Citations and the Zipf–Mandelbrot Law

Z. K. Silagadze
Budker Institute of Nuclear Physics,
630 090, Novosibirsk, Russia

A curious observation was made that the rank statistics of scientific citation numbers follows the Zipf–Mandelbrot law. The same power-like behavior is exhibited by some simple random citation models. The observed regularity indicates not so much the peculiar character of the underlying (complex) process, but more likely than it is usually assumed, its more stochastic nature.

1. Introduction

We begin with an explanation of Zipf's law. If we assign ranks to all words of some natural language according to their frequencies in some long text (e.g., the Christian *Bible*), then the resulting frequency-rank distribution follows a very simple empirical law

$$f(r) = \frac{a}{r^{\gamma}} \tag{1}$$

with $a \approx 0.1$ and $\gamma \approx 1$. This was observed by G. K. Zipf for many languages more than 50 years ago [1, 2]. More modern studies [3] also confirm a very good accuracy of this rather strange regularity.

In his attempt to derive Zipf's law from information theory, Mandelbrot [4, 5] produced a slightly generalized version of it:

$$f(r) = \frac{p_1}{(p_2 + r)^{p_3}}, \tag{2}$$

p_1, p_2, and p_3 all being constants.

The same inverse power-law statistical distributions were found in embarrassingly different situations (e.g., [6, 7]). In economics, it was discovered by Pareto [8] over 100 years ago and states that incomes of individuals or firms are inversely proportional to their rank. In less formal words [9], "most success seems to migrate to those people or companies who already are very popular." In demography [2, 10, 11], city sizes (populations) also are power-like functions of their ranks. The same regularity reveals itself in the distributions of areas covered by satellite cities and villages around huge urban centers [12].

Remarkably enough, as is claimed in [13], in countries such as the former USSR and China, where natural demographic processes were significantly distorted, city sizes do not follow Zipf's law!

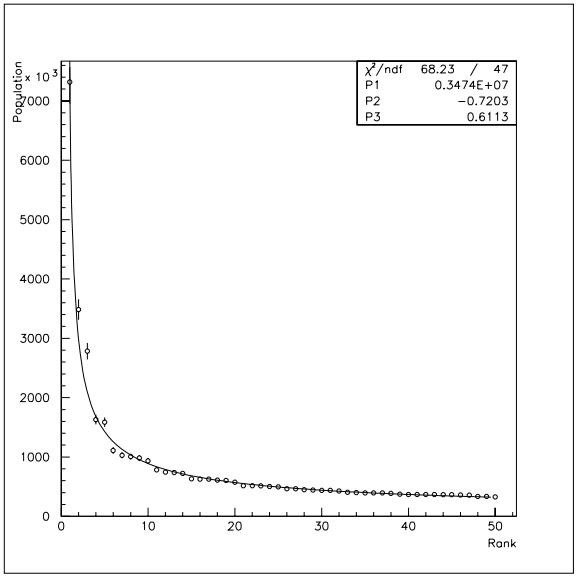

Figure 1. Distribution of the 50 largest cities (USA) according to their rank.

Other examples of zipfian behavior are encountered in chaotic dynamical systems with multiple attractors [14], in biology [15], ecology [16], social sciences, and so forth [17].

The most recent examples of Zipf-like distributions are related to the World Wide Web surfing process [18, 19].

Does all this sound like a joke and seem improbable? I thought so when I became aware of this weird law from M. Gell-Mann's book *The Quark and the Jaguar* [20]. Figure 1 shows the distribution of the 50 largest cities (USA) according to their rank [21], fitted by equation (2). The actual values of fitted parameters depend on the details of the fit. I assume (rather arbitrarily) 5% error in data.

More empirical evidence may be needed to accept improbable things. Figure 2 shows another instance, the list of the most populated countries [22] fitted by the Mandelbrot formula of equation (2). An even simpler zipfian a/r parameterization will work in this case fairly well!

2. Fun with citations

All of this was known long ago. Of course it is exciting to check the correctness of the Zipf–Mandelbrot law personally. But more exciting is to find whether this rule still holds in a new area. The SPIRES database provides an excellent possibility to check scientific citations against Zipf–Maldebrot's regularity.

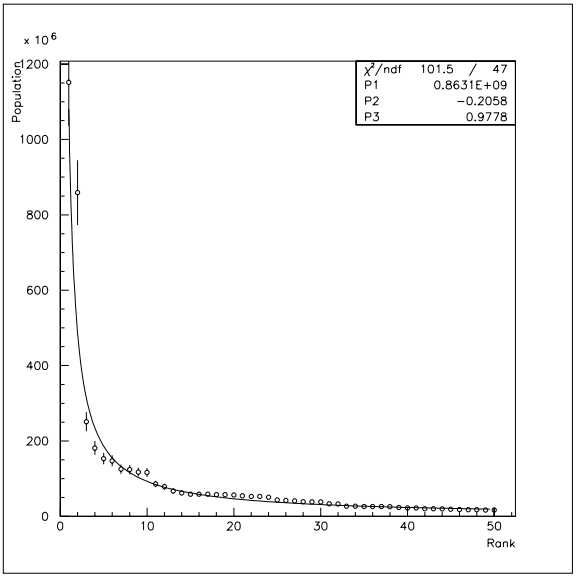

Figure 2. Rank distribution of the most populated countries.

As I became involved with these matters because of M. Gell-Mann's book, my first try naturally was his own citations. The results were encouraging, as shown in Figure 3.

But maybe M. Gell-Mann is not the best choice for this goal. SPIRES is a rather novel phenomenon, and M. Gell-Mann's many important papers were written long before its creation. So they are purely represented in the database. Therefore, let us try a present day citation favorite, E. Witten. Figure 4 shows his 160 most cited papers according to SPIRES [23]. Note once more that the values of fitted parameters may depend significantly on the details of the fit. In this and the previous case I chose $\sqrt{N}$ as an estimate for data errors, so as not to ascribe too much importance to data points with small numbers of citations. In other occasions I assume 5% errors. Needless to say, both choices are arbitrary.

You have probably noticed very big values of the prefactor p_1. Of course this is related to the rather big values of the other two parameters. We can understand a big value of the p_2 parameter as follows. The data set of an individual physicist's papers are a subset of data about all physicists. So we can think of p_2 as being an average number of papers from other scientists between two given papers of the physicists under consideration. Whether right or not, this explanation gains some empirical support if we consider the top cited papers in SPIRES [24] (review of particle physics is excluded) shown in Figure 5. As can be seen, p_2 is fairly small.

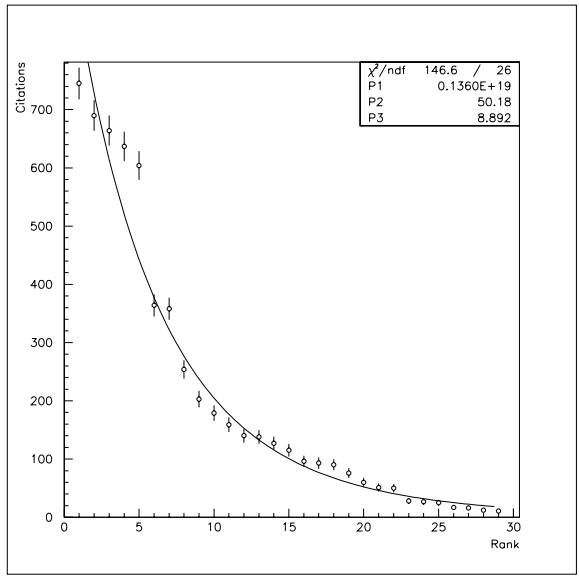

Figure 3. Citations on M. Gell-Mann's papers from the SPIRES database.

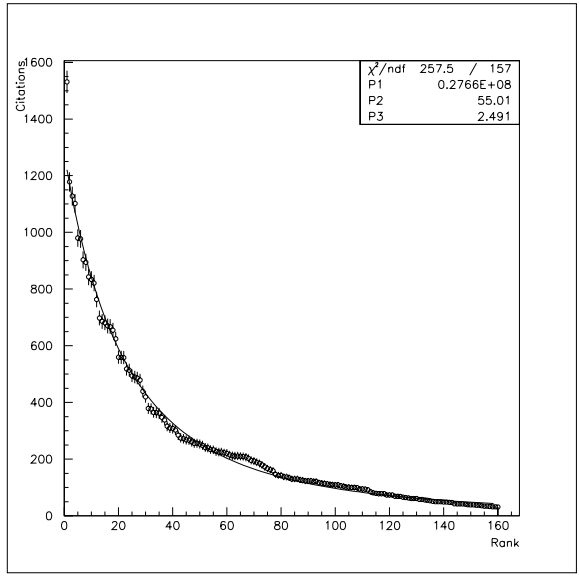

Figure 4. Rank distribution of the most cited papers by E. Witten according to the SPIRES database.

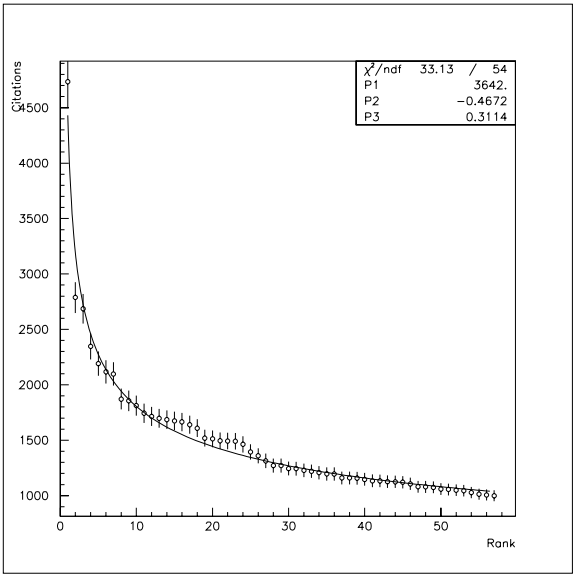

Figure 5. Top cited papers in the SPIRES database.

It is also possible to find the list of 1120 most cited physicists (not only from high energy physics) on the World Wide Web [25]. Again, equation (2) with $p_1 = 3.81 \cdot 10^4$, $p_2 = 10.7$, and $p_3 = 0.395$ gives an excellent fit. For the bulk of the data, Mandelbrot's curve gives better than 5% precision!

You may wonder now why p_2 is relatively high. I really do not know. Maybe the list is still incomplete for the lower ranked papers. In any case, if you take just the first 100 entries from this list, the fit results in $p_1 = 2.1 \cdot 10^4$, $p_2 = -0.09$, and $p_3 = 0.271$. This example also shows that the Mandelbrot curve with constant p_1, p_2, and p_3 is not actually as good an approximation as one might judge from the previously given histograms. This is because different parts of data prefer different values of the Mandelbrot parameters.

3. Any explanation?

The general character of the Zipf–Mandelbrot law is hypnotizing. Several wildly different areas have already been mentioned where it was encountered. Can it be considered as some universal law for complex systems? And if so, what is the underlying principle which unifies all of these seemingly different systems? What kind of principle can be common for natural languages, individual wealth distribution in some society, urban development, scientific citations, and female first name frequencies distribution? The latter is reproduced with data from [26] in Figure 6.

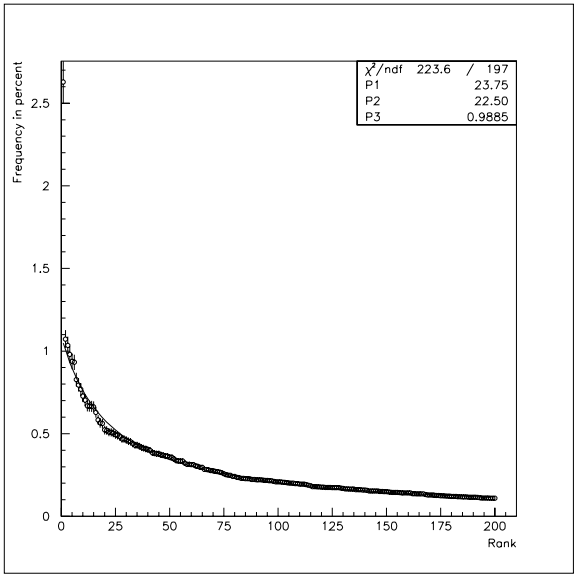

Figure 6. Rank distribution of female first names.

Another question is: Do the Mandelbrot parameters p_2 and p_3 tell us anything about the (complex) process which triggered the corresponding Zipf–Mandelbrot distribution? For this goal an important issue is how to perform the fit (least square, χ^2, method of moments [19], or something else). I do not have any answer to this question now. However, compare the parameters for the female first name distribution (Figure 6) and male first name distribution (Figure 7), data was taken from [26]. In both cases the χ^2 fit was applied with 5% errors assumed for each point. The power-counting parameter p_3 is the same for both distributions, although the p_2 parameter has different values.

If you are fascinated by a possibility that very different complex systems can be described by a single simple law, you may be disappointed (as was I) to learn that some simple stochastic processes can lead to the very same zipfian behavior. Say, what profit will you have from knowing that some text exhibits Zipf's regularity, if this gives you no idea that the text was written by Shakespeare or by a monkey? Alas, it was shown [4, 27–29] that random texts ("monkey languages") exhibit Zipf's-law-like word frequency distribution. So Zipf's law seems to be at least "linguistically very shallow" [5] and "is not a deep law in natural language as one might first have thought" [28].

Two different approaches to the explanation of Zipf's law are very well summarized in G. Miller's introduction to the 1965 edition of Zipf's book [1]:

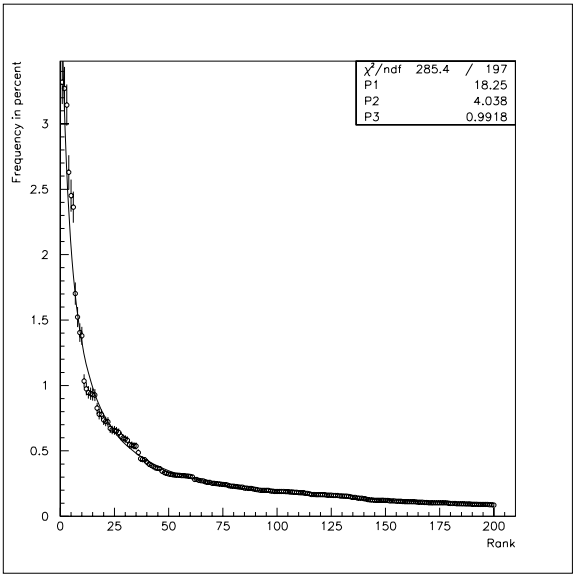

Figure 7. Rank distribution of male first names.

> Faced with this massive statistical regularity, you have two alternatives. Either you can assume that it reflects some universal property of human mind, or you can assume that it reflects some necessary consequence of the laws of probabilities. Zipf chose the synthetic hypothesis and searched for a principle of least effort that would explain the apparent equilibrium between uniformity and diversity in our use of words. Most others who were subsequently attracted to the problems chose the analytic hypothesis and searched for a probabilistic explanation. Now, thirty years later, it seems clear that the others were right. Zipf's curves are merely one way to express a necessary consequence of regarding a message source as a stochastic process.

Were "others" indeed right? Even in the realm of linguistics the debate is still not over after another 30 years have passed [30]. In the case of random texts, the origin of Zipf's law is well understood [31, 32]. In fact such texts exhibit no zipfian distribution at all, but log-normal distribution, the latter giving in some cases a very good approximation to Zipf's law. So there is no doubt that simple stochastic (Bernoulli or Markov) processes can lead to a zipfian behavior. No dynamically nontrivial property (interactions and interdependence) is required at all from the underlying system. But it was also stressed in the literature [13, 33] that this fact does not preclude more complex and realistic systems from exhibiting zipfian behavior because of underlying nontrivial dynamics. In this case, we can hope that the Zipf–Mandelbrot

parameters will be meaningful and can tell something about the system properties. Let us note that the rank-frequency distribution for complex systems is not always zipfian. For example, if we consider the frequency of occurrence of letters, instead of words, in a long text, the empirical universal behavior, valid over 100 natural languages with alphabet sizes ranged between 14 and 60, is logarithmic [34]

$$f(r) = A - B \ln r$$

where A and B are constants. This fact, of course, is interesting by itself. It is argued in [34] that both regularities (zipfian and logarithmic) can have the common stochastic origin.

An interesting example of Zipf–Mandelbrot's parameters being useful and effective is provided by ecology [35, 36]. The exponent p_3 is related to the evenness of the ecological community. It has higher values for "simple" and lower values for "complex" systems. The parameter p_2 is related to the "diversity of the environment" [36] and serves as a measure of the complexity of initial preconditions.

The other pole in the explanation of Zipf's law seeks some universal principle behind it, such as "least effort" [2], "minimum cost" [4], "minimum energy" [37], or "equilibrium" [38]. The most impressive and, as the ecological example shows, fruitful explanation is given by B. Mandelbrot [5, 39] and is based on fractals and self-similarity.

The suggested explanations are almost as numerous as the observed manifestations of this universal power-like behavior. This probably indicates that some important ingredient in this regularity is still not being grasped. As M. Gell-Mann concludes [20] "Zipf's law remains essentially unexplained."

4. The almighty chance

If monkeys can write texts they can make citations too! So let us imagine the following random citation model.

- At the beginning there is one seminal paper.
- Every sequential paper makes at most 10 citations (or cites all preceding papers if there are less than 10).
- All preceding papers have an equal probability of being cited.
- Multiple citations are excluded. So if some paper is selected by chance as a citation candidate more than once, the selection is ignored (in this case the total number of citations in a new paper will be less than 10).

I don't know about monkeys but it is simple to use a computer to simulate such a process. Figure 8 shows the result of a simulation for 1000 papers. An apparent power-like structure can be seen, although with staircase behavior. This stepwise structure is exptected to disappear

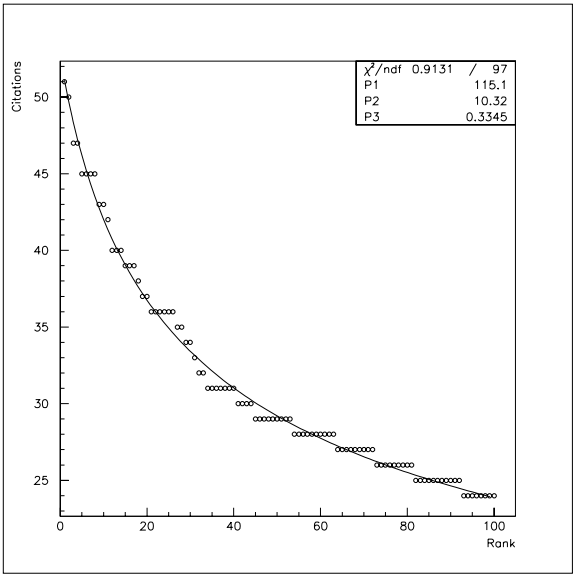

Figure 8. Rank distribution of random citations for the equal probability model.

if the democracy between papers is eliminated and some papers are made more probable to be cited.

Note that even the value of exponent p_3 is reasonably close to what was really observed for the most cited papers. But this can be merely an accident and I do not like to make some far-fetched conclusion about the nature of the citation process from this fact.

In reality, "success seems to attract success" [9]. Therefore, let us try to see what happens if the equal probability axiom is changed by perhaps a more realistic one.

- The probability for a paper to be cited is proportional to $n + 1$, where n is the present total citation number for the paper.

It is still assumed that all preceding papers compete to be cited by a new paper, but with probabilities as follows from the given law. The result for 1000 papers is shown in Figure 9.

The fit seems not so good now, nevertheless you can notice some resemblance with the case of individual scientists. Again I refrain from premature conclusions. Although it is not entirely surprising that the more well-known a given paper of a certain author is, the more probable becomes its citation in a new paper.

5. Discussion

Scientific citations (leaving aside first name frequencies) provide one more example of the Zipf–Mandelbrot regularity. I do not know

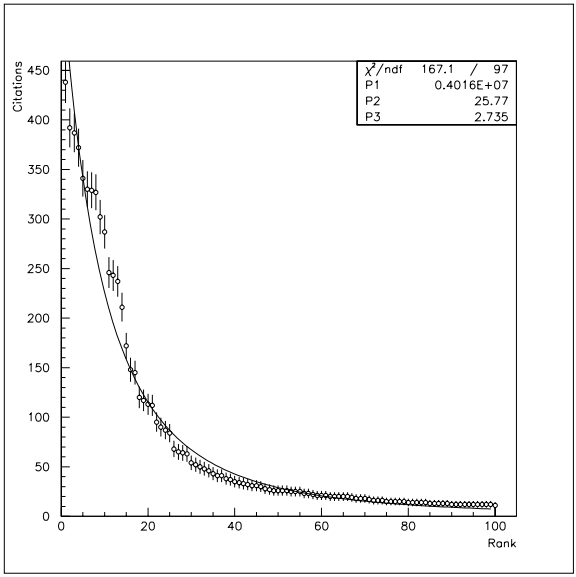

Figure 9. Rank distribution of random citations without equal probabilities.

whether this fact indicates only the significant stochastic nature of the process or something else. In any case SPIRES, and the World Wide Web in general, gives us an excellent opportunity to study the characteristics of the complex process of scientific citations.

I do not know either whether Mandelbrot's parameters are meaningful in this case, or if they can tell us something nontrivial about the citation process.

The very generality of the Zipf–Mandelbrot regularity can make it rather "shallow." But remember, that the originality of answers on the question of whether there is something serious behind the Zipf–Mandelbrot law depends on how restrictive a framework we assume for the answer. A shallow framework will probably guarantee shallow answers. But if we do not restrict our imagination from the beginning, answers can turn out to be quite nontrivial. For example, fractals and self-similarity are certainly great and not shallow ideas.

Acknowledgments

This work was done while the author was visiting Stanford Linear Accelerator Center. I'm grateful to Helmut Marsiske and Lynore Tillim for kind hospitality.

Note added

After this paper was completed, I learned that the Zipf's distribution in scientific citations was discovered in fact earlier by S. Redner [40]. He also cites some previous studies on citations, which were unknown to me. In particular, in [41] it is argued that the citation distribution of the most cited physicists can be fitted by a stretched exponential curve.

I also became aware of G. Parisi's interesting contribution [42] from Dr. S. Juhos.

I thank S. Redner and S. Juhos for their correspondence.

References

[1] G. K. Zipf, *The Psycho-biology of Language: An Introduction to Dynamic Philology* (Houghton Mifflin Company, 1935; MIT Press, 1965).

[2] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Cambridge, MA, 1949; Addison-Wesley, 1965).

[3] H. Kučera and W. N. Francis, *Computational Analysis of Present-Day American English* (Brown University, Providence, 1967).

[4] B. Mandelbrot, "An Informational Theory of the Statistical Structure of Language," in *Communication Theory*, edited by Willis Jackson (Butterworths, London, 1953).

[5] B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).

[6] B. J. West, *An Essay of the Importance of Being Non-linear*, *Lecture Notes in Biomathematics*, **62**, (Springer-Verlag, Berlin, 1985).

[7] E. W. Montroll and M. F. Shlesinger, "Maximum Entropy Formalism, Fractals, Scaling Phenomena, and 1/f Noise: A Tale of Tails," *Journal of Statistical Physics*, **32** (1983) 209.

[8] V. Pareto, *Cours d'economie politique* (Lusanne, F. Rouge, 1896).

V. Pareto, *The New Theories of Economics*, http://melbecon.unimelb.edu.au/het/pareto/theories.

[9] J. van Till, *Fractanomics—The Issue of Scale in the Network Economy*, http://huizen.dds.nl/~vantill/fractanomics.html.

[10] F. Auerbach, "Das Gesetz der Bevölkerungskonzentration," *Petermans Mitteilungen*, **59** (1913) 74.

[11] D. H. Zanette and S. C. Manrubia, "Role of Intermittency in Urban Development: A Model of Large-Scale City Formation," *Physical Review Letters*, **79** (1997) 523.

[12] H. A. Makse, S. Havlin, and H. E. Stanley, "Modelling Urban Growth Patterns," *Nature*, **377** (1995) 608.

[13] M. Marsili and Yi-Cheng Zhang, "Interacting Individuals Leading to Zipf's Law," *Physical Review Letters*, **80** (1998) 2741.

[14] J. S. Nicolis and I. Tsuda, "On the Parallel between Zipf's Law and 1/f Processes in Chaotic Systems Possessing Coexisting Attractors," *Progress of Theoretical Physics*, **82** (1989) 254.

[15] J. C. Willis, *Age and Area* (Cambridge University Press, Cambridge, 1922).

[16] D. R. Margalef, "Information Theory in Ecology," in *Memorias de la Real Academia de Ciencias y Artes de Barcelona*, **23** (1957) 373.

S. Frontier, "Application of Fractal Theory to Ecology," in *Developments in Numerical Ecology*, edited by P. Legendre and L. Legendre, *NATO ASI Series*, **14** (Springer-Verlag, Berlin, Heidelberg, 1987).

L. Aleya and J. Devaux, "The Concept of Seasonal Succession Theory Applied to Phytoplankton through the Coupling use of Diversity Index and Rank-frequency Diagrams in an Eutrophic System," *Internationale Revue der gesamten Hydrobiologie*, **77** (1992) 579.

[17] *Studies on Zipf's Law*, edited by H. Guiter and M. V. Arapov (Studienverlag Dr. N. Brockmeyer, Bochum, Germany, 1982).

[18] C. R. Cunha, A. Bestavros, and M. E. Crovella, *Characteristics of WWW Client-based Traces*, 1995, http://cs-www.bu.edu/faculty/crovella/paper-archive/TR-95-010/paper.html.

J. Nielsen, *Zipf Curves and Website Popularity*, 1997, http://www.useit.com/alertbox/zipf.html.

[19] B. A. Huberman, P. L. T. Pirolli, J. E. Pitkow, and R. M. Lukose, "Strong Regularities in World Wide Web Surfing," *Science*, **280** (1998) 95.

[20] M. Gell-Mann, *The Quark and the Jaguar* (W. H. Freeman and Company, New York, 1994).

[21] *1992 Information Please Almanac*, http://www.kempf.com/~mrk/stats/c-ct-50.txt.

[22] *1992 Information Please Almanac*, http://www.kempf.com/~mrk/stats/nations.txt.

[23] *SPIRES Database*, http://www-spires.slac.stanford.edu/find/hep.

[24] *SPIRES Database*, http://www.slac.stanford.edu/library/topcites/top40all.1997.html.

[25] *ISI's 1120 Most Cited Physicists, 1981–June 1997*, http://pcb4122.univ-lemans.fr/1120physiciens.html.

[26] *U. S. Census Bureau*, http://www.census.gov/ftp/pub/genealogy/names/.

[27] G. Miller, "Some Effects of Intermittent Silence," *American Journal of Psychology*, **70** (1957) 311.

G. Miller and N. Chomsky, in *Handbook of Mathematical Psychology II*, edited by R. Luce, R. Bush, and E. Galanter (Wiley, New York, 1963).

[28] W. Li, "Random Texts Exhibit Zipf's-law-like Word Frequency Distribution," *IEEE Transactions on Information Theory*, **38** (1992) 1842.

W. Li, "Comments to 'Bell Curves and Monkey Languages'," (letter to the editor), *Complexity*, **1** (6) (1996) 6.

[29] J. Nicolis, *Chaos and Information Processing* (World Scientific, 1991).

[30] A. A. Tsonis, C. Schultz, and P. A. Tsonis, "Zipf's Law and the Structure and Evolution of Languages," *Complexity*, **2** (5) (1997) 12.

W. Li, "Comments On Zipf's Law and the Structures and Evolution of Natural Languages," *Complexity*, **3** (5) (1998) 9.

[31] R. Perline, "Zipf's Law, the Central Limit Theorem, and the Random Division of the Unit Interval," *Physical Review E*, **54** (1996) 220.

[32] G. Troll and P. beim Graben, "Zipf's Law Is Not a Consequence of the Central Limit Theorem," *Physical Review E*, **57** (1998) 1347.

[33] R. Günther, L. Levitin, B. Schapiro, and P. Wagner, "Zipf's Law and the Effect of Ranking on Probability Distributions," *International Journal of Theoretical Physics*, **35** (1996) 395.

[34] I. Kanter and D. A. Kessler, "Markov Processes: Linguistics, Zipf's Law and Long-range Correlations," *Physical Review Letters*, **74** (1995) 4559.

[35] S. Frontier, "Diversity and Structure in Aquatic Ecosystems," *Oceanography and Marine Biology Annual Review*, **23** (1985) 253.

[36] S. Juhos and L. Vörös, "Structural Changes During Eutrophication as Revealed by the Zipf–Mandelbrot Model in Lake Balaton, Hungary," *Hydrobiologia*, **370** (1998), 237; http://www.blki.hu/~ndgy/www/zm1.html.

[37] Yu. A. Shreider, "Theoretical Derivation of Text Statistical Features," *Problemy Peredachi Informatsii*, **3** (1967) 57.

[38] J. K. Orlov, "Ein Modell der Häufigkeitsstruktur des Vokabulars," in [17].

[39] B. Mandelbrot, *Fractals and Scaling in Finance* (Springer-Verlag, Berlin, 1997).

[40] S. Redner, "How Popular Is your Paper? An Empirical Study of the Citation Distribution," *European Physical Journal*, **B4** (1998) 131; cond-mat/9804163.

[41] J. Laherrere and D. Sornette, "Stretched Exponential Distributions in Nature and Economy: 'Fat Tails' with Characteristic Scales," *European Physical Journal*, **B2** (1998) 525; cond-mat/9801293.

[42] G. Parisi, "On the Emergence of Tree-like Structures in Complex Systems," in *Field Theory, Disorder, and Simulations* (World Scientific, *Lecture Notes in Physics*, **49** (1992) 298).