International Journal of Theoretical Physics, Vol. 35, No. 2, 1996

Zipf's Law and the Effect of Ranking on Probability Distributions

R. Günther,[1] L. Levitin,[1,2] B. Schapiro,[1] and P. Wagner[1,3]

Received June 8, 1995

Ranking procedures are widely used in the description of many different types of complex systems. Zipf's law is one of the most remarkable frequency–rank relationships and has been observed independently in physics, linguistics, biology, demography, etc. We show that ranking plays a crucial role in making it possible to detect empirical relationships in systems that exist in one realization only, even when the statistical ensemble to which the systems belong has a very broad probability distribution. Analytical results and numerical simulations are presented which clarify the relations between the probability distributions and the behavior of expected values for unranked and ranked random variables. This analysis is performed, in particular, for the evolutionary model presented in our previous papers which leads to Zipf's law and reveals the underlying mechanism of this phenomenon in terms of a system with interdependent and interacting components as opposed to the "ideal gas" models suggested by previous researchers. The ranking procedure applied to this model leads to a new, unexpected phenomenon: a characteristic "staircase" behavior of the mean values of the ranked variables (ranked occupation numbers). This result is due to the broadness of the probability distributions for the occupation numbers and does not follow from the "ideal gas" model. Thus, it provides an opportunity, by comparison with empirical data, to obtain evidence as to which model relates to reality.

1. INTRODUCTION

Many empirical relationships observed in complex systems of remarkably different nature imply ranking procedures. Perhaps the most famous one is the frequency–rank relationship known in linguistics as Zipf's law (Zipf, 1935), which was first found by Pareto (1897) in economics and appears with astonishing invariability in physics (Nicolis and Tsuda, 1989), biology

[1] Naturwissenschaftliches und Medizinisches Institut (NMI), D-72762 Reutlingen, Germany.
[2] Boston University, College of Engineering, Boston, Massachusetts 02215.
[3] Zentrum für Paralleles Rechnen, Mathematisches Institut der Universität zu Köln, 50923 Köln, Germany.

0020-7748/96/0200-0395$09.50/0 © 1996 Plenum Publishing Corporation

(Willis, 1922), demography (Auerbach, 1913), social sciences, 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

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

where c 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 (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 (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 Schapiro (1994), Günther *et al.* (1992), and Levitin and Schapiro (1993). 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.

However, this model calls for a deeper analysis which would go beyond the expected values, since here is the point where the ranking procedure comes into play. Consider any set of integer-valued random variables. Obviously, the distribution of new random variables defined by rank, i.e., the joint distribution of the largest values in all realizations (we mean by realizations of the statistical ensemble sets of sampled values of all random variables), second largest values, etc., up to the smallest values taken from all realizations is always different from the distribution of the original random variables. In particular, any realization of the ranked variables is, by definition, a monotonically decreasing function of the rank, while this is not the case for the original random variables. This difference is the more significant the broader the distribution of the original random variables. In fact, the distributions of the ranked random variables are much narrower than the original distributions, and their expected values may obey a law different from that for the original random variables.

As in the case of Zipf's law, the ranking of observed data according to their frequency is a widespread empirical procedure. It should be borne in mind that, as a result, we obtain just *a single realization of the set of ranked random variables*. Due to the narrowness of the ranked distributions, this realization can better represent the entire statistical ensemble of the *ranked* random variables. But at the same time, it can differ drastically from the typical representatives of the original statistical ensemble. Overlooking this fact may lead to curious artifacts and erroneous conclusions, as shown in Section 4. To our knowledge, up to now there has been no general description of the effects of ranking. This article provides such a description for a few cases which are solved analytically. The results enable us to understand better the consequences of ranking.

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. THE EVOLUTIONARY MODEL

Let us recall the model of the development of an evolutionary system presented in Schapiro (1994), Günther *et al.* (1992), and 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, 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 Nth 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:

$$\Pr\{N_k(N+1) = n_k + 1 \,|\, N_k(N) = n_k\}$$

$$= P_{k,N+1}(n_k + 1 \,|\, n_k) = (1 - c(N))\frac{n_k}{N} \tag{2}$$

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

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

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

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

and for any N

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

Formulas (2)–(4) define a branching Markov process. For the purpose of this paper we assume $c(N) = c = \text{const}$ and $c << 1$. For instance, in linguistics c rarely reaches 0.1 (Günther and Wagner, (1995)) and usually does not exceed 0.05. Empirical data support also the assumption $c = \text{const}$, at least in the case of news texts analyzed in Sharman (1989). [Note, however, that in the case of literature as analyzed in Günther and Wagner (1995) more

complicated $c(N)$ arises which can be understood as a power-law decrease $c(N) \propto N^{-q}$ for $N >> 1$.] It has been found, for instance (Sharman, 1989), that the number of different words grows with almost constant rate when the total number of words N in the text changes from 2×10^5 to 10^6.

As shown in Schapiro (1994), Günther *et al.* (1992), and Levitin and Schapiro (1993), this model leads to Zipf's law for the expected values of the populations of the species. Namely, if the species are numbered by the order of their appearance, then the frequency of species s_k for $N >> 1$, $c << 1$ is asymptotically equal to

$$f_k(N) = \frac{c^{1-c}N^{-c}}{k^{1-c}} \tag{5}$$

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

However, it would be naive to assume that the behavior of the expected values is 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 ($N >> 1$):

$$p_{k,N}(n_k) = \Pr\{N_k(N) = n_k\} = a_k(1 - a_k)^{n_k - 1} \tag{6}$$

where

$$a_k = \left(\frac{k}{cN}\right)^{1-c} \tag{7}$$

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 $\tilde{N}_r$, 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 Günther *et al.* (1992) (see also Fig. 3 below). 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.) Obviously, the curve

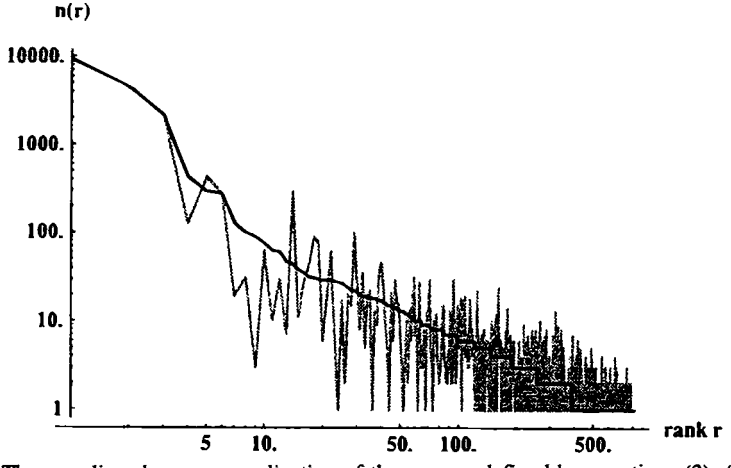

Fig. 1. The gray line shows one realization of the process defined by equations (2)–(4), with the parameters $c = 0.02$ and $N = 40{,}000$. The x axis is the species number as determined by the creation time of this species. The black curve shows the same realization, but now ranked according to the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

for the expected values of the ranked variables is, in general, steeper than that for the unranked ones.

The effect of ranking may depend on both the probability distributions of the random variables involved and the behavior of their expected values. One can conjecture that the effect is minimal if the probability distributions are narrow for the unranked variables from the beginning (in particular, the effect vanishes if the random variables are in fact constants: their ranked sample values coincide with ranked expected values). However, if the distributions are broad [as given by (6)], the ranked expected values may follow a law different from that for the unranked variables.

To our knowledge, the effect of ranking has not been analyzed previously. The present paper considers that problem for the special case of exponential-type probability distributions with emphasis on the results that follow from the evolutionary model given above. Our goal is to show that the mechanism suggested by the model leads indeed to an empirically observable Zipfian behavior. However, we consider other cases which seemingly have empirical counterparts as well.

3. PROBABILITY DISTRIBUTIONS FOR RANKED VARIABLES

Consider an ensemble of N particles (individuals) of A different classes (species) with a joint probability distribution of occupation numbers (populations) N_k, $k = 1, 2, \ldots, A$:

$$\Pr\{N_1 = n_1, \ldots, N_A = n_A\} = p(n_1, \ldots, n_A) \qquad (n_k = 0, 1, 2, \ldots)$$

Let us call $\mathbf{n} = (n_1, n_2, \ldots, n_A)$ the species state vector or the realization of the system of random variables $\{N_k\}$.

Define a new set of A random variables $\{\tilde{N}_r\}$ $(r = 1, 2, \ldots, A)$ in the following way. For each realization $(n_1, n_2, \ldots, n_A)$ reorder the components of the state vector in decreasing order, thereby obtaining a rank state vector $\tilde{\mathbf{n}} = (\tilde{n}_1, \tilde{n}_2, \ldots, \tilde{n}_A)$, where $\tilde{n}_1 \geq \tilde{n}_2 \geq \cdots \geq \tilde{n}_A$. The joint probability distribution for the new random variables $\{\tilde{N}_r\}$ (the rank populations) is given by

$$\begin{aligned}\Pr\{\tilde{N}_1 = \tilde{n}_1, \ldots, \tilde{N}_A = \tilde{n}_A\} &= \tilde{p}(\tilde{n}_1, \ldots, \tilde{n}_A) \\ &= \sum_{n_1 \cdots n_A \in M} p(n_1, \ldots, n_A) \qquad (8)\end{aligned}$$

where the summation is taken over all species state vectors $\mathbf{n}$ that belong to the equivalence class M of the rank state vector $\tilde{\mathbf{n}}$ according to ranking, i.e., that produce the same rank state vector $\tilde{\mathbf{n}}$. Note that only realizations that correspond to permutations of unequal components $n_1, \ldots, n_A$ contribute to the sum (8). For example, when $A = 2$, we obtain

$$\tilde{p}(\tilde{n}_1, \tilde{n}_2) = p(\tilde{n}_1, \tilde{n}_2) + p(\tilde{n}_2, \tilde{n}_1) - p(\tilde{n}_1, \tilde{n}_2)\delta_{\tilde{n}_1, \tilde{n}_2} \qquad (9)$$

where δ_{ij} is Kronecker's symbol: $\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ otherwise. The last term in (9) prevents the species state vector $(\tilde{n}_1, \tilde{n}_1)$ from being counted twice in the calculation of probability $\tilde{p}(\tilde{n}_1, \tilde{n}_1)$ of the rank state vector $(\tilde{n}_1, \tilde{n}_1)$.

A tedious calculation yields the general formula for the joint probability distribution of the rank populations $\{\tilde{N}_r\}$ in terms of joint probabilities of the species populations $\{N_k\}$:

$$\begin{aligned}&\Pr\{\tilde{N}_1 = \tilde{n}_1, \ldots, \tilde{N}_A = \tilde{n}_A\} \\ &= \tilde{p}(\tilde{n}_1, \ldots, \tilde{n}_A) \\ &= \sum_{\{\pi\}} p(\pi(n_1), \ldots, \pi(n_A)) \\ &\quad \times \left(1 + \sum_{m=2}^{A} \frac{m^2 - 3m + 1}{m!} \sum_{j=1}^{A-m+1} \delta_{\pi(n_j)\pi(n_{j+1})\cdots\pi(n_{j+m-1})}\right) \qquad (10)\end{aligned}$$

where $\{\pi\}$ is the set of all possible $A!$ permutations of values $n_1, \ldots, n_A$. Formula (10) can be proved by induction.

The random variables N_k in the model of Section 2 are, strictly speaking, not independent. In particular, they obey the equality $\sum_{k=1}^{A} N_k = N$. However,

the dependence vanishes for $N - A >> 1$, $A >> 1$, which is the case of interest to us. We assume henceforth that

$$p(n_1, \ldots, n_A) = \prod_{i=1}^{A} p_i(n_i) \tag{11}$$

Moreover, we consider the special case of a geometric (exponential) distribution for each N_k:

$$\Pr\{N_k = n_k\} = p_k(n_k) = (1 - x_k)x_k^{n_k - 1} \tag{12}$$

where $k = 1, \ldots, A$, and $n_k = 1, 2, \ldots$.

As discussed in Appendix A, the probability distributions of rank variables in this case have the form

$$\Pr\{\tilde{N}_r = n\} = \tilde{p}_r(n)$$
$$= \sum_{m=r}^{A} (-1)^{m-r} \binom{m-1}{r-1} \sum_{\{\pi_m^A\}} \left(1 - \prod_{i=1}^{m} x_{k_i}\right)\left(\prod_{i=1}^{m} x_{k_i}\right)^{n-1} \tag{13}$$

where $\{\pi_m^A\}$ is the set of all possible $\binom{A}{m}$ choices of m elements $x_{k_1}, \ldots, x_{k_m}$ out of A elements $x_1, \ldots, x_A$.

The expected value $E(\tilde{N}_r)$ and the variance $V(\tilde{N}_r)$ of the rank variables are given by the expressions

$$E(\tilde{N}_r) = \sum_{m=r}^{A} (-1)^{m-r} \binom{m-1}{r-1} \sum_{\{\pi_m^A\}} \left(1 - \prod_{i=1}^{m} x_{k_i}\right)^{-1} \tag{14}$$

and

$$V(\tilde{N}_r) = \sum_{m=r}^{A} (-1)^{m-r} \binom{m-1}{r-1} \sum_{\{\pi_m^A\}} \left[\left(1 - \prod_{i=1}^{m} x_{k_i}\right)^{-2} \prod_{i=1}^{m} x_{k_i}\right] \tag{15}$$

Formulas (13)–(15) will be the starting point of our further analysis. Before returning to the model of Section 2, we consider a simpler example which demonstrates clearly the effect of ranking on the expected values and the width of the probability distributions.

4. UNIFORM JOINT PROBABILITY DISTRIBUTION ON A SIMPLEX

Suppose random variables N_k ($k = 1, 2, \ldots, A$) have a uniform distribution on a simplex $\Sigma_{k=1}^{A} N_k = N$. In other words, probabilities of all state vectors $(n_1, n_2, \ldots, n_A)$ such that $\Sigma_{k=1}^{A} n_k = N$ are equal and given by

$$p(n_1, \ldots, n_A) = \binom{N-1}{A-1}^{-1} \tag{16}$$

where $\binom{N-1}{A-1}$ is the total number of different state vectors, i.e., the number of ways in which N particles can be distributed among A boxes so that each box contains at least one particle.

The distribution for one component N_k is obtained in the same way: fix the number n_k of particles in the kth box and count the number of all remaining partitions. Hence

$$\Pr\{N_k = n_k\} = p_k(n_k) = \binom{N-1-n_k}{A-2} \Big/ \binom{N-1}{A-1} \tag{17}$$

For $N - A >> 1$, $A >> 1$ it leads to a geometric distribution:

$$p_k(n_k) \approx \frac{A-1}{N-1}\left(1 - \frac{A-1}{N-1}\right)^{n_k - 1} \tag{18}$$

This is exactly the procedure one uses in statistical mechanics to derive the canonical ensemble from a microcanonical one. For the case when the energy of a subsystem ("box") is proportional to the number of particles in it (ideal gas!), the Gibbs distribution turns into expression (18).

The asymptotic expressions for the expected values and standard deviations of ranked populations can be derived by use of (14) and (15) (see Appendix B):

$$E(\tilde{N}_r) \approx 1 + \frac{\log(r/A)}{\log(1 - A/N)} \tag{19}$$

$$\sigma_r = [V(\tilde{N}_r)]^{1/2}$$

$$\approx \frac{\log[1 - (1/r - 1/A)^{1/2}] - \log[1 + (1/r - 1/A)^{1/2}]}{\log(1 - A/N)} \tag{20}$$

It follows from (19) and (20) that the ratio $\sigma_r/E(N_r)$ decreases with increase of A as $1/\log(A)$ for $r << A$ and as $1/\sqrt{A}$ for $r/A = \text{const}$.

This demonstrates the effect of the narrowing of distributions due to the ranking. Numerical examples illustrating the rank dependence of the expected values and standard deviations are given in Fig. 2, where instead of the approximations (19) and (20), numerical summations of the exact expression (B2) were used. Figure 3 shows the probability distributions of the unranked, (17), and various ranked probability distributions, where again the exact formula (B1) is used.

As shown in Appendix B, formula (19) gives a very rough approximation of the values $E(\tilde{N}_r)$ which is applicable to the entire curve only if $A^2/N \leq 2$.

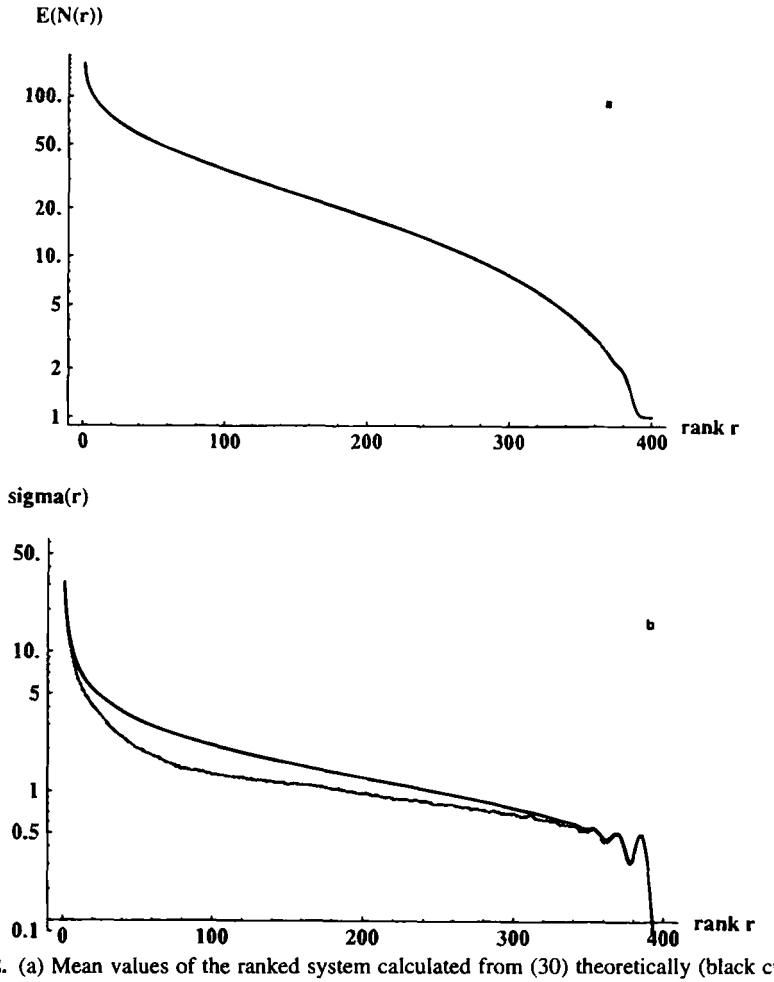

Fig. 2. (a) Mean values of the ranked system calculated from (30) theoretically (black curve) and compared with a numerical simulation of the corresponding system (gray curve). Simulation data are from a Markov model (2)–(4) with $c(N) = 0$ and $n_k = 1\ \forall k = 1, \ldots, A$, leading to the same probability distribution (18) for all species. Deviations between theory and simulation are smaller than the width of the plotted curves. Parameters used are $A = 400$ and $N = 10{,}000$, and 1000 systems for the numerical determination of the mean values. (b) As in (a), but now for the standard deviations of the ranked system. Deviations are larger than in part (a) due to our assumption (11), which is only poorly satisfied for $A = 400$ species. However, the agreement is still satisfactory.

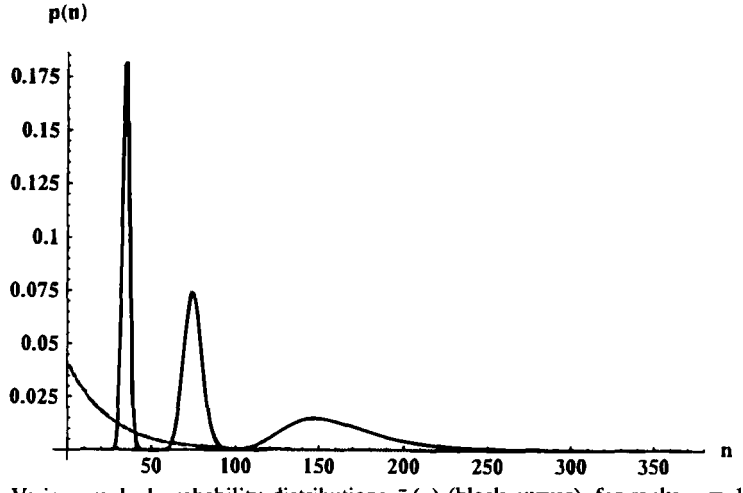

Fig. 3. Various ranked probability distributions $\bar{p}_r(n)$ (black curves), for ranks $r = 100$, 20, and 1 (from left to right), together with $p_k(n)$ (gray), which is the same for all species. Equation (B1) was used for calculating $\bar{p}_r(n)$; parameters used are $A = 400$ and $N = 10{,}000$.

When the opposite inequality is valid, the expected values display a remarkable "staircase" behavior for larger ranks, such that

$$r \geq \frac{4AN^2}{4N^2 + A^3} \tag{21}$$

The length of the steps and the width of the "steep" intervals are given by (B9) and (B10). It should be emphasized that this staircase behavior of the expected values appears as a result of ranking and stems from the breadth of the distribution (18) for the unranked variables. It has nothing to do with the steps in single empirical observations caused by the fact that the random variables considered take on only integer values.

Distributions of mean values of the form (19) can be found in empirical examples. They stem typically from systems where one has only a very limited number of features, e.g., if one considers ranking of letters, ranking of DNA triplets (Borodovsky and Gusein-Zade, 1989), or the ranking of small numbers appearing in texts [therefore it is sometimes called the "log-law of numbers" in Brokes (1982).

We stress the fact that in this case the ranking procedure produces a structure, even though the original system is completely void of it. Of course, there is some minimal structure in the single-species probability distribution, but no structure in the mean values. This structure has been generated by the ranking procedure itself. As a further example of this we have investigated a model where all single-species probability distributions are given by $p(n)$

$= 1/N$ for $n = 1, \ldots, N$. This leads to a different behavior; here we found $E(\tilde{N}_r) = N(A - r + 1)/(A + 1)$. Again we see that ranking introduces structures where there is none in the unranked (original) system.

5. RANKING IN ZIPF'S LAW MODEL

We return now to the case of the evolutionary model described in Section 2. We assume that the original random variables (the population numbers of different species) have exponential distributions (6) with unequal parameters a_k given by (7). Their expected values obey Zipf's law:

$$E(N_k) = a_k^{-1} = \left(\frac{cN}{k}\right)^{1-c} \tag{22}$$

As before, we consider now the set of new random variables $\{\tilde{N}_r\}$ which are populations of ranks.

As shown in Appendix C, where we derive an expression for a more general case of $E(N_k) = \beta k^{-\gamma}$, when the exponent in Zipf's law for the species population expected values $E(N_k)$ is very close to 1, i.e., when $c \to 0$ and $E(N_k) = A/k$, the expected values of ranked populations follow the same law for $r >> 1$:

$$E(\tilde{N}_r) \approx \left\lfloor \frac{A - 1}{r} \right\rfloor \tag{23}$$

but, in contrast with expected values for the unranked variables, $E(\tilde{N}_r)$ demonstrate a "staircase" behavior (similar to that of Section 4). More careful analysis leads to the conclusion that, in fact, the curve remains smooth up to larger ranks. The "steps" become visible if $r \geq 3A^{2/3}$ [cf. (C13)].

Departing from the model of Section 2, one can consider a situation when $E(N_k)$ remains large even for values of k close to A. Such a case seems to occur empirically, for instance, in the distribution of populations of cities and towns in a country [for a recent example, see Frankhauser (1991)], or in the case of texts with very limited vocabulary (e.g., Katsikas and Nicolis, 1990). The results derived in Appendix C can be used even in this case; however, we can get the $E(\tilde{N}_r)$ curves in these cases only numerically.

In such situations one should expect a steep decline of the rank expected values for highest rank $r \leq A$. This decline follows a logarithmic law:

$$E(\tilde{N}_k) \approx \log\left(\frac{A}{r}\right) \tag{24}$$

Calculations performed with the use of expressions (C5) and (C9) showed excellent agreement with empirical data; see Figs. 4 and 5.

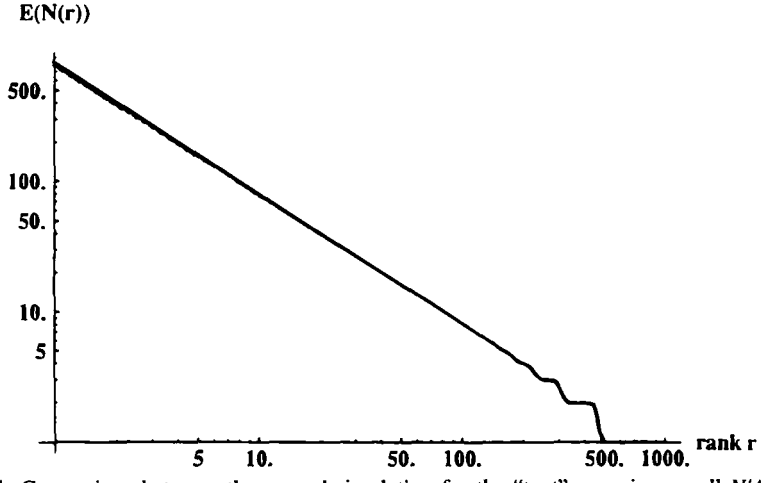

Fig. 4. Comparison between theory and simulation for the "text" case, i.e., small N/A ratio, where steps in the $E(\tilde{N}_r)$ curve are expected. Parameters used are $N = 6000$, $A = 1000$, and $\gamma = 0.95$. Shown are the $E(\tilde{N}_r)$ curve (black) as determined by equations (C5) and (C9) and an ensemble average (gray) of a numerical simulation. The numerical simulation does not use the Markov process (2)–(4), but a system where A independent exponential distributions with $E(N_k) = \beta k^{-\gamma}$ have been generated.

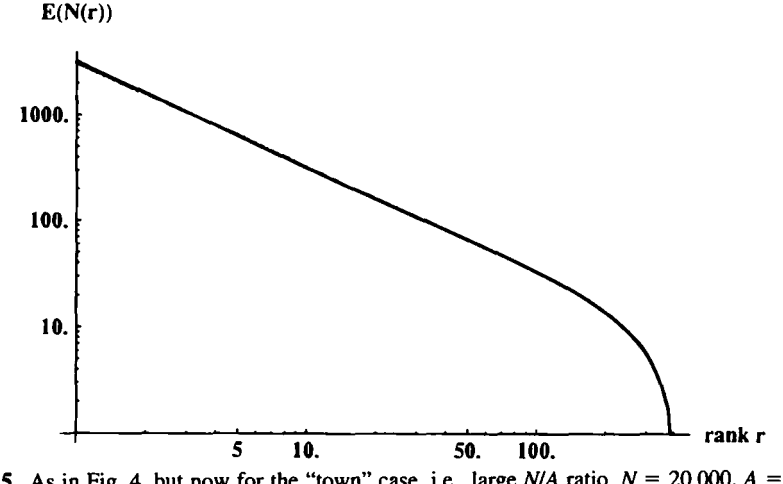

Fig. 5. As in Fig. 4, but now for the "town" case, i.e., large N/A ratio, $N = 20{,}000$, $A = 400$, and $\gamma = 0.95$. Plotted is the theoretical $E(\tilde{N}_r)$ curve (black) obtained by equations (C5) and (C9), compared to a numerical simulation (gray). The numerical data are obtained as discussed in the caption of Fig. 4.

6. SUMMARY OF THE RESULTS

In this article, we have investigated the effect of ranking procedures on probability distributions. We have found that ranking at its best is a way of reducing random fluctuations and narrowing broad probability distributions in order to get meaningful quantitative measurements in systems which are evolving, nonstationary, and exist in one realization only.

We have seen that in some cases one should be very careful in applying ranking to a given system. For instance, the system in Section 4, which lacks any structure, assumes a logarithmic dependence of the mean values on rank and a "staircase" behavior. An example of such a system is given by investigations of DNA triplets (Borodovsky and Gusein-Zade, 1989), where some authors (Katsikas and Nicolis, 1990) erroneously recognized Zipf's law, which deviates from the empirical results in Borodovsky and Gusein-Zade (1989). Because the DNA triplets are very similar to letters of an alphabet, not to words of a given text, the "log-law of numbers" is in agreement with our expectations for such a system. In general, based on the results derived in Section 3, one understands that such a logarithmic distribution of means after ranking is obtained when all possible combinations of sequences have equal probability. This was shown by Borodovsky and Gusein-Zade (1989), and seen also by Brokes (1982), who states, "In fact, I have come to regard conformity with this law as a test of homogeneity in social contexts."

Let us point out that the mean values arise from distributions which are given by equation (13) and which are shown to be sharply peaked, so that the observation of the logarithmic decrease is made possible. Thus no fundamental structure in the unranked means is needed to explain the result of frequencies of base sequences of DNA. The same argument applies also to frequencies of letters (as opposed to words).

In systems where one has a fundamental Zipfian structure of the mean values, we have shown ranking to be a meaningful procedure. As above, it produces sharp distributions out of diffuse ones, in this way amplifying the structure which exists in the original system. However, it works best only in the class of systems whose typical representative is the ranking of words in a given text. It turns out that in this case almost half of the ranks are filled with frequencies corresponding to appearances of words only once in the whole text! In this case the argumentation leading to equation (23) applies, and the Zipfian behavior of the original system is observed in the whole domain.

In all other cases, whose typical representative is given by the example of ranking of towns according to their number of inhabitants, equation (24) shows that at high ranks the Zipfian behavior is replaced by a logarithmic one, thus leading to a sharp dropoff in the usual log–log plot. Of course, this behavior is a consequence of the fact that mean numbers of inhabitants

are very large: there is simply not enough room at the end for the ranked means of higher ranks to lie on the Zipf curve, because all low ranks are shifted somewhat to higher values—the sum of those relatively small deviations suffices in absolute value to contain so much of the population that there is not as much left as would be needed for the amount necessary for the Zipf curve at the higher ranks.

In addition, there are relations among these three classes, especially between the log-law of numbers and the pure Zipfian. If the Zipf exponent in the original system is varied from one to zero, the logarithmic regime grows, finally leading to (for exponent zero) the log-law of numbers.

7. CONCLUSIONS

In the case of the Markov model and its extensions (Schapiro, 1994; Günther *et al.*, 1992, 1993; Levitin and Schapiro, 1993), the ranking procedure leads to a sharpening of the original broad distributions, which may serve as an example of a meaningful application of ranking. In related work (Günther *et al.*, 1993) we have introduced more general transition probabilities but still linear ones, which again lead to Zipf's law, so we have some evidence that there is a large set of transition probabilities in the space of all possible ones which can generate Zipf's law.

To sum up, one may suppose that the ubiquitous appearance of Zipf's law is based on two (independent) effects: First, the fact that very general transition probabilities lead to Zipf's law, a statement which gets additional support from the models of Shreider (1967), Mandelbrot (1983), and Li (1992), where Markov models of zeroth order were shown to be able to generate Zipf's law. This has to be compared to our models (Schapiro, 1994; Günther *et al.*, 1992, 1993; Levitin and Schapiro, 1993), which use first-order linear and nonlinear Markov processes for this purpose. Let us reiterate that our model 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. 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.

This robustness has been pointed out already in Mandelbrot (1983). Let us note at this point that ranking is not such a far-fetched procedure as might seem at first glance. We believe that this procedure is very important, for example, in animal and human perception. Note in this context Kohonen (1982), where it is shown that even very simple neural nets are able to perform a ranking procedure on a local basis, which leads after a number of iterations to the global ranking considered in this article. This opens an

additional line of thought, which is connected to the fact that the ranking procedure used in this article is a global transformation which reaches its fixed point in one step. We think that it would be very interesting to consider local transformations which act on nearest neighbors only and which need several steps to reach the fixed point of a ranked configuration.

This proposed local ranking procedure is reminiscent of well-known examples of ranking hierarchies in animal groups like apes or chickens. Exchanges in hierarchy take place on a local basis, not on a global one. The example of the animal group may in addition serve for the following analogy: restricted to the feature "rank," an animal's observation and social interaction has been sharpened to its place in the hierarchy alone, essentially independent of the wealth of features that are also present because of the individual character of each animal. It is essential for cooperativity in the group that any animal recognizes correctly its place in the hierarchy. In addition, this recognition leads to a sharpening of an animal's behavior, in that it has only to obey a restricted set of rules, the ones which are connected with its actual rank in the hierarchy.

Let us come back to the question of perception. We think that ranking gives animals and humans the ability to structure their perceptions, of course at the risk of observational artifacts. But this constraint must favor those distributions of observables which are not changed essentially by the ranking procedure. It is an important result of this work in our view that the empirically observed relation for mean values after ranking, the Zipf distribution, is an example of a distribution which satisfies this constraint: power-law distributions are stable under the procedure of ranking. Therefore one may speculate that there is a deeper level underlying these considerations:

- Structures like populations and texts to a high degree evolve in a cycle of creation and observation, re-creation and reobservation.
- Features which are favored from the point of view of observation, e.g., which are stable under procedures which are natural for, or at least constitute, observation, are likely to appear and be amplified during the (co)evolution of these structures.
- Zipf structures exactly fulfill the constraint of enhancing the structure and identifiability of features, as shown above.
- Thus a Zipf structure may evolve because basic mechanisms of observation favor exactly such a structure in an evolving complex system.

Therefore one can conclude that the appearance of Zipf's law is not caused by pure accident, but can be understood on a level which considers the interplay between laws in complex systems, which, unlike physical laws, may be evolving.

As pointed out in the example of the hierarchy in animal groups, in general one may speak of a coevolution of laws (behavioral ones) and objects (animals). We believe that for a further understanding of complex systems it will be necessary to take into account this interplay.

APPENDIX A

Let us derive the probability distributions for the rank variables $\tilde{N}_r$ (the size of the population of rank r; $r = 1, \ldots, A$) under the assumption of independence (11). We have

$$\begin{aligned}
\tilde{p}_r(n) &= \Pr\{\tilde{N}_r = n\} \\
&= \Pr\{\tilde{N}_1 \geq n, \ldots, \tilde{N}_{r-1} \geq n, \tilde{N}_r = n, \tilde{N}_{r+1} \leq n, \ldots, \tilde{N}_A \leq n\} \\
&= \sum_{i=0}^{r-1} \sum_{j=0}^{A-r} \Pr\{\tilde{N}_1 \geq n+1, \ldots, \tilde{N}_{r-i-1} \geq n+1, \tilde{N}_{r-i} = \cdots = \tilde{N}_r \\
&\qquad = \cdots = \tilde{N}_{r+j} = n, \tilde{N}_{r+j+1} \leq n-1, \ldots, \tilde{N}_A \leq n-1\} \\
&= \sum_{i=0}^{r-1} \sum_{j=0}^{A-r} \sum_{\{\pi^A_{r-i-1}\}} \sum_{\{\pi^{A-r+i+1}_{i+j+1}\}} \prod_{q=1}^{r-i-1} \Pr\{N_{k_q} \geq n+1\} \\
&\quad \times \prod_{s=r-i}^{r+j} \Pr\{N_{k_s} = n\} \prod_{t=r+j+1}^{A} \Pr\{N_{k_t} \leq n-1\} \qquad \text{(A1)}
\end{aligned}$$

where $\{\pi^l_m\}$ is the set of all possible choices of m elements from l elements.

We turn now to the special case (12) of exponential distributions for the species populations N_k:

$$\Pr\{N_k = n\} = p_k(n) = (1 - x_k)x_k^{n-1} \qquad \text{(A2)}$$

Then

$$\Pr\{N_k \geq n+1\} = x_k^n \qquad \text{(A3)}$$

$$\Pr\{N_k \leq n-1\} = 1 - x_k^n \qquad \text{(A4)}$$

Substituting (A2)–(A4) into (A1), after a series of manipulations, we arrive at the expression (13). Formula (13) can also be conveniently proved by induction. Note that the rank variables $\tilde{N}_r$ are far from being independent, even though we assume the species variables N_k are independent. This can be easily discerned from (10).

Since distribution (13) is nothing but a linear combination of geometric distributions, the expressions (14) and (15) for the expected values and the variances of variables $\tilde{N}_r$ can be derived from (13) straightforwardly.

APPENDIX B

Consider the case when all random variables N_k, $k = 1, 2, \ldots, A$, have the same probability distribution (18). Then the general formula (13) yields

$$\begin{aligned}
\tilde{p}_r(n) &= \Pr\{\tilde{N}_r = n\} \\
&= \sum_{m=r}^{A} (-1)^{m-r} \binom{m-1}{r-1} \binom{A}{m} (x^{m(n-1)} - x^{mn}) \\
&= \frac{A!}{(r-1)!\,(A-r)!} \sum_{s=0}^{A-r} (-1)^s \binom{A-r}{s} \frac{1}{s+r} (x^{(n-1)(s+r)} - x^{n(s+r)}) \\
&= \frac{A!}{(r-1)!\,(A-r)!} \sum_{s=0}^{A-r} (-1)^s \binom{A-r}{s} \int_{x^n}^{x^{n-1}} z^{s+r-1}\, dz \\
&= \frac{A!}{(r-1)!\,(A-r)!} \int_{x^n}^{x^{n-1}} z^{r-1}(1-z)^{A-r}\, dz \\
&= \frac{A!}{(r-1)!\,(A-r)!} (B_{x^n}{}^{n-1}(r, A-r+1) - B_{x^n}{}^{n}(r, A-r+1)) \qquad \text{(B1)}
\end{aligned}$$

where $B_t(r, A - r + 1)$ is Euler's incomplete beta function, and $x = (N - A)/N$.

Hence, the expected values $E(\tilde{N}_r)$ are

$$\begin{aligned}
E(\tilde{N}_r) &= \frac{A}{(r-1)!\,(A-r)!} \sum_{i=0}^{\infty} \int_0^{x^i} z^{r-1}(1-z)^{A-r}\, dz \\
&= \frac{A!}{(r-1)!\,(A-r)!} \sum_{i=0}^{\infty} B_{x^i}(r, A-r+1) \qquad \text{(B2)}
\end{aligned}$$

Expression (B2) cannot be reduced exactly to a closed form. However, the behavior of the expected values of rank populations can be analyzed taking into account that the integrand in (B2) has a sharp maximum at $z = (r - 1)/(A - 1)$. The function

$$f(z) = \frac{A!}{(r-1)!\,(A-r)!} z^{r-1}(1-z)^{A-r} \qquad \text{(B3)}$$

can be viewed as a probability density function of a random variable Z, such that

$$E(Z) = \frac{r}{A+1} \qquad \text{and} \qquad V(Z) = \frac{r(A-r+1)}{(A+1)^2(A+2)} \qquad \text{(B4)}$$

The width of the distribution (B3) vanishes in the worst case $r \approx A/2$ as $A^{-1/2}$ with the increase of A. Thus, for large A,

$$\frac{A!}{(r-1)!\,(A-r)!}\,B_{x^i}(r, A-r+1) \approx \begin{cases} 0 & \text{if } x^i < \dfrac{r-1}{A-1} \\ 1 & \text{if } x^i > \dfrac{r-1}{A-1} \end{cases} \tag{B5}$$

It follows that the contribution to $E(\tilde{N}_r)$ is given almost exclusively by the terms in (30) for which

$$x^i > \frac{r-1}{A-1} \qquad \text{or} \qquad i > \frac{\log[(r-1)/(A-1)]}{\log x} \tag{B6}$$

with each term contributing almost exactly one. This leads to the expression (19) for the expected values:

$$E(\tilde{N}_r) \approx \sum_{i=0}^{\log[(r-1)/(A-1)]/\log x} 1 \approx 1 + \frac{\log(r/A)}{\log(1 - A/N)}$$

The approximation (20) for the variance $V(\tilde{N}_r)$ can be derived from (B1) in a similar way.

However, the approximation (19) is quite rough: it gives only the smoothed outline of the distribution of the expected values. Indeed, since the values x^i ($i = 0, 1, 2, \ldots$) form a discrete sequence of points, it follows from (B1) and (B5) that $E(\tilde{N}_r)$ should remain constant for all values of r such that $x^{i+1} < (r-1)/(A-1) < x^i$ and then experience a sudden jump (equal to 1) when $(r-1)/(A-1)$ becomes smaller than x^{i+1}. Thus, the expected values should display a characteristic "staircase" behavior. This "staircase" shape of the curve for $E(\tilde{N}_r)$ 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 $\{\tilde{N}_r\}$, due to the fact that the distribution (18) is very broad.

The length of the ith step of the "staircase" (counting from right to left) is given by $(\Delta r)_i = r_{\max}(i) - r_{\min}(i)$, where $r_{\max}(i)$ and $r_{\min}(i)$ are defined as follows:

$$\begin{aligned} \frac{r_{\max}-1}{A-1} < x^i < \frac{r_{\max}}{A-1} \\ \frac{r_{\min}-1}{A-1} < x^{i+1} < \frac{r_{\min}}{A-1} \end{aligned} \tag{B7}$$

Thus, the ith jumps occurs at the point

$$r_{\max}(i) = 1 + \lfloor x^i(A-1) \rfloor \tag{B8}$$

The length of the ith step is equal to

$$(\Delta r)_i = \lfloor x^i(A-1) \rfloor - \lfloor x^{i+1}(A-1) \rfloor \tag{B9}$$

In particular, the zeroth step {between ranks $r_{\max}(0) = A$ and $r_{\min}(0) = 1 + \lfloor [(N-A)/(N-1)](A-1) \rfloor$} has a length $(\Delta r)_0 = \lfloor (A-1)^2/(N-1) \rfloor$.

However, this "second approximation" is also inaccurate, since it ignores the fact that the width of the distribution (B3) is finite (and not infinitesimal). Due to this fact, the expected values curve will have no abrupt jumps, but rather intervals of steep change. The width of such intervals can be estimated from (B4) as

$$2A\sigma = 2A[V(Z)]^{1/2} = 2A\left[\frac{r(A-r+1)}{(A+1)^2(A+2)}\right]^{1/2} \approx \left[\frac{r(A-r)}{A}\right]^{1/2} \tag{B10}$$

This "staircase" picture disappears for lower ranks when the length of the step given by (B9) becomes approximately equal to the width of the interval of growth (B10). Thus, the steps are observed for ranks such that

$$r \geq \frac{4AN^2}{4N^2 + A^3} \tag{B11}$$

Hence, the steps disappear and the entire curve becomes smooth when

$$\frac{A^2}{N} \leq 2 \tag{B12}$$

APPENDIX C

Consider expression (13) for the probability distribution $\tilde{p}_r(n)$ of the ranked populations:

$$\tilde{p}_r(n) = \sum_{m=r}^{A} (-1)^{m-r} \binom{m-1}{r-1} \sum_{\{\pi_m^A\}} \left(1 - \prod_{i=1}^{m} x_{k_i}\right)\left(\prod_{i=1}^{m} x_{k_i}\right)^{n-1}, \tag{C1}$$

In this appendix we will derive approximate formulas for the case of the distribution (6), but with more general x_k than (7); we use instead

$$x_k = 1 - a_k = 1 - \frac{1}{E(N_k)} = 1 - \frac{r^\gamma}{\beta} \tag{C2}$$

with γ not necessarily ≈ 1. Of course, we can specify the parameters γ and β to match the ones imposed by the Markov model (2)–(4).

To calculate from (C1) and (C2) the moments of the distributions, especially the means and the variances, we have to introduce an approximation

to get rid of the sum over choices included above. Multiplying out the product in the sum in (C1), we can write down the distribution in the following way:

$$\tilde{p}_r(n) = F_r(n-1) - F_r(n) \tag{C3}$$

$$F_r(n) = \sum_{m=r}^{A} \binom{m-1}{r-1} (-1)^{m-r} \sum_{k_1<\cdots<k_m}^{A} \prod_{i=1}^{m} x_{k_i}^n \tag{C4}$$

The expectation value and the second moment can be expressed as

$$E(\tilde{N}_r) = \sum_{n=0}^{\infty} F_r(n) \tag{C5}$$

$$V(\tilde{N}_r) = \sum_{n=0}^{\infty} (2n+1) F_r(n) \tag{C6}$$

We can proceed further by approximating the multiple sum in the definition of $F_r(n)$ as

$$\sum_{k_1,\ldots,k_m}^{A} \prod_{i=1}^{m} x_{k_i}^n \le \binom{A}{m} \frac{(\sum_{k=1}^{A} x_k^n)^m}{A^m} = \binom{A}{m} (\Phi(n))^m \tag{C7}$$

where we have defined

$$\Phi(n) = \frac{1}{A} \sum_{k=1}^{A} x_k^n \tag{C8}$$

Note that while this approximation in itself is an upper bound for the sum, it turns out that for the whole expression $F_r(n)$ this is not so, because the definition of $F_r(n)$ involves an *alternating* sum over these terms.

Inserting this approximation, we arrive at

$$F_r(n) = \int_0^{\Phi(n)} d\xi \, A \binom{A-1}{r-1} \xi^{r-1} (1-\xi)^{A-r} \tag{C9}$$

which is just the definition of the incomplete beta function $B_{\Phi(n)}(r, A-r+1)$.

We point out that this expression is exact in the case of equal unranked mean values, i.e.; if $E(N_k) = N/A \ \forall k$, and leads to results derived in Appendix B.

For the special case of the Markov model (2)–(4), in the limit case $c \to 0$, $\beta = A$, we can find an approximate expression for $E(\tilde{N}_r)$:

$$E(\tilde{N}_r) \approx \left\lfloor \frac{A-1}{r} \exp\left(-\frac{1}{r}\right) \right\rfloor \tag{C10}$$

The exponential factor in (C10) affects the small ranks only, and thus, for $r >> 1$

$$E(\tilde{N}_r) \approx \left\lfloor \frac{A-1}{r} \right\rfloor \tag{C11}$$

[It looks paradoxical that (C10) gives smaller values for small ranks than (C11). This is an artifact of our "all-or-nothing" approximation.]

Expression (C11) displays the Zipfian behavior of the expected values of rank variables in the "smooth" approximation. However, the floor function in (C11) recalls the "staircase" phenomenon. The length of the ith step now is

$$(\Delta r)_i \approx (A-1)\left(\frac{1}{i} - \frac{1}{i+1}\right) = \frac{A-1}{i(i+1)} \tag{C12}$$

Taking into account (B10), one can see that the steps appear, in fact, only for larger ranks, such that

$$r \geq 3A^{2/3} \tag{C13}$$

ACKNOWLEDGMENTS

It is a pleasure to thank Prof. O. E. Rössler and Prof. W. Ebeling for fruitful discussions.

REFERENCES

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

Borodovsky, M. Yu., and Gusein-Zade, S. M. (1989). A general rule for ranked series of codon frequencies in different genomes, *Journal of Biomolecular Structure and Dynamics*, **6**, 1001.

Brokes, B. C. (1982). Qualitative analysis in the humanities: The advantage of ranking techniques, in *Studies on Zipf's Law*, H. Guiter and M. V. Arapov, eds., Studienverlag Dr. N. Brockmeyer, Bochum, Germany.

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, Bochum, Germany.

Günther, R., and Wagner, P. (1995). Analysis of real texts, NRI Internal Report.

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

Günther, R., Schapiro, B., and Wagner, P. (1993). Critical specific complexity: Recent results, NMI Internal Report.

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 $1/f$ 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].

Schapiro, B. (1994). An approach to the physics of complexity, *Chaos, Solitons and Fractals,* **4**, 115–123.

Sharman, R. A. (1989). Observational evidence for a statistical model of language, IBM UKSC Report 205, September 1989.

Shreider, Yu. A. (1967). Theoretical derivation of text statistical features, *Problemy Peredachi Informatsii,* **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.