Phase transition in a random fragmentation problem with applications to computer science

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **35** (2002) L501–L507 PII: S0305-4470(02)37797-7

LETTER TO THE EDITOR

Phase transition in a random fragmentation problem with applications to computer science

David S Dean and Satya N Majumdar

IRSAMC, Laboratoire de Physique Quantique (UMR 5626 du CNRS), Université Paul Sabatier,
118 route de Narbonne, 31062 Toulouse Cedex 04, France

Received 30 May 2002
Published 2 August 2002
Online at stacks.iop.org/JPhysA/35/L501

Abstract

We study a fragmentation problem where an initial object of size x is broken into m random pieces provided $x > x_0$ where x_0 is an atomic cut-off. Subsequently, the fragmentation process continues for each of those daughter pieces whose sizes are bigger than x_0. The process stops when all the fragments have sizes smaller than x_0. We show that the fluctuation of the total number of splitting events, characterized by the variance, generically undergoes a nontrivial phase transition as one tunes the branching number m through a critical value $m = m_c$. For $m < m_c$, the fluctuations are Gaussian where as for $m > m_c$ they are anomalously large and non-Gaussian. We apply this general result to analyse two different search algorithms in computer science.

PACS numbers: 02.50.−r, 05.40.−a, 89.20.−a

Fragmentation is a widely studied phenomenon [1] with applications ranging from conventional fracture of solids [2] and collision induced fragmentation in atomic nuclei and aggregates [3] to seemingly unrelated fields such as disordered systems [4] and geology [5]. In this paper, we consider a problem where an object of initial size (or length) x is first broken into m random pieces of sizes $x_i = r_i x$ with $\sum_{i=1}^{m} r_i = 1$ provided the initial size $x > x_0$ where x_0 is a fixed 'atomic' threshold. At the next stage, each of those m pieces with sizes bigger than x_0 is further broken into m random pieces and so on. Clearly, the process stops after a finite number of fragmentation or splitting events when the sizes of all the pieces become less than x_0. This problem and its close cousins have already appeared in numerous contexts including the energy cascades in turbulence [6], rupture processes in earthquakes [7], stock market crashes [8], binary search algorithms [9–11], stochastic fragmentation [12] and DNA segmentation algorithms [13]. It therefore comes as somewhat of a surprise that there is a nontrivial phase transition in this problem as one tunes the branching number m through a critical value $m = m_c$.

In this letter, we study analytically the statistics of the total number of fragmentation events $n(x)$ up to the end of the process as a function of the initial size x. We show that, while the average number of events $\mu(x)$ always grows linearly with x for large x, the asymptotic behaviour of the variance $\nu(x)$, characterizing the fluctuations, undergoes a phase transition at a critical value $m = m_c$,

$$\nu(x) \sim \begin{cases} x & m < m_c \\ x^{2\theta} & m \geqslant m_c \end{cases}. \tag{1}$$

The exponent θ is nontrivial and increases monotonically with m for $m \geqslant m_c$ starting at $\theta\ (m = m_c) = 1/2$ and the amplitude of the leading $x^{2\theta}$ term has log-periodic oscillations for $m \geqslant m_c$. This signals unusually large fluctuations in $n(x)$ for $m > m_c$. The full distribution of $n(x)$ also changes from being Gaussian for $m < m_c$ to non-Gaussian for $m > m_c$. This phase transition is rather generic for any fragmentation problem with an 'atomic' threshold. However, the critical value m_c and the exponent θ are nonuniversal and depend on the distribution function of the random fractions r_i. In this letter, we establish this generic phase transition and then calculate explicitly m_c and θ for two special cases with direct applications in computer science. Indeed, it is in the context of the computer science problem of counting the number of nodes of an m-ary search tree that the first special case of this rather generic phase transition was discovered [15].

In this fragmentation problem with a fixed lower cut-off x_0, one first breaks the initial piece of length x provided $x > x_0$ into m pieces of sizes $x_i = r_i x$. The sizes of each of these 'daughters' are then examined. Only those pieces whose sizes exceed x_0 are considered 'active' and those with sizes less than x_0 are considered 'frozen'. Each of the active pieces is then subsequently broken into m pieces and so on. The fractions r_i characterizing a splitting event are considered to be independent from one event to another but are drawn each time from the same joint distribution function $\eta_m(r_1, r_2, \ldots, r_m)$. As the splitting process conserves the total size, the fractions r_i satisfy the constraint $\sum_{i=1}^{m} r_i = 1$. In addition, we consider the splitting process to be isotropic, i.e., all the m daughters resulting from a splitting event are statistically equivalent. This indicates that the marginal distribution of any one of the r_i is independent of i and is given by

$$\eta_1(r) = \int \eta_m(r, r_2, \ldots, r_m) \prod_{i=2}^{m} \mathrm{d}r_i. \tag{2}$$

We will henceforth denote the average over the whole history of the splitting procedure (until the end of the process) by $\overline{\cdots}$ and the average over the r_i associated with a single splitting event by $\langle\cdot\rangle$. The conservation law along with the isotropy implies that $\langle r\rangle = \int \eta_1(r) r\, \mathrm{d}r = 1/m$.

Clearly, the total number of splitting events $n(x) = 0$ if $x < x_0$. On the other hand, if $x > x_0$ there will be at least one splitting and it is easy to write a recursion relation for $n(x)$,

$$n(x) = 1 + \sum_{i=1}^{m} n(r_i x). \tag{3}$$

Using the isotropy of the splitting distribution and taking the average over equation (3), we find that $\mu(x) \equiv \overline{n(x)}$ satisfies the recursion for $x > x_0$,

$$\mu(x) = 1 + m\langle\mu(rx)\rangle = 1 + m \int_{x_0/x}^{1} \mathrm{d}r\, \eta_1(r)\mu(rx) \tag{4}$$

where the lower limit in the above integral comes from the condition $n(x) = 0$ for $x < x_0$. Without any loss of generality we set $x_0 = 1$, i.e., we measure all sizes in units of the atomic size. Since $1 \leqslant x < \infty$ in equation (4), it is convenient to make a change of variable $x = \mathrm{e}^{\alpha}$ so that $0 \leqslant \alpha < \infty$ and write $\mu(\mathrm{e}^{\alpha}) = F(\alpha)$. The resulting equation for $F(\alpha)$ is solved by taking the Laplace transform of equation (4) and one finds that $\tilde{F}(s) = \int_0^{\infty} \mathrm{d}\alpha\, F(\alpha)\, \mathrm{e}^{-s\alpha}$ is given by

$$\tilde{F}(s) = \frac{1}{s[1 - mw(s)]} \tag{5}$$

where $w(s) = \langle r^s \rangle = \int_0^1 \mathrm{d}r\, \eta_1(r) r^s$. Assuming that $\tilde{F}(s)$ has simple poles at $s = \lambda_k$, the Laplace transform in equation (5) can be inverted to obtain $\mu(x) = a_0 + \sum_k a_k x^{\lambda_k}$ with $a_0 = 1/(1-m)$ (coming from the pole at $s = 0$) and $a_k = -1/[m\lambda_k w'(\lambda_k)]$. From the conservation law $\langle r \rangle = 1/m$, one finds that $s = 1$ is always a pole of $\tilde{F}(s)$. Besides, since $0 \leqslant r \leqslant 1$, the pole at $s = 1$ is also the one with the largest real part and hence will dominate the large x behaviour of $\mu(x)$. Let λ and λ^* denote the pair of complex conjugate poles with the next largest real part. Then keeping only the leading corrections to the asymptotic behaviour one finds

$$\mu(x) \approx a_1 x + a_2 x^{\lambda} + a_2^* x^{\lambda^*} \tag{6}$$

where $a_1 = -1/[mw'(1)] = -1\big/\big[m \int_0^1 \mathrm{d}r\, \eta_1(r) r \ln(r)\big]$.

We now turn to the variance $\nu(x)$ of the total number of splittings $n(x)$:

$$\nu(x) = \overline{(n(x) - \mu(x))^2}. \tag{7}$$

By squaring equation (3) and after some straightforward algebra we find the recursion relation

$$\nu(x) = f(x) + m \int_{1/x}^{1} \mathrm{d}r\, \eta_1(r) \nu(rx) \tag{8}$$

where $f(x) = \big\langle \big(\sum_{i=1}^m [\mu(r_i x) - \langle \mu(rx) \rangle]\big)^2 \big\rangle$. Once again the change of variable $x = \mathrm{e}^{\alpha}$ followed by a subsequent Laplace transform with respect to α, $\tilde{\nu}(s) = \int_0^\infty \mathrm{d}\alpha\, \nu(\mathrm{e}^{\alpha})\, \mathrm{e}^{-s\alpha}$ yields

$$\tilde{\nu}(s) = \frac{\tilde{f}(s)}{s[1 - mw(s)]}. \tag{9}$$

Using the asymptotic expression of $\mu(x)$ from equation (6) in the expression for $f(x)$ one finds that the leading term of $f(x)$ for large x is given by

$$f(x) \approx b_1 x^{2\lambda} + b_2 x^{2\lambda^*} + b_3 x^{(\lambda+\lambda^*)} \tag{10}$$

where the b_j are constants. These behaviours indicate that $\tilde{f}(s)$ has poles at $s = 2\lambda$, $s = 2\lambda^*$ and $s = \lambda + \lambda^*$. Thus when $\mathrm{Re}(\lambda) < 1/2$, these poles occur to the left of $s = 1$ in the complex s plane. From equation (9) it follows that the asymptotic large x behaviour of $\nu(x)$ will then be controlled by the $s = 1$ pole arising from the denominator $[1 - mw(s)]$ and $\nu(x) \sim x$ for large x. On the other hand, when $\mathrm{Re}(\lambda) > 1/2$, the dominant poles governing the large x behaviour are the three poles of $\tilde{f}(s)$ with the real part $2\mathrm{Re}(\lambda) > 1$. Hence in that case, $\nu(x) \sim x^{2\theta}$ where $\theta = \mathrm{Re}(\lambda)$. Note also that for $\mathrm{Re}(\lambda) \geqslant 1/2$, the amplitude of the leading term $x^{2\theta}$ in $\nu(x)$ will have log-periodic oscillations due to the nonzero imaginary parts of the poles λ and λ^*. This phase transition will always occur whenever one can tune the pole λ continuously through the critical value $\mathrm{Re}(\lambda) = 1/2$. In the following two examples we show explicitly that this can be achieved, in a natural way, by tuning the branching number m.

The m-ary search tree. It is well known that one of the most efficient ways of sorting the incoming data to a computer is to organize the data on a tree [14]. Consider the sorting of an incoming data string consisting of N distinct elements labelled by the sequence $1, 2, \ldots, N$. Consider a particular random sequence of arrival of these N elements (see figure 1). An m-ary search tree stores this sequence on a growing tree structure where each node of the tree can contain at most $(m-1)$ data points [9, 11]. A node if filled branches into m leaves. The first $(m-1)$ elements are stored in the root of the tree in an ordered sequence $w_1 < w_2 < \cdots < w_{m-1}$. Any subsequent element must belong to one of the m sets of numbers $A_1 = [1, w_1)$, $A_i = (w_i, w_{i+1})$ with $1 \leqslant i \leqslant m-2$ and $A_m = (w_{m-1}, N]$. To each of these sets or intervals we associate a leaf of the tree leading to a new node. A new data

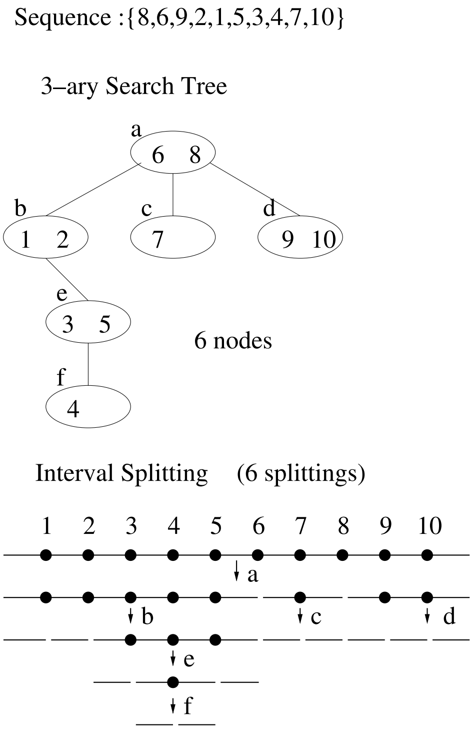

Figure 1. Top: the construction of the 3-ary tree for a sequence of numbers between 1 and 10. Bottom: the induced random interval splitting. The nodes and corresponding splittings are labelled a, b, c, d, e and f. An interval can only split if it contains a •.

point w, arriving subsequently, is sent down the leaf corresponding to the set A_i if $w \in A_i$ and is stored in a daughter node at the base of that leaf. Once a daughter node is filled with $m - 1$ numbers it, in turn, gives rise to m new leaves and so on. An example for a 3-ary search tree with $N = 10$ is shown in figure 1.

Each sequence of the incoming data will give rise to a different m-ary tree configuration. If the incoming data are random, all the trees occur with equal probability. It is easy to see that the total number of occupied nodes M (each containing at least one element) is a random variable as it varies from one tree configuration to another, except for $m = 2$ where $M = N$. The statistics of M was recently studied by computer scientists using rather involved combinatorial analysis and it was found that while $\overline{M} \sim N$ for large N, the variance $\nu \sim N$ for $m < 26$ and as $\sim N^{2\theta}$ for $m > 26$ [15]. We show below that this strange result is just a special case of the general phase transition in the fragmentation problem discussed here.

The construction of the m-ary search tree can be mapped exactly onto the splitting of the interval $[1, N]$ [9, 11]. It is easy to see that the incoming elements $w_1, w_2, \ldots, w_{m-1}$ split the initial interval into m parts A_i. If all the $N!$ possible sequences arrive with equal probability then the points $w_1, w_2, \ldots, w_{m-1}$ are distributed uniformly on $[1, N]$ (these are the numbers stored in the first node). We split the interval $[1, N]$ into m subintervals corresponding to the A_i introduced above. If a subinterval A_i is empty (i.e. has no • in figure 2) then no data points can go down the corresponding leaf and hence such an interval (of length < 2) will not split further. If the subset A_i contains only one •, the arrival of the corresponding single data point still splits the interval into m parts (some of the intervals so created may be of length 0).

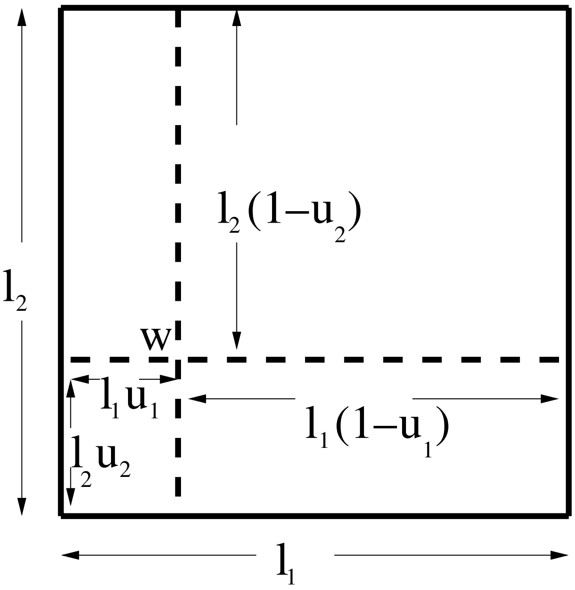

Figure 2. An example of the splitting of a rectangle into four daughter rectangles about the point $\boldsymbol{w} = [l_1u_1, l_2u_2]$. The same process is continued on each daughter until the area of the daughter becomes less than x_0.

This corresponds to the atomic threshold $x_0 = 2$ in our general problem and $x = N/2$ corresponds to the initial size in units of the atomic size.

The crucial point is that the number of occupied nodes M in the m-ary search tree is identical to the number of splittings n $(x = N/2)$ in this fragmentation problem. For large N one can pass to a continuum limit and use the known marginal probability density function $\eta_1(r) = (m-1)(1-r)^{m-2}$ [10, 11] for the continuum interval splitting problem in our general formula. We get $\overline{M} = \mu(x = N/2) \approx a_1N/2$ for large N where $a_1 = 1\big/\big[\sum_{k=2}^{m} 1/k\big]$. Also $w(s) = \langle r^s\rangle = (m-1)B(s+1, m-1)$ where $B(m, n)$ is the standard Beta function. Therefore, the poles of $\tilde{F}(s)$ in equation (5) occur at the roots of the equation $m(m-1)B(s+1, m-1) = 1$. It is easy to check using *Mathematica* that one can arrive at the critical condition $\mathrm{Re}(\lambda) = 1/2$ by tuning m through the value $m = m_c \approx 26.0461\ldots$. Therefore, from our general theory, we find that the variance $\nu \sim N$ for $m < m_c$ and $\nu \sim N^{2\theta}$ for $m > m_c$. The exponent $\theta = \mathrm{Re}(\lambda)$ where λ is the root of $m(m-1)B(s+1, m-1) = 1$ that is closest (to the left) to $s = 1$ when $m > m_c$.

Cuboid splitting. In the previous problem we have considered the sorting of a data string where each element is a scalar. A natural generalization is when each element is a D-dimensional vector $\boldsymbol{w}$ whose kth component $w_k \in [0, l_k]$ for $1 \leqslant k \leqslant D$ [16]. The first element of the data string is then assigned to the point $\boldsymbol{w}$ in the cuboid of edge lengths l_k. If the first element is random, then its components $w_k = u_kl_k$, where the u_k are independent random variables uniformly distributed on [0, 1]. Once this first element is stored, it splits the original cuboid into 2^D sub-cuboids obtained by drawing D lines perpendicular to each of the faces of the cuboid (see figure 2). When the second vector arrives, one compares its components with those of the first element and places it in one of the 2^D sub-cuboids (and thereby splits that sub-cuboid) and the process continues. After each splitting event, the dimensions of the $m = 2^D$ new sub-cuboids can be represented by $l_k'(\sigma) = l_ku_k(1+\sigma_k)/2+l_k(1-u_k)(1-\sigma_k)/2$,

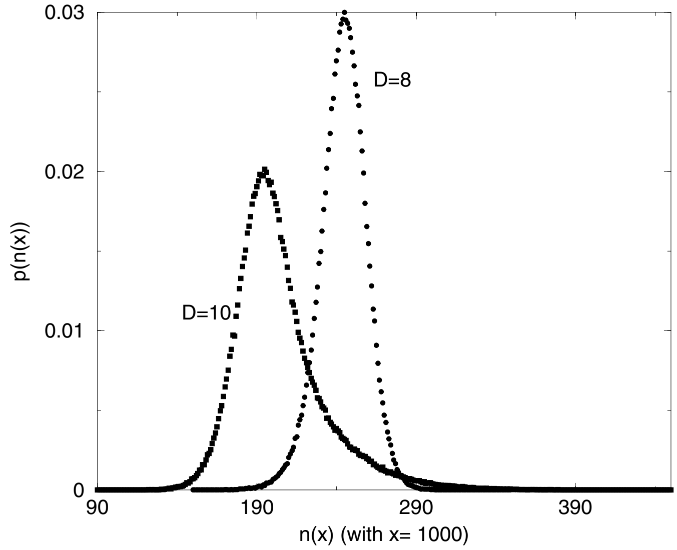

Figure 3. The distribution $p(n(x))$ of the number of splittings of a cuboid of original volume $x = 1000$ for $D = 8$ (filled circles) and for $D = 10$ (filled squares). The distribution is Gaussian for $D = 8$, but has a non-Gaussian skewness for $D = 10$. The histogram was formed by numerically splitting 5×10^5 samples in each case.

where the σ_k are Ising spins. An example with $D = 2$ is shown in figure 2. We know that the random fragmentation of a cuboid was studied in the physics literature, though the questions addressed were different [17].

The volume of any of the sub-cuboids upon splitting the cuboid of volume x is $x' = xr(\sigma)$, where $r(\sigma) = \prod_{k=1}^{D} (u_k(1+\sigma_k)/2 + (1-u_k)(1-\sigma_k)/2)$. Hence, in this problem, one has $m = 2^D$ and the marginal distribution of a given $r(\sigma)$ can be shown to be

$$\eta_1(r) = \frac{[-\ln(r)]^{D-1}}{(D-1)!} \qquad 0 \leqslant r \leqslant 1. \tag{11}$$

From equation (5) with this marginal distribution one finds that

$$\tilde{F}(s) = \frac{1}{s\left(1 - \frac{2^D}{(s+1)^D}\right)}. \tag{12}$$

One then finds $\mu(x) \approx 2x/D$ for large x. The function $\tilde{F}(s)$ has a total of $(D+1)$ poles: one at $s = 0$ and the others at $s = -1 + 2\mathrm{e}^{2\pi\mathrm{i}n/D}$ with $n = 0, 1, \ldots, (D-1)$. The poles closest to the left of $s = 1$ are the complex conjugate pair $\lambda = -1 + 2\mathrm{e}^{2\pi\mathrm{i}/D}$ and $\lambda^* = -1 + 2\mathrm{e}^{-2\pi\mathrm{i}/D}$. Thus $\mathrm{Re}(\lambda) = -1 + 2\cos(2\pi/D)$. From our general theory, it follows that by tuning m or equivalently D, it is possible to encounter the critical point $\mathrm{Re}(\lambda) = 1/2$ at $D = D_c = \pi/\sin^{-1}(1/2\sqrt{2}) = 8.69\ldots$. Hence the variance $\nu(x) \sim x$ for $D < D_c$, and for $D \geqslant D_c$, $\nu(x) \sim x^{2\theta}$ where $\theta = \mathrm{Re}(\lambda) = 2\cos(2\pi/D) - 1$.

We have verified the above predictions by numerically carrying out the splitting procedure on a large number of samples with atomic cut-off $x_0 = 1$. The analytical predictions for the mean and the variance are well verified though an accurate measurement of the exponent θ is difficult due to statistical fluctuations and finite size corrections. We have also measured the histogram of the number of splittings. For $D < D_c$ this distribution is Gaussian; however, for $D > D_c$ the distribution becomes skewed towards large values of $n(x)$ having an anomalous tail. Shown in figure 3 is the distribution measured for $D = 8$ and $D = 10$. The difference is clearly visible. The non-Gaussian behaviour is also visible for the case $D = 9$ but less pronounced as $\mathrm{Re}(\lambda)$ is quite close to $1/2$. While we can rigorously prove that the distribution is indeed Gaussian in the sub-critical regime, we have not been able to calculate the full distribution in the supercritical regime. Qualitatively, it is clear however that as D increases the volume of the cuboid becomes more concentrated about its surface and hence the splitting point $\boldsymbol{w}$ for large D is generically closer to the surfaces. This means that the splitting procedure will tend to cut the cuboid into more unequal pieces than at lower dimensions, a mixture of *blocks* of larger volume and *slices* of smaller volume. It is thus the blocks which are sliced rather than split in their middle which contribute to the long tail in the distribution of $n(x)$.

In conclusion we have shown that a fragmentation process with an atomic threshold can undergo a nontrivial phase transition in the fluctuations of the number of splittings at a critical value of the branching number m. The calculation of the full probability distribution of the number of splittings remains a challenging unsolved problem. We have provided applications of our general results in two computer science problems. The mechanism of this transition is remarkably simple and therefore one expects it to be rather generic with broad applications since many random processes can be mapped to the type of fragmentation model considered here.

References

[1] For a general review of fragmentation, see Redner S 1990 *Statistical Models for the Fracture of Disordered Media* ed H J Herrmann and S Roux (New York: Elsevier)
[2] Lawn B R and Wilshaw T R 1975 *Fracture of Brittle Solids* (Cambridge: Cambridge University Press)
[3] Campi X, Krivine H, Sator N and Plagnol E 2000 *Eur. Phys. J.* D **11** 233
[4] Derrida B and Flyvbjerg H 1987 *J. Phys. A: Math. Gen.* **20** 5273
Flyvbjerg H and Kjaer N J 1988 *J. Phys. A: Math. Gen.* **21** 1695
Higgs P G 1995 *Phys. Rev.* E **51** 95
Derrida B and Jung-Muller B 1999 *J. Stat. Phys.* **94** 277
[5] Turcotte D L 1986 *J. Geophys. Res.* **91** 1921
[6] Greiner M, Eggers H C and Lipa P 1998 *Phys. Rev. Lett.* **80** 5333
[7] Newman W I and Gabrielov A M 1991 *Int. J. Fract.* **50** 1
Newman W I, Gabrielov A M, Durand T A, Phoenix S L and Turcotte D L 1994 *Physica* D **77** 200
[8] Sornette D and Johansen A 1998 *Physica* A **261** 581
[9] Devroye L 1986 *J. ACM* **33** 489
[10] Krapivsky P L and Majumdar S N 2000 *Phys. Rev. Lett.* **85** 5492
[11] Majumdar S N and Krapivsky P L 2002 *Phys. Rev.* E **65** 036127
[12] Krapivsky P L, Ben-Naim E and Grosse I 2001 *Preprint* cond-mat/0108547
[13] Bernaola-Galván P, Román-Roldán R and Oliver J L 1996 *Phys. Rev.* E **53** 5181
Li W 2001 *Phys. Rev. Lett.* **86** 5815
[14] Knuth D E 1988 *The Art of Computer Programming, Sorting and Searching* vol 3, 2nd edn (Reading, MA: Addison-Wesley)
[15] Mahmoud H M and Pittel B 1989 *J. Algorithms* **10** 52
Chern H-H and Hwang H-K 2001 *Random Struct. Algorithms* **19** 316
[16] Finkel R A and Bentley J L 1974 *Acta Inform.* **4** 1
[17] Krapivsky P L and Ben-Naim E 1994 *Phys. Rev.* E **50** 3502