison between theory and simulation. Parameters used are $N = 6000$, $A = 1000$, and $\gamma = 0.95$. Shown are the theoretical $E(\widetilde{N}_r)$ (black) and an ensemble average (gray) of a numerical simulation. The "staircase" behavior is observed for sufficiently large ranks.

5. Conclusions

One may conclude that the ubiquitous appearance of Zipf's law is based on two independent effects. The first is the fact that very general transition probabilities lead to Zipf's law. The second reason why Zipf's law is found so often is probably based on the ranking procedure, which makes Zipf structures empirically observable because they are robust under its application.

Let us reiterate that the model from [Günther *et al.* 1996] is the first one that not only leads to the overall Zipfian behavior, but predicts a new, verifiable phenomenon: the deviations from the "ideal" Zipf law in the form of the "staircase" behavior of the expected values. Thus, our model suggests the emergence of a second-tier structure of "superclasses" – groups of classes with almost equal populations.

Acknowledgments

The author would like to express his most cordial thanks to K. Aizikov (Boston University) for his indispensable assistance in preparation of this paper.

References

Auerbach, F., 1913, Das Gesetz der Bevökerungskonzentration [The law of population concentration], *Petermans Mitteilungen,* **59**, 74.

Frankhauser, P., 1991, The Pareto-Zipf-distribution of urban systems as stochastic process, in *Models of Selforganization in Complex Systems,* W. Ebeling, M. Peschel, and W. Weidlich, eds., Akademie Verlag, Berlin.

H. Guiter and M. V. Arapov, eds., 1982, *Studies on Zipf's Law,* Studienverlag Dr. N. Brockmeyer,

Günther, R., Schapiro, B., and Wagner, P., 1992, Physical complexity and Zipf's law, *International Journal of Theoretical Physics,* **31**, 525-543.

Günther, R., Levitin, L., Schapiro, B., and Wagner, P., 1996, Zipf's Law and the Effect of Ranking on Probability Distribution. *Intern. Journal of Theoretical Physics*, **35**, No.2, 395-417.

Katsikas, A. A., and Nicolis, J. S., 1990, Chaotic dynamics of generating Markov partitions and linguistic sequences mimicking Zipf's law, *Nuovo Cimento,* **12D**, 177.

Kohonen, T., 1982, Analysis of a simple self-organizing process, *Biological Cybernetics,***44,** 135-140.

Levitin, L. B., and Schapiro, B., 1993, Zipf's law and information complexity in an evolutionary system, in *Proceedings IEEE International Symposium on Information Theory,* San Antonio, Texas, p. 76.

Li, W., 1992, Random texts exhibit Zipf's-law-like word frequency distributions, *IEEE Transactions on Information Theory,* **38**, 1842.

Mandelbrot, B. B.,1953, An information theory of the statistical structure of language, in *Communication Theory,* W. Jackson, ed., London, pp. 486-502.

Mandelbrot, B. B.,1983, The *Fractal Geometry of Nature,* Freeman, New York.

Nicolis, J. S., and Tsuda, I., 1989, On the parallel between Zipf's law and l/f process in chaotic systems possessing coexisting attractors, *Progress of Theoretical Physics,* **82**, 254-274.

Orlov, J. K., 1982, Ein Modell der Häufigkeitsstruktur des Vokabulars, in *Studies on Zipf's Law,*H. Guiter and M. V. Arapov, eds., Studienverlag Dr. N. Brockmeyer, Bochum, Germany.

Pareto, V., 1897, *Cour d'Economie Politique,* Lausanne and Paris [reprinted in *Oevre Completes,*Genf Droz].

Shreider, Yu. A., 1967, Theoretical derivation of text statistical features, *Problemy Peredachi lnformatsii(Problems of Information Transmition),* **3**, 57-63.

Willis, J. C., 1922, *Age and Area,* Cambridge University Press, Cambridge.

Zipf, G. K., 1935, *The Psychobiology of Language,* Houghton-Miflin, Boston.

Zipf, G. K., 1949, *Human Behavior and the Principle of Least Effort,* Addison-Wesley, Cambridge, Massachusetts.