P H Y S I C A L   R E V I E W   L E T T E R S

Finite-Size Scaling in Extreme Statistics

G. Györgyi, N. R. Moloney, K. Ozogány, and Z. Rácz

Institute for Theoretical Physics—HAS, Eötvös University, Pázmány sétány 1/a, 1117 Budapest, Hungary

We study the deviations from the limit distributions in extreme value statistics arising due to the finite size (FS) of data sets. A renormalization method is introduced for the case of independent, identically distributed (iid) variables, showing that the iid universality classes are subdivided according to the exponent of the FS convergence, which determines the leading order FS shape correction function as well. It is found that, for the correlated systems of subcritical percolation and $1/f^\alpha$ stationary ($\alpha < 1$) noise, the iid shape correction compares favorably to simulations. Furthermore, for the strongly correlated regime ($\alpha > 1$) of $1/f^\alpha$ noise, the shape correction is obtained in terms of the limit distribution itself.

Extreme value statistics (EVS) has been much studied in engineering [1], finance [2], and environmental sciences [3] where extreme events may have disastrous consequences. Recently, interest in EVS has also been growing in physics, e.g., in spin glasses [4], interface fluctuations [5,6], and front propagations [7]. Unfortunately, the use of EVS is hampered by the cost of acquiring good quality statistics: EVS is derived from the extremes of subsets of a data set, requiring abundant data for reasonable statistics. Data analysis is further complicated by the fact that while the EVS limit distribution may be known, the convergence with increasing sample size is slow. Clearly, a detailed finite-size (FS) analysis providing the *convergence rate* and *shape corrections* to the limit distribution is much needed. While for independent, identically distributed (iid) variables FS studies exist in the mathematical literature [8], for correlated systems the convergence rate and shape corrections are known only in a few cases, such as Brownian motion [9] and random matrix problems (see, e.g., [10]).

In this Letter we use analytic and phenomenological approaches, combined with simulations, to investigate FS scaling in EVS. First, we develop a renormalization group (RG) method, in which the limit distribution is a fixed point of the flow in function space of the finite-sample EVS distributions. Applied to iid variables, the approach provides an intuitive and accessible summary of the mathematical results for the leading FS correction, including the explicit forms of the shape corrections (scaling functions). Next, we consider two systems with correlated variables, namely percolation and $1/f^\alpha$ signals. We numerically study the distribution of the largest clusters in subcritical percolation. While the limit distribution is known to be an iid problem [11,12], we find that even the FS correction fits the iid prediction well. In the case of the maximum statistics of $1/f^\alpha$ signals, $0 \leq \alpha < 1$ corresponds to the weakly correlated regime, with an iid limit distribution [13]. Our simulations indicate that the FS properties are very close to the iid case for $0 \leq \alpha \lesssim 0.5$, but deviations appear for $0.5 \lesssim \alpha < 1$. For $\alpha > 1$, however, the convergence be-

comes fast (power law) and we can show that, under a mild assumption, the FS shape correction is given in terms of the limit distribution and, furthermore, the order as well as the shape of the correction strongly depends on the way the distribution is scaled.

The case of iid variables has been extensively studied [14], and we begin our FS study by a reinterpretation of the original derivation of the extreme limit distributions [15]. Consider random variables $z_1, z_2, \ldots, z_N$ with parent density $\rho(z)$ and integrated distribution $\mu(z) = \int_{-\infty}^{z} \rho(s)ds$. The maximum of the z_i has the integrated distribution $\mu^N(z)$, and the basic observation [15] is that if, after an appropriate scale change $z = a_N x + b_N$, $\mu^N(z)$ tends to a limit distribution $M(x)$ as $N \rightarrow \infty$, then the same limit should be reproduced for another $N' = pN$. This requirement can be cast in the form

$$M(x) = [\hat{R}_p M](x) \equiv M^p(a_p x + b_p), \qquad (1)$$

where the right-hand side defines $\hat{R}_p$, which can be interpreted as a RG operator based on the analogy with critical phenomena. Indeed, the operation of raising to power $p > 1$ and shifting by b_p eliminates the irrelevant small argument part of the parent distribution, a_p rescales the relevant "degrees of freedom," and the fixed point condition, Eq. (1), determines the limit distribution.

Equation (1) is solved by $a_p = p^\gamma$, $b_p = \gamma^{-1}(p^\gamma - 1)$, and $M(x) = \exp[-(1 + \gamma x)^{1/\gamma}]$, where the final scale of x and the position of the distribution are set by the standardization $M(0) = M'(0) = 1/e$. We thus have a line of fixed points parametrized by γ, and $M(x)$ is the generalized extreme value distribution. The traditional universality classes are called Fréchet (power decay of parent at infinity), Fisher-Tippett-Gumbel (FTG, faster than power decay), and Weibull (power decay at a finite cutoff), and correspond to $\gamma > 0, = 0, < 0$, respectively.

In the RG picture, the FS behavior is determined by the action of the RG transformation on the neighborhood of the fixed point $M(x)$. Thus, we consider distributions as $M_\epsilon(x) = M((x + \epsilon \psi(x)))$, assuming ϵ is small. Standard-

ization implies $\psi(0) = \psi'(0) = 0$, and the scale of ϵ is set by $\psi''(0) = 1$. Our central observation is that the large N behavior corresponds to the eigenvalue problem

$$M_{\epsilon'}(x) = [\hat{R}_p M_\epsilon](x) = M_\epsilon^p(a_{p,\epsilon} x + b_{p,\epsilon}), \quad (2)$$

where linearization in ϵ is understood, $a_{p,\epsilon}$, $b_{p,\epsilon}$ differ from the fixed point values a_p, b_p determined above to $O(\epsilon)$, and the eigenvalue is $\lambda = \epsilon'/\epsilon$. Using the fixed point relation (1), we obtain $(\lambda/a_p)\psi''(x) = \psi''(a_p x + b_p)$ whose solution with p-independent $\psi(x)$ reads as

$$\psi(x) = \frac{(1 + \gamma x)^{\gamma'/\gamma+1} - (\gamma' + \gamma)x - 1}{\gamma'(\gamma' + \gamma)} \quad (3)$$

$$\lambda = p^{\gamma'}. \quad (4)$$

Thus we see that, for a given universality class parametrized by γ, a new parameter γ' emerges characterizing the eigenvalues and eigenfunctions of Eq. (2). Note that shape corrections equivalent to ψ have been obtained by direct methods in the mathematical literature [8,16].

In order to link the RG result with the N dependence, we write $\epsilon = \epsilon_N$ and use (4) to find $\epsilon' = \epsilon_{pN} = p^{\gamma'}\epsilon_N$. Assuming a power form one obtains

$$\epsilon_N \propto N^{\gamma'}, \quad (5)$$

or, more precisely, $\frac{d\ln|\epsilon_N|}{d\ln N} \to \gamma'$. Thus γ' is the FS convergence rate (stability implies $\gamma' \le 0$). To find ϵ_N and thus γ' for a given parent, $\mu(z)$, we study the integrated distribution function $\mu^N(z)$ with the shift and scale parameters $b_N = h(\ln N)$, $a_N = h'(\ln N)$ expressed through $h(y) = \mu^{-1}(e^{-e^{-y}})$. Close to the fixed point one has

$$M_N(x) = \mu^N(a_N x + b_N) \approx M(x + \epsilon_N \psi(x)), \quad (6)$$

and differentiating $-\ln[-\ln M_N(x)]$ twice at $x = 0$ gives, to leading order, $da_N/db_N \to \gamma$, and at next order

$$\epsilon_N = \gamma - da_N/db_N \sim N^{\gamma'}. \quad (7)$$

The convergence rate γ' is now determined and the perturbation function $\psi(x)$ follows from (3). This gives a practical meaning to the results from RG theory.

For data analysis it is convenient to represent the FS correction with zero mean $\langle x \rangle$ and unit variance σ_x^2 (finite for $\gamma < 1/2$) by using the variable $y = (x - \langle x \rangle)/\sigma_x$. Here we consider the FTG class ($\gamma = 0$) with limit distribution $M^{(0)}(y) = e^{-e^{-(ay+b)}}$ where $a = \pi/\sqrt{6}$, $b = \gamma_E$, the latter being Euler's constant. Writing $M_N(y) \approx M^{(0)}(y) + \epsilon_N M^{(1)}(y)$, we have for the correction

$$M^{(1)}(y) = P^{(0)}(y)[e^{\gamma'(ay+b)} + \alpha y + \beta]/a\gamma'^2, \quad (8)$$

where $P^{(0)}(y) = M^{(0)}(y)'$, $\alpha = \Gamma(1 - \gamma')\frac{b+\Psi(1-\gamma')}{a}$, with $\Psi(z) = \Gamma'(z)/\Gamma(z)$ and $\beta = -\Gamma(1 - \gamma')$. For $\gamma' = 0$, Eq. (8) becomes [$\zeta(z)$ denotes Riemann's ζ function]

$$M^{(1)}(y) = P^{(0)}(y)[a^3(y^2 - 1) - 2\zeta(3)y]/2a^2. \quad (9)$$

We illustrate the above results on FTG class parents with the commonly found asymptote $1 - \mu(z) \propto e^{-z^\delta}/z^\theta$ ($\delta > 0$). Using Eq. (7), for $\delta \ne 1$ we have $\epsilon_N \approx (\delta - 1)/(\delta \ln N)$, so $\gamma' = 0$. For $\delta = 1$, $\theta = 0$ (exponential distribution) one finds $\epsilon_N \approx 1/2N$ and $\gamma' = -1$, while for $\delta = 1$ and $\theta \ne 0$, we have $\epsilon_N \approx -\theta/\ln^2 N$, so again $\gamma' = 0$. Thus generically $\gamma' = 0$, with FS shape correction given by (9) and the perturbation decaying logarithmically. Faster, $1/N$, convergence is seen only for $\delta = 1$, $\theta = 0$ where Eq. (8) applies with $\gamma' = -1$.

As an application, we studied the FS corrections to the distribution of the largest cluster size on a square lattice in subcritical site percolation, $p < p_c$, where $p_c \approx 0.592\ldots$ is the critical occupation probability. Because of the finite correlation length, clusters in a large system are nearly independent. Then, as shown by an intuitive RG picture of [11] and by other methods in [12], the size distribution of the largest cluster obeys FTG, provided the inherent discreteness of the problem is treated appropriately. It remains an open question, however, whether the FS corrections can also be described by iid theory. To answer this question, we first note that the asymptote of the distribution of the cluster size s is $s^{-1}\exp(-s/s_\xi)$ [17], where s_ξ is the cutoff size. This asymptote corresponds to $\delta = \theta = 1$ in the example of the previous paragraph, so (9) gives the FS scaling function and $\epsilon_N \approx -1/\ln^2 N$. To compare theory with simulation, we collected statistics for the largest cluster in systems of sizes $L = 500$ and 1000 in an ensemble of $\approx 10^7$ runs, resulting in a relatively smooth histogram. The shape correction was then obtained by subtracting the empirical histogram from the FTG distribution, multiplied by $\ln^2 N$, where N is the average number of clusters. The result is compared with the iid theory in Fig. 1. As can be seen the curves match surprisingly well, suggesting that the iid theory is also relevant for the FS corrections.

Next, we treat correlated time signals, $h(t)$, and study the FS effects on the distribution of their maxima h_m. To have control on correlations, $1/f^\alpha$ noise is chosen, where the Fourier amplitudes are independent Gaussian variables, with variance $f^{-\alpha}$, and uniform, random phase [18]. For $\alpha = 0$, the process is white noise, for $0 < \alpha < 1$ it is stationary with correlations decaying as $t^{\alpha-1}$, while for $\alpha > 1$ the fluctuations of the signal diverge as $t \to \infty$. The distribution of h_m has recently been studied intensively; see [19] and references therein. The main features to be recalled here are that the FTG distribution applies for $0 \le \alpha < 1$ [13] while, for $\alpha > 1$, nontrivial distributions emerge whose shape depends on boundary conditions and the reference point from where the maximum is measured. Here we concentrate on the FS correction of the distribution of maxima, and to be specific, the maximum is measured from the mean of a periodic signal. First, consider $\alpha < 1$. Figure 2 shows the FS shape correction for
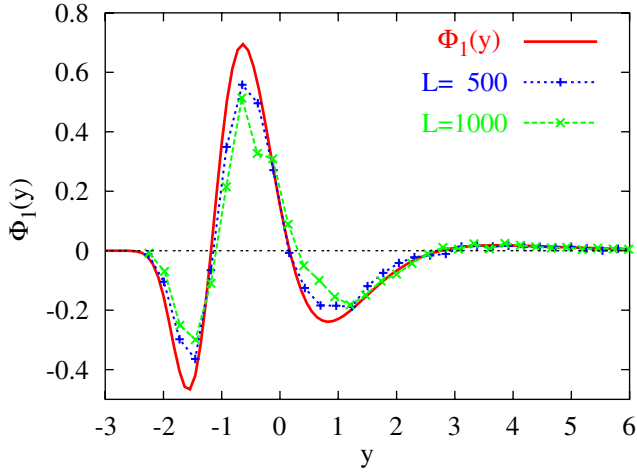
FIG. 1 (color online). Finite-size analysis for the largest clusters in subcritical percolation. The prediction from iid theory (solid line) is $\Phi_1(y) = M^{(1)}(y)'$ from Eq. (9). The simulation results (dotted line) were obtained with occupation probability $p = 0.25$ and system sizes $L = 500, 1000$.

$\alpha = 0.4$ together with the iid prediction. The latter comes from a Gaussian parent, thus $\epsilon_N \approx 1/2 \ln N$, $\gamma' = 0$. The theoretical curve is $\Phi_1(y) = -M^{(1)}(y)'/2$ by Eq. (9), corresponding to a $\ln N$ magnification factor in the simulation curves which visibly approach $\Phi_1(x)$. A similar approach can be seen for all $\alpha \lesssim 0.5$. The inset, showing the skewness for finite systems, also suggests that the leading FS correction may be described by the iid theory for $\alpha \lesssim 0.5$. Beyond 0.5, the deviation from the iid correction grows, so while the limit is FTG, the question of whether the correction is of iid type remains open. Note that EVS in a model
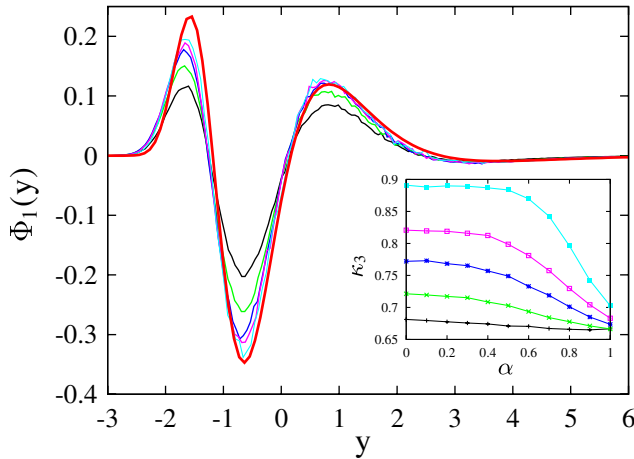

FIG. 2 (color online). FS shape correction to the maximum distribution in $1/f^\alpha$ noise with $\alpha = 0.4$. The solid line is the iid theory $\Phi_1(y) = -M^{(1)}(y)'/2$ from Eq. (9), while the others are simulations for system sizes $N = 2^5, 2^7, 2^9, 2^{11}, 2^{15}$, with minimum decreasing in this order. The inset shows the skewness of the maximum distribution for various sizes slowly converging to the FTG value $\kappa_3^0 \approx 1.14$.

of quantum chaos [10] provides an example where the correlations are weak enough to yield a FTG limit distribution, but the FS corrections are distinct from the iid result.

We now turn to $\alpha > 1$, where $\langle h_m \rangle$ diverges as $\langle h_m \rangle \sim N^{(\alpha-1)/2}$ [20], and the approach to the limit distribution improves from logarithmic to power-law. This effect can be seen in our simulations as well as in the exact results [6,21] for random walks ($\alpha = 2$). The limit distribution for $\alpha = 2$ is given by the Airy distribution $\Phi_{Ai}(z)$ with $z = h_m/\sqrt{N}$ while the first correction to scaling is [9]

$$\Phi(z) \approx \Phi_{Ai}(z) + \Phi'_{Ai}(z)/\sqrt{2N}. \qquad (10)$$

The simplicity of the above result calls for a simple explanation. Indeed, Eq. (10) follows from the assumption that the shape of the distribution relaxes faster than its position. To see this for arbitrary $\alpha > 1$, we note that the nth cumulant of h_m scales for large N as $\kappa_n \sim N^{n\theta}$ with $\theta = (\alpha - 1)/2$ [20]. Next, we write the corrections to scaling of κ_n-s as

$$\kappa_n = N^{n\theta}(\kappa_n^0 + \kappa_n^1 N^{-\omega_n} + \cdots), \qquad (11)$$

and assume that $\omega_n > \omega_1$ for $n > 1$. This assumption implies that the shape of the function relaxes faster than its position. Introducing now the scaled variable $z = h_m/N^\theta$ and expanding the cumulant generating function of h_m in $N^{-\omega_1}$ yields the scaled distribution function $\Phi_N(z) = N^\theta P_N(N^\theta z)$ to first order as

$$\Phi_N(z) \approx \Phi(z) - \kappa_1^1 \Phi'(z) N^{-\omega_1}, \qquad (12)$$

where $\Phi(z)$ is the scaling function in the $N \to \infty$ limit. For $\alpha = 2$ one has $\omega_1 = 1/2$ and $\kappa_1^1 = -1/\sqrt{2}$ [9]; thus, Eq. (10) is recovered.

The choice of scaling variable may change both the rate and amplitude of the correction term. For example, a natural choice is to scale by the average [$x = h_m/\langle h_m \rangle$, $\Phi_N(x) = \langle h_m \rangle P_N(\langle h_m \rangle x)$]. It yields the same rate of convergence, but it does alter the scaling function of the first order correction. Indeed, using $x = h_m/\langle h_m \rangle$ and expanding the cumulant generating function in $1/N^{\omega_1}$ results in

$$\Phi_N(x) \approx \Phi(x) - (\kappa_1^1/\kappa_1^0)[(x-1)\Phi(x)]' N^{-\omega_1}. \qquad (13)$$

The limiting function $\Phi(x)$ for $\alpha = 2$ and its correction $\Phi_1(x)$ with an amplitude 0.1 are shown in the upper panel of Fig. 3. Since $\Phi_1(x)$ has two zeros, $\Phi_N(x)$ is nearly fixed at two points, so the main correction comes from the central weight shifted to the tails of the distribution.

If the main FS effect is due to $\langle h_m \rangle$ then the use of the scaling variable $y = (h_m - \langle h_m \rangle)/\sqrt{\kappa_2}$ eliminates these corrections and, as seen in Fig. 3, improves convergence dramatically. The shape correction may be calculated by assuming that $\omega_n > \omega_2$ for all $n > 2$. This means that the leading FS correction comes from κ_2 and the cumulant generating function to first order in $1/N^{\omega_2}$ yields
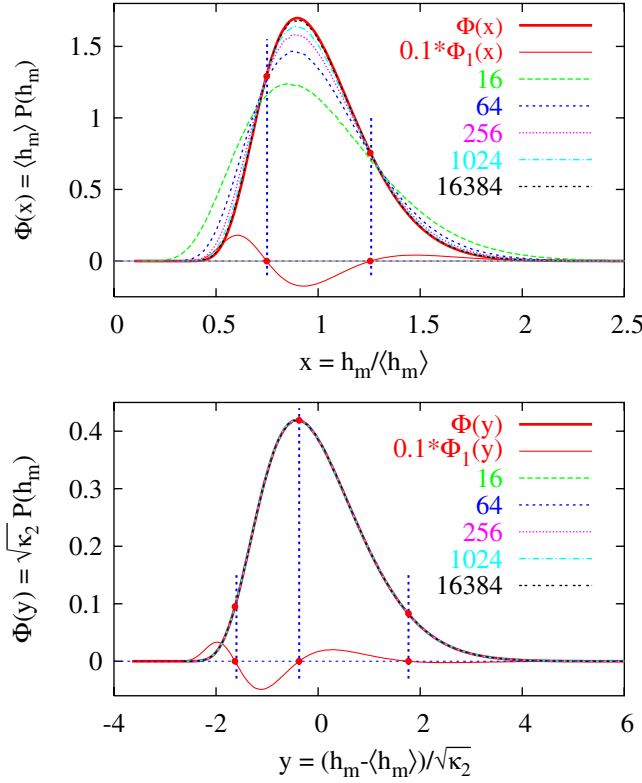
FIG. 3 (color online). Demonstration of the increasing speed of convergence to the limit distribution Φ, for $\alpha = 2$. Results for system sizes $N = 16, \dots, 16\,384$ are shown using scaling variables $x = h_m/\langle h_m \rangle$ and $y = (h_m - \langle h_m \rangle)/\sqrt{\kappa_2}$ in the upper and lower panels, respectively. Φ_1 is the shape correction function.

$$\Phi_N(y) \approx \Phi(y) - (\kappa_2^1/2\kappa_2^0)[y\Phi(y) + \Phi'(y)]'N^{-\omega_2}. \quad (14)$$

As seen, the shape correction can again be expressed in terms of the limit distribution. The scaling function $\Phi_1(y)$ displayed in Fig. 3 has three zeros which restrict possible deviations from the limit distribution to higher order. In addition, we found numerically $\omega_2 \approx 3/2$, resulting in such a fast convergence that the curves with various N-s cannot be distinguished within linewidth.

We can make an intuitive proposition for the ω_n-s, based on an analogy with the cumulants $\kappa_n(w_2)$ of the roughness w_2 of $1/f^\alpha$ signals [18]. There, the large N asymptote can be rewritten as $\kappa_n(w_2) \sim \langle w_2 \rangle^n (1 - b_n \langle w_2 \rangle^{(1-n\alpha)/(\alpha-1)})$. Now assuming the same exponent $(1 - n\alpha)/(\alpha - 1)$ for $\kappa_n(h_m)$, and reverting to the N dependence we get $\omega_n(\alpha) = (n\alpha - 1)/2$ for the FS exponent. For Brownian motion we recover $\omega_2(2) = 3/2$, in accordance with simulations. Furthermore, the criterion $\omega_n < \omega_{n+1}$ is satisfied, so $\omega_2(\alpha) = \alpha - 1/2$ is the candidate for the FS exponent. Since $\omega_n(\alpha)$ increases with α, the convergence is expected

to improve for larger α. Indeed, our simulations for $\alpha = 4$ show that the same convergence as in the lower panel of Fig. 3 can already be obtained in the $x = h_m/\langle h_m \rangle$ scaling since $\omega_1(4) = \omega_2(2)$.

Let us conclude with two remarks. First, higher than leading FS corrections may also be of importance. While next to leading order results for the iid case are available [16], we actually worked out an algorithm to arbitrary orders, which will be presented elsewhere. Second, we expect that the RG approach will be useful for sorting out the universal aspects of EVS in correlated systems.

[1] E. J. Gumbel, *Statistics of Extremes* (Dover, New York, 1958).
[2] P. Embrecht, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance* (Springer, Berlin, 1997).
[3] R. W. Katz, M. B. Parlange, and P. Naveau, Adv. Water Resour. **25**, 1287 (2002).
[4] J.-P. Bouchaud and M. Mézard, J. Phys. A **30**, 7997 (1997).
[5] G. Györgyi *et al.*, Phys. Rev. E **68**, 056116 (2003).
[6] S. N. Majumdar and A. Comtet, Phys. Rev. Lett. **92**, 225501 (2004).
[7] P. L. Krapivsky and S. N. Majumdar, Phys. Rev. Lett. **85**, 5492 (2000).
[8] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction* (Springer, New York, 2006).
[9] G. Schehr and S. N. Majumdar, Phys. Rev. E **73**, 056103 (2006).
[10] A. Lakshminarayan, S. Tomsovic, O. Bohigas, and S. N. Majumdar, Phys. Rev. Lett. **100**, 044103 (2008).
[11] M. Z. Bazant, Phys. Rev. E **62**, 1660 (2000).
[12] R. van der Hofstad and F. Redig, J. Stat. Phys. **122**, 671 (2006).
[13] S. M. Berman, Ann. Math. Stat. **35**, 502 (1964).
[14] J. Galambos, *The Asymptotic Theory of Extreme Value Statistics* (John Wiley & Sons, New York, 1978).
[15] R. Fisher and L. Tippett, Proc. Cambridge Philos. Soc. **24**, 180 (1928).
[16] L. de Haan and S. Resnick, Ann. Probab. **24**, 97 (1996).
[17] D. Stauffer and A. Aharony, *Introduction To Percolation Theory* (Taylor and Francis, London, 1994).
[18] T. Antal *et al.*, Phys. Rev. E **65**, 046140 (2002).
[19] T. W. Burkhardt *et al.*, Phys. Rev. E **76**, 041119 (2007).
[20] G. Györgyi, N. R. Moloney, K. Ozogány, and Z. Rácz, Phys. Rev. E **75**, 021123 (2007).
[21] S. Majumdar and A. Comtet, J. Stat. Phys. **119**, 777 (2005).