

Statistics of largest loops in a random walk

Deniz Ertaş¹ and Yacov Kantor²

¹*Department of Physics, Harvard University, Cambridge, Massachusetts 02138*

²*School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69 978, Israel*

(Received 5 September 1996)

We report further findings on the size distribution of the largest neutral segments in a sequence of N randomly charged monomers [D. Ertaş and Y. Kantor, Phys. Rev. E **53**, 846 (1996)]. Upon mapping to one-dimensional random walks (RW's), this corresponds to finding the probability distribution for the size L of the largest segment that returns to its starting position in an N -step RW. We focus primarily on the large N , $\ell=L/N \ll 1$ limit, which exhibits an essential singularity. We establish analytical upper and lower bounds on the probability distribution, and numerically probe the distribution down to $\ell \approx 0.04$ (corresponding to probabilities as low as 10^{-15}) using a recursive Monte Carlo algorithm. We also investigate the possibility of singularities at $\ell = 1/k$ for integer k . [S1063-651X(97)09301-X]

PACS number(s): 02.50.-r, 05.40.+j

I. INTRODUCTION

It has recently been shown that ground state conformations of polyampholytes, a particular type of heteropolymers built with a random mixture of positively and negatively charged groups along their backbone, are extremely sensitive to their total (excess) charge Q . A detailed study of the Q dependence of the radius of gyration R_g [1,2] determined that a reasonable compromise between stretching (which minimizes the electrostatic energy) and remaining compact (which gains in condensation energy) is for the polyampholyte to form a *necklace* of weakly charged blobs connected with highly charged “necks,” by taking advantage of the charge fluctuations along the chain. The results of Monte Carlo [1] and exact enumeration [2] studies qualitatively support such a picture.

While the exact treatment of electrostatic interactions is not possible, we can pose a simplified problem which, we hope, captures some essential features of this necklace model. For example, we may ask what the typical size of the largest neutral (or weakly charged) segment in a random sequence of N charges will be. In order to answer this question, we investigated the size distribution of the largest neutral segments in polyampholytes with N monomers (N -mers). This problem can be mapped to a one-dimensional random walk (RW): the sequence of charges $\{q_i\}$ ($i=1, \dots, N; q_i = \pm 1$) corresponds to an N -step walk $\omega \equiv \{q_1, \dots, q_N\}$ with the same sequence of unit steps in the positive or negative directions along an axis, where the probability of going up or down is equal to $1/2$ at each step. Figure 1 depicts an example of such a sequence and the corresponding path, where $S_i(\omega) = \sum_{j=1}^i q_j$ is the position of the path at index i . [$S_0(\omega) \equiv 0$.] A segment of L monomers with zero total charge thus corresponds to an L -step loop inside the RW. In this paper, we further investigate properties of the probability $P_N(L)$ that the *largest* loop in an N -step RW has length L , or, equivalently, the probability $Z_N(L) = \sum_{L'=0}^{L-1} P_N(L')$ that all loops in an N -step RW are shorter than L . Earlier results about a generalized version of this and other related problems can be found in Refs. [3,4].

In the continuum ($N \rightarrow \infty$) limit, it is more convenient to work with the *probability density*

$$p(\ell) \equiv \frac{N}{2} [P_N(L) + P_N(L+1)] \quad (1)$$

and

$$z(\ell) = \int_0^\ell d\ell' p(\ell'), \quad (2)$$

where $\ell = L/N$ is the appropriate scaling variable for this problem.

There is an apparent simplicity of the formulation of the problem, i.e., it is similar (and related) to the classical RW problems [5], such as the problem of first passage times or the problem of last return to the starting point, for which probability distributions can be computed exactly by using the method of reflections [6], and obey the same scaling in the continuum limit. However, the search for the *longest* loop of the RW, among all possible starting points, creates a more complicated problem. In its essence, the problem is more related to the statistics of self-avoiding, rather than regular, random walks. This relation becomes more transparent in the $\ell \rightarrow 1$ and $\ell \rightarrow 0$ limits. The former limit had been extensively studied in Ref. [4], and the latter will be discussed in Sec. III. The “self-interacting nature” of the problem can be seen even more clearly in its generalizations to arbitrary space dimension d , where many analogies between this problem and the self-avoiding walks exist.

Our earlier investigations revealed remarkable properties of the probability density $p(\ell)$: It diverges as $p(\ell) \sim 1/\sqrt{1-\ell}$ for $\ell \rightarrow 1$, and has a discontinuous derivative at $\ell = 1/2$. Furthermore, it has an essential singularity at $\ell = 0$ of the form $p(\ell) \sim \exp(-B/\ell)$. An analytical solution in this limit still remains elusive. We had not been able to determine $p(\ell)$ even numerically below $\ell \approx 0.15$ due to the very small probabilities involved near $\ell = 0$, severely limiting a straightforward Monte Carlo approach. Because of these difficulties, the existence and precise form of this singularity (including possible power law prefactors, etc.) was

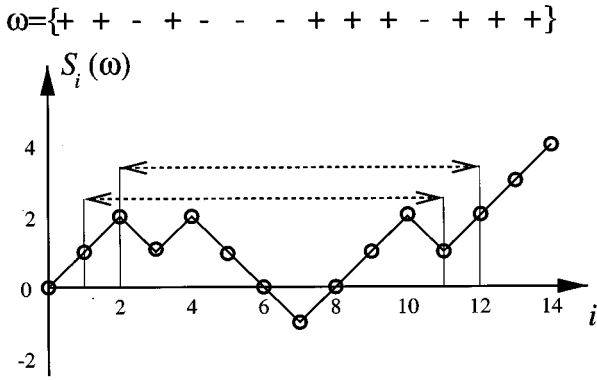


FIG. 1. Example of a sequence ω with $N=14$ charges, and the corresponding walk depicted by $S_i(\omega)$. In this case, the longest loops have lengths $L=10$ (dotted lines).

not well established. Since the publication of that work, we have developed an improved Monte Carlo algorithm that is capable of probing significantly smaller values of ℓ numerically. Combined with strict analytical bounds on $z(\ell)$, the results strongly favor the existence of this singularity, and the proper form of the $\ell \rightarrow 0$ limit can be determined with high precision. In this paper, we report the results of these complementary findings.

It should be noted that similar behavior is exhibited by extremal properties of a number of random processes, such as a one-dimensional random cutting process [7] (which can be generalized to higher dimensions [8]) and return times in a random walk [8]. These models exhibit singularities at $\ell = 1/k$, which become progressively weaker as the integer k is increased, leading to an essential singularity at $\ell = 0$. Although it was claimed that our problem falls into the same category and therefore should exhibit singularities at $\ell = 1/2, 1/3, 1/4, \dots$ [8], we believe that it differs from these models in a way that undermines the reasoning for this claim, as we shall discuss in Sec. IV. In particular, we have numerically verified that the suggested singularity at $\ell = 1/3$ does not exist, unless it has a very small prefactor.

The rest of the paper is organized as follows: First, we establish upper and lower bounds on $z(\ell)$. We then describe an efficient Monte Carlo algorithm that enables us to determine $z(\ell)$ down to very small values, and present results from its implementation. Finally, we discuss the possible relevance of other random models with similar characteristic properties.

II. UPPER AND LOWER BOUNDS

In this section, we establish rigorous upper and lower bounds on the probability distribution $z(\ell)$, both of which have the same functional form. The existence of these bounds significantly restricts possible asymptotic forms of $z(\ell)$ in the $\ell \rightarrow 0$ limit.

The main strategy is similar for establishing both upper and lower bounds. Walks whose largest loops are much smaller than their overall length are typically very biased in one direction, and sections of the walk that are separated by more than the largest loop size are very weakly correlated. For a given (small) value of ℓ , let us divide each walk into

roughly $1/\ell$ segments of similar size. There are *necessary* conditions that each segment must satisfy independently for the overall walk to contribute to $z(\ell)$. If the probability for a random segment to satisfy these conditions is p_n , then $z(\ell) < p_n^{1/\ell}$. Similarly, each segment can be designed to satisfy certain conditions that are *sufficient* to ensure that the overall walk contributes to $z(\ell)$. If the corresponding probability for these conditions is p_s , then $z(\ell) > p_s^{1/\ell}$. The rest of this section is devoted to establishing a set of necessary and sufficient conditions and calculating the corresponding probabilities.

Let us first investigate necessary conditions. Let ω be an N -step walk whose largest loop is less than L -steps long, and has $S_N(\omega) > 0$. We shall focus on the cases where $m = N/L$ is an integer for now. Let us split ω into m mutually exclusive segments $\{\omega_1, \dots, \omega_m\}$ of length L where $\omega_i = \{q_{(i-1)L+1}, \dots, q_{iL}\}$. It is easy to see that ω satisfies the inequalities

$$S_{iL}(\omega) > S_{(i-1)L}(\omega), \quad 0 < i \leq m, \quad (3)$$

or, equivalently,

$$S_L(\omega_i) > 0, \quad 0 < i \leq m, \quad (4)$$

i.e., each of the m segments need to have a positive displacement. The probability for this is just $p_n = 1/2$, and therefore $Z_N(N/m) < 2^{1-m}$ (the additional factor of 2 comes from RWs with $S_N < 0$). Consequently, $Z_N(L) < 2^{2-(N/L)}$ for any value of N and L . This establishes a strict upper bound, which is significant for small values of ℓ :

$$z(\ell) < 4 \exp(-\ln 2 / \ell). \quad (5)$$

It is possible to further improve on this upper bound, and we will next demonstrate such an improvement which is by no means final. Consider a pair of adjacent segments (e.g., ω_1 and ω_2) described above, with $S_L(\omega_1), S_L(\omega_2) > 0$. Let i be the *smallest* index where $S_i(\omega_1) = S_L(\omega_1)$, and j the *largest* index where $S_j(\omega_2) = 0$. In that case, the segment from i to $L+j$ (on ω) is a loop, and therefore $i > j$ since ω cannot have a loop larger than L . For two randomly selected segments, this condition is satisfied with probability $1/2$, which can be calculated from the known probability distribution of “last return to the origin” [4,6]. Since there are $m/2$ statistically independent adjacent pairs, this observation further suppresses the upper bound on the probability distribution by a factor of $2^{-m/2}$, improving the overall upper bound to

$$z(\ell) < 4 \sqrt{2} \exp\left(-\frac{3 \ln 2}{2\ell}\right), \quad (6)$$

which makes the best (so far) analytical lower bound on the exponential factor $B > 3 \ln 2 / 2 \approx 1.03972$.

In order to find a lower bound on the probability distribution, let us again consider the sequence ω and its m pieces $\{\omega_i\}$ of length L each. We would like to construct each ω_i independently in such a way as to guarantee that the resulting walk ω does not have loops larger than L . This can again be done in many different ways, and the following is by no means optimal. The quality of the bound usually depends on how complicated the specifications of each piece are, and the

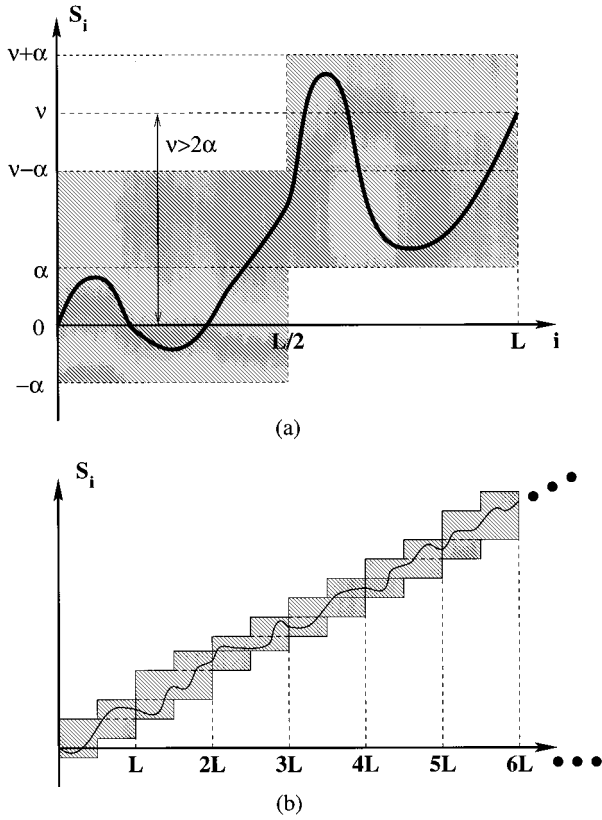


FIG. 2. (a) Example of a walk that satisfies the conditions in Eq. (7). Each such walk remains entirely within the shaded area. (b) When such walks are joined together, the resulting walk does not have loops that are larger than or equal to L , since such loops cannot fit in the shaded area.

limiting factor seems to be the analytical tractability of the associated probabilities. The following represents the best bound we have been able to establish analytically.

The specifications of each piece are as follows:

$$\begin{cases} -\alpha < S_i < S_L - \alpha, & 0 < i \leq L/2, \\ \alpha < S_i < S_L + \alpha, & L/2 < i \leq L. \end{cases} \quad (7)$$

Figure 2(a) shows these specifications graphically. Clearly, $S_L > 2\alpha$ is required. Figure 2(b) shows how the joining of such pieces results in a sequence ω that has no loops larger than L .

The probability p_s of meeting the stated specifications can be evaluated numerically to high accuracy using the method of reflections [6] and summing over all possible values of $S_{L/2}$ and S_L for a given α . The largest value for the probability yields the tightest lower bound on $z(\ell)$, so it is desirable to tune α in order to optimize the bound. We pick $\alpha = 0.5\sqrt{L}$, which is very close to the optimal value. In that case, the probability for a RW to satisfy the requirements (7) for large L is $p_s \approx 0.031585$. This yields

$$z(\ell) > 2p_s \exp(-\ln p_s / \ell) \approx 0.06317 e^{-3.455/\ell}. \quad (8)$$

Clearly, neither the upper nor the lower bounds we have established are very tight, and they do not rule out the possibility of a power-law prefactor. However, there is very con-

vincing numerical evidence that there is no power law prefactor in $z(\ell)$, i.e., that $\lim_{\ell \rightarrow 0} z(\ell) = C \exp(-B/\ell)$, where C and B are constants that are determined in the following section.

III. NUMERICAL WORK

In this section, we present numerical studies to determine $p(\ell)$ and $z(\ell)$ in the $\ell \ll 1$ limit. As stated earlier [3,4], a standard Monte Carlo method of determining $p(\ell)$ from a random sample of all possible walks is ineffective at probing $\ell \leq 0.15$, since the probabilities become very small. A similar problem arises when it is necessary to randomly sample very large self-avoiding walks (SAWs) in two and three dimensions: The probability of generating a SAW is exponentially small in its overall length, i.e., the probability of picking a SAW out of RWs of length $N \ll 1$ scales as $P_{\text{SAW}}(N) \sim N^\gamma e^{-aN}$, where a and γ are constants that depend only on the dimensionality of the SAW. A common way to circumvent this problem is to build large SAWs recursively by joining smaller SAWs. This method significantly reduces the number of operations needed by completely eliminating its dependence on the leading exponential factor: The probability of creating a SAW of length N by joining two randomly selected SAWs of length $N/2$ scales only as $N^{-\gamma}$, and the number of operations needed to generate a randomly sampled SAW grows as $e^{\gamma(\log_2 N)^2/2}$ instead of e^{aN} . Of course, creating SAWs in one dimension is trivial, but the extension of this method to one-dimensional walks is still very useful for our problem, since creating RWs with very small loops is similar to creating SAWs [in fact $P_{\text{SAW}}(N) = Z_N(1)$], and can be used to sample $z(\ell)$ efficiently at small ℓ .

In this implementation of the algorithm, we start from pairs of RWs of length L (with nonzero total displacement) and join them, keeping only resultant walks whose largest loops are smaller than L . At the first level, this creates walks that contribute to $Z_{2L}(L)$, with equal probability. We then iterate this process by pairing the resultant walks at each level. After the n th level, we end up with a representative sample of all walks that contribute to $Z_{2^n L}(L)$, which can then be used to determine a histogram for the probability distributions for $0 < \ell < 2^{-n}$.

We also need to keep track of the probability of success R_n at each level, which is given by

$$R_n(L) \equiv \frac{Z_{2^n L}(L)}{[Z_{2^{n-1} L}(L)]^2}, \quad (9)$$

in order to determine the overall normalization of the probability distributions. One big advantage of studying one-dimensional walks is that the probability of success $R_n(L)$ actually becomes *independent of n* , i.e., in the continuum limit

$$z(\ell) = R[z(2\ell)]^2, \quad \ell \ll 1, \quad (10)$$

where $R = \lim_{L \rightarrow \infty} \lim_{n \rightarrow \infty} R_n(L)$ is a nonzero constant. (For the one-dimensional SAW, the probability of success is just $1/2$.) Typically, variations in $R_n(L)$ were within statistical fluctuations (0.1% to 0.3%) for $n \geq 3$. When $R_n(L)$ is inde-

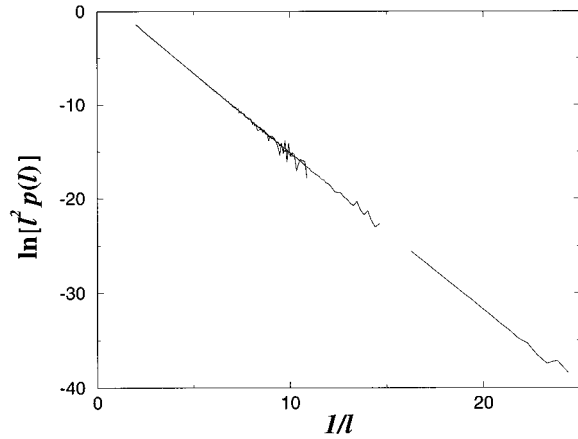


FIG. 3. The probability density $p(\ell)$ for $0.04 < \ell < 1/2$ confirms the suggested form (11) down to probabilities as low as 10^{-15} . The overall walk size is $N=2048$. Four (partially overlapping) plots were generated from runs that terminated after recursion levels 1 through 4.

pendent of n , the number of operations needed to sample a representative walk that contributes to $z(\ell)$ is only polynomial in ℓ^{-1} , which speeds up the algorithm enormously. Furthermore, this implies that for $\ell \ll 1$,

$$z(\ell) = C \exp\{-B/\ell\}, \quad (11)$$

$$p(\ell) = \frac{BC}{\ell^2} \exp\{-B/\ell\}, \quad (12)$$

where $C=R^{-1}$ and B are constants; there are no power-law prefactors in $z(\ell)$. This result can be verified numerically by looking at the results of the described recursive algorithm: Fig. 3 confirms the functional form (12) over about twelve decades in the probability density $p(\ell)$, probing values of ℓ down to 0.04 [9].

The constants C and B in the continuum limit can be determined accurately by plotting their dependence on walk length. C is simply the inverse of the success probability R as mentioned earlier, whereas B is given by the slope of the graph in Fig 3. Figure 4 shows these plots, which yield

$$C = 4.57 \pm 0.01, \quad (13)$$

$$B = 1.73 \pm 0.02. \quad (14)$$

IV. RELATED PROBLEMS

Behavior that is strikingly similar to those of $p(\ell)$ are exhibited by probability distributions of extremal properties in certain random systems. One simple example is a one-dimensional random cutting process [7,8]: A unit interval is cut at a randomly selected point (with uniform probability), and the same cutting process is repeatedly applied to the interval that remains to the *right* of the latest cut, ad infinitum. The probability distribution $p'(\ell)$ for the size of the *largest* interval that remains at the end of the cutting process exhibits singularities of the form $|\ell - 1/k|^{k-1}$ at each value of k , which become progressively weaker as the integer k is increased, leading to an essential singularity at $\ell=0$. The

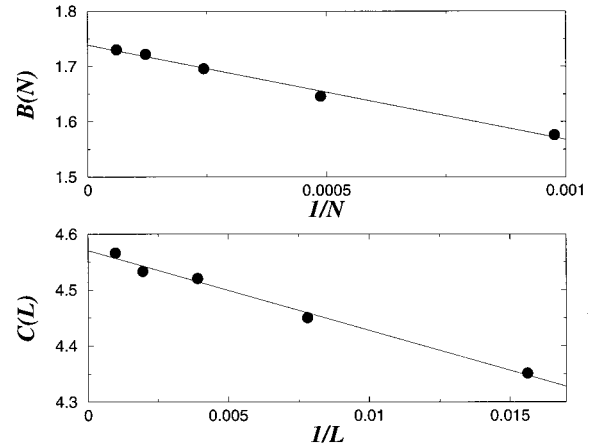


FIG. 4. Size dependence of the constants that appear in the $L/N \ll 1$ limit of $Z_N(L)$ and $P_N(L)$. Top: The exponential constant $B(N)$ determined from plots of $\ell^2 p(\ell)$ as a function of total walk length N . Statistical errors are smaller than symbol sizes. Bottom: The prefactor $C(L)$ determined from success probabilities $R_4(L)$ as a function of largest loop size L . Statistical errors are roughly the size of symbols.

origin of these singularities can be traced to the fact that the pieces (among which the largest one is chosen) constitute a *partition* of the entire interval, which implies that the sum of the sizes of all pieces equals the size of the initial interval, which is 1. Consequently, any piece that is larger than $1/2$ is necessarily the largest, and in general there can be at most $k-1$ pieces that are larger than $1/k$. This causes singular behavior in $p'(\ell)$ at $\ell=1/k$ for all k . Similar ‘‘sum rules’’ apply to all the other systems that are discussed in Ref. [8]. However, this property is not satisfied by our problem, since loops can and do overlap. We have numerically examined the vicinity of $\ell=1/3$, and conclude that there are no singularities in the first and second derivatives of $p(\ell)$ with a prefactor of $O(1)$. Although we cannot rule out the possibility of weaker singularities or unusually small prefactors, the evidence seems to suggest that they do not exist.

V. CONCLUSION

With the help of an efficient Monte Carlo algorithm and analytical upper and lower bounds, we have clarified some of the issues surrounding the behavior of the probability density $p(\ell)$ for small values of its argument, and we have been able to better understand and characterize the essential singularity at $\ell=0$. In this limit, the connection of this problem to SAWs becomes much more transparent, and it is likely that this connection can be further exploited.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, by the MRSEC program through Grant No. DMR-9400396, and through Grants No. DMR-9106237, No. DMR-9417047, and No. DMR-9416910, and by the Israel Science Foundation founded by The Israel Academy of Sciences under Grant No. 246/96.

- [1] Y. Kantor and M. Kardar, *Europhys. Lett.* **27**, 643 (1994); *Phys. Rev. E* **51**, 1299 (1995).
- [2] Y. Kantor and M. Kardar, *Phys. Rev. E* **52**, 835 (1995).
- [3] Y. Kantor and D. Ertaş, *J. Phys. A* **27**, L907 (1994).
- [4] D. Ertaş and Y. Kantor, *Phys. Rev. E* **53**, 846 (1996).
- [5] S. Chandrasekhar, *Rev. Mod. Phys.* **15**, 1 (1943).
- [6] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. (John Wiley and Sons, New York, 1968), Vol. 1.
- [7] B. Derrida and H. Flyvbjerg, *J. Phys. A* **20**, 5273 (1987).
- [8] L. Frachebourg, I. Ispilatov, and P. L. Krapivsky, *Phys. Rev. E* **52**, R5727 (1995).
- [9] In principle, it is straightforward to probe much smaller values of ℓ since the algorithm is only polynomial in ℓ^{-1} . Our numerical results demonstrate what can be achieved with runs that typically last a few days on a workstation such as SUN SPARCstation 5, and were sufficient for the purposes of this study.