Rank distributions of words in correlated symbolic systems and the Zipf law

K. E. Kechedzhi, O. V. Usatenko,* and V. A. Yampol'skii

A. Ya. Usikov Institute for Radiophysics and Electronics, Ukrainian Academy of Science, 12 Proskura Street, 61085 Kharkov, Ukraine

(Received 21 June 2005; published 28 October 2005)

The binary many-step Markov chain with the steplike memory function is considered as a model for the analysis of rank distributions of words in correlated stochastic symbolic systems. We prove that this distribution obeys the power law with the exponent of the order of unity in the case of rather strong persistent correlations. The Zipf law is shown to be valid for the rank distribution of words with lengths about and shorter than the correlation length in the Markov sequence. A self-similarity in the rank distribution with respect to the decimation procedure is observed.

DOI: 10.1103/PhysRevE.72.046138 PACS number(s): 02.50.Ga, 87.10.+e, 05.40.−a

I. INTRODUCTION

The rank distributions in the stochastic systems attract the attention of specialists in the physics and many other fields of science because of their universal power-law character, $W \propto R^{-\alpha}$ (the so-called Zipf law [1]). Here W is the frequency of some occurring object and R stands for the rank of this object. Discovered originally for the rank distribution of words in the natural languages, the Zipf law was later observed in the rank distributions of other objects, such as distributions of "words" in the DNA sequences [2], PC codes [2], seismology [3], capitals of stock market players [4] (in economics, the Zipf law in slightly different form is known as Pareto's principle or the 80-20 rule [5]), the population of cities, areas occupied by countries, masses of cosmic objects, etc. (see Refs. [6,7]). In spite of numerous attempts to describe this phenomenon analytically, a deep insight into the problem has not so far been gained. In the strict sense, the Zipf law implies the closeness of the exponent α to unity. Besides, the Zipf law should be valid within the fitting range of about 2–3 decades. From this point of view, the rank distributions in some above-mentioned real systems do not comply with these criteria. Therefore, we will use the term "Zipf's-like law" for the power rank distributions if α differs noticeably from unity.

To define the rank distribution $W(R)$ of some objects in a definite sequence, it is necessary to establish a correspondence between objects and their frequencies of appearing in a sequence and to arrange the objects in ascending order of these frequencies. Analyzing the rank distribution in a dynamic system one could find oneself faced with two possibilities. The first one is characterized by the power-law rank distribution as a result of some kind of probabilistic game established by *a priori* defined nonequivalence of words. The Mandelbrot models [8,9], Li model [10], and many other models are constructed on the basis of the choice of the *a priori* nonequivalent competitors, and specifically this nonequivalency is a reason for obeying the rank distributions to the Zipf law. For example, the nonequivalency of the words in literary texts is caused by their different lengths and, consequently, by their different statistical weight that is defined by a number of characters in the word and the capacity of an alphabet. To illustrate this we will consider an elementary example below. In the second possibility, the power-law rank distribution appears as a result of correlations in a system. In this case, the competitive objects are considered as *a priori* equivalent, i.e., having "the same rights" in the competition [11]. The rank distribution of the overlapping triplets in the DNA sequences can serve as a vivid example of the real systems for which this approach is essential (see Ref. [2]).

We suggest an analytically solvable model of the additive many-step Markov chain where the rank distribution of different L words (L is the consequent symbols in the chain) is examined. In order to increase the statistical sampling of the words, we consider all L words in the chain, that could overlap each other. A special numerical analysis shows that such a type of sampling does not effect on the rank distribution of the L words. Since the words are equal in their length, the rank distribution in this system occurs as a result of the correlations. In other words, our model is based specifically on the second approach to the choice of the competitors in the sequence.

We have analytically studied the rank distributions of words of the length L in the Markov chains. If the Markov chain possesses the steplike memory function considered in Ref. [14], this distribution is shown to be from the staircase form. For strong correlations, the envelope curve of the rank distribution obeys the power law with the exponent of the order of unity, i.e., the distribution is described by the Zipf's-like law. The obtained results allow us to express our view upon the origin of Zipf's law. In particular, we have made sure that the correlations of symbols within the competitive words are sufficient for the appearance of the Zipf's-like law in their rank distributions. Contrary to the speculations about the connection of the Zipf's-like law to the *long-range correlations* (with the scale of about the text length) that were expressed clearly by a number of authors [12,13], we have demonstrated that the *short-range correlations* (with the scale of about the word length L) can result in such a rank distribution.

The suggested approach to the problem of the Zipf law is expedient because we are provided with the theoretical parameters that affect both the character of correlations and the rank distribution of words occurring in the Markov chain.

*Electronic address: usatenko@ire.kharkov.ua

Due to this circumstance, we examine the relationship between the rank distributions and the correlation properties of the system.

The paper is organized as follows. In Sec. II, we examine an elementary example of the system possessing Zipf's-like rank distribution due to the difference in the length of competitive words. Further, in Sec. III, we introduce the formalism of the additive many-step Markov chains. In Sec. IV, we derive the Zipf's-like rank distribution of the words of equal lengths in the many-step Markov chain. In Subsec. IV A, we study the rank distribution for words shorter than the memory length of the Markov chain. We observe a remarkable property of scaling of the distribution in this case in Sec. IV B. In Sec. IV C we present the results of numerical simulations for the rank distribution of long words, longer than the memory length. We conclude with some general remarks in Sec. V.

II. ELEMENTARY EXAMPLE

This example is actually similar to the model developed by Li in Ref. [10]. Let us consider a homogeneous stationary unbiased binary sequence of symbols, $a_i=\{0,1\}$. The L word is defined as a set of sequential symbols of definite length L being taken at the arbitrary point of a sequence. The rank distribution of words is a relationship connecting the probability W of occurring a certain word to the corresponding rank. The words are ordered in ascending rank order, $W(1) \geqslant W(2) \cdots \geqslant W(2^L)$. This model represents an example of a system where the competitors have *a priori* nonequivalent "rights."

Consider a set of words of arbitrary lengths in a noncorrelated binary sequence including all one-symbol words, all two-symbol words, all three-symbol words, and so on up to some finite length $L_0 \gg 1$. The frequency of the appearance a concrete L word in the noncorrelated sequence is evidently proportional to 2^{-L},

$$W(L) \propto 1/2^L \tag{1}$$

and does not depend on the arrangements of the symbols in the word. This means that the frequency-rank plot is degenerated, i.e., it consists of steps, each of them corresponding to the set of words with the fixed length L. Let us derive a function $W(R)$ corresponding to the envelope curve passing through the right edges of each step on the plot. It is obvious that the length of the step corresponding to the L words equals 2^L, so the rank of the right edge point of this step is described by the equation

$$R(L) = \sum_{i=0}^{L} 2^i. \tag{2}$$

One can easily derive the function $W(R)$ from Eqs. (1) and (2),

$$W(R) = \frac{(\mathrm{const})}{R+1}, \tag{3}$$

i.e., the Zipf distribution.

In the next section, to avoid this trivial situation, we consider the ranking of the words of equal length L, i.e., to perform the so-called n-tuple Zipf-analysis [2] of a binary system. It is obvious that in the case of uncorrelated sequences the distribution is flat, i.e., all words have equal rank frequencies. Therefore, the correlations must be introduced into the systems under consideration.

III. MODEL

We examine the binary *N-step Markov chain*, which is determined by the *conditional probability* $P(a_i|a_{i-N},a_{i-N+1},\ldots,a_{i-1})$ of occurring the definite symbol a_i (for example, $a_i=0$) after symbols $a_{i-N},a_{i-N+1},\ldots,a_{i-1}$. We define this quantity by the equation

$$P(a_i=0|a_{i-N},a_{i-N+1},\ldots,a_{i-1}) = \sum_{j=i-N}^{i-1} f(a_j). \tag{4}$$

The value of N is referred to as the memory length of the Markov chain. Here $f(a_j)$ is the contribution of symbol a_j to the conditional probability. This contribution is assumed to be independent of the distance $i-j$ between the symbols if $i-j \leqslant N$. From here it follows that in this simple case the conditional probability depends only on the number k of unities among symbols $a_{i-1},a_{i-2},\ldots,a_{i-N}$ preceding the generated one, a_i,

$$P(a_i=0|a_{i-N},a_{i-N+1},\ldots,a_{i-1}) = P_k = 1/2 + \mu(1-2k/N). \tag{5}$$

Here $\mu=f(0)-1/2$ is the strength of correlations in the sequence, $-1/2<\mu<1/2$. The case with $\mu=0$ corresponds to the noncorrelated random sequence of symbols. At $\mu \to 1/2$, the correlations are extremely strong and the sequence consists of large blocks containing only unities or zeros with relatively small "domain walls" between them [15]. The positive (negative) values of μ correspond to the persistent (antipersistent) correlation [the attraction (repulsion) of symbols of the same kind]. We will consider the case of $\mu \geqslant 0$ only, the formalism in the opposite case $\mu<0$ has some distinctions.

Let us introduce a probability $b_N(a_1a_2\ldots a_N)$ of the occurrence of the certain N-word $(a_1,a_2,\ldots,a_N)$. One could easily verify that b_N satisfies the condition of compatibility for the Chapmen-Kolmogorov equation (see, for example, Ref. [16]):

$$b_N(a_1\ldots a_N) = \sum_{a=0,1} b_N(aa_1\ldots a_{N-1})P(a_N|a,a_1,\ldots,a_{N-1}). \tag{6}$$

Here we have 2^N homogeneous algebraic equations for 2^N probabilities b_N of occurring the N words. Besides, the normalization equation $\Sigma b_N=1$ should be satisfied.

In the case under study, the set of equations can be substantially simplified owing to the following statement. It was proved [15] that the probability $b_N(a_1a_2\ldots a_N)$ depends on the number k of unities in the N word only, i.e., it is independent of the arrangement of symbols in the word

$(a_1, a_2, \dots, a_N)$. This statement is valid owing to the chosen simple model (5) of the Markov chain.

In Ref. [15], Eqs. 6 were solved and then the distribution function of words with arbitrary length $L \leqslant N$ was derived

$$b_L(k) = b_L(0)\frac{\Gamma(n+k)\Gamma(n+L-k)}{\Gamma(n)\Gamma(n+L)}, \quad 0 \leqslant k \leqslant L \leqslant N, \tag{7}$$

with

$$b_L(0) = \frac{4^n}{2\sqrt{\pi}}\frac{\Gamma(1/2+n)\Gamma(n+L)}{\Gamma(2n+L)}.$$

This formula defines the probability of the occurring L-word containing k unities with a definite arrangement of symbols. This probability depends on the number k but does not depend on the arrangement of symbols in the word. The parameter $n > 0$,

$$n = \frac{N(1-2\mu)}{4\mu}, \tag{8}$$

governs the strength of the persistent correlations. The case $n \to \infty$ corresponds to the absence of correlations. Under this situation, the probability b_L does not depend on k. At $n \to \infty$, Eq. (7) yields $b_L = 1/2^L$ that coincides with the frequency (1) of the occurrence of the L words in the noncorrelated sequence. The case $n \ll 1$ conforms to the strong persistence. In the next section, we make use of Eq. (7) to derive the rank distribution of L words in the Markov chain under consideration.

IV. ZIPF'S-LIKE DISTRIBUTION

A. Distribution for $L \leqslant N$

At finite values of $n > 0$, the probability b_L depends on k. A simple analysis shows, that the plot $b_L(k)$ has a concave form because of the persistence ("attraction" of symbols to each other) in the system. The words in which unities (or zeros) are prevalent occur more frequently than the words with an approximately equal number of different symbols. Specifically this property provides the nonflat rank distribution of words with equal lengths in the Markov chain.

In order to clarify the form of the rank distribution, we note that the specific degeneration of the frequency of the words occurrence takes place. The probability $b(k)$ (here and below we omit the subscript L) does not depend on the arrangement of symbols within the L word. This means that the rank distribution of the L words with $L \leqslant N$ should be of a staircase form. Yet another kind of degeneration results from the nonbias property of the sequence: the probability $b(k)$ is symmetric with respect to the change $k \leftrightarrow (L-k)$, $b(k) = b(L-k)$. Thus, as a result of these two symmetries, $2\mathrm{C}_L^k = 2L!/k!(L-k)!$ different words occur with the same frequency $b(k)$. Each of the steps can be labeled by the number $k \leqslant L/2$ of unities (or zeros) within them and is characterized by the length equal to the degeneracy multiplicity $2\mathrm{C}_L^k$. The right edge of the kth step corresponds to the rank $R_{max}(k)$,

$$R_{max}(k) = 2\sum_{i=0}^{k} \mathrm{C}_L^i. \tag{9}$$

The ranks of all words containing k unities belong to the interval

$$R_{max}(k-1) + 1 \leqslant R \leqslant R_{max}(k). \tag{10}$$

Equations (7) and (9) being considered as a parametrically defined function $W(R) = b(k(R))$ yields the envelope curve passing through the right edges of the steps in the rank distribution.

Using the Stirling formula for the gamma functions [which is valid at $L, k, (L-k) \gg 1$] and replacing the summation in Eq. (9) by integration, one can easily obtain the asymptotic expression for the dependence $R(W)$,

$$R = 2^L \sqrt{\frac{\zeta}{\pi}}\left(\frac{W}{B}\right)^{-\frac{1}{\zeta}} ln^{-1/2}\left(\frac{W}{B}\right) \tag{11}$$

with

$$\zeta = \frac{1}{1+2n/L},$$

$$B = 4^n\sqrt{\pi}\frac{\Gamma(1/2+n)}{\Gamma(n)\Gamma(2n+L)}\left(n + \frac{L}{2} - 1\right)^{2n+L-1} \times \exp(-2n-L+2). \tag{12}$$

The distribution Eq. (11) differs from the power-law form by the logarithmic multiplier. If one neglects this weak logarithmic dependence, the Zipf's-like law for the rank distribution would be obtained from Eq. (11),

$$W \propto R^{-\zeta}. \tag{13}$$

Actually, the change (9) of variable, $k \to R(k)$, i.e., stretching the abscissa axis, slows down the decay of the frequency W with an increase of k and results in the power law Eq. (13). The similar result was achieved in Sec. II by the change of variable $L \to R(L)$ Eq. (2) for the rank distribution of the words with different lengths L in the noncorrelated sequence. Such a procedure was first outlined in Ref. [10].

The obtained result is demonstrated in Fig. 1. The dotted line shows the plot of the rank distribution obtained from Eqs. (7), (9), and (10) at $L=14$, $N=15$, $\mu=15/46$, $n=4$. This plot passes closely to the solid line, which demonstrates the results from the numerical simulations of the rank distribution of words having the length $L=14$ in the Markov chain generated using the same parameters N and μ [17]. The dash-dotted line in this figure is the envelope curve described by Eq. (11).

The exponent in the Zipf's distribution (13) is governed by the ratio n/L. In the case of weak correlations, at $n \to \infty$, the value of ζ tends to zero and the Zipf distribution appears to be destroyed, i.e., all the words of the length L occur with almost the same probabilities $W = 2^{-L}$. The opposite situation, $n/L \to 0$, corresponds to the strong correlations in the Markov chain. Equation (12) shows that in this case the exponent

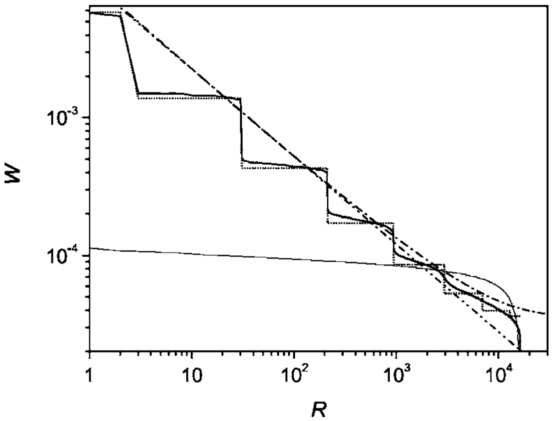

FIG. 1. The rank distribution $W(R)$ of the words with length $L=14$ in the Markov chain determined by the conditional probability Eq. (5) at $N=15$, $\mu=15/46$. The solid line is for the numerical simulations, the dotted line describes the exact distribution obtained from Eqs. (7), (9), and (10), the dash-dotted line is the plot of envelope asymptotics (11), the dash-dot-dotted line describes the Zipf's-like asymptotics Eq. (13) with $\zeta=7/11$. The thin solid line obtained numerically at $\mu=-15/46$, $N=15$, $L=14$ is for the antipersistent correlations.

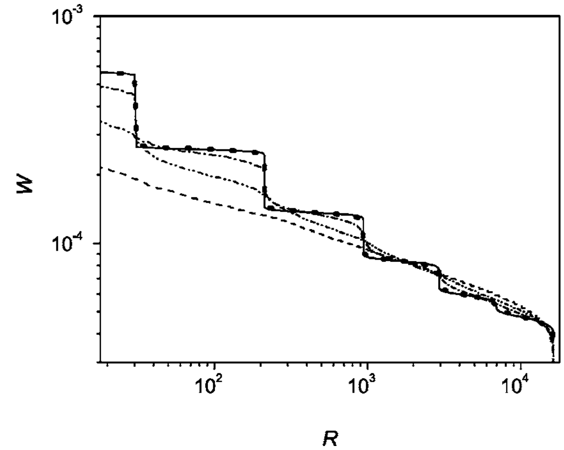

FIG. 2. Rank distributions of the L words with $L=14$ in the N-step Markov sequences reduced by randomly removing some symbols. The solid line is for the initial sequence with $N=32$, $n=15$. The symbols corresponding to the decimation parameter $\lambda=2$ ($N^*=16$) lie almost on the solid one. Other lines are for decimated sequences with $N^*<L$. Specifically, the dash-dotted, dash-dot-dotted, and dashed lines correspond to $N^*=8$, $N^*=4$, and to $N^*=2$, respectively.

ζ tends to unity. The plot of Zipf's distribution (13) with $\zeta=7/11$ is demonstrated by the dash-dot-dotted line in Fig. 1.

B. Scaling property of the rank distribution

The Zipf's-like law is frequently associated with the property of scale invariance [13], i.e., the invariance of the exponent in the Zipf law with respect to a certain decimation procedure. An analogous property referred to as the self-similarity appears in the frame of the model presented above. Let us reduce the N-step Markov sequence by regularly (or randomly) removing some symbols and introduce the decimation parameter $\lambda<1$ which represents the fraction of symbols kept in the chain. As is shown in Ref. [15], the reduced chain possesses the same statistical properties as the initial one but is characterized by the renormalized memory length, N^*, and the persistence parameter, μ^*,

$$N^* = N\lambda, \quad \mu^* = \mu\frac{\lambda}{1-2\mu(1-\lambda)}. \tag{14}$$

Indeed, the conditional probability P_k^* of the occurrence of the symbol zero after k unities among the preceding N^* symbols in the reduced chain is described by Eq. (5) where N and μ should be replaced by renormalized parameters N^* and μ^*. Considering the Zipf law, we are interested in the invariance of the Zipf plot with respect to the decimation procedure. According to Eqs. (11) and (12), the slope of the Zipf plot depends on the parameter n only. This parameter remains unchanged after the transformation $N\to N^*$, $\mu\to\mu^*$ [see Eqs. (8) and (14)]. As a result, the Zipf plots for rank distributions of the L words with $L<N$, obtained from the initial and decimated sequences, coincide. This self-similarity property is illustrated in Fig. 2.

The self-similarity is often observed in the correlated symbolic systems. However, this property is not necessary for the appearance of Zipf's-like rank distribution. Section II and papers [10] provide examples of such systems.

C. Rank distribution for $L>N$: Discussion

Now, let us look into the rank distribution of L words with $L>N$. This problem is not amenable to analytic calculations, and, therefore, numerical simulations are applied. In this case, the above-mentioned degeneration of the probability of word occurrence does not exist. So, smearing of the steps in the rank distribution takes place at $L>N$. This smearing occurs gradually as the word length L increases, and the steps appear to be completely smoothed away at sufficiently high values of L (see curves in Fig. 3). This means that contrary to

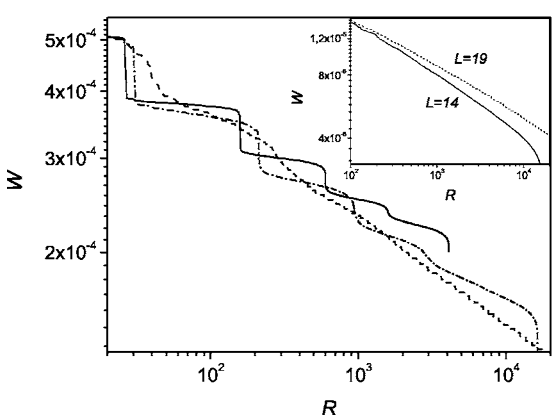

FIG. 3. The Zipf plots for L words with $L>N$. The solid, dash-dotted, and dashed lines are for $L=12,14,18$, respectively. The parameters of the Markov chain are $N=12$, $\mu=0.1$. The phenomenon of step smearing is noted with a growth in the word length L. In the inset: the Zipf plots for L words with $N=4$, $\mu=0.1$ at $L>L_{cr}$ (the lengths of words are shown near the curves).

case $L \leqslant N$, the rank distribution itself has the Zipf's-like form.

We would like to draw attention to the nonmonotonous behavior of the Zipf slope ζ with an increase in the word length L. As is seen from Eq. (12), this parameter increases at $L < N$. This growth continues at $L > N$ as well but only up to a certain value of $L = L_{cr} > N$. The maximum of ζ appears at $L = L_{cr} > N$ and then, at $L > L_{cr}$, ζ starts to decrease. This decrease is demonstrated in the inset to Fig. 3.

It is necessary to note that the position $L = L_{cr}$ of the maximum in the $\zeta(L)$ dependence is strongly related to the characteristic correlation length $2l_c + N$ in the Markov chain being studied. According to Ref. [15], the symbols correlate with each other not only within the memory length N but within the enlarged region $2l_c + N$ where l_c represents the characteristic attenuation length of the fluctuations. Thus, the highest value of the exponent ζ in the rank distribution of words in the Markov chain is achieved if the size of the competitive words is close to the correlation length.

If the words are shorter than the correlation length, $L < N + 2l_c$, the specific features of the correlations become apparent in the rank distribution that results in deviations from the power law. In the system that is considered in this paper, the deviations from the power law manifest themselves in the occurrence of the steps in the rank distribution and in the additional weak logarithmic multiplier in Eq. (11). Moreover, at $N \sim 1$ the rank distribution deviates significantly from the power law and assumes the exponential shape at $N = 1$ [13]. In the opposite limiting case, at $L \gg N + 2l_c$, the correlations over the whole word length disappear, and the rank distribution tends to a constant.

Note that specifically the persistent correlations in the Markov chain (that correspond to the attraction between the same symbols) lead to the pronounced Zipf law in the rank distribution of words. Indeed, the thin solid line in Fig. 1 demonstrates the very weak $W(R)$ dependence for the case of the antipersistent correlations, at $\mu = -15/46$.

V. CONCLUSIONS

We have presented a model of symbolic stochastic dynamical system with nonzero correlations possessing the power-law rank distribution with the exponent of the order of unity. A direct analytic connection between the correlation parameters of the Markov chain and the power-law rank distribution is established. The remarkable property of invariance with respect to the decimation procedure for words shorter than the memory length is observed. It is shown that the Zipf's-like law occurs if the length of the competitive words is of the order of the correlation length.

The obtained results allow us to suggest the following physical picture of the manifestation of the Zipf's-like law in the correlated systems. The correlations should be presented in the system but the noise of sufficient intensity should be imposed on these correlations over the length of competitive words. Within the considered stochastic system, this noise is provided by sufficiently strong fluctuations observed on the scales of the word length. The role of the noise consists of concealing the specific peculiarities of the correlations. Owing to the fluctuations, the concrete shape of the correlations does not appear to be very important. Accordingly, the Zipf law is a consequence of the rather rapidly damping persistent correlations of a quite arbitrary form, i.e., the global correlations in the system are not necessary. The Zipf law is a manifestation of the inner microstructure of the system being a result of the attraction between building blocks of the same kind.

We would like to emphasize that both circumstances, the difference in the length of words and the correlations between symbols, could be of great significance in forming the Zipf law in the rank distribution of words in correlated symbolic systems. The special investigations are necessary to clarify the role of these factors in the rank distributions in concrete real systems.

ACKNOWLEDGMENT

We acknowledge Dr. S. S. Denisov for the helpful discussions.

[1] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, New York, 1949).
[2] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **52**, 2939 (1995).
[3] D. Sornette, L. Knopoff, Y. Y. Kagan, and C. Vanneste, J. Geophys. Res. **101**, 13883 (1996); e-print arXiv: cond-mat/9510035.
[4] G. Caldarelli, M. Marsili, and Y.-C. Zhang, Europhys. Lett. **40**, 479 (1997).
[5] V. Pareto, *Le Cour d'Economie Politique* (Macmillan, London, 1896).
[6] http://myhome.hanafos.com/philoint/phd-data/Zipf's-Law-2.htm
[7] B. A. Trubnikov, Priroda **11**, 3 (1993) (in Russian).
[8] B. B. Mandelbrot, in *Proceedings of the Symposium on Applied Math*, edited by Roman Jacobson (AMS, New York, 1961), Vol. 12.
[9] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983).
[10] W. Li, IEEE Trans. Inf. Theory **38**, 1842 (1992); Complexity **1**, 6 (1996).
[11] I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).
[12] S. Denisov, Phys. Lett. A **235**, 447 (1997).
[13] A. Czirok, R. N. Mantegna, S. Havlin, and H. E. Stanley, Phys. Rev. E **52**, 446 (1995).
[14] O. V. Usatenko and V. A. Yampol'skii, Phys. Rev. Lett. **90**, 110601 (2003).
[15] O. V. Usatenko, V. A. Yampol'skii, K. E. Kechedzhy, and S. S. Mel'nyk, Phys. Rev. E **68**, 061107 (2003).

[16] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, Springer Series in Synergetics Vol. 13 (Springer-Verlag, Berlin, 1985).
[17] The first stage in the construction of the N-step Markov chain is a generation of the initial noncorrelated N symbols, zeros, and unities, identically distributed with equal probabilities $(1/2)$. Each consequent symbol was then added to the chain with the conditional probability determined by the previous N symbols in accordance with Eq. (5).