Extremal properties of random trees

E. Ben-Naim[1], P. L. Krapivsky[2], Satya N. Majumdar[3,4]
[1] *Theoretical Division and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[2] *Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA*
[3] *CNRS, IRSAMC, Laboratoire de Physique Quantique, Universite' Paul Sabatier, 31062 Toulouse, France*
[4] *Tata Institute of Fundamental Research, Homi Bhabba Road, Mumbai-400005, India*

We investigate extremal statistical properties such as the maximal and the minimal heights of randomly generated binary trees. By analyzing the master evolution equations we show that the cumulative distribution of extremal heights approaches a traveling wave form. The wave front in the minimal case is governed by the small-extremal-height tail of the distribution, and conversely, the front in the maximal case is governed by the large-extremal-height tail of the distribution. We determine several statistical characteristics of the extremal height distribution analytically. In particular, the expected minimal and maximal heights grow logarithmically with the tree size, N, $h_{\rm min} \sim v_{\rm min} \ln N$, and $h_{\rm max} \sim v_{\rm max} \ln N$, with $v_{\rm min} = 0.373365\ldots$ and $v_{\rm max} = 4.31107\ldots$, respectively. Corrections to this asymptotic behavior are of order $\mathcal{O}(\ln \ln N)$.

PACS numbers: 02.50.-r, 05.40.-a, 89.20.Ff

Random trees play an important role in data storage and retrieval algorithms in computer science [1–6]. They also arise in physical situations such as collision processes in gases [7], random fragmentation processes [8,9], and diffusion-limited aggregation [10]. In each case, extremal characteristics such as the maximal or the minimal height of the tree, namely, the maximal [3] or the minimal [4] number of bonds separating the tree root from a node are of interest. In data storage algorithms, these distances yield the best-case or the worst-case-scenario performances. In kinetic theory, the largest Lyapunov exponent is related to the maximum height problem [7].

In this article, we study extremal properties of randomly generated binary trees using rate equation theory. Techniques developed in aggregation processes are well suited for treating random trees since the tree merger process is simply an aggregation process. We study the distributions of extremal (both minimal and maximal) heights of a tree. In both cases, the average extremal tree height grows logarithmically with the number of leaves and the cumulative distribution of extremal tree heights approaches a traveling wave solution. The logarithmic growth prefactors equal the traveling wave velocities, which are set by a velocity selection principle. These velocities can be alternatively obtained using a simpler (and independent) intuitive argument. Interestingly, the wave front in the minimal case is determined by the small-height tail of the distribution, while in the maximal case it is determined by the large-height tail of the distribution.

Let us introduce the tree generation model. Initially, the system consists of an infinite number of trivial (single-leaf) trees. Then, two trees are picked at random and attached to a common root. This merging process is repeated indefinitely with rate set to 2 without loss of generality. Let $c(t)$ be the number density of trees at time t. Initially, $c(0) = 1$, and since this quantity evolves according to $dc/dt = -c^2$ one has

$$c(t) = \frac{1}{1+t}. \tag{1}$$

Mass conservation implies that N, the average number of leaves in a tree, grows linearly with time, $N = c^{-1} = 1+t$. While the corresponding mass distribution has been extensively studied in coalescence processes [11,12], we are interested here in the distribution of extremal characteristics such as the minimal and maximal number of bonds between the tree root and its nodes. The leading behavior of the average of these distributions can be obtained from the following intuitive argument.

The distribution of tree heights, namely, of the distances between the tree root and the nodes can be obtained immediately. Let $P_n(t)$ be the probability that the distance between a randomly chosen leaf and the root of the parent tree equals n at time t. As the tree generation process is random, $P_n(t)$ obeys Poisson statistics

$$P_n(t) = \frac{[h(t)]^n}{n!} e^{-h(t)}, \tag{2}$$

with $h(t)$ the average tree height. Consider a leaf in the system. Each time its corresponding tree merges with another tree, the distance to the root is augmented by one. This process occurs with rate 2 and hence, $dh/dt = 2c$. Integrating this equation subject to the initial condition $h(0) = 0$ yields $h(t) = 2\ln(1+t) = 2\ln N$. One anticipates that the expected minimal number grows logarithmically as well, $h_{\rm min} \simeq v_{\rm min} \ln N$. To estimate $h_{\rm min}$ we sum the small-n tail of the normalized height distribution $c^{-1}P_n$ to unity, $\sum_{n=0}^{h_{\rm min}} c^{-1} P_n = 1$. Substituting Eqs. (1)–(2), the relation $h_{\rm min} \simeq v_{\rm min} \ln N$, and the Stirling formula $\ln n! \sim n\ln n - n$ into the above relation we obtain the transcendental equation

$$v \ln \frac{2e}{v} = 1. \tag{3}$$

This equation has two solutions with the lower (higher) velocity corresponding to the growth of the average minimal (maximal) tree height. Indeed, repeating the above steps for the maximal height using $\sum_{n=h_{\max}}^{\infty} c^{-1}P_n = 1$ again leads to the same equation. Solving Eq. (3) yields

$$\begin{aligned} h_{\min} &\simeq v_{\min} \ln N, & v_{\min} &= 0.373365; \\ h_{\max} &\simeq v_{\max} \ln N, & v_{\max} &= 4.31107. \end{aligned} \tag{4}$$

This probabilistic argument correctly predicts both velocities, and additionally, it demonstrates that the two extremal statistics are intimately related. We note that the latter maximal height value has emerged from quite different calculations in studies of collision processes in gases [7] and fragmentation processes [8,9].

We now turn to studying the entire distribution of extremal characteristics. Rather than considering the two extremal height distributions separately, we study a more general model which interpolates between the two cases. In this model, each tree carries an extremal height k. The result of a merger between trees with extremal heights k_1 and k_2, is a new tree with extremal height k given by

$$k = \begin{cases} \min(k_1, k_2) + 1 & \text{with prob.} \;\; p, \\ \max(k_1, k_2) + 1 & \text{with prob.} \;\; 1 - p\,. \end{cases} \tag{5}$$

Here, p is a mixing parameter whose limits $p = 1$ and $p = 0$ correspond to the minimal and the maximal heights problems, respectively.

The number density of trees with extremal height k, $c_k(t)$, evolves according to the master equation

$$\frac{dc_k}{dt} = c_{k-1}^2 - 2cc_k + 2pc_{k-1}\sum_{j=k}^{\infty} c_j + 2(1-p)c_{k-1}\sum_{j=0}^{k-2} c_j. \tag{6}$$

Here $c = \sum_{j\geq 0} c_j$ is the total tree density and one can verify that it indeed evolves according to $dc/dt = -c^2$. The master equation (6) should be solved subject to the initial condition $c_k(0) = \delta_{k,0}$. It proves useful to introduce the cumulative fractions

$$A_k = c^{-1}\sum_{j=k}^{\infty} c_j, \tag{7}$$

and a new time variable

$$T = \int_0^t d\tau\, c(\tau) = \ln(1+t). \tag{8}$$

These variables recast Eqs. (6) into

$$\frac{dA_k}{dT} = -A_k + 2(1-p)A_{k-1} + (2p-1)A_{k-1}^2, \tag{9}$$

which should be solved subject to the step function initial conditions, $A_k(0) = 1$ for $k \leq 0$ and $A_k(0) = 0$ otherwise.

In the long time limit, $A_k(T)$ approaches a traveling wave form, $A_k(T) \to A(k - vT)$, with $A(x)$ being a solution of the nonlinear difference-differential equation

$$vA'(x) = A(x) - 2(1-p)A(x-1) - (2p-1)A^2(x-1), \tag{10}$$

subject to the boundary conditions $A(-\infty) = 1$ and $A(\infty) = 0$. Fortunately, the velocity v can be determined without solving the nonlinear nonlocal equation (10) exactly. To determine v, it is enough to analyze the asymptotic behavior of the front at one of its two tails. Different considerations apply in the regions $p \leq 1/2$ and $p \geq 1/2$. We first consider the case $1/2 \leq p \leq 1$. Here, Eq. (10) admits an exponential solution in the small-k tail, $A(x) \to 1 - e^{\lambda x}$ as $x \to -\infty$. Substituting this form in Eq. (10), we find that the yet to be determined velocity v and decay exponent λ are related via

$$v = \frac{1 - 2pe^{-\lambda}}{\lambda}. \tag{11}$$

While a class of velocities is in principle possible, the extremum value is selected for compact initial conditions. This behavior is similar to velocity selection occurring for example in the classic Fisher reaction-diffusion equation [13,14]. Evaluating this extremum yields a generalization of the transcendental equation (3)

$$v \ln\frac{2ep}{v} = 1, \qquad\qquad p > \frac{1}{2}. \tag{12}$$

In particular, $v \to 1$ when $p \to 1/2$. In the minimal height case ($p = 1$) we recover the aforementioned value $v_{\min} = 0.373365$. The selected decay coefficient satisfies $2p(1+\lambda) = e^{\lambda}$ and in the minimal case $\lambda = 1.67835$.

Let us now turn to the complementary $p < 1/2$ case where contrary to the $p > 1/2$ case, the large-extremal-height tail admits an exponentially decaying solution $A(x) = e^{-\mu x}$, as $x \to +\infty$. Here, the yet to be determined velocity and decay coefficient are related via

$$v = \frac{2(1-p)e^{\mu} - 1}{\mu}. \tag{13}$$

Again, applying the velocity selection principle implies that the minimal possible velocity is selected. The selected decay coefficient satisfies $2(1-p)(1-\mu) = e^{-\mu}$ and the selected velocity obeys

$$v \ln\frac{2e(1-p)}{v} = 1, \qquad\qquad p < \frac{1}{2}. \tag{14}$$

In the maximal case ($p = 0$) one recovers the velocity $v_{\max} = 4.31107$ and additionally, the decay coefficient is $\mu = 0.768039$. While we have not proved this selection principle, our numerical integration of Eq. (9) supports the findings in the minimal and the maximal cases. We confirmed that a traveling wave solution is indeed approached, and that the selected velocities fall within 0.1% of the theoretical values. Curiously, this is a unique case where velocity selection is also supported by an independent physical argument.

Interestingly, different mechanisms drive the front in the regions $p < 1/2$ and $p > 1/2$. In the region $p > 1/2$ which includes the minimal case, the small-extremal-height tail of the distribution dictates the velocity and in fact, the entire distribution is enslaved to this exponential tail. The opposite is true for $p < 1/2$ which includes the maximal case. Here, the large-extremal-height tail of the distribution governs the wave velocity and the wave form. These behaviors are physical, especially when the limiting cases are considered: the distribution of minimal (maximal) tree heights is governed by extremely small (large) fluctuations. The point $p_c = 1/2$ can be regarded as a critical point. In particular, the different velocity equations (12) and (14) imply non-analytic behavior at $p_c = 1/2$ where $v_c = 1$, and analysis of the leading behavior in the vicinity of this point yields

$$|v - v_c| \simeq 2|p - p_c|^{1/2}, \quad \text{when} \quad p \to p_c. \tag{15}$$

Exact analysis of the special case $p = 1/2$ is given below.

Asymptotically, while the wave front advances at a constant rate v, there is a slow T^{-1} correction in the leading order, resulting in a logarithmic correction to the front position. A similar correction was first derived by Bramson in the context of reaction-diffusion equations [14], and was subsequently generalized [15–17]. We now calculate the leading correction employing the approach of Ref. [16]. Let us first consider the $p > 1/2$ case. Substituting $A_k(T) = 1 - a_k(T)$ into Eq. (9), and ignoring the quadratic term a_{k-1}^2, yields

$$\frac{da_k}{dT} = -a_k + 2pa_{k-1}. \tag{16}$$

Substituting the scaling solution

$$a_k(T) = T^\alpha G(x\,T^{-\alpha})e^{\lambda x}, \quad x = k - vT - w(T). \tag{17}$$

into (16) shows that different leading orders are compatible provided that the exponent $\alpha = 1/2$ and that the correction to the front location is $w(T) = \beta \ln T$. The former constraint reflects a hidden diffusive scale and the latter gives the aforementioned logarithmic correction with yet undetermined amplitude β. Substituting these behaviors in Eq. (16) we get

$$\begin{aligned} 0 &= T^{1/2}\left(2pe^{-\lambda} - 1 + \lambda v\right)G(z) + \left(v - 2pe^{-\lambda}\right)G'(z) \\ &+ T^{-1/2}\left[e^{-\lambda}G''(z) + \frac{1}{2}\,zG'(z) + \frac{2\lambda\beta - 1}{2}\,G(z)\right] + \ldots \end{aligned}$$

The terms of the leading $\mathcal{O}(T^{1/2})$ order cancel when $2pe^{-\lambda} + \lambda v = 1$, i.e., when the velocity v and the decay exponent λ are related through Eq. (11). The terms of order $\mathcal{O}(1)$ cancel when $2pe^{-\lambda} = v$. Remarkably, this relation together with Eq. (11) hold only in the extremal point where v and λ are given by Eqs. (12) and (11), respectively. Finally, the terms in the lowest $\mathcal{O}(T^{-1/2})$ order cancel when the scaling function $G(z)$ satisfies the parabolic cylinder equation

$$v\frac{d^2G}{dz^2} + z\,\frac{dG}{dz} + (2\beta\lambda - 1)G(z) = 0. \tag{18}$$

This differential equation should be solved subject to the appropriate boundary conditions: (i) $G(z) \to 0$ for $z \to -\infty$ as $B_k(T)$ must vanish when $T \to \infty$, and (ii) $G(z) \sim z$ for $z \to 0$ to ensure that $B_k(T)$ is independent of k for large k. The former boundary condition selects one of the two possible solutions, $G(z) = C\,e^{-y^2/4}\,D_\nu(y)$, where $y = z/\sqrt{v}$, $\nu = 2(\beta\lambda - 1)$, and $D_\nu(y)$ is the parabolic cylinder function with index ν. The second boundary condition fixes the index, $\nu = 1$, implying $\beta = 3/(2\lambda)$. Hence, we find that the expected extremal tree height, $\langle k\rangle = c^{-1}\sum_k kc_k$, grows with N as

$$\langle k\rangle = v\ln N + \frac{3}{2\lambda}\,\ln\ln N, \tag{19}$$

where the relation $T = \ln N$ was used. Obviously, the latter log-log correction can not be obtained from the heuristic argument presented earlier. The same analysis can be carried out for the $p < 1/2$ case where we find,

$$\langle k\rangle = v\ln N - \frac{3}{2\mu}\,\ln\ln N, \tag{20}$$

with v and μ given by Eqs. (14), and (13), respectively. For completeness, we merely quote results of a more sophisticated approach [17] which allows calculation of the second leading correction to the growth rate

$$\frac{d\langle k\rangle}{dT} = \begin{cases} \frac{3}{2\lambda}\,T^{-1} + \frac{3}{\lambda^2}\sqrt{\frac{\pi}{2v}}\,T^{-3/2}, & p > 1/2, \\ -\frac{3}{2\mu}\,T^{-1} + \frac{3}{\mu^2}\sqrt{\frac{\pi}{2v}}\,T^{-3/2}, & p < 1/2. \end{cases} \tag{21}$$

We now return to the critical case $p = 1/2$. The critical behavior is simpler because Eq. (9) becomes linear, and it can be easily solved by the generating function method. From Eq. (9) we find that the generating function $Q(z,T) = \sum_{k\geq 1} A_k z^k$ satisfies the differential equation $dQ/dT = z + (z-1)Q(z,T)$, subject to the initial condition $Q(z,0) = 0$. This equation admits the solution $Q(z,T) = z[1 - e^{-(1-z)T}]/(1-z)$, and expanding in powers of z gives the cumulative distribution

$$A_k(T) = e^{-T}\sum_{m=k}^{\infty}\frac{T^m}{m!}. \tag{22}$$

Using Eq. (7) and $c(T) = e^{-T}$, we find that the extremal height distribution is proportional to a Poisson distribution, $c_k(T) = e^{-2T}T^k/k!$. Asymptotically, the normalized height distribution approaches a Gaussian, $c^{-1}c_k(T) \to \exp[-(k-T)^2/2T]/\sqrt{2\pi T}$, and the cumulative fractions easily follow

$$A_k(T) \to \frac{1}{2}\mathrm{erf}\left[\frac{k-T}{\sqrt{2T}}\right]. \tag{23}$$

Hence, both tails are Gaussian, consistent with the fact that $\lambda = \mu = 0$ when $p = 1/2$. Interestingly, the hidden diffusive scale becomes pronounced, and the wave front broadens indefinitely with a width of the order $\sqrt{T}$.

Thus far, we have considered closed systems where only trivial trees are initially present. However, in many applications such as data storage algorithms, as well as in physical situations such as river networks [18] and fragmentation processes [19], there may be a constant input into the system. We therefore consider the natural case where an initially empty system is subject to a uniform input of trivial trees. In this case, we must add an additional input term $\delta_{k,0}$ into the right-hand side of Eq. (6). The initial condition now reads $c_k(0) = 0$. The overall density evolves according to $dc/dt = 1 - c^2$, i.e., $c(t) = \tanh(t)$. Hence, the system eventually reaches a steady state with density $c = 1$.

We restrict our attention to the steady state distributions which can be obtained by equating the time derivatives in the master equations (6) to zero. Again, we introduce the cumulative densities $B_k = \sum_{j=k}^{\infty} c_j$ which at the steady state satisfy $B_0 = c = 1$ and

$$B_k = (1-p)B_{k-1} + (p-1/2)B_{k-1}^2 \tag{24}$$

for $k \geq 1$. Three different behaviors arise depending on whether $p = 1$, $0 < p < 1$, or $p = 0$. For $p = 1$, solving (24) recursively gives $B_k = 2^{-(2^k-1)}$. Therefore, $c_k = 2^{-(2^k-1)}\left(1 - 2^{-2^k}\right)$, implying that the minimal height distribution decays as an unusual double-exponential. For $0 < p < 1$, one can neglect the nonlinear term in Eq. (24) in the large k limit. Thence, $B_k \sim (1-p)^k$ implying a generic exponential decay of the distribution c_k at large k. The critical behavior disappears and the only notable feature of the $p = 1/2$ case is that it is exactly solvable, $B_k = 2^{-k}$. For the maximal height problem ($p = 0$), the recursion (24) simplifies (in the large k limit) to a differential equation $dB/dk = -B^2/2$ which is solved to give $B_k \simeq 2k^{-1}$. Thus, the maximal height distribution exhibits a power-law decay, $c_k \simeq 2k^{-2}$ for large k. To summarize, we quote (up to a numeric prefactor) the three leading large-k behaviors

$$c_k \sim \begin{cases} 2^{-2^k}, & p = 1; \\ (1-p)^k, & 0 < p < 1; \\ k^{-2}, & p = 0. \end{cases} \tag{25}$$

Thus, input dramatically alters the height distributions with a wide array of possible outcomes including double exponential, exponential, and algebraic decays.

In summary, we have shown that extremal properties of random trees can be obtained by analyzing the corresponding nonlinear evolution equations. The cumulative distributions of extremal tree heights approach a traveling wave solution. The mean extremal values grow logarithmically with the tree size and there is an additional weak double logarithmic correction. The corresponding growth velocities were obtained from an elementary probabilistic argument and from an extremum selection criteria on the traveling wave solution. Interestingly, while the traveling wave velocity and form in the minimal case is determined by extremely small height fluctuations, the opposite holds for the maximal case. The transition between these two behaviors is marked by a sharp phase transition in a model which interpolates between the two extremal characteristics. Additionally, we have showed that the presence of input may lead to double exponential, exponential, or even algebraic decays of the extremal height distribution. It will be interesting to apply rate equation theory to a closely related random tree characteristics, e.g., the bifurcation ratio and the rank [10,18].

We are thankful to M. B. Hastings and Z. Toroczkai for interesting discussions, and to the DOE (W-7405-ENG-36) and NSF(DMR9978902) for financial support.

[1] J. M. Robson, Austr. Comput. J. **11**, 151 (1979).
[2] B. Pittel, J. Math. Anal. Appl. **103**, 461 (1984).
[3] L. Devroye, J. ACM **33**, 489 (1986).
[4] P. Noguiera, Disc. Appl. Math. **109**, 253 (2001).
[5] H. M. Mahmoud, *Evolution of Random Search Trees* (John Wiley, New York, 1992).
[6] D. E. Knuth, *The Art of Computer Programming, vol. 3, Sorting and Searching* (Addison-Wesley, Reading, 1998).
[7] R. van Zon, H. van Beijeren, and Ch. Dellago, Phys. Rev. Lett. **80**, 2035 (1998).
[8] T. Hattori and H. Ochiai, preprint (1998).
[9] P. L. Krapivsky and S. N. Majumdar, Phys. Rev. Lett. **85**, 5492 (2000).
[10] I. Yekutieli and B. B. Mandelbrot, J. Phys. A **27**, 285 (1994); T. C. Halsey, Europhys. Lett. **39**, 43 (1997).
[11] M. V. Smoluchowski, Z. Phys. Chem. **92**, 215 (1917).
[12] For a recent review, see D. Aldous, Bernoulli **5**, 3 (1999).
[13] J. D. Murray, *Mathematical Biology* (Springer-Verlag, New York, 1989).
[14] M. Bramson, *Convergence of Solutions of the Kolmogorov Equation to Travelling Waves* (American Mathematical Society, Providence, R.I., 1983).
[15] W. van Saarloos, Phys. Rev. A **39**, 6367 (1989).
[16] E. Brunet and B. Derrida, Phys. Rev. E **56**, 2597 (1997).
[17] U. Ebert and W. van Saarloos, Phys. Rev. Lett. **80**, 1650 (1998); Physica D **146**, 1 (2000).
[18] P. S. Dodds and D. H. Rothman, Phys. Rev. E **59**, 4865 (1999).
[19] E. Ben-Naim and P. L. Krapivsky, Phys. Lett. A **275**, 48 (2000).