International Journal of Theoretical Physics, Vol. 31, No. 3, 1992

Physical Complexity and Zipf's Law

R. Günther,[1] B. Schapiro,[1] and P. Wagner[1]

Received October 1, 1991

This article deals with a measure of the complexity of a physical system recently proposed by Schapiro and puts it into the context of other recently discussed measures of complexity. We discuss this new measure in terms of a simple Markovian evolution model, extending and specifying the model given by Schapiro, which has the advantage of being analyically tractable. We find that the proposed complexity measure leads to interesting results: there exists a kind of phase transition in this system with a vanishing value of the probability c of generating a new species. This phase transition is related to a specific complexity of about 3 bits. By investigating decreasing c ($c \sim N^{-q}$, N the total number of individuals), we find that the complexity per species grows monotonically with q, diverging logarithmically with N as q goes to infinity.

1. INTRODUCTION

The last few years have seen a growing interest in what can be called complex systems. It seems that the theoretical tools of physics developed so far are only poorly able to treat really complex systems, and that new concepts are probably needed. To understand the nature of the problem, it is appropriate to start with the simplest examples of sufficiently complex behavior. It is commonly believed that the properties of chaotic, nonlinear systems represent the first fingerprint of complexity in physical systems. To be more specific: systems which display complex behavior hold a very delicate balance between order and disorder. In a recent investigation in the theory of cellular automata (CA), Langton (1990) found that the most complex behavior in a certain class of cellular automata is found "at the edge of chaos," i.e., between the class I and II CA on the one hand, and the chaotic class III CA on the other. In the CA, the role of stochastic or disordered behavior is played by the deterministic chaos which class III CA

[1] Naturwissenschaftliches und Medizinisches Institut an der Universität Tübingen in Reutlingen, D-7410 Reutlingen, Germany.

seem to display. We consider this line of thought in a related article (Günther *et al.*, 1991).

The first thing to do in trying to develop a physical theory for a new class of phenomena is to look for an observable which characterizes the properties in which one is interested.

Based on work by Kolmogorov (1965) and Shannon and Weaver (1959), who were the first to use the difficulty of forecasting the behavior of a given system as a measure of complexity, there have been many proposed measures of complexity; see, for instance, Chaitin (1966), Lempel and Ziv (1976), Wolfram (1984), Grassberger (1986), Hogg and Huberman (1986), Kaspar and Schuster (1987), Peliti and Vulpiani (1988), Crutchfield and Young (1989), Badii (1990), D'Allessandro and Politi (1990), Auerbach and Proccacia (1990), Badii *et al.* (1991), and Urías (1991).

In order to get a clear picture of the requirements such a measure should satisfy, we cite the following points made by Badii *et al.* (1991), who claimed that such a measure: (1) should be relative of the observer's ability; (2) should lie between order and disorder, i.e., should be zero if the system is totally ordered or if it is totally disordered; (3) should not be extensive; and (4) should not increase if one takes direct products of independent systems.

There are some remarks we would like to add to this list: The first point, the dependence of the measured complexity on the observer's abilities, has to be strengthened: the measure of complexity is not only relative to the observer's abilities, we conjecture that it is relative to her own complexity. This conjecture is based on the following considerations: It is known that observation always influences the quantity under observation and that as long as one deals with classical objects this influence can be made arbitrarily small. But the following subtleties arise if one tries to measure complexity.

Any observation consists in some process during which the observer interacts with the observed system. The measuring process itself is characterized by a certain complexity; this means that in contrast to conventional situations, the very quantity under study describes the process of studying. In the case of complexity, we want to characterize a property of the system which is also a property of the process of measurement itself.

Suppose, for instance, that the observer is a computer, who tries—equipped with some program—to analyze a given time series. The program on this computer has a certain complexity. It is obvious that the internal state of the computer changes in the course of the calculation. What we have in fact is a system composed of the computer program and the data set. The chances of this joint system are in a certain sense representations of the structures of the system under observation.

Thus, the measuring apparatus must be able to support a process of observation which is at least as complex as the process to be studied; if it cannot support such a process, the complexity of the measuring process will be smaller and hence the measured complexity will be smaller than the "true" complexity of the system at hand. Thus, the computer program which gives rise to structures of the demanded complexity must itself possess a certain minimal complexity sufficient to build up such structures during the interaction with the observed system.

Finally, complexity is *not* a local, additive quantity in the sense that measuring only parts of the system and adding the respective contributions will lead to a correct value for the complexity of the whole, as in a length measurement where we can use a small rule of 30 cm to measure arbitrary distances.

The second point of Badii *et al.*'s list is commonly believed. In the light of our considerations above we only add a note: if the observer knows exactly all about a system, then even for the very complex-looking behavior of chaotic systems or of cellular automata, the measure of complexity should be zero, too! We claim that the measure of complexity should be zero for deterministic systems and an arbitrary complex observer, which in Badii *et al.*'s terminology is an observer with arbitrarily high abilities (Badii *et al.* were aware of this fact, which they mentioned in a footnote).

The first two points are fulfilled by the measure proposed by Schapiro (1991), which is just the mutual information between two successive states of the system under consideration in the course of time and which will be called "physical complexity":

$$\begin{aligned} {}^{\mathrm{P}}C(t) &= \sum_{\{S_t, S_{t-1}\}} p(S_t, S_{t-1}) \ln \frac{p(S_t, S_{t-1})}{p(S_t) p(S_{t-1})} \\ &= H_{\mathrm{marg}}(t) - H_{\mathrm{cond}}(t \mid t-1) \qquad (1) \end{aligned}$$

This concept of complexity is based on the image of time itself as a sort of channel through which the successive states of the system communicate. If the state at time t corresponds of the "source," then the state at time $t+1$ corresponds to the "receiver" in Shannon and Weaver's terminology. This concept realizes the original intention of Kolmogorov to interpret complexity as a measure of the unpredictability of the "next" state in the generation process of a symbolic sequence.

The last two points of Badii *et al.*'s list have to be discussed further in the light of (1). In the model we consider below, we find that the complexity measure (1) is an extensive variable if considered as an ensemble average over such Markovian systems. However, if we normalize the entropies which constitute the physical complexity ${}^{\mathrm{P}}C(t)$ of the number of species in our simple evolution process, or equivalently to the number of features, we get an intensive variable which displays a very interesting behavior. In the case of the evolution model it is easy to speak about features or species. One can imagine, however, that there exist systems in which there is a large amount of arbitrariness in deciding what constitutes a "feature." This is again a point where the observer comes into play. The problem in the case of a conventional physical measurement is not to find features, because any observation is based on the *a priori* definition of observables and this definition induces a "minimal" definition of what constitutes a feature; in a general context, however, it may be difficult to find or to define observables. But in many cases at least the choice of the "resolution" may be at the observer's disposal, and can determine the value of the measured complexity. Consider, for example, a length measurement: the observable in this case is obviously length, but what is arbitrary is the resolution at which the system has to be considered; and both the observable and the resolution determine the feature. This is, in our opinion, a property which cannot be circumvented for any reasonable measure of complexity. But the variation of the observed value depending on the resolution may itself lead to an interesting characterization of a complex system.

We now turn to the discussion of the proposed evolution model, and look at what one can learn from the behavior of a simple toy model for a complex system.

2. THE MODEL

We consider as a simple mathematical model for a complex system the following nonstationary Markov process. A population of A species, each consisting of n_i individuals, evolves according to the following transition probabilities:

1. The probability $c(N)$ for a new species to be generated is

$$P(A; N \rightarrow A+1; N+1) = c(N) \tag{2}$$

2. The probability for an already existing species to get a new individual is

$$P(n_i; N \rightarrow n_i + 1; N+1) = \frac{[1-c(N)]n_i}{N} =: \frac{\gamma(N)n_i}{N} \tag{3}$$

To ensure that at each time step there is exactly one individual born, we choose the boundary condition

$$n_i(1) = \delta_{i1} \Rightarrow \sum_{i=1}^{N} n_i = N \tag{4}$$

Thus, we have at time step N exactly N individuals in our population.

To calculate the complexity of the process, we need the total probability distribution $P(n_1, n_2, \ldots, n_N; N)$. This expression contains all possible information about our system and is difficult to evaluate. We note, however, that the different species of the population interact only by means of the boundary condition. If the mean number of species is large, but much smaller than the total number of individuals, this interaction can be neglected and the total probability distribution factorizes:

$$P(n_1, n_2, \ldots, n_N; N) \approx \prod_{i=1}^{N} P(n_i; N) \tag{5}$$

The probability distributions for a single species and for the number of species at a given time step N can be derived by solving the master equations

$$P(A+1; N+1) = c(N)P(A; N) + \gamma(N)P(A+1; N) \tag{6a}$$

$$P(n_i+1; N+1) = \frac{\gamma(N)n_i}{N} P(n_i; N) + \left(1 + \gamma(N)\frac{n_i+1}{N}\right)P(n_i+1; N) \tag{6b}$$

This is done in Appendix A. One gets for the probability distributions of species and for the distribution of individuals within species

$$P(A; N) = \frac{1}{(A-1)!}\, \partial_s^{A-1} \prod_{k=1}^{N-1} [1 + c(k)(s-1)]|_{s=0} \tag{7a}$$

$$P(n_i; N) = \sum_{t_o=k-1}^{N-1} P(k-1; t_o) \sum_{l=1}^{n_i-1} \binom{n_i-1}{l-1} (-)^{l-1} \prod_{k=t_o+1}^{N-1} \left(1 - l\frac{\gamma(k)}{k}\right) \tag{7b}$$

where $P(k, t_o)$ represents the probability that there are exactly k species at time t_o.

To investigate the model further, we must specify the generating parameter c. In general the probability to create a new species will be a function of time that has values in the unit interval. Values near one will lead to a system with almost as many species as individuals, whereas values near zero correspond to a system with a mean number of species much smaller than the total number of individuals.

We will consider two cases or the functional dependence of c on time:

1. A constant rate of production of new species.
2. A decreasing rate proportional to N^{-q}.

3. THE MODEL AT CONSTANT SPECIES PRODUCTION RATE

Despite the very complicated explicit expression for the general probability distribution, there are some quantities which can be evaluated quite simply. Thus, it is possible to calculate the expectation values for the number of species and the number of individuals within each species without getting into the trouble of explicitly handling complicated probability distributions. We have solved the master equations for the process in terms of the generating functions, so that we can use (A3b) of Appendix A to calculate the moments of the distribution. One gets in the simplest case of the first moment

$$\langle n_k(N)\rangle = \sum_{t=k-1}^{N-1} cP(k-1,t)\prod_{j=t+1}^{N-1}\left(1+\frac{\gamma(j)}{j}\right) \tag{8}$$

If one approximates now the first sum by taking the expectation value of the time at which species k is generated, one gets

$$\langle n_k(N)\rangle = \left(\frac{N}{T_k}\right)^{1-c} \tag{9}$$

Using the relation for the mean number of species at time N, which can be derived from equation (A3a),

$$\langle A\rangle = 1+(N-1)c \tag{10}$$

we obtain the expected time of generation of species k

$$T_k = 1+\frac{k-1}{c} \tag{11}$$

Inserting this value into (9), we get for the mean number of individuals

$$\langle n_i(N)\rangle = \left(\frac{cN}{i-1+c}\right)^{1-c} \tag{12}$$

This relation between rank (= species number) and number of elements (= number of individuals) is known as Zipf's law and is found in a wide range of contexts, from frequencies of words in a large text (Guiter and Arapov, 1982), to listings of towns according to inhabitants, and many more examples. For a list of references see Schapiro (1991).

Having solved the master equation for the evolution process, we can calculate the probability distribution. The probability distribution for species with number higher than one is approximated by (see Appendix A)

$$\begin{aligned} P_k(n_k;N) &= a_k(1-a_k)^{n_k-1} \\ a_k &= \left(\frac{k-1}{cN}\right)^{1-c} \end{aligned} \tag{13}$$

Using these results, we can calculate the complexity of the evolution process. Starting with the expression for the entropy of species k, which can be evaluated exactly to give

$$H_k = -\sum_{n_k=1}^{N-1} a_k(1-a_k)^{n_k-1} \ln[a_k(1-a_k)^{n_k-1}]$$

$$= -\left[\ln a_k + \frac{1-a_k}{a_k}\ln(1-a_k)\right] \tag{14}$$

and summing over k, one obtains in the limit of large N and small c

$$\frac{H_{\text{marg}}}{\langle A\rangle} = \frac{(1-c)N}{1+(N-1)c}\left[\psi\left(1+\frac{c}{1-c}\right) - \psi(1)\right] \tag{15}$$

where $\psi(x)$ is the polygamma function $d\ln[\Gamma(x)]/dx$.

To calculate the complexity, we must add the conditional entropy, which is given by

$$H_{\text{cond}} = \sum_{k=1}^{N-1} H_k^{\text{cond}}$$

$$H_k^{\text{cond}} = \sum_{n_k=1}^{N-1} P(n_k; N)[p_{n_k}\ln p_{n_k} + (1-p_{n_k})\ln(1-p_{n_k})] \tag{16}$$

$$p_{n_k} = (1-c)\frac{n_k}{N}$$

Taking again $N \to \infty$ and $c \ll 1$, this can be evaluated to give

$$\frac{H^{\text{cond}}}{\langle A\rangle} = \frac{1-c}{cN}\ln N \tag{17}$$

That means that the conditional entropy vanishes for constant c in the limit of large N. The complexity is then given by the first term in (1).

We note that the maximal complexity, according to (15), regarded as a function of the parameter c is attained at $c = 1/\sqrt{N}$. The numerical value at the maximum is $\pi^2/6$, which is about 2.37 bits per species. This value is an upper bound.

4. THE MODEL AT DECREASING SPECIES PRODUCTION RATE

We now turn to the case of decreasing c. The actual form of c as a function of time is of marginal interest—we choose it to be $c(N) = bN^{-q}$, where q varies between 0 and ∞. If $q = 0$, we obtain the already discussed case of constant production rate. We will be interested in the behavior of the complexity for higher values of q. Intuitively one expects that if q grows unbounded, there will be only a small number of species being produced. In the extreme case of only one species surviving, one should expect the complexity to vanish. For $q=0$ we have already obtained the result of nonvanishing complexity. It will be of interest to see if the complexity increases without bound if q grows, or if there is some maximum at a certain value of q beyond which the complexity decreases.

We first calculate the expectation value of the numbers of individuals n_i. In the same way as before, we get

$$\langle n_k\rangle = N\left(\frac{b}{k(1-q)}\right)^{1/(1-q)} \exp\left[-\frac{b}{qN^q}\left(\frac{<A>}{k}\right)^{-q/(1-q)}\right] \tag{18}$$

We again find a Zipf-like behavior, but this time with the exponent $1/(1-q)$ slightly larger than unity. This is a more conventional case for actual examples, such as the rankings of texts (Guiter, 1982). We note that the frequencies n_i/N do not vanish in the limit of large N, in contrast to the case of constant c; we interpret this behavior as the ability of the system to conserve some kind of a "minimal structure."

Turning to the evaluation of the complexity, we use the solution of the master equations (6). We approximate the distribution of generation times by the mean value:

$$\langle A\rangle \approx \frac{b}{1-q} N^{1-q} \quad \Rightarrow \quad <t_o> \approx \left((1-q)\frac{k}{b}\right)^{1/(1-q)} \tag{19}$$

Inserting this into the solution (7b), we finally obtain for the probability distribution of species k

$$\begin{aligned} P_k(n_k\,; N) &= a_k(1-a_k)^{n_k-1} \\ a_k &= \left(\frac{k}{\langle A\rangle}\right)^{1/(1-q)} \exp\left[-\frac{b}{qN^q}\left(\frac{\langle A\rangle}{k}\right)^{q/(1-q)}\right] \end{aligned} \tag{20}$$

To evaluate the complexity, we need the marginal and the conditional part. As before [see (16)], we can derive the conditional part:

$$\frac{H_k^{\text{cond}}}{\langle A\rangle} = \frac{\ln N}{N^{1-q}}\left(\frac{(1-q)^{(1-2q)/(1-q)}}{q}\right) \tag{21}$$

This equation, which is valid for $q \leq 1/2$, means that in the limit of large N this does not contribute to the complexity.

The marginal part can be evaluated analytically under the assumption that the exponent in a_k does not vary much in the region of interest for k, which seems reasonable, since it can be written

$$\frac{b}{qN^q}\left(\frac{\langle A\rangle}{k}\right)^{q/(1-q)} \approx \frac{b^{1+q}k^{q/(q-1)}}{q(1-q)^q} \tag{22}$$

If k is varying, for instance, from 10 to ∞, this expression varies between 0.1 and 0, so that the exponent in (20) has its minimum at 1.0 and its maximum at 1.1. (Note that because we sum over infinitely many k values, a finite number of terms with k smaller than a given value do not contribute.)

Summing over all n and all allowed values of k, we find for the complexity per species

$$\frac{C}{\langle A\rangle} = \frac{H}{\langle A\rangle} \approx \frac{\Psi(1)-\Psi(1-q)}{q\ln 2} \tag{23}$$

The numerical values for the physical complexity ${}^{\mathrm{P}}C$ given by this formula vary from ${}^{\mathrm{P}}C=2.37$ bits for $q=0$ to ${}^{\mathrm{P}}C=4.0$ bits at $q=0.5$. This means that this time we find the value of 2.37 as a lower bound for the complexity of the evolution model. As we mentioned above, the model with time-dependent c is able to conserve a minimal structure (the ranking hierarchy) in the course of time; this has to be compared with the constant-c model, where the Zipf structure is lost for large times.

Because we obtained the value of 2.37 bits as an upper bound in this system and the same value as a lower bound in the structure-conserving system, one may interpret this specific complexity as a critical value necessary to conserve any structure, which is, for example, a condition for a system to be able to do computation.

We are interested in the behavior of the physical complexity as a function of q. Note that for $q>1$ the expectation value of the species number $\langle A\rangle$ remains finite. With the form for $c(N)$ as chosen above, in the limit $q\to\infty$ the value of $\langle A\rangle$ approaches 2. In this case the two species are generated in the first two time steps; the model consisting of two species can be solved exactly (see Appendix B); we find that in this case the specific marginal entropy grows proportional to $\ln(N)$, while the specific conditional entropy is just 1 bit. In this case even the specific entropy becomes infinite with $N\to\infty$. So this model does not show the kind of phase transition which can be found in dynamical systems such as CA and the logistic map (Crutchfield and Young, 1989), where the complexity is largest at intermediate values of the entropy. We will mention this point in the last section, where we discuss possible implications.

5. NUMERICAL RESULTS

In order to check the analytical results obtained above, we performed numerical simulations of an ensemble of such Markovian systems. In most cases we used an ensemble which consists of 1000 systems whose evolution we followed for $N = 20{,}000$ time steps. It is straightforward to calculate from such simulations the expectation values of the number of individuals per species, i.e., Zipf's law. The results are shown in Figure 1 for a system with constant $c(N)$, compared with the theoretical curve given by (12). In Figure 2 we simulate a system with nonconstant $c(N)$ and compare the results with (18). It can be seen that the numerical results are well described by Zipf's law, i.e., (12) and (18). The numerical values turned out to be slightly larger than the theoretical ones, which is easily explained by the finiteness of the simulations.

It is also easy to calculate the probability distributions for the individual numbers of a single species, which is shown in Figure 3 for the distribution for the first species and in Figure 4 for the fifth species in the case of constant $c(N)$. Also included in both figures for comparison are the theoretical curves given by (A4) and (13). Again the agreement between theory and numerical experiment is excellent.

It is more difficult to calculate the complexity. If the species number is high enough, the probability distributions are well described by only a few different individual numbers for these species. Therefore one obtains the entropies just by summation over the probability distribution so obtained. However, for low species number, because of the low number of systems

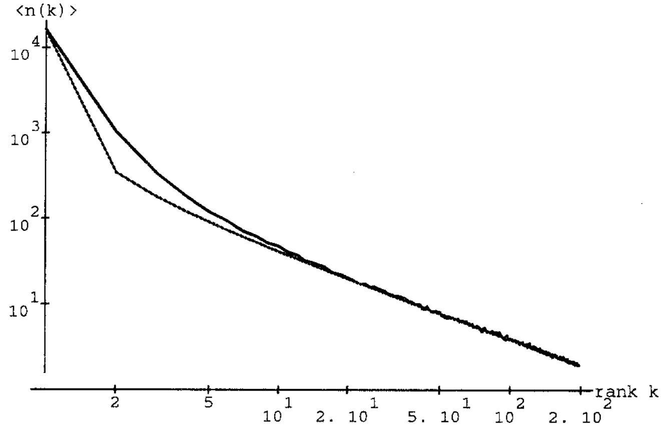

Fig. 1. Comparison between the numerically determined frequency-rank relation (solid curve) and the theoretical one (light curve), as given by equation (15). We choose $c(N) = c = 0.02$, $N = 20{,}000$.

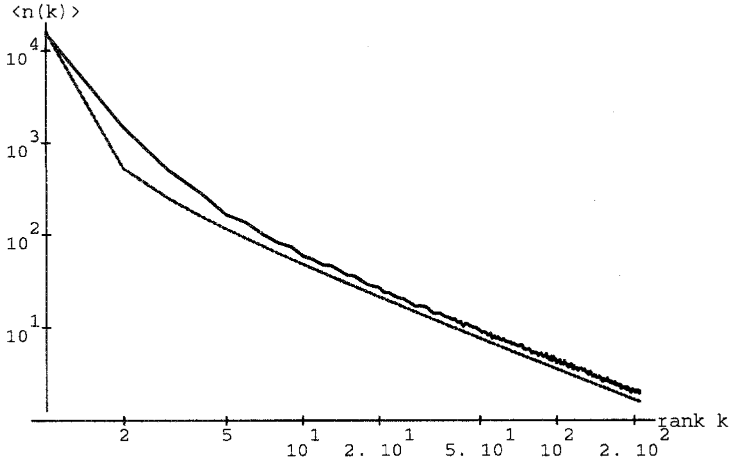

Fig. 2. As in Figure 1, but now with $c(N) = bN^{-q}$, where $b = 0.04$, $q = 0.08$, $N = 20{,}000$. The slope in the log-log plot is $\gamma = 1.09$.

and the large number of realizations, one always has to be aware of difficulties with the calculation of the probability distributions. The method we choose in the numerical simulations to approximate those distributions is to bin the obtained individual numbers. In this way, the values of the entropies depend on the number of bins chosen to approximate the probability distributions. To avoid effects due to bad statistics in each individual bin, one cannot choose an arbitrarily large number of bins. Just summing up the probabilities into the bins leads to an entropy which is always less than the "true" entropy one would have obtained in the limit of infinitely many

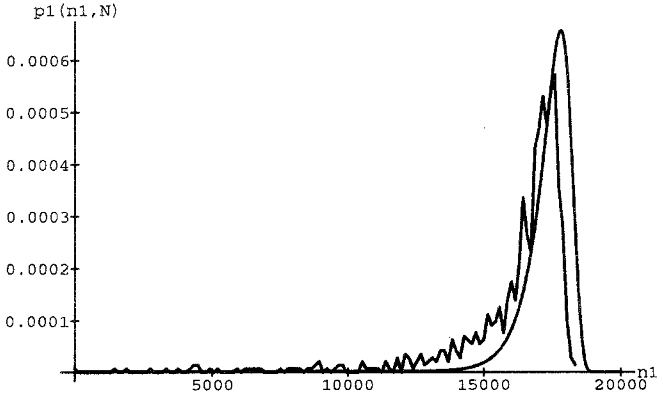

Fig. 3. Probability distribution for the first species, compared with the theoretical function (16) (smooth curve). Parameters are $c(N) = c = 0.02$, $N = 20{,}000$, and number of systems 1000.

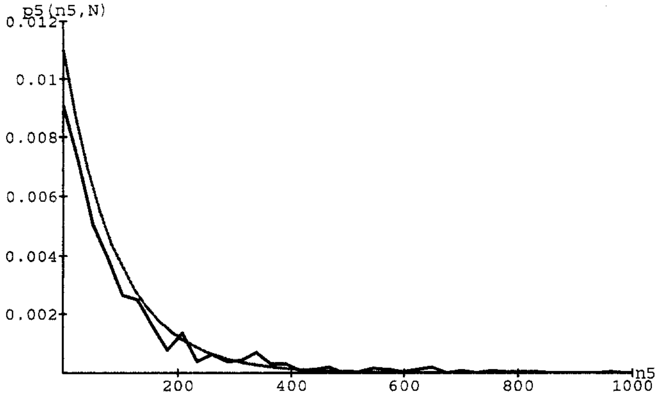

Fig. 4. As in Figure 3, but now for the fifth species, compared with the theoretical function (17). Parameters are as in Figure 3.

systems. One can correct for this effect if one assumes that within one bin one has a uniform distribution, thus adding to the entropies obtained by summing over bins a factor logarithm of the number of possible realizations/bin, which leads to an overestimation of the entropies. In Figure 5 we plot the numerically determined specific complexity for several numbers of bins; except for the overestimated case, the values are always below the theoretical value. However, it seems that the behavior of the numerically determined

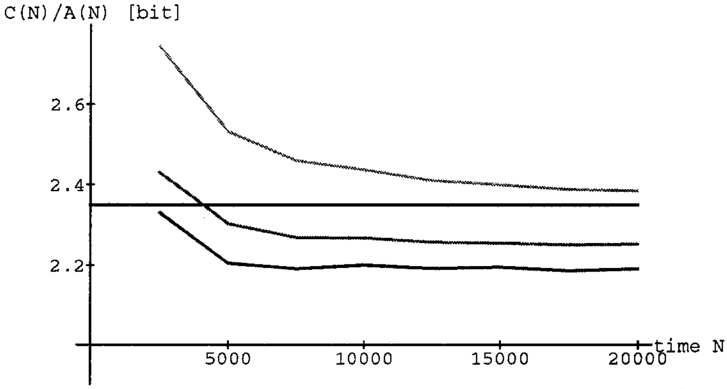

Fig. 5. Numerically determined specific complexity, calculated with 64 bins to approximate the probability distribution (lowest curve), with 128 bins, and under the assumption of equiprobability into the bins (upper curve), compared to the theoretical specific complexity (straight line) with $N \to \infty$. Parameters are as in Figure 4.

complexity is in good agreement with the analytically determined value. One can do the same procedure for the time-dependent $c(N)$, which is shown in Figure 6.

As a preliminary numerically obtained result we add that this model displays one further interesting behavior: as can be seen, for instance, in Figure 4, the probability distribution for the individual numbers for one species is very broad. In fact it can be shown that the standard deviation of the individual number grows proportional to the mean value. In systems like texts, for instance, this is a totally unexpected behavior, because an experimental verification of Zipf's law would not be possible if the distributions were that broad; it would be very improbable to get a correct mean value (which is necessary to find Zipf's law) by just considering one realization, as is done in investigations of texts. In the case of the model under consideration, we have "ranked" the individual numbers according to the creation time of the species. However, this may not be a good way to obtain Zipf's law. If one just sorts the systems under consideration according to their real, actual individual number, regardless of their creation time, then one again obtains Zipf's law (now with an exponent larger than unity—as already mentioned, the more conventional case), which can be seen in Figure 7. In addition, as one calculates the probability distribution for a system ranked according to the individual numbers, then the probability distribution of each rank becomes a very narrow one, as can be seen in Figure 8 for species number 25 versus rank number 25. Thus, the ranking has a severe effect on the probability distributions. The analytical results concerning this effect and also a more complete discussion will be published elsewhere (Wagner *et al.*, 1991).

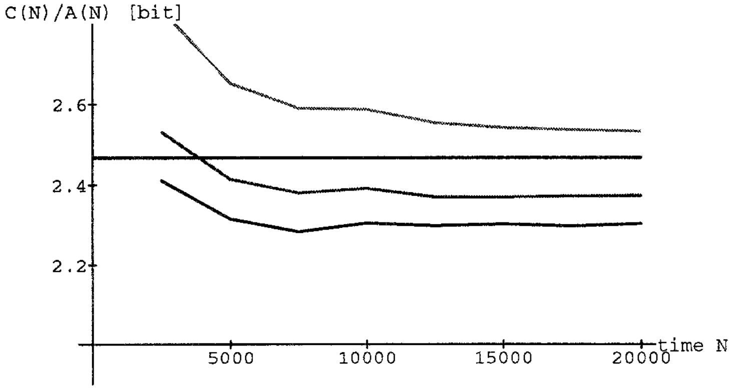

Fig. 6. As in Figure 5, but with $c(N)$ given by $c(N)=bN^{-q}$, with $b=0.04$, $q=0.08$, $N=20{,}000$.

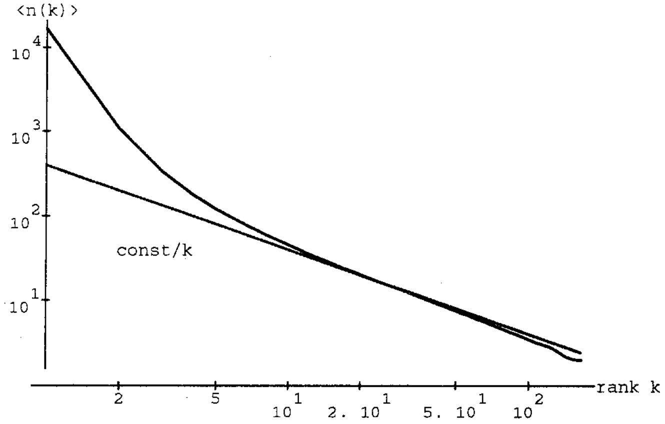

Fig. 7. Plot of the frequency-rank distribution for the system with constant $c(N)$, ranked according to number of individuals. The straight line represents a curve with slope exactly 1.0, while the numerically determined (solid) line has a slope significantly greater than 1; in this case one gets approximately $\gamma = 1.08$.

6. CONCLUSIONS

In this paper we used a very simple Markovian evolution model to get acquainted with the properties of the mutual information between two temporally successive states of a process as a measure of complexity.

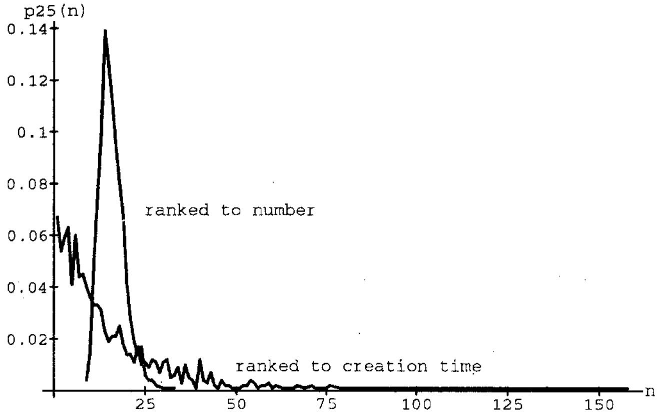

Fig. 8. Comparison between the probability distribution for species 5 (according to the creation time) and the probability distribution for rank five, which is obtained by sorting each system into the ensemble according to the individual numbers. Note that the distribution has become very narrow.

The analytical results for this system are new and interesting in their own right, but we focused our attention on the proposed measure of complexity. The parameter which determines the behavior of the system is the probability for the creation of a new species $c(N)$. We considered two cases: a constant value which in addition is taken to be much smaller than one (Section 3), and c as a decreasing function of the number of individuals or equivalently of time (Section 4). The most important results are the following:

1. Our model shows in both cases of constant and of decreasing generating parameter a relation between the mean number of individuals and the number of the corresponding species which is known as Zipf's law. This means that the number of individuals of a given species is essentially inversely proportional to the rank of that species. This behavior can be observed in a wide range of contexts [for a list of references see Schapiro (1991)]. In contrast to the quite general nature of our model, which does not specify what properties the species or features must have, most of the attempts to explain Zipf-like behavior use more stringent assumptions (Guiter and Arapov, 1982).

2. The complexity at a fixed time, regarded as a function of c, has a maximum at $c = 1/\sqrt{N}$. Because the complexity stays finite at its maximum and vanishes identically at $c \equiv 0$, the function $C(c)$ is, in the limit $N \to \infty$ (which in our model is the analogue of the "thermodynamic" limit), not analytical at $c = 0$. This nonanalyticity is due alone to the marginal entropy of the system, and thus reminds one of the Bose–Einstein condensation in thermodynamics. Indeed, this system can be described in terms of boson creation and annihilation operators (Günther, 1991), which establishes the above-mentioned similarity to Bose–Einstein condensation. We note in addition that the number of species at the maximum of the complexity is just the square root of the number of individuals.

3. In the system with constant c, the Zipf structure is not preserved in the thermodynamic limit, which has to be compared with the result for time-dependent $c(N)$, where this structure *is* preserved. This can be seen in (18). One may speculate that the only systems able to exhibit really complex behavior are those in which there is some deceleration in the increase of the number of features. This speculation is confirmed by our next result:

4. The complexity of the system with decreasing $c(N)$ is *always* above the complexity of the system with $c(N) = \text{const}$. To be more specific, we found a critical value for the specific complexity in the following sense. For the model with constant c, the critical value of 2.37 bits is always an upper bond, while for the model with decreasing species production rate it is always a lower bond for the specific complexity of the system. We interpret this fact

in connection with a process of primitive learning: the deceleration of the production of new features (species) means the progressive deformation of the weight structure upon the existing features. Learning in this sense is the building of a reproducible structure of such a deformation, i.e., the building of a structure over structure. It is remarkable that the system *with* learning is, according to our model and to this interpretation, always more complex than that without learning.

5. Concerning the case of decreasing generation parameter, we have a second result, which is more difficult to understand: If we investigate the behavior of the system in the limit of very large values of q, we find a complexity increasing with the logarithm of time, just as is the case for the complexity of a simple Brownian particle (Schapiro, 1989), a result which is counterintuitive at first glance. But if we remember that the proposed complexity measure is a measure of the difficulty of forecasting, it makes sense: although Brownian motion is a very simple process, and so is the behavior of the Markovian evolution model for $q \to \infty$, the concrete realization in the limit of very large times is very hard to forecast, in spite of the fact that the immediate future of the system can be estimated quite correctly, as the small contribution of the conditional entropy suggests.

Therefore, it may be useful to look more closely at the structure of the kind of phase transitions mentioned in the introduction, where the complexity has a large value at intermediate values of the entropy. In the system we considered here, the contribution to the complexity in the limit of large N is only due to the marginal entropy. If one considers systems where the conditional entropy plays a more dominant role, then we expect a behavior similar to what has been observed in Crutchfield and Young (1989) and Langton (1990). It is not difficult to construct a model similar to that investigated in this paper, where one can understand the interplay between marginal and conditional entropy (Günther *et al.*, 1991).

However, this example shows that there are many conceptual questions which remain to be discussed in terms of models and real systems about the properties of complex systems.

Let us make a final remark. The use of Markovian processes also leads to a connection between this model and the theory of nonlinear dynamical systems, which can be described by Markov chains of finite but arbitrary order. It would also be very interesting to look at such systems and see what can be learned about them in the light of processes of the type considered in this paper. As a first approach, we mention Nicolis and Nicolis (1990) and Katsikas and Nicolis (1990), who succeeded in mimicking Zipf's law with a generating partition for the logistic map. Further investigations in this field are a possible line of research.

APPENDIX A. THE PROBABILITY DISTRIBUTIONS FOR DIFFERENT SPECIES

Introducing the generating functions

$$G_A(s; N) = \sum_{A=1}^{N-1} S^A P(A; N) \tag{A1a}$$

$$G_i(s; N) = \sum_{n_i=1}^{N-1} S^{n_i} P(n_i; N) \tag{A1b}$$

we have the corresponding master equations [cf. (6a), (6b)]

$$G_A(s; N) = [1 + (s-1)c(N)] G_A(s; N) \tag{A2a}$$

$$G_{n_i}(s; N) = G_{n_i}(s; N) + \frac{s(s-1)}{N} \gamma\, \partial_s G_{n_i}(s; N) + scP(k-1; N-1) \tag{A2b}$$

These equations are solved quite easily and one gets for the generating functions

$$G_A(s; N) = S \prod_{k=1}^{N-1} [1 + c(k)(s-1)] \tag{A3a}$$

$$G_k(s; N) = \sum_{t=k-1}^{N-1} cP(k-1; t) \sum_{l=0}^{\infty} \left(\frac{1-s}{s}\right)^l (-1)^l \prod_{j=t+1}^{N-1} \left(1 + \frac{l\gamma(j)}{j}\right) \tag{A3b}$$

$$= \sum_{t=k-1}^{N-1} cP(k-1, t) \sum_{l=1}^{\infty} \left(\frac{s}{1-s}\right)^l (-1)^{l-1} \prod_{j=t+1}^{N-1} \left(1 - \frac{l\gamma(j)}{j}\right) \tag{A3c}$$

Because of the starting condition, species number one has a behavior very different from the rest of the population. One finds for the probability that this species has n_1 individuals

$$P_1(n_1; N) = \frac{1}{c\sqrt{\pi}} \kappa (\kappa n_1)^{(1/2c)-1} \exp[-(\kappa n_1)^{1/c}]$$
$$\kappa = \frac{1}{(N-1)^{1-c}} \tag{A4}$$

This holds for $c \ll 1$. This result shows that the limit $c \to 0$ cannot be derived by a perturbation-theoretic approach, because (A4) is not analytical at $c = 0$. The distribution for higher species is given by (13).

APPENDIX B. THE LIMIT $q \to \infty$

In this limit there can be at most two species in the population. If we choose the constant b in $c = bN^{-q}$ equal to unity, then we have just a new boundary condition; after two time steps we have two species, each having one individual. Because the probability to create a new individual is proportional to the number of those already generated, the probability distribution can be written in a simple form.

We first notice that, according to the definition (3), at the generation of the (n_i+1)th individual of a species we get a factor n_i. Thus, after the n_kth individual is created we have a total factor of $(n_k-1)!$. Comparing this with the property of the two-dimensional integral in complex space

$$\int [d\xi\, d\bar{\xi}]\, e^{-\xi\bar{\xi}}\xi^k\bar{\xi}^l = \delta_{kl}k! \tag{B1}$$

we can write the probability distribution for the system with two species as

$$P(n_1, n_2; N) = \int \prod_{k=1}^{2} [d\xi\, d\bar{\xi}] \left[\exp\left(-\sum_{k=1}^{2} \xi_k\bar{\xi}_k\right)\right] \prod_{j=1}^{2} \xi_j^{n_j-1} \frac{(\bar{\xi}_1+\bar{\xi}_2)^{N-2}}{(N-1)!} \tag{B2}$$

With the Gaussian measure we have the following relations:

$$\begin{gathered} [\partial_{\xi_k}, \xi_l] = \delta_{kl} \\ \int [d\xi\, d\bar{\xi}]\, e^{-\xi\bar{\xi}}\xi f(\xi) g(\bar{\xi}) = \int [d\xi\, d\bar{\xi}]\, e^{-\xi\bar{\xi}} f(\xi)\, \partial_{\bar{\xi}} g(\bar{\xi}) \end{gathered} \tag{B3}$$

Using this as the expression, we can evaluate (B2):

$$P(n_1, n_2; N) = \frac{1}{(N-1)}\, \delta_{n_1+n_2, N} \tag{B4}$$

This means we simply have a uniform distribution, which exists on the line $n_1+n_2=N$.

ACKNOWLEDGMENTS

It is a pleasure to thank Prof. L. Levitin for many enlightening discussions on the topics covered here. Discussions with A. Hoff are also gratefully acknowledged.

REFERENCES

Auerbach, D., and Procaccia, I. (1900). *Physical Review A*, **41**, 6602–6614.

Badii, R. (1990). Unfolding complexity in nonlinear dynamical systems, in *Measures of Complexity and Chaos*, N. B. Abraham *et al.*, eds., Plenum Press, New York.

Badii, R., Finardi, M., and Broggi, G. (1991). Unfolding complexity and modelling asymptotic scaling behaviour, in *Chaos, Order and Patterns*, P. Cvitanović, ed., Plenum Press, New York.

Chaitin, G. J. (1966). *Journal of the Association for Computing Machinery*, **13**, 547–560.

Crutchfield, J. P., and Young, K. (1989). *Physical Review Letters*, **63**, 105.

D'Alessandro, G., and Politi, A. (1990). *Physical Review Letters*, **64**, 1609–1612.

Grassberger, P. (1986). *International Journal of Theoretical Physics*, **25**, 907.

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

Günther, R. (1991). NMI Internal Report No 4/1991.

Günther, R., Schapiro, B., and Wagner, P. (1991). To be published.

Hogg, T., and Huberman, B. A. (1986). *Physica D*, **22**, 376.

Kaspar, F., and Schuster, H. G. (1987). *Physical Review A*, **36**, 843–848.

Katsikas, A. A., and Nicolis, J. S. (1990). *Nuovo Cimento D*, **12**, 177–195.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information, *Problems in Information Transmission*, **1**, 1–7.

Langton, C. G. (1990). *Physica D*, **42**, 12–37.

Lempel, A., and Ziv, J. (1976). *IEEE Transactions Information Theory*, **22**, 75.

Nicolis, G., and Nicolis, C. (1990). *Physica A*, **163**, 215–231

Peliti, L., Vulpiani, A., eds. (1988). *Measures of Complexity*, Springer-Verlag, Berlin.

Schapiro, B. (1989). Unpublished work.

Schapiro, B. (1991). An approach to the physics of complexity, *Journal of Nonlinear Biology*, to appear.

Shannon, C. E., and Weaver, W. (1959). *The Mathematical Theory of Communication*, University of Illinois Press.

Urias, J. (1991). *Physica D*, **47**, 498–508.

Wagner, P., Schapiro, B., and Günther, R. (1991). To be published.

Wolfram, S. (1984). *Communications in Mathematical Physics*, **96**, 15 (1984).