Possible origin of power-law behavior in *n*-tuple Zipf analysis

András Czirók,[1,2,*] H. Eugene Stanley,[1] and Tamás Vicsek[2]
[1]*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215*
[2]*Department of Atomic Physics, Eötvös University, Budapest, Puskin utca 5-7, 1088 Hungary*
(Received 27 October 1995)

In *n*-tuple Zipf analysis, ''words'' are defined as strings of *n* digits, and their normalized frequency of occurrence ω is measured for a given ''text'' (sequence of digits). In the case of various non-Markovian sequences, the probability density of the frequencies $\mathcal{P}(\omega)$ has a power-law tail. Here we argue that a broad class of unbiased binary texts exhibiting a *nonexponential* distribution of cluster sizes can indeed yield a power-law behavior of $\mathcal{P}(\omega)$, where we define clusters to be strings of identical digits. We support this result by numerical studies of long-range correlated sequences generated by three different methods that result in nonexponential cluster-size distribution: inverse Fourier transformation, Lévy walks, and the expansion-modification system. Our calculations shed light on the possible connection between the Zipf plot and the non-Markovian nature of the text: as the long-range correlations become dominant, the probability of the appearance of long clusters is increased, leading to the observed ''scaling'' in the Zipf plot. [S1063-651X(96)08006-3]

PACS number(s): 87.10.+e, 02.50.−r

Recently Zipf and related *n*-gram entropy analysis [1–3] have been applied to various complex signals or ''texts'' [4–10]. This kind of measurement was originally introduced in the context of natural languages and is performed by calculating the normalized frequency of occurrence ω of each word in a given text. By sorting the words according to their frequency, a rank R can be assigned to each word, with $R=1$ being the most frequent, $R=2$ the second most frequent, and so on.

For natural languages the $\omega(R)$ function can be well approximated by a power law with an exponent ζ close to one:

$$\omega \sim R^{-\zeta}. \tag{1}$$

The Zipf function $\omega(R)$ is closely related to a quantity that does not require the concept of ranking, the probability density of the word frequencies $\mathcal{P}(\omega)$, where $\mathcal{P}(\omega)d\omega$ is proportional to the number of words occurring with a frequency in the interval $[\omega, \omega+d\omega]$. For a given frequency ω, the rank R can be calculated as the total number of words occurring more frequently than ω:

$$R(\omega) \sim \int_{\omega}^{\infty} \mathcal{P}(\omega')d\omega'. \tag{2}$$

Thus the observation of power-law Zipf plot is equivalent to the presence of a power-law tail in $\mathcal{P}(\omega)$ [11].

If—unlike in the case of natural languages—the basic units are not defined, it is of interest to modify the Zipf analysis by defining a ''word'' to be an *n*-digit string of the text. To carry out this ''*n*-tuple Zipf'' analysis, a window of length *n* is moved along the sequence, one character at a step, and we record the occurrence of each *n*-tuple. In the case of long, unbiased sequences where each symbol is an independent random variable, each *n*-tuple is expected to occur with the same frequency, so $\omega(R)$ is constant and ζ vanishes. However, in many cases of interest a power-law Zipf plot (sometimes called a ''linguistic feature'' [1,5]) is observed. So far, the exact origin of this ''scaling'' behavior has been puzzling, although there have been indications that it can be related to the presence of long-range correlations in the texts [11]. On the other hand, the power-law Zipf plot cannot be *equivalent* to the presence of long-range correlations, as mixing the order of the nonoverlapping *n*-tuples eliminates the correlations on length scales larger than *n* but does not change the frequencies of these nonoverlapping words. According to numerical tests, this mixing procedure also does not alter ζ if we count the words by shifting the window by single digits, as we defined above.

To interpret results obtained by Zipf and other related statistical methods such as *n*-gram entropy, the general difficulty is that we must take into account complex features of the text such as long-range correlations [12]. Thus the basic approaches that characterize the sequence by either symbol frequencies or first order Markovian probabilities ρ_{xy} (denoting the conditional probability that a digit y follows a digit x) are often not sufficient. One possibility to better capture the complexity is to work with higher order conditional probabilities, but they are rather difficult to obtain both empirically and theoretically [13].

Here we introduce a *simpler* alternative. Suppose that the alphabet (set of symbols in the sequence) consists of only two elements. Then the text can be considered as built up by consecutive *clusters* of zeros and ones, where the clusters are defined as identical consecutive digits [14]. An *unbiased* sequence can be characterized by a single cluster-size distribution function P_k, where P_k denotes the probability that a randomly selected cluster (consisting of either type of symbol) consists of k digits ($\Sigma_{k>0}P_k=1$). In the more general case of a biased sequence, for each kind of digit a different distribution function can be assigned, which complicates the calculations but does not change our conclusions. P_k contains considerably more information than the ρ_{00} and ρ_{01}

*Electronic address: czirok@hercules.elte.hu

conditional probabilities [15], but as we do not take into account the correlations among different clusters, still does not have the ''full'' characterization of the sequence. Here we point out that in many cases the Zipf plot is strongly related to P_k: for a broad class of texts the nonexponential distribution of the cluster sizes can account for the observed power law $\omega(R)$.

The cluster-size distribution function P_k enables us to estimate the frequency ω of a given n-tuple consisting of m clusters (m-cluster word) with lengths $\ell_1,\ell_2,\ldots,\ell_m$ ($\Sigma_{i=1}^m \ell_i = n$). Since we assume independence of the clusters, the probability of the event that a cluster of length j follows a cluster of length k is given by $P_j P_k$. Thus for large n and m the frequency of a given m-cluster word is estimated by

$$\ln\omega = \sum_{i=1}^{m} \ln P_{\ell_i}, \tag{3}$$

where we neglected boundary effects: the actual length of the first and last cluster can be longer than ℓ_1 and ℓ_m, respectively [16].

First, we consider two cases: (i) If the digits of an unbiased sequence are independent random variables, then $P_k = 1/2^k$, as all the last $k-1$ digits must be identical to the first digit of the cluster, and the $(k+1)$th digit must be different. (ii) For a first order Markov process, P_k is given by $\rho_{11}^{k-1}\rho_{10} = \rho_{00}^{k-1}\rho_{01}$. In both cases P_k is *exponential*, and the tail of the probability distribution function $\mathcal{P}(\omega)$ decays faster than a power law [11].

Second, we investigate a more general situation with *nonexponential* cluster-size distribution P_k. Suppose that $\ln P_k$ can be written in the form of a Taylor expansion,

$$\ln P_k \approx A_0 + A_1 k + \frac{A_2}{2}k^2 + \frac{A_3}{6}k^3 + \cdots. \tag{4}$$

We will show that the quadratic term of the expansion above (for $A_2 > 0$) can yield a power-law tail in $\mathcal{P}(\omega)$. We make the following ansatz, consistent with $1 \ll m < n$ and $|A_0| \ll |A_1 n|$ [17]:

$$\ln P_k = Ak + Bk^2, \tag{5}$$

where A is determined by the normalization condition $\Sigma_{k=1}^n P_k = 1$. Substituting this into Eq. (3) yields

$$\ln\omega = An + B\sum_{i=1}^{m} \ell_i^2, \tag{6}$$

where we used the fact that $\Sigma_{i=1}^m \ell_j = n$.

Under these assumptions the relation between the frequencies of two words X and X' is independent of the parameters describing the functional form of P_k: $\omega_X > \omega_{X'}$ holds if $\Sigma_{i=1}^m \ell_i^2 > \Sigma_{i=1}^{m'} \ell_i'^2$ and these latter quantities are determined by the structure of the words (the number and length of the clusters they contain) only. Since— by definition — the rank of the word X is given by the total number of such words X' for which $\omega_X < \omega_{X'}$, we can see that the rank R of a given word is independent of A and B.

The above argument leads to the result that *if* the Zipf plot is indeed a power law, then the exponent ζ must be proportional to B. To see this, let us consider the Zipf plots of two sequences displaying different cluster size distribution characterized by A,B and $\widetilde{A},\widetilde{B}$, respectively. The rank R of a given word is the same in both cases according to the argument above, but the frequencies are different. Let us denote by $\omega(R)$ and $\widetilde{\omega}(R)$ the frequency of the given word in the two sequences. These values are determined by Eq. (6). Eliminating $\Sigma_{i=1}^m \ell_i^2$ from the equations we have

$$\ln\widetilde{\omega}(R) = \frac{\widetilde{B}}{B}\ln\omega(R) + \text{const} \tag{7}$$

for any R. Substituting $\omega \sim R^{\zeta}$ and $\widetilde{\omega} \sim R^{\widetilde{\zeta}}$ into Eq. (7) yields

$$\zeta \sim B. \tag{8}$$

Now we will focus on the emergence of the power law (1). According to Eq. (6), the power law in $\mathcal{P}(\omega)$ is equivalent to an exponential decay in the density distribution $\mathcal{P}(\Sigma_{i=1}^m \ell_i^2)$ considering all possible n-tuples. This is already a ''universal'' number theoretical property and can be calculated for each value of n. According to the numerical evaluations $\mathcal{P}(\Sigma_{i=1}^m \ell_i^2)$ decays as $\exp(-\Sigma_{i=1}^m \ell_i^2/n)$, see Fig. 1. In

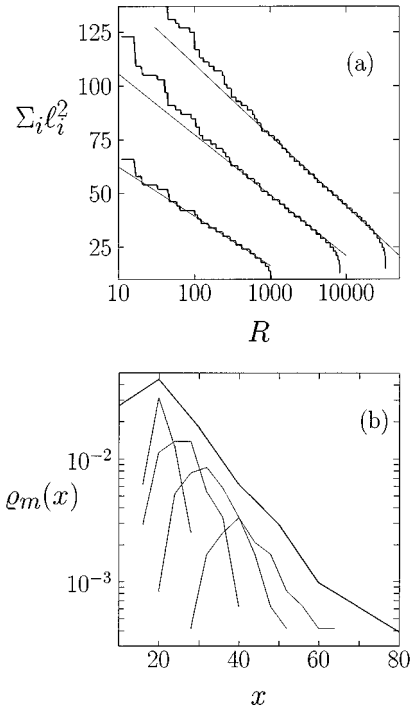

FIG. 1. (a) Numerically calculated Zipf plots based on the *ansatz* (5) for three different word lengths $n = 15,13,10$ (from top to bottom). For a given n-tuple of digits [0,1], we plotted the dependence on rank R of the quantity $\Sigma_{i=1}^m \ell_i^2$ (which is essentially $\ln\omega$, neglecting a normalizing factor that would only shift the entire curve parallel to the vertical axis.) (b) The numerically calculated probability density function $\rho(x)$ of the logarithmic word frequencies for $n = 10$ (upper curve), where $x \equiv \ln\omega$. Note the exponential decay, as indicated by the linearity of the top curve. We also display the corresponding distributions of the m-cluster words, for $m = 3,4,5,6$ (from right to left), showing that a characteristic frequency $\widetilde{x}_m$ and a probability density $\widetilde{\rho}_m$ can be assigned for each value of m. More details are in the Appendix.

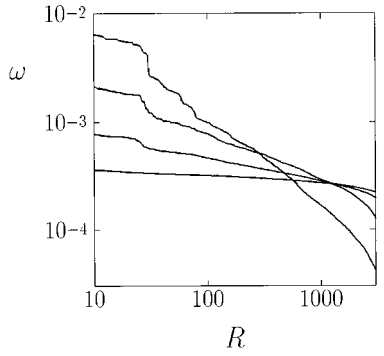

FIG. 2. Measured Zipf plots of long-range correlated sequences of length $N=10^7$ generated by the IFT method, for four values of the autocorrelation exponent: $\alpha=0.8$, 0.7, 0.6, and 0.5 (from top to bottom). All the sequences studied consisted of 10^7 digits. The small nonzero slope of the $\alpha=0.5$ curve is an artifact caused by the finite sample size.

the Appendix we also give an approximate derivation of this behavior. Since, apart from multiplicative factors, $\mathcal{P}(\Sigma_{i=1}^{m}\ell_i^2)$ is the density distribution of the logarithm of the word frequencies in a sequence characterized with $B=1$, this result means that for $B=1$ we have $\zeta=n$. Taking (8) into account gives

$$\zeta=nB. \tag{9}$$

Figure 1(a) shows Zipf plots calculated numerically with the ansatz (5) for $B=1$ and various n. Indeed, the scaling regime extends for increasing n, and the exponent ζ is in good agreement with (9). The corresponding logarithmic probability densities are plotted in Fig. 1(b) for $n=10$.

We tested the above explanation of the origin of the power-law Zipf plots (Fig. 2) on all the test sequences studied in Ref. [11], which were generated by three different methods: inverse Fourier transformation (IFT) [18–21], Lévy walks [22], and the expansion-modification system [23].

Typical cluster-size distribution functions for various α parameters of the IFT sequences are displayed in Fig. 3. Note that when the long-range order disappears (for the value

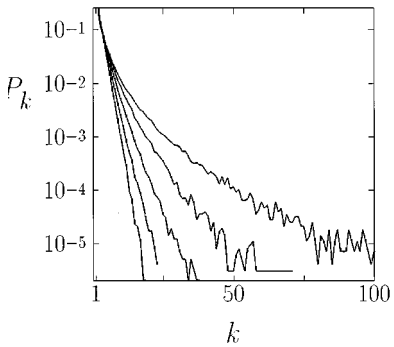

FIG. 3. Log-linear plot of the measured cluster-size distribution functions of the IFT sequences for $\alpha=0.5$, 0.6, 0.7, 0.8, and 0.9 (from left to right). P_k denotes the probability that a randomly selected cluster consists of k digits. For $\alpha=0.5$ the distribution is exponential, consistent with the Markovian nature (absence of long-range correlations) of the sequence.

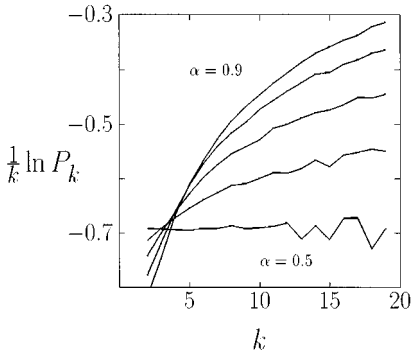

FIG. 4. The coefficient B in ansatz (5) can be fit by a linear regression of $\ln P_k/k$ over a range $0<k<n$. The curves are calculated from the cluster-size distribution functions shown in Fig. 3. In the case of a Markovian sequence (lower curve, $\alpha=0.5$), B vanishes.

0.5 of the long-range correlation exponent α), P_k decays exponentially. In all other cases the decay is slower than exponential, but not a power law in general. The coefficient B of the quadratic term in (5) can be fit by a linear regression of $(\ln P_k)/k$ vs k (Fig. 4), for $0<k<N$. For all the sequences investigated, a close relationship was found between B and ζ (Fig. 5), which is quite close to (9) predicted by our simple arguments. This result indicates that in all cases we studied the nonexponential cluster-size distribution gives the dominant contribution to the power-law behavior of the Zipf plot.

The approach presented above is limited to cases where the correlations between the consecutive clusters are negligible. However, in the case of alphabets consisting of many symbols the Zipf plot can be influenced by the correlations among the various clusters. These effects can be treated by a generalization of Eq. (3): First (or higher) order conditional probabilities can be introduced giving the probability P_{kl} of the event that a cluster of length l follows a cluster of length k as $P_{kl}\equiv P_k P_l(1+\epsilon_{kl})$, where ϵ_{kl} satisfies $\Sigma_l \epsilon_{kl} P_l=0$. Now, (3) can be written as

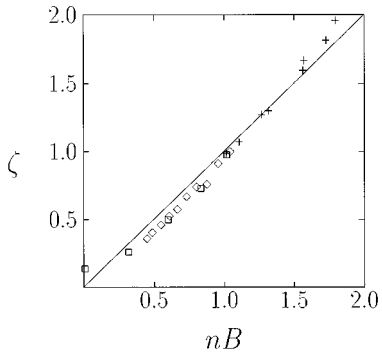

FIG. 5. The Zipf exponent ζ vs the quadratic coefficient B, which was calculated from the cluster-size calculations for the three methods, IFT (□), Lévy (◇), and expansion-modification system (EMS) (+). Note that the universal behavior $\zeta\approx nB$ is supported by these calculations. For low values of ζ, the effects due to the finite size of the sequence dominate the Zipf analysis. In the case of the EMS sequences, the power-law behavior of the Zipf plots is less good, yielding large (up to ± 0.1, depending on the fitting regime) uncertainties in ζ (see Ref. [8] for more details). In the other two cases, the error in ζ is in the order of ± 0.05.

$$\ln\omega=\sum_{i=1}^{m}\ln P_{\ell_i}+\sum_{i=1}^{m-1}\ln(1+\epsilon_{\ell_i\ell_{i+1}}). \tag{10}$$

In the $m\gg 1$ and $\epsilon\ll 1$ limits $\Sigma_i\ln(1+\epsilon_{\ell_i\ell_{i+1}})\approx\Sigma_i\epsilon_{\ell_i\ell_{i+1}}\approx m\Sigma_{kl}P_{kl}\epsilon_{kl}$. Taking into account the definition of ϵ_{kl}, $\Sigma_{kl}P_{kl}\epsilon_{kl}=\Sigma_{kl}P_kP_l\epsilon_{kl}^2$. Thus the second correction term in Eq. (10) is of the order of $m\epsilon^2P^2$, where ϵ and P denote the typical magnitude of ϵ_{kl} and P_k. This means that the simple approximation (3) neglecting the correlations should give satisfactory results even for nonvanishing correlations if $|\epsilon^2P^2|\ll|\ln P|$.

In summary, for a broad class of texts that can be considered as built up by consecutive clusters of identical digits, the Zipf plot can be explained by Eq. (6). This result also sheds light on the connection between the Zipf plot and the non-Markovian nature of the text: as the long-range correlations becomes dominant, the probability of the appearance of long clusters is increasing. As a consequence, the tail of the cluster size distribution function will decay slower than the exponential predicted by the Markovian approximation, and this nonexponential tail yields the observed ''scaling'' in the Zipf plot.

We have benefited from the discussions with S. V. Buldyrev, S. Havlin, and R. N. Mantegna. This work was supported by the U.S.-Hungarian Joint Fund Contract No. 352.

APPENDIX: DISCUSSION OF EQ. (9)

In this Appendix, we present the discussion of the relation $\zeta=nB$, which was obtained in the text by numerical evaluation of the distribution of $x\equiv\Sigma_{i=1}^{m}\ell_i^2$ for all possible n-tuples. We analyze the probability density function $\rho(x)$, which is strongly related to $\mathcal{P}(\omega)$, since apart from an additive constant, x is identical with $\ln\omega$. $\rho(x)$ is built up as a sum of density distributions of various m-cluster words as

$$\rho(x)=\sum_{m=1}^{n}\rho_m(x), \tag{A1}$$

where $\rho_m(x)dx$ is given by the number of m-cluster words satisfying $\Sigma_{i=1}^{m}\ell_i^2\in[x,x+dx]$ divided by the total number of words.

Assume that we can ''parametrize'' $\rho(x)$ by m: for most values of x a specific m^* can be selected so that the dominant contribution to $\rho(x)$ is given by the m^*-cluster words. In other words this means that for each m^* we can assign a characteristic value of $\tilde{x}_{m^*}$ and a characteristic density $\tilde{\rho}_{m^*}$. Hence we expect

$$\rho(\tilde{x}_{m^*})\approx\tilde{\rho}_{m^*}. \tag{A2}$$

A simple estimation for these characteristic values can be the following: let $\tilde{\rho}_{m^*}$ be proportional to the number of m^*-cluster words N_{m^*}, and let the characteristic frequency be proportional to $\omega_{\max}(m^*)$, which is the frequency of the most frequent m^*-cluster word.

The latter quantity, $\omega_{\max}(m^*)$, can be calculated as follows. Increasing the length ℓ_i of the longest cluster in an m-cluster word at the expense of a smaller cluster with length ℓ_j in such a way that the number of clusters does not change (e.g., 0001100→0000100) yields a more frequently occurring word, since according to (6) $\omega(\ell_1,\ldots,\ell_i,\ell_j,\ldots,\ell_m)<\omega(\ell_1,\ldots,\ell_i+1,\ell_j-1,\ldots,\ell_m)$, if $\ell_j\leqslant\ell_i$. Thus ω (hence x) is maximal if one of the clusters is $(n-m-1)$ digits long and all the other clusters consist of a single digit, because in this case the step described above cannot be performed. Hence

$$x_{\max}=An+B[m^*-1+(n-m^*+1)^2]\approx An+B(n-m^*)^2. \tag{A3}$$

Now N_{m^*} is given by simple combinatorics (we must place m^*-1 separators into $n-1$ possible positions)

$$\ln N_{m^*}=\ln\binom{n-1}{m^*-1}\approx n\ln\frac{n}{n-m^*}+m^*\ln\frac{n-m^*}{m^*}$$
$$=-n[\mu\ln\mu+(1-\mu)\ln(1-\mu)], \tag{A4}$$

where $\mu\equiv m^*/n$. For moderate values of μ ($\mu\approx 1/4$), the above equation can be approximated by

$$\ln N_{m^*}=n[\ln 2-2(\mu-1/2)^2]. \tag{A5}$$

Linearizing both Eqs. (13) and (15) in μ and substituting into (A2) yields

$$\ln\rho(2\mu Bn^2+\text{const})\approx\text{const}-2n\mu, \tag{A6}$$

resulting in Eq. (9).

[1] G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Reading, MA, 1949).
[2] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **30**, 50 (1951).
[3] L. Brillouin, *Science and Information Theory* (Academic, New York, 1956).
[4] M. Damashek, Science **267**, 843 (1995).
[5] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994); Phys. Rev. E **52**, 2939 (1995).
[6] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons, and H. E. Stanley (unpublished).
[7] M. H. R. Stanley, S. V. Buldyrev, S. Havlin, R. Mantegna, M.A. Salinger, and H. E. Stanley, Eco. Lett. **49**, 453 (1995).
[8] S. Wolfram, Commun. Math. Phys. **96**, 15 (1984).
[9] P. Grassberger, IEEE Trans. Inf. Theory **35**, 669 (1989).
[10] C.T. Meadow, J. Wang, and M. Stamboulie, J. Inf. Sci. **19**, 247 (1993).
[11] A. Czirók, R. N. Mantegna, S. Havlin, and H. E. Stanley, Phys. Rev. E **52**, 446 (1995).

[12] A. Schenkel, J. Zhang, and Y.-C. Zhang, Fractals **1**, 47 (1993); M. Amit, Y. Shmerler, E. Eisenberg, M. Abraham, and N. Shnerb, *ibid.* **2**, 7 (1994); W. C. Ebeling and A. Neiman, Physica A **215**, 233 (1995).

[13] T. Schürmann and P. Grassberger (unpublished).

[14] As an example, the 10-tuple ''0100011100'' consists of five clusters, two of length 3, one of length 2, and two of length 1.

[15] The Markov probabilities characterizing the sequence can be obtained from P_k in the following manner. As an example, $\rho_{11}=\omega_{11}/\omega_1$, where we denoted by ω_s the normalized frequency of the substring s. In other words ω_s gives the probability of the event that the given substring occurs at a certain position in the text. ω_{11} can be calculated as the product of the probabilities that a selected digit is 1, and the cluster containing this digit also contains the next digit of the sequence, yielding

$$\omega_{11}=\omega_1\sum_{k>0} P_k\left(\frac{k-1}{k}\right).$$

[16] To calculate precisely the frequency ω of a given m-cluster word, we must take into account that the *first* cluster containing the first ℓ_1 digits of the word must start at a well-defined position. Thus

$$\ln\omega=\ln\left(\frac{P_{\ell_1}}{\ell_1}+\frac{P_{\ell_1+1}}{\ell_1+1}+\ldots\right)+\sum_{i=2}^{m-1}\ln P_i$$
$$+\ln(P_{\ell_m}+P_{\ell_m+1}+\cdots).$$

[17] In general, we cannot neglect the A_0 constant in the Taylor expansion (4), e.g., in the case of Markovian sequences this is the only term that determines the word frequencies, as the linear term always yields a constant. However, if higher-order terms are also present, then for large enough n the effects of A_0 can be negligible.

[18] C. K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature **356**, 168 (1992).

[19] H. Makse, S. Havlin, H. E. Stanley, and M. Schwartz, Chaos, Solitons, and Fractals **6**, 295 (1995); Phys. Rev. E **53**, 5445 (1996).

[20] S. Prakash, S. Havlin, M. Schwartz, and H. E. Stanley, Phys. Rev. A **46**, R1724 (1992).

[21] S. Havlin, R. Blumberg-Selinger, M. Schwartz, H. E. Stanley, and A. Bunde, Phys. Rev. Lett. **61**, 1438 (1988).

[22] G. Zumofen, A. Blumen, J. Klafter, and M. F. Shlesinger, J. Stat. Phys. **54**, 1519 (1989); S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 4514 (1993).

[23] W. Li, Phys. Rev. A **43**, 5240 (1991).