The Largest Cluster in Subcritical Percolation

Martin Z. Bazant
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA
(April 4, 2000)

The statistical behavior of the size (or mass) of the largest cluster in subcritical percolation on a finite lattice of size N is investigated (below the upper critical dimension, presumably $d_c = 6$). It is argued that as $N \to \infty$ the cumulative distribution function converges to the Fisher-Tippett (or Gumbel) distribution $e^{-e^{-z}}$ in a certain weak sense (when suitably normalized). The mean grows like $s_\xi^* \log N$, where $s_\xi^*(p)$ is a "crossover size". The standard deviation is bounded near $s_\xi^* \pi/\sqrt{6}$ with persistent fluctuations due to discreteness. These predictions are verified by Monte Carlo simulations on $d = 2$ square lattices of up to 30 million sites, which also reveal finite-size scaling. The results are explained in terms of a flow in the space of probability distributions as $N \to \infty$. The subcritical segment of the physical manifold ($0 < p < p_c$) approaches a line of limit cycles where the flow is approximately described by a "renormalization group" from the classical theory of extreme order statistics.

PACS: 64.60.Ak,02.50.-r

I. INTRODUCTION

In the latter half of this century, percolation has become the canonical model of quenched spatial disorder [1]. Among its many areas of application are polymer gelation, hopping conduction in semiconductors and flow in porous media [2]. Percolation has also attracted the attention of mathematicians because it offers challenging problems in probability theory of relevance to statistical physics [3,4]. Since rigorous results are often not easily obtained, however, computer simulation has played a central role in the motivation and testing of new theoretical ideas [5].

Most analytical and numerical studies have examined the critical point ($p = p_c$) where the correlation length $\xi(p)$ diverges, but here we focus on subcritical percolation ($p < p_c$) characterized by $\xi < \infty$. In this case, it is known that the cluster-size distribution $n_s(p)$, the number of clusters of size (or mass) s per site of an infinite hypercubic lattice of coordination z, decays exponentially for all $p < 1/(z-1) < p_c$ [6,7]

$$\log n_s \asymp -s \quad \text{as} \quad s \to \infty, \tag{1}$$

where $a_n \asymp b_n$ means "a_n scales like b_n", or more precisely

$$0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \le \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty. \tag{2}$$

(The quantity $P_s = n_s s$, which is the probability that the origin is part of a cluster of size s, is also sometimes called the "cluster-size distribution" [6,7].) The total number of finite clusters per lattice site at the critical point $n_c \equiv \sum_{s=1}^{\infty} n_s(p_c)$ is known analytically for $d = 2$ bond percolation [8,9] and numerically for site and bond percolation for various lattices in $d = 2$ and $d = 3$ [12]. Universal finite-size corrections to n_c have also been studied extensively [10–13].

Beyond the rigorous result (1), it is believed that the cluster-size distribution decays exponentially for all $p < p_c$ with a characteristic size $s_\xi(p)$ and a power-law prefactor

$$n_s \asymp s^{-\theta} e^{-s/s_\xi} \quad \text{as} \quad s/s_\xi \to \infty. \tag{3}$$

where the exponent θ is supposed to be independent of p with $\theta = 1$ for $d = 2$ and $\theta = 3/2$ for $d = 3$, respectively [1]. The quantity s_ξ in (3) is called the "crossover size" since large clusters ($s \gg 1$) of size much smaller than s_ξ behave "critically", while much larger clusters behave "subcritically", as explained below. Because large clusters are fractal objects, the crossover size and the correlation length are related by $s_\xi \propto \xi^D$, where $D < d$ is the fractal dimension of the infinite cluster at $p = p_c$.

In contrast to the cluster-size distribution, relatively little is known about the size of the *largest* cluster $S_{(N)}$ in a *finite* system of size $N = L^d$ for $p < p_c$, with the notable exception of the recent work of Borgs et al. [14]. (Our notation for the random variable $S_{(N)}$ is explained below.) It is widely believed that the mean largest-cluster size $\mu_N \equiv \mathrm{E}[S_{(N)}]$ scales like $\mu_N \propto s_\xi \log N$ for $p < p_c$. This follows from the heuristic argument $N n_\mu \approx 1$, which supposes that the largest cluster can be placed independently at any site in the lattice [1]. (This useful idea is extended significantly in section II below.) Recently, from certain scaling axioms verified for $d = 2$ and believed to hold for $d \le d_c = 6$, Borgs *et al.* have proved the somewhat weaker statement $\mu_{L^d}/\xi'^D \asymp \log(L/\xi')$ as $L/\xi' \to \infty$, or equivalently

$$\mu_N / s'_\xi \asymp \log(N/s'_\xi) \quad \text{as} \quad N/s_\xi \to \infty \tag{4}$$

where $\xi'(p)$ is another correlation length defined in terms of "sponge-crossing probabilities" and $s'_\xi \equiv \xi'^D$ is a corresponding crossover size [14]. (Note that $d \leq d_c$ is assumed throughout this paper.)

In applications $S_{(N)}$ provides a measure of the maximum connectivity of a random medium, which is of fundamental interest in the subcritical regime. From a theoretical point of view, the "strength" (or concentration) of the largest cluster $S_{(N)}/N$ plays the role of an order parameter since its expected value in the "thermodynamic limit"

$$P_\infty(p) = \lim_{N\to\infty} \mu_N / N \tag{5}$$

has a discontinuous slope at $p = p_c$ with $P_\infty(p \leq p_c) = 0$ and $P_\infty(p > p_c) > 0$. Beyond the limiting behavior of the mean μ_N, however, a much more complete understanding of the percolation transition is contained in the cumulative distribution function (c.d.f.) of the largest-cluster size

$$F_N(s) \equiv \text{Prob}(S_{(N)} \leq s) \tag{6}$$

which also describes all size-dependent fluctuations of the order parameter. In this sense, the behavior of $F_N(s)$ near the critical point fully describes the "birth of the infinite cluster" [14]. Beginning with the same scaling axioms as in deriving (4), Borgs et al. have also proved that $F_N(s)$ varies significantly only on the scale of the mean for $p < p_c$

$$\lim_{\epsilon\to 0} \liminf_{N\to\infty} \left[F_N\left(\epsilon^{-1} s'_\xi \log(N/s'_\xi)\right) - F_N\left(\epsilon s'_\xi \log(N/s'_\xi)\right) \right] = 1. \tag{7}$$

It is believed that (4) and (7) would also hold with the usual definition of ξ as the decay length of the pair correlation function [14], so we expect $\xi' \propto \xi$ and $s'_\xi \propto s_\xi$.

Although (4) and (7) provide important rigorous justification for the logarithmic scaling of the mean μ_N, the shape of the distribution $F_N(s)$ and scaling of the variance $\sigma_N^2 \equiv \text{Var}[S_{(N)}]$ appear not to have been studied (either numerically or analytically) before this work. Moreover, no connections have yet been made between subcritical percolation and the classical limit theorems of probability theory. Such fruitful connections, which are known to explain Gaussian fluctuations away from the critical point in thermal phase transitions [15], would presumably come from the statistical theory of extremes [16–20].

The article is organized as follows. First, in order to build the reader's physical intuition, simple approximations are made in section II to derive the asymptotic behavior of $F_N(s)$ and propose finite-size scaling laws for μ_N and σ_N. In section III, these predictions are verified for the $d = 2$ square lattice with computer simulations, which also provide empirical functional forms and numerical parameters for the scaling laws. Finally, in section IV the preceding results are explained in terms of a "subcritical renormalization group".

II. SIMPLE ARGUMENTS

A. Connection with Extreme Order Statistics

Consider site percolation on a periodic, hypercubic lattice of $N = L^d$ sites. Since any cluster can be uniquely identified with the site nearest to its center of mass (of lowest index, if there is more than one such site), we can define a set of N independent, identically distributed (i.i.d.) random variables (r.v.) $\{S_i\}$ such that $S_i = s$ if the largest cluster centered at site i has size s and $S_i = 0$ if no cluster is centered there. Clearly, the most probable value of S_i is zero, since the number of clusters is always much less than the number of sites, and it is exceedingly rare to have more than one cluster centered at the same site, e.g. when one cluster encircles another.

We seek the c.d.f. $F_N(s)$ of the "extreme order statistic" [19,20]

$$S_{(N)} \equiv \max_{1\leq i\leq N} S_i \tag{8}$$

in $N \to \infty$ with $p < p_c$ fixed. Extreme order statistics have many classical applications, such as the fracture strength of solids, the occurrence of manufacturing defects and the frequency of extreme weather [19]. More recently in statistical physics, extreme order statistics have been applied to glassy relaxation on fractal structures [21], the dynamics of elastic manifolds in random media [22,23], the random energy model [24,25], decaying Burgers turbulence [24], dispersive transport in amorphous materials [26] and random sequential adsorption [27]. In such applications, extreme order statistics are used to describe the most important features of a random *energy* landscape, e.g. lowest activation energy barriers.

In this work, we show that the theory of extreme order statistics also has relevance for the *geometric* features of random systems. In one dimension, the largest cluster in percolation bears some resemblance to the longest increasing subsequence of a random permutation, which is known to exhibit similar limiting statistics (see Ref. [28] for a recent review), although the former problem is much simpler [29]. Of course, the interesting cases of percolation, however, are in higher dimensions, which we address here.

B. A First Approximation Based on Independence

The main difficulty in the percolation problem for $S_{(N)}$, aside from the complexity of the parent distribution, is that the r.v. $\{S_i\}$ are correlated. Much is known about order statistics of i.i.d. r.v. [19], but dependent r.v. have been studied mostly in cases much simpler than percolation [20]. Nevertheless, considerable insight is gained by neglecting correlations in deriving an asymptotic form of $F_N(s)$, which will be justified below in section IV. As one might expect, correlations in the subcritical regime are too weak to have an effect in the thermodynamic limit.

Whenever $N \gg s_\xi$, which holds in the limit $p \to 0$ for fixed $N \gg 1$, cluster sizes comparable to the system size are exponentially rare according to (1). Since correlations between the r.v. $\{S_i\}$ arise due to excluded volume effects (see below), $\mathrm{Cov}[S_i, S_j]$ is exponentially small for most pairs of sites (i, j) in this limit. Therefore, as a natural first approximation we assume N independent selections from a continuous parent distribution with exponential decay

$$\mathrm{Prob}(S_i \le s) \sim 1 - e^{-s/s_\xi^*} \quad \text{as} \quad s \to \infty \tag{9}$$

where $s_\xi^*(p)$ is an effective crossover size (see below). Note that the asymptotic distribution of the maximum of i.i.d. r.v. is entirely determined by the tail of the parent distribution [17,18], so the complicated behavior of S_i for small sizes is irrelevant. From the method of Cramér [19] applied to (9), we quickly find

$$F_N(s) \sim \left(1 - e^{-s/s_\xi^*}\right)^N = \left(1 - \frac{e^{-(s - s_\xi^* \log N)/s_\xi^*}}{N}\right)^N \tag{10}$$

which implies

$$\lim_{N\to\infty} G_N(z) = e^{-e^{-z}} \tag{11}$$

where

$$G_N(z) \equiv F_N(s_\xi^* z + s_\xi^* \log N) = \mathrm{Prob}\left(S_{(N)}/s_\xi^* \le z + \log N\right) \tag{12}$$

is a normalized c.d.f. Therefore, in this simple approximation the largest-cluster size is sampled from the Fisher-Tippett distribution [30] with c.d.f. $e^{-e^{-z}}$, mean $\gamma = 0.5772\ldots$ (Euler's constant) and variance $\pi^2/6$ [17]; the mean largest-cluster size grows logarithmically $\mu_N/s_\xi^* \sim \log N + \gamma$, while the standard deviation converges to a certain constant $\sigma_N/s_\xi^* \to \pi/\sqrt{6}$. Comparing with (4), we can view the leading-order asymptotic behavior of the mean

$$\mu_N \sim s_\xi^* \log N \quad \text{as} \quad N \to \infty \tag{13}$$

as defining the effective crossover size s_ξ^* (should it exist), which is presumably proportional to the others introduced above $s_\xi^* \propto s'_\xi \propto s_\xi$.

C. Corrections Due to Discreteness

There appears to be a problem with (11) for percolation on a lattice: A discrete c.d.f. (which is a piecewise constant function) cannot converge to a continuous function when scaled by a bounded standard deviation. In fact, since s in (9)–(10) is restricted to integer values, the limit in (11) does not exist. Instead, if we replace s by $[s]$ (the nearest integer to s) in (9), then the normalized c.d.f. $G_N(z)$ defined by (12) approaches a quasi-periodic sequence of piecewise constant functions with period roughly $1/s_\xi^*$ in $\log N$,

$$G_N(z) = \left(1 - \frac{-e^{-z + \delta_N(z)/s_\xi^*}}{N}\right)^N \sim e^{-e^{-z+\delta_N(z)/s_\xi^*}} \quad \text{as} \quad N \to \infty \tag{14}$$

where

$$\delta_N(z) \equiv s_\xi^*(z + \log N) - [s_\xi^*(z + \log N)]. \tag{15}$$

(The limiting sequence is strictly periodic only when e^{1/s_ξ^*} is an integer.) The piecewise constant functions in (14) converge weakly in the sense that as $N \to \infty$ the "step edges" periodically trace out two continuous functions

$$\overline{G}(z) \equiv \limsup_{N\to\infty} G_N(z) = e^{-e^{-z-1/(2s_\xi^*)}} \tag{16a}$$

$$\underline{G}(z) \equiv \liminf_{N\to\infty} G_N(z) = e^{-e^{-z+1/(2s_\xi^*)}} \tag{16b}$$

which define a stationary "envelope" of width $1/s_\xi^*$ about the Fisher-Tippett distribution. If we let a be the lattice spacing (which we take to be unity), then the envelope width would be a^d/s_ξ^*, showing that the lack of convergence is controlled by the relative importance of discreteness on the scale of the crossover size. Note that the continuous distribution (11) is recovered in the limit $p \to p_c$ (taken after the limit $N \to \infty$)

$$\lim_{p\to p_c} \overline{G}(z) = \lim_{p\to p_c} \underline{G}(z) = e^{-e^{-z}}, \tag{17}$$

as the crossover size diverges and hence the envelope width vanishes.

For $s_\xi^* < \infty$ ($p < p_c$), the continuum result for the scaling of the mean (13) still holds, but the standard deviation has persistent fluctuations due to discreteness

$$\sigma_N/s_\xi^* \sim \pi/\sqrt{6} + \epsilon_N \quad \text{as} \quad N \to \infty \tag{18}$$

where ϵ_N is periodic in $\log N$ with period $1/s_\xi^*$. Because the limiting sequence (14) fluctuates periodically about a certain fixed distribution, it can be viewed as a "limit cycle" in some appropriate Banach space (see below). Intuitively, the distribution conforms asymptotically to the Fisher-Tippett distribution as closely as possible within the constraints imposed by discreteness.

D. Corrections due to Correlations

The simple derivation of (14) should be valid whenever $s_\xi \ll 1$ (or $s'_\xi \ll 1$ or $s_\xi^* \ll 1$) because then even a single site qualifies as a large cluster. If $s_\xi \approx 1$, however, non-negligible correlations among the r.v. $\{S_i\}$ arise because a cluster of size s_ξ excludes on the order of s_ξ nearby sites from being part of any other cluster. If $s_\xi \gg 1$, on the order of $\xi^d \propto s_\xi^{d/D} \gg s_\xi$ sites are excluded by such a cluster since it engulfs many smaller, exterior regions due to its fractal geometry ($D < d$). Therefore, correlations can be included heuristically by replacing N with N/s_ξ^α in (14) which simply shifts the mean by a constant $\Delta\mu/s_\xi = -\alpha \log s_\xi$ without affecting the leading-order scaling behavior (13), where $\alpha = 0$ if $s_\xi \ll 1$, $\alpha = 1$ if $s_\xi \approx 1$ and $\alpha = d/D$ if $s_\xi \gg 1$. Note that the effect of correlations is negligible for $N \gg s_\xi$. Correlations do, however, control the finite-size scaling at smaller values of N.

E. Finite-Size Scaling

There are only two relevant length scales in percolation, the correlation length ξ and the lattice spacing a (normalized to unity), or equivalently two mass scales, the crossover size s_ξ (or s'_ξ or s_ξ^*) and the volume of a lattice cell a^d (also normalized to unity). If $s_\xi \gg a$, then discrete lattice effects on "large" clusters with sizes on the order of s_ξ or larger become negligible, and the system has only one relevant mass scale s_ξ. As a consequence of the single scale s_ξ in the limit $p \to p_c$, any function of N and s_ξ is expected to collapse into a self-similar form interpolating between a critical power-law in N valid for $1 \ll N \ll s_\xi$ and a subcritical function of N/s_ξ^α (for some constant α) valid for $1 \ll s_\xi \ll N$. For example, because $\mu_N(p)$ and $\sigma_N(p)$ have the dimensions of s_ξ, we have

$$\mu_N/s_\xi = \Phi\left(N/s_\xi^\alpha\right) \tag{19a}$$

$$\sigma_N/s_\xi = \Psi\left(N/s_\xi^\alpha\right) \tag{19b}$$

for some universal functions $\Phi(x)$ and $\Psi(x)$ which do not depend on p. In the critical regime $N \ll s_\xi$, it is expected that $\mu_N \propto L^D = N^{D/d}$ and that both μ_N and σ_N are asymptotically independent of s_ξ, which implies $\alpha = d/D$ and $\Phi(x) \propto \Psi(x) \propto s^{D/d}$ as $s \to 0$. From (7) and (14), we also expect $\Phi(x) \asymp \log x$ and $\Psi(x) \asymp 1$ as $x \to \infty$.

The classical idea behind the finite-size scaling ansatz (19) can be understood as follows. A large subcritical cluster (on an infinite lattice) intersected with a finite box of side L exhibits a crossover from "critical scaling" at small scales $a \ll L \ll \xi$ (where a portion of it typically spans the box) to "subcritical scaling" at large scales $L \gg \xi$ (where it is entirely contained within the box). Note that the lattice spacing a is irrelevant as long as $\xi \gg a$; all systems with the same ratio L/ξ should have equivalent statistics, up to small corrections of order a/ξ due to discreteness. Of course, as $p \to 0$ the finite-size scaling ansatz breaks down, and discrete effects eventually dominate over correlation effects, as explained above.

III. NUMERICAL RESULTS

A. Methods

In order to test the predictions of the previous section, numerical simulations are performed for site percolation on periodic $d = 2$ square lattices of sizes $N = 5^2$, 13^2, 31^2, 74^2, 129^2, 175^2, 415^2, 982^2, 2324^2 and 5500^2 with $p = 0.05, 0.1, 0.15, \ldots 0.5$ [31]. Note that the value $p_c = 0.592\ 7460 \pm 0.000\ 0005$ has been determined numerically in this case [32]. For each (N, p), between $M = 2 \times 10^5$ and $M = 10^8$ samples are generated, and clusters are identified by a recursive "burning" algorithm [5,33]. With these methods, trillions of clusters are counted in several months of CPU time on Silicon Graphics R-10,000 processors.

In performing such large-scale simulations, special attention must be paid to the choice of (pseudo)random-number generator [32,34]. With the standard 32-bit generator `rand()`, the largest observed cluster sizes tend to come in multiples of integers ≥ 2 (after accumulating data from a very large number of "random" realizations), which indicates that the periodicity of the generator is having an artificial effect. In all the simulations reported here, however, the 48-bit generator `drand48()` is used, and the numerical cluster-size distributions $n_s(p)$ appear to be free of any systematic errors.

B. Largest-Cluster Distributions

The measured largest-cluster distributions are in very close agreement with the predictions of (14)–(16) for all $p < p_c$, as shown in Fig. 1 for the case $p = 0.15$. In order to check the shape of the c.d.f. against $e^{-e^{-z}}$, the distributions are normalized to have mean γ and variance $\pi^2/6$, which differs somewhat from the normalization given above in (12). As predicted by (16) the discrete c.d.f.s in Fig. 1(a) lie almost perfectly within a continuous envelope between two Fisher-Tippett distributions. Likewise, the discrete probability density functions (p.d.f.) shown in Fig. 1(b) for $p = 0.15$, which are simply the step heights in Fig. 1(a), exhibit the expected small fluctuations about the Fisher-Tippett p.d.f. $e^{-z-e^{-z}}$ due to discreteness. Using the value $s^*_\xi(0.15) = 1.313$ (determined independently below), the width of the envelope is seen to be very close to $1/s^*_\xi$. Note that the c.d.f.s in Fig. 1(a) are shifted slightly outside the envelope by $\epsilon_N\sqrt{6}/(\pi s^*_\xi)$ because sizes have been scaled by $\sigma_N\sqrt{6}/\pi$ rather than by s^*_ξ. Overall, the agreement between (14)–(16) and the simulation results is excellent for all the values of p considered here, thus lending some credence to the approximations of the previous section.

C. Cluster-Size Distributions

In order to test the finite-size scaling laws (19), numerical values of the crossover size $s_\xi(p)$ are obtained by fitting the cluster-size distributions $n_s(p)$ to (3). When compiling these distributions, unwanted finite-size effects are minimized by requiring that $N^{1/d}$ exceed the largest observed cluster size (for a given value of p). With this restriction, a single cluster cannot directly see the periodic boundary conditions. Motivated by (3), the tails of the cluster-size distributions are fit to

$$\log n_s = C - \theta \log s - s/s_\xi \quad \text{for} \quad s > s_{\text{tail}}. \tag{20}$$

Fitting to such an asymptotic form requires some care: The starting point of the fit s_{tail} must be large enough that the asymptotic behavior is dominant but also small enough that the fit is not degraded by statistical fluctuations.

In this work s_{tail} is systematically chosen where $|ds_\xi/ds_{\text{tail}}|$, $|d\theta/ds_{\text{tail}}|$ and χ^2 are minimal ($\chi^2 \approx 1$). The fit is deemed reliable when the value thus obtained at s_{tail} is contained within all the other confidence intervals for fits with $s'_{\text{tail}} > s_{\text{tail}}$. Because the raw distributions have bin counts ranging from over 10^{11} at size 1 down to 0 and 1 in the large-size tails, the fitting cannot be done by the usual least-squares method, which assumes normally distributed errors. Instead, the parameters (C, θ, s_ξ) are fit to the $n_s(p)$ data by Poisson regression, which properly handles the discrete, rare events in the tail (using the package X-Lisp-Stat [35]).

The fitting results are given in Table I. Fitting errors grow as $p \to 0$ because less data is available to accurately resolve the tail of $n_s(p)$ and also as $p \to p_c$ due to critical slowing down. Although the results for s_ξ should be reliable, the results for θ (not needed in this work) could change somewhat if different corrections to scaling were considered [5]. Therefore, the observed small deviation of θ from its conjectured value [1] of 1 (for all $0 \le p < p_c$) may only be an artifact of the fit.

D. Scaling of the Mean and Variance

As shown in Figs. 2–3, the collapse of the mean and standard deviation of the largest-cluster size plotted as μ_N/s_ξ and σ_N/s_ξ versus $N/s_\xi^{d/D}$ (using $D = 91/48$ [1]) is nearly perfect for $p \ge 0.30$. As discrete-lattice effects become important at smaller values of p, however, the data drifts off the universal curves, and the tiny oscillations predicted by (18) begin to be visible in the standard deviation. These effects are most pronounced when $s_\xi < 1$ ($p \le 0.10$) since then the interpretation of $s_\xi^{d/D}$ as an excluded volume is meaningless and the second length scale $a = 1$ cannot be ignored. Indeed, when μ/s_ξ and σ/s_ξ are plotted versus $N/\max\{1, s_\xi^{d/D}\}$, as shown in Fig. 4, the data for $s_\xi < 1$ lies much closer to the universal curves, consistent with the heuristic arguments given above in section II D.

From the simulation results with $s_\xi \gg 1$, the universal scaling functions in (19) for the $d = 2$ square lattice can be determined numerically. For $p \ge 0.30$, the scaling function $\Phi(x)$ for the mean is fit to the empirical form:

$$\Phi(x) = \left[a_2 + \frac{a_3}{(a_4 + x)^{a_5}}\right] \log\left[1 + \left(\frac{x}{a_1}\right)^{D/d}\right] \tag{21}$$

where the best parameter values (in the least squares sense) are $a_1 = 8.1 \pm 0.5$, $a_2 = 0.954 \pm 0.005$, $a_3 = 3.3 \pm 0.2$, $a_4 = 1.0 \pm 0.3$ and $a_5 = 0.61 \pm 0.2$. The collapsed data in Fig. 2 shows a smooth crossover between the expected critical and subcritical scaling laws, $\Phi(x) \sim 30.3 x^{D/d}$ as $x \to 0$ and $\Phi(x) \sim a_2 \log(1 + (x/a_1)^{D/d}) \sim (a_2 D/d) \log x = 0.90 \log x$ as $x \to \infty$, respectively. The simulation result $\mu_N \sim 0.90 s_\xi \log N$ justifies our definition of the effective crossover size s_ξ^* in (13) and for the case of the square lattice relates it to the crossover size s_ξ in (3) via $s_\xi^* = 0.90 s_\xi$.

Although the standard deviation appears to be bounded from the data shown in Fig. 3, we can only safely conclude $\sigma_N = o(\log\log N)$ because the subcritical portion of the data only spans five decades in $N/s_\xi^{d/D}$ (due to memory restrictions). Following the derivation in the next section, however, it can be proved [43] that $\sigma_N = O(1)$ follows from very reasonable assumptions related to (4) and (7). Therefore, for $p \ge 0.30$ the scaling function $\Psi(x)$ for the standard deviation is fit to the empirical form:

$$\Psi(x) = b_2 \left[1 - \frac{1}{1 + b_3 \log\left(1 + (x/b_1)^{D/d}\right)}\right] \tag{22}$$

where $b_1 = 8.4 \pm 0.8$, $b_2 = 1.23 \pm 0.01$ and $b_3 = 1.5 \pm 0.1$. Once again, as shown in Fig. 3, the collapsed data for $s_\xi \gg 1$ fits closely the expected scaling laws $\Psi(x) \sim 0.25 x^{D/d}$ as $x \to 0$ and $\Psi(x) \sim b_2 = 1.23$ as $x \to \infty$. Note that $\sigma/s_\xi^* \sim 1.23/0.90 = 1.36$ for $s_\xi \gg 1$, which differs from $\pi/\sqrt{6} = 1.2825\ldots$ by only 6.5%.

IV. SUBCRITICAL RENORMALIZATION

A. Flow in the Space of Distributions

There is a profound connection between renormalization-group (RG) concepts from the theory of critical phenomena [36,37] and the limit theorems of probability theory through what one might call "renormalization of the order parameter" (as opposed to "renormalization of the coupling constant" [38]). For many second-order phase transitions, the appropriate order parameter is a sum or average of identical, correlated r.v. indexed by the sites of a lattice,

e.g. the total magnetization in the Ising model. In such cases the central limit theorem for i.i.d. r.v. describes the behavior of the order parameter away from the critical point, where correlations are unimportant, and the mathematical concept of a "stable distribution" [40,42,41] amounts to a fixed point of an RG in the space of probability distributions of the order parameter [15,41].

In the case of percolation, although the appropriate order parameter is not the sum but rather the maximum of certain r.v., RG concepts can still be applied. Consider the c.d.f. of the largest-cluster size $F_N(s)$, which we normalize (or rather, successively "renormalize") as follows

$$G_N(z) \equiv F_N(\sigma_N z + \mu_N) = \text{Prob}\,(Z_N \leq z)\,, \tag{23}$$

where

$$Z_N \equiv \frac{S_{(N)} - \mu_N}{\sigma_N} \tag{24}$$

is a r.v. with zero mean and unit variance. Note that since $S_{(N)}$ assumes only integer values, $G_N(z)$ is a piecewise constant function of $z \in \Re$ with discontinuities at a countable set of points $\{(s-\mu_N)/\sigma_N | s \in \aleph\}$ with equal spacing $1/\sigma_N$.

The discrete mapping $G_N(z)$ can be viewed as a flow with increasing N (in some appropriate Banach space, e.g. L^p) which advects distributions towards various possible limiting behaviors. The subcritical portion of the flow is depicted in Fig. 5. For each $N \in \aleph$, the set of normalized distributions $\{G_N\}$ parameterized by $0 \leq p \leq 1$ forms a one-dimensional manifold, which we call the "physical manifold". The ends of the physical manifold corresponding to $p = 0$ and $p = 1$ are pinned at trivial fixed points, which are unit step functions centered at $x = 0$ and $x = N$, respectively (before normalization). Although these fixed points affect the nearby flow, every trajectory with $0 < p < 1$ eventually escapes toward one of three possible limiting behaviors for sufficiently large N: subcritical ($0 < p < p_c$), critical ($p = p_c$) or supercritical ($p_c < p < 1$). The latter two cases will be considered elsewhere; here we focus on subcritical behavior.

According to the heuristic arguments in section II and the simulation results in section III, the subcritical segment of the physical manifold is advected into a line of limit cycles (12)–(18) around the Fisher-Tippett distribution once the system size exceeds the crossover size $N \gg s_\xi$, or more precisely, $N \gg s_\xi^{d/D}$ ($L \gg \xi$). The envelope manifolds $\underline{G}$ and $\overline{G}$ for $0 < p < p_c$ defined in (16) enclose the limit cycles. As sketched in Fig. 5, the "radius" of each limit cycle grows as $p \to 0$ like $1/s_\xi(p)$, which reflects the influence of the $p = 0$ fixed point representing discreteness. In the opposite limit $p \to p_c$ (in the subcritical regime $s_\xi^{d/D} = o(N)$), the envelope manifolds meet at a fixed point corresponding to the continuous Fisher-Tippett distribution.

The approach to a fixed point is generally characterized by self-similarity, which holds "universally" for all trajectories leading to it. In the present case of a lattice-based system, this asymptotic self-similarity can described by a real-space RG which relates the c.d.f. $G_N(z)$ for a system of size $N = mn$ to the c.d.f. for each of n identical, contiguous cells (or blocks) of size m

$$G_{mn} = R_n G_m \tag{25}$$

in the limit $m \to \infty$ with n fixed, as shown in Fig. 6. As usual, the renormalization operators form an Abelian semigroup under composition $R_{mn} = R_m \circ R_n = R_n \circ R_m$. These kinds of arguments are typically applied to a coupling constant in the vicinity of a critical fixed point, where they capture the effect of long-range correlations [37]. They apply equally well, however, to the order-parameter distribution at a subcritical fixed point, where correlations disappear.

In a system exhibiting a phase transition, there is a *different* RG of the form (25) valid near each of the various fixed points. As shown in Figure 5, subcritical trajectories with $s_\xi \ll 1$ pass by the $p = 0$ fixed point and quickly become ensnared in the subcritical limit cycles, which are described by a RG given below. Such trajectories never feel much influence from the critical fixed point because correlation effects are dominated by discrete-lattice effects, due to proximity of the $p = 0$ fixed point. For larger values of $p < p_c$ such that $1 \ll s_\xi < \infty$, however, subcritical trajectories at first approach the critical fixed point ($1 \ll N \ll s_\xi^{d/D}$) before crossing over to the subcritical limit cycles ($N \gg s_\xi^{d/D}$). This crossover behavior was demonstrated for the mean and variance above in Figs. 2–3, but it also holds for the shape of the distribution.

In the vicinity of the critical fixed point ($1 \ll N \ll s_\xi^{d/D}$), trajectories obey a different RG reflecting the dominance of long-range correlations. The critical fixed point is unstable in the sense that subcritical and supercritical trajectories eventually crossover to a different limiting behavior along the direction of an unstable manifold. One such "crossover manifold" shown in Fig. 5, which connects the critical and subcritical fixed points, corresponds to the limits $N \to \infty$

and $p \to p_c$ with $N/s_\xi^{d/D} \to c$ for some constant $c > 0$. Likewise, the stable manifold converging to the critical fixed point corresponds to the limit $N \to \infty$ with $p = p_c$.

B. The Subcritical Renormalization Group

More than seventy years ago, Fréchet [16] and Fisher and Tippett [17] deduced the possible limiting distributions for extremes for i.i.d. r.v. with the following ingenious argument: If one partitions $N = mn$ i.i.d. r.v. into n disjoint subsets containing m r.v. each, then the largest of the mn outcomes is equal to the largest of the n largest outcomes in each subset of size m

$$S_{(mn)} = \max_{1 \le i \le n} S_{(m)}^i \tag{26}$$

where $S_{(m)}^i$ is the largest outcome in the ith subset. Since the $S_{(m)}^i$ are themselves i.i.d. r.v., the c.d.f. $F_N(s)$ of $S_{(N)}$ obeys the exact recursion

$$F_{mn}(s) = F_m(s)^n \tag{27}$$

for all m and n ($m^{1/d}, n^{1/d} \in \aleph$). In terms of the normalized distribution (23), the recursion takes the form

$$G_{mn}(z) = G_m\left(\frac{\sigma_{mn} z + \mu_{mn} - \mu_m}{\sigma_m}\right)^n, \tag{28}$$

which is essentially the subcritical RG for the normalized largest-cluster size distribution in percolation, but we must also address correlations and discreteness. In going from (27) to (28) we have defined a "renormalized" order-parameter distribution for percolation valid near the subcritical fixed point, in much the same way that the Kadanoff-Wilson block-spin construction defines a renormalized coupling constant for the Ising model valid near the critical fixed point [37].

The power of the cell-renormalization approach is that it provides a natural way to bound correlations and show that the subcritical limit cycles in percolation are described by the same RG as in the case of independent random variables (except for the subtle, persistent fluctuations due to discreteness described earlier). This is demonstrated rigorously in Ref. [43], but here we simply explain the basic ideas of these authors. The strategy of the proof (inspired by Fisher and Tippett) is to fix the number of cells $n > 1$ and let the size of each cell m diverge. Since correlations decay exponentially with distance in the subcritical regime, it seems plausible that the "renormalized" cell random variables $S_{(m)}^i$ would become uncorrelated (as the surface-to-volume ratio vanishes) in the limit $m \to \infty$, at least if the dimension were not too high ($d \le d_c$).

Precise bounds on the intercell correlations can be obtained as follows [43]. If the cells were independent (with free boundary conditions) we would have $F_{mn}(s) = F_m(s)^n$ as above, but due to correlations we have instead the upper bound

$$F_{mn}(s) \le F_m(s)^n \tag{29}$$

because joining the n cells together (and thus allowing clusters to connect and grow) can only increase the size of the largest cluster (and thus decrease the probability that the largest cluster has size $\le s$). A lower bound can be obtained by considering a set of "supercells" (again with free boundary conditions) formed by appending a "skin" of *linear* width $s/2$ to each of the original cells, as shown in Fig. 6. If the mass of the largest cluster intersected with each of these overlapping supercells were independently $\le s$, then the largest cluster overall would also have mass $\le s$ (because even a linear chain of length s would necessarily be completely contained in one supercell), which yields [45]

$$F_{(m^{1/d}+s)^d}(s)^n \le F_{mn}(s) \le F_m(s)^n. \tag{30}$$

These inequalities, which are valid for any dimension d, are the analogs of the Fréchet-Fisher-Tippett "RG" (27) for subcritical percolation, and from them the Fisher-Tippett behavior of the subcritical limit cycles can be established [43]. Heuristically, it is quite plausible that if the "typical" largest cluster size, say within z standard deviations of the mean

$$s_{mn}(z) = \mu_{mn} + z\sigma_{mn}, \tag{31}$$

does not grow too fast, i.e. $s_{mn}(z) \ll m^{1/d}$) as $m \to \infty$ with n and z fixed, then (30) should reduce asymptotically to (27). Given the results of Borgs et al. (4)–(7), we actually expect the much stronger bound $s_{mn}(z) = O(\log m)$. As explained below, this logarithmic scaling selects the Fisher-Tippett distribution $e^{-e^{-z}}$ from among the possible fixed points of (27).

C. The Subcritical Fixed Point

The subcritical fixed point is described by the classical theory for extremes of i.i.d. r.v. [19,20]. Following Fisher and Tippett [17], let us assume for now that a continuous fixed point of (28) exists pointwise for all z, i.e. $G_N(z) \to G(z)$ as $N \to \infty$ and $p \to p_c$ such that $s_\xi = o(N)$. In this case, there must exist finite constants $a_n > 0$ and b_n defined by

$$\lim_{m\to\infty} \frac{\sigma_{mn}}{\sigma_m} = a_n \tag{32a}$$

$$\lim_{m\to\infty} \frac{\mu_{mn} - \mu_m}{\sigma_m} = b_n \tag{32b}$$

such that the limiting distribution $G(z)$ obeys the equation [17]

$$G(z) = G(a_n z + b_n)^n \tag{33}$$

which was first discovered by Fisher and Tippett [44]. This functional equation has exactly three solutions, given in Table II, up to trivial translations and rescalings of z by constants. In the case of i.i.d. r.v. the basins of attraction of these three fixed points, which depend only on the tail of the parent distribution, were first characterized by Gnedenko [18]. In the case of percolation, we have argued above that the appropriate parent distribution has an exponential tail, which suggests that the Fisher-Tippett distribution is indeed the subcritical fixed point (again, ignoring discreteness).

Still assuming that a continuous limiting distribution $G(z)$ exists, let us make the following additional assumptions

$$\mu_N \sim s_\xi^* \log N \tag{34a}$$

$$\sigma_N - \sigma_{N-1} = o(1/N) \tag{34b}$$

which are clearly supported by our numerical simulations and are consistent with the rigorous results (4) and (7). These scaling axioms are expected to hold for all $d \le d_c$. Note that Eq. (34b) implies $\sigma_N = O(\log N) = O(\mu_N)$; with the fact that σ_N must be an increasing sequence, it also implies $a_n = 1$ for all $n \in \aleph$ (see [43]). From (34a) and (32b), we have

$$\frac{\mu_{mn} - \mu_m}{\sigma_m} \sim \frac{s_\xi^* \log n}{\sigma_m} \to b_n. \tag{35}$$

There are two possibilities: $\sigma_m \to \infty$ and $\sigma_m \to a$ for some constant $a > 0$. In the former case, we have $b_n = 0$ and hence $a_n \neq 1$ (see Table II), which is a contradiction. In the latter case, $b_n = (s_\xi^*/a) \log n$. Without loss of generality we can set $a = s_\xi^*$ (since this simply amounts to rescaling z) and obtain the equation

$$G(z) = G(z + \log n)^n \tag{36}$$

whose only nontrivial solution is $e^{-e^{-z}}$. This also implies that the standard deviation converges to a constant proportional to the crossover size, $\sigma_N \to s_\xi^* \sqrt{\pi}/6$.

D. The Subcritical Limit Cycles

Of course, the assumption of pointwise convergence to a continuous limiting distribution is wrong (e.g. see Fig. 1). Nevertheless, the conclusions of our simple derivation are not very different from those of a rigorous analysis including correlations and discreteness [43]. Note that although a limiting distribution $G(z) = \lim G_N(z)$ does not exist, the envelope functions $\underline{G}(z) = \liminf G_N(z)$ and $\overline{G}(z) = \limsup G_N(z)$ do exist. Assuming that $\underline{G}(z)$ and $\overline{G}(z)$ are continuous (although $G_N(z)$ is not), it can be shown from (34) that the envelope functions have the Fisher-Tippett form

$$\overline{G}(z - z_1) = \underline{G}(z - z_2) = e^{-e^{-z}} \tag{37}$$

for some constants $-\infty < z_1 \le z_2 < \infty$ and that the variance is bounded on the scale of the crossover size $\sigma_m / s_\xi^* = O(1)$. The latter result supports our assumption above in fitting the simulation data to (22). The reader is referred to Ref. [43] for a detailed proof of (37), which follows the RG strategy outlined here. The heuristic arguments and simulation results in sections II and III also lead us to conjecture that the "envelope width" $z_2 - z_1$ is simply set by the "strength" of the discreteness, i.e. the ratio of the lattice cell volume ($a^d = 1$) to the crossover size

$$z_2 - z_1 = \frac{1}{s_\xi^*}, \tag{38}$$

which vanishes in the limit $p \to p_c$.

V. CONCLUSION

In this article, a heuristic theory of the finite-size scaling of the largest-cluster size in subcritical percolation is presented and supported by numerical simulations. As expected away from a critical point, correlations are weak enough that a classical limiting distribution from the theory of extremes of independent random variables is recovered once the system size greatly exceeds the correlation length. This behavior can be easily understood via a cell-renormalization scheme, which also provides a suitable framework for rigorous analysis. Work is underway to extend this work to the supercritical case, where another classical limiting distribution arises, and the critical case, which involves a new universality class.

ACKNOWLEDGMENTS

The author is grateful to C. Borgs, J. T. Chayes, R. Kotecký and D. B. Wilson for stimulating discussions, E. Kaxiras and W. C. Carter for computer resources and A. Connor-Sax for help with X-Lisp-Stat. This work was supported by an NSF infrastructure grant.

---

[1] D. Stauffer and A. Aharony, *Introduction to Percolation Theory* (Taylor & Francis, London, 1992).
[2] M. Sahimi, *Applications of Percolation Theory* (Taylor & Francis, London, 1994).
[3] H. Kesten, *Percolation Theory for Mathematicians* (Birkhauser, Boston, 1982).
[4] G. Grimmett, *Percolation* (Springer, New York, 1989).
[5] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics* (Springer, New York, 1992).
[6] H. Kunz and B. Souillard, J. Stat. Phys. **19**, 77 (1978); Phys. Rev. Lett. **40**, 133 (1978).
[7] M. Schwartz, Phys. Rev. B **18**, 2364 (1978).
[8] H. N. V. Temperley and E. H. Lieb, Proc. R. Soc. London A **322**, 251 (1971).
[9] R. J. Baxter, H. N. V. Temperley and S. E. Ashley, Proc. R. Soc. London A **358**, 535 (1978).
[10] R. M. Ziff, S. R. Finch and V. S. Adamchik, Phys. Rev. Lett. **79**, 3447 (1997).
[11] P. Kleban and R. M. Ziff, Phys. Rev. B **57**, R8075 (1998).
[12] C. D. Lorenz and R. M. Ziff, J. Phys. A **31**, 8147 (1998).
[13] R. M. Ziff, C. D. Lorenz and P. Kleban, Physica A, to appear (1999).
[14] C. Borgs, J. T. Chayes, H. Kesten and J. Spencer, unpublished.
[15] M. Cassandro and G. Jona-Lasinio, Adv. Phys. **27**, 913 (1978).
[16] M. Fréchet, Ann. Soc. Polonaise de Math. **6**, 93 (1927).
[17] R. A. Fisher and L. H. C. Tippett, Proc. Cambridge Phil. Soc. **24**, 180 (1928).
[18] B. V. Gnedenko, Ann. Math. **44** (2), 423 (1943).
[19] E. J. Gumbel, *Statistics of Extremes* (Columbia Univ. Press, New York, 1958).
[20] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, second edition (Robert Krieger, Malabar, Florida, 1987).
[21] R. Rammal, J. Physique **46**, 1837 (1985).
[22] V. M. Vinokur, M. C. Marchetti, L.-W. Chen, Phys. Rev. Lett. **77**, 1845 (1996).
[23] J.-P. Bouchaud and M. Mézard, Physica D **107**, 174 (1997).
[24] J.-P. Bouchaud and M. Mézard, J. Phys. A **30**, 7797 (1997).
[25] D. B. Saakian, preprint. http://xxx.lanl.gov/abs/cond-mat/9906013
[26] K. W. Kehr, K. P. N. Murthy and H. Ambaye, Physica A **253**, 9 (1998).
[27] R. M. Ziff, J. Phys. A **27**, L657 (1994).
[28] D. Aldous and P. Diaconis, Bull. Amer. Math. Soc. **36**, 413 (1999).
[29] In one dimension, the largest cluster is simply the longest sequence of only "heads" in N flips of a fair coin.
[30] The c.d.f. $e^{-e^{-z}}$ is often called the "Gumbel distribution" in honor of Gumbel's pioneering work on applications [19], but here we acknowledge Fisher and Tippett who first discovered it [17].
[31] The code was originally written as a course project for Applied Physics 297r "Computational Approaches in Many-Body Physics" taught by E. Kaxiras at Harvard University in 1995.
[32] R. M. Ziff, Phys. Rev. Lett. **69**, 2670 (1992).
[33] In spite of its suboptimal performance near the critical point [1], the recursive burning algorithm efficiently identifies the relatively small clusters generated in subcritical simulations.
[34] R. M. Ziff, Computers in Physics **12**, 385 (1998), and references therein.

[35] L. Tierney, *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics* (Wiley, New York, 1990).
[36] M. E. Fisher, Rev. Mod. Phys. **70**, 653 (1998) and references therein.
[37] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, New York, 1992).
[38] P. J. Reynolds, H. E. Stanley and W. Klein, Phys. Rev. B **21**, 1223 (1980) and references therein.
[39] J.-P. Hovi and A. Aharony, Phys. Rev. E **53**, 235 (1996) and references therein.
[40] P. Levy, *Calcul des Probabilités* (Gauthier-Villars, Paris, 1926).
[41] J.-P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990).
[42] W. Feller, *Introduction to Probability Theory and Its Applications*, Vol. 2 (Wiley, New York, 1971).
[43] M. Z. Bazant, C. Borgs, J. T. Chayes, R. Kotecký, D. W. Wilson, unpublished.
[44] Fréchet derived the same equation with $b_n = 0$ a year earlier [16]. The year before that, Lévy published his pioneering work on stable distributions for sums of random variables [40].
[45] This is technically an application of the FKG inequality [4] as explained in Ref. [43].

p	s_ξ	θ
0.05	0.603 ± 0.005	1.0 ± 0.1
0.10	0.976 ± 0.001	0.97 ± 0.05
0.15	1.459 ± 0.001	0.99 ± 0.02
0.20	2.156 ± 0.001	1.03 ± 0.01
0.25	3.226 ± 0.001	1.075 ± 0.005
0.30	4.987 ± 0.002	1.109 ± 0.004
0.35	8.156 ± 0.005	1.129 ± 0.005
0.40	14.63 ± 0.03	1.13 ± 0.03
0.45	31.4 ± 0.1	1.20 ± 0.03
0.50	91.5 ± 0.2	1.20 ± 0.03

TABLE I. The measured correlation size $s_\xi(p)$ and exponent $\theta(p)$ for site percolation on the $d = 2$ square lattice.

Name	$G(z)$	Range	a_n	b_n	Basin of Attraction
Fréchet	$e^{-z^{-\alpha}}$	$(0, \infty)$	> 1	0	power-law tails
Weibull	$e^{-(-z)^{\alpha}}$	$(-\infty, 0)$	< 1	0	finite tails
Fisher-Tippett	$e^{-e^{-z}}$	$(-\infty, \infty)$	1	$\log n$	exponential tails

TABLE II. Summary of solutions to the Fisher-Tippett equation. In the last column, parent probability distributions $p_1(x)$ in the basins of attraction of each fixed point are (roughly) described by their decay at large x. (See Refs. [18–20] for more details.)

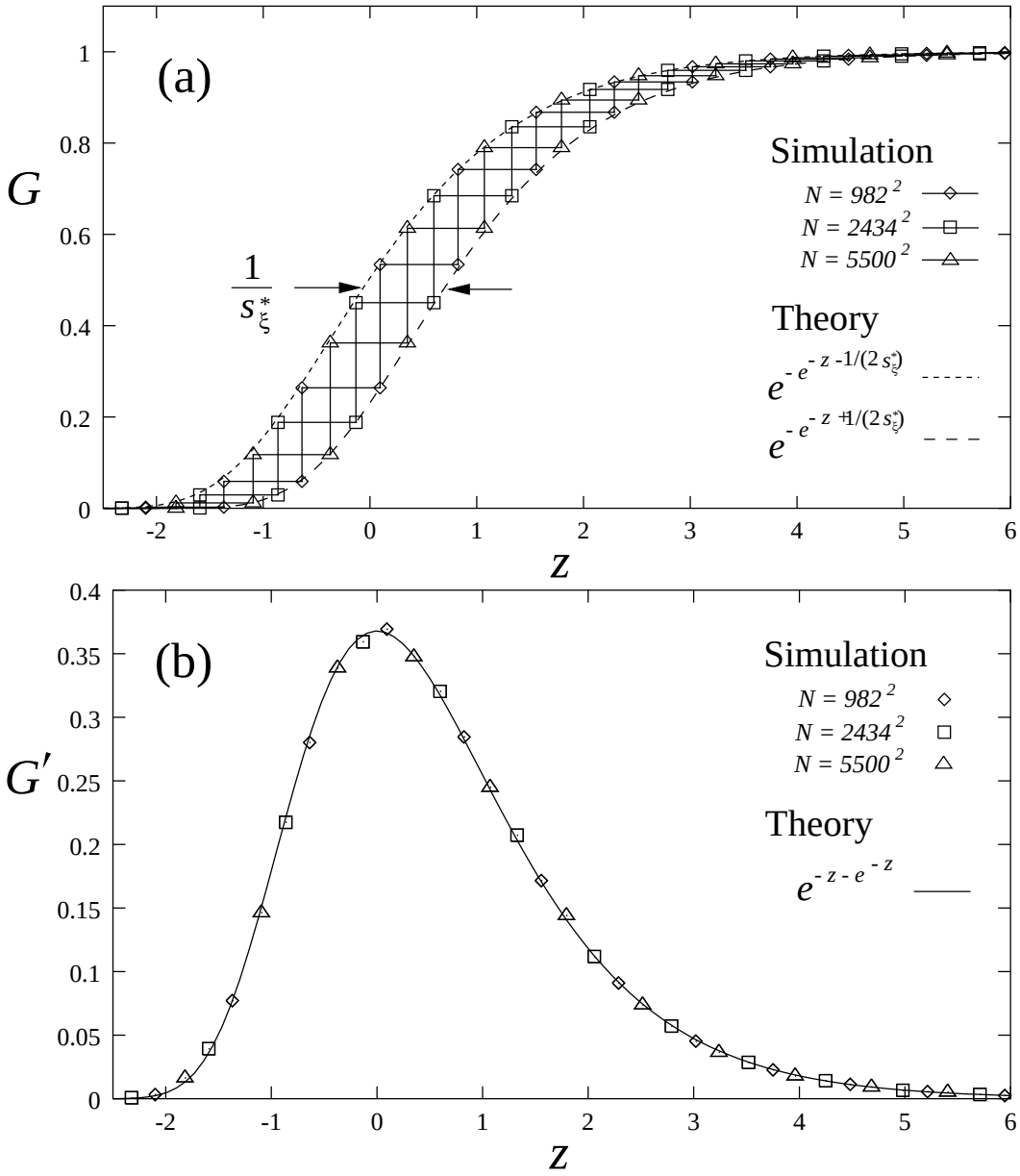

FIG. 1. The discrete c.d.f. in (a) and p.d.f. in (b) of the largest-cluster size for $p = 0.15$ and $N = 982^2$, 2434^2, 5500^2, normalized to have mean γ and variance $\pi^2/6$. The c.d.f.s in (a) are compared with (16), where $s^*_\xi = 0.90 \cdot s_\xi(0.15) = 1.313$.

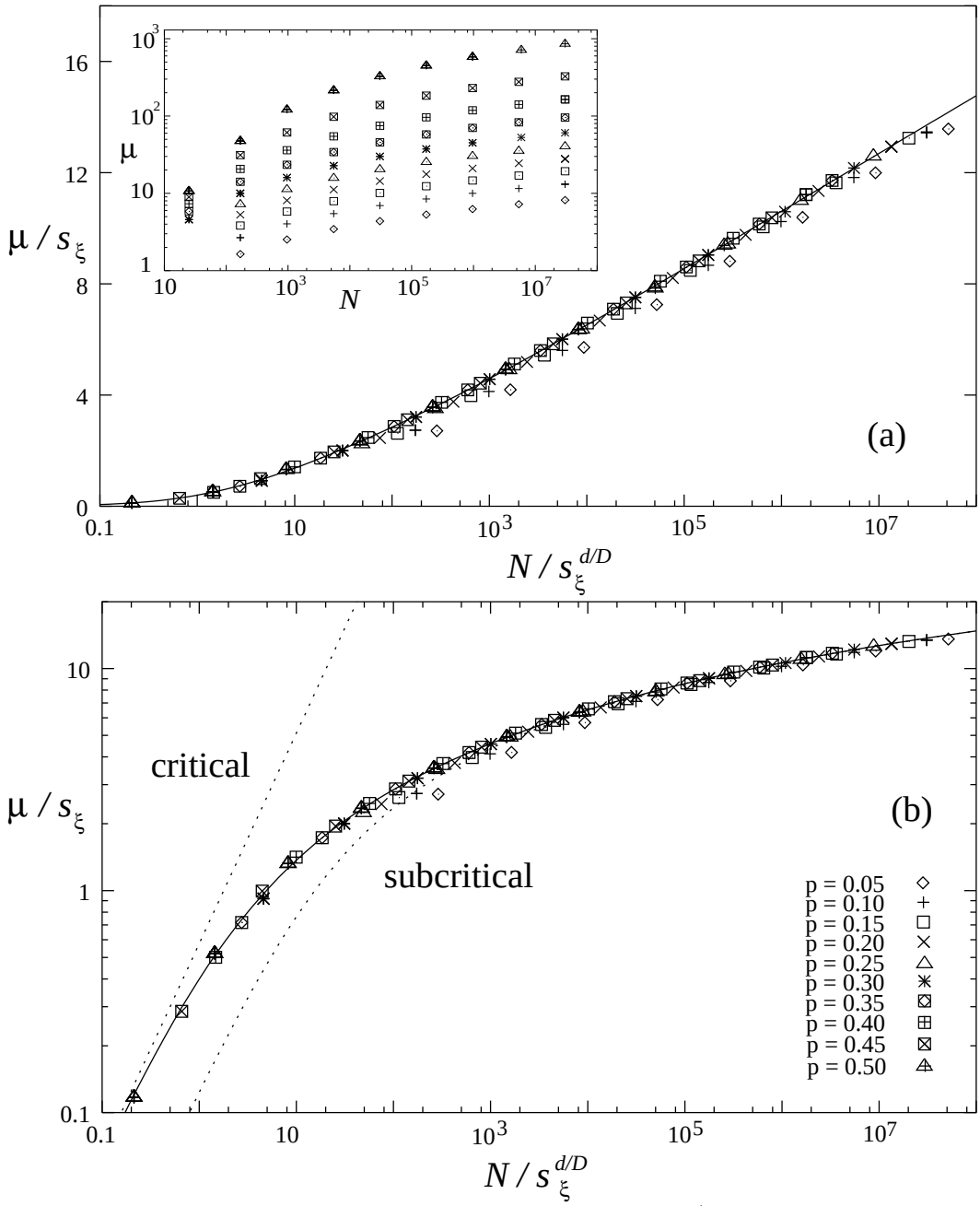

FIG. 2. The mean largest cluster size plotted as μ/s_ξ versus $N/s_\xi^{d/D}$ on a log-linear plot in (a) and a log-log plot in (b). The solid line fits the $p \geq 0.30$ data to Eq. (21) with asymptotic forms given by the dotted lines. The raw data is in the inset of (a); the legend in (b) applies throughout.

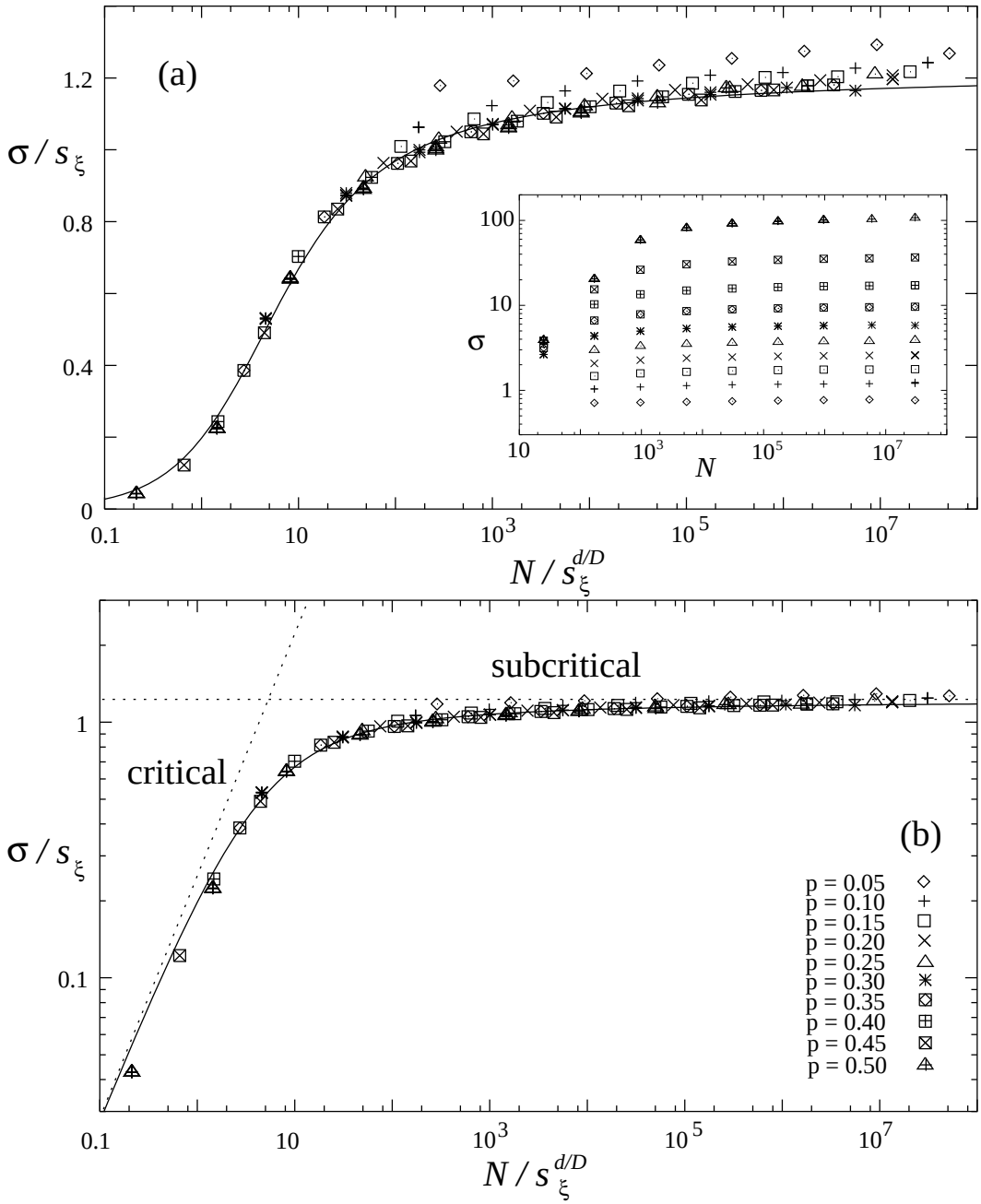

FIG. 3. The standard deviation of the largest cluster size plotted exactly as in Fig. 2. In this case, the $p \geq 0.30$ data is fit to Eq. (22).

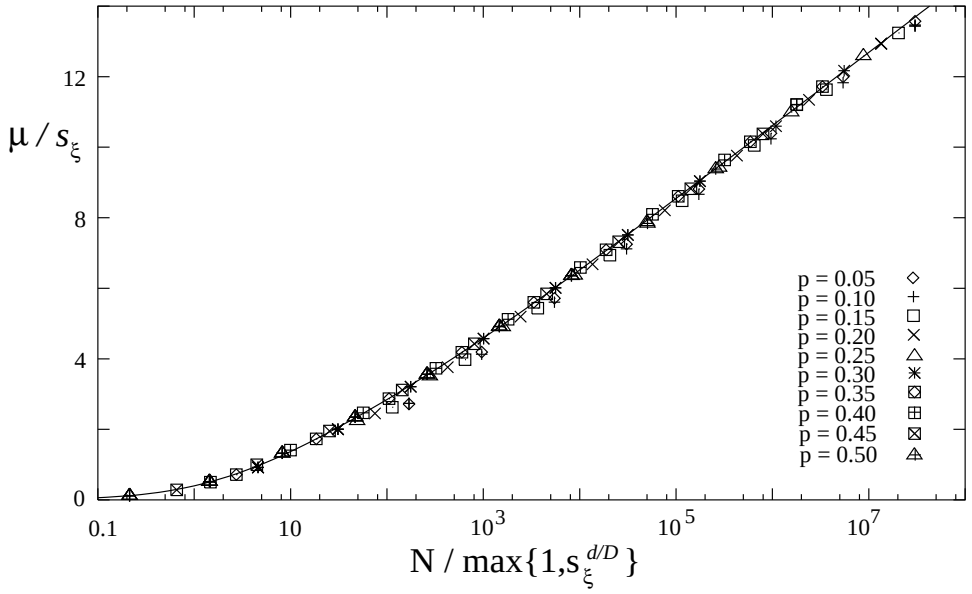

FIG. 4. The mean largest-cluster size plotted as μ/s_ξ versus $N/\max\{1, s_\xi^{d/D}\}$.

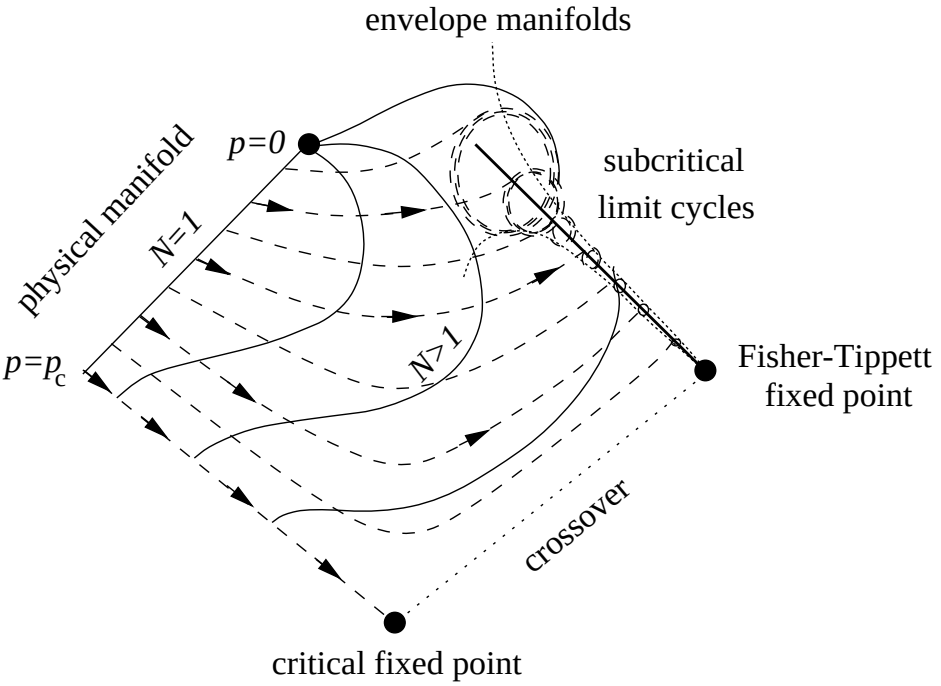

FIG. 5. Sketch of the trajectories (dashed lines) of the normalized largest-cluster size c.d.f. $G_N(z)$ (in some appropriate function space) for $p \leq p_c$. Arrows indicate directions of flow as $N \to \infty$. The physical manifold is shown for $N = 1$ and three larger values of N (solid lines). Also shown are three fixed points (thick dots), the subcritical envelope manifolds (short dotted lines) and the crossover manifold (long dotted line).

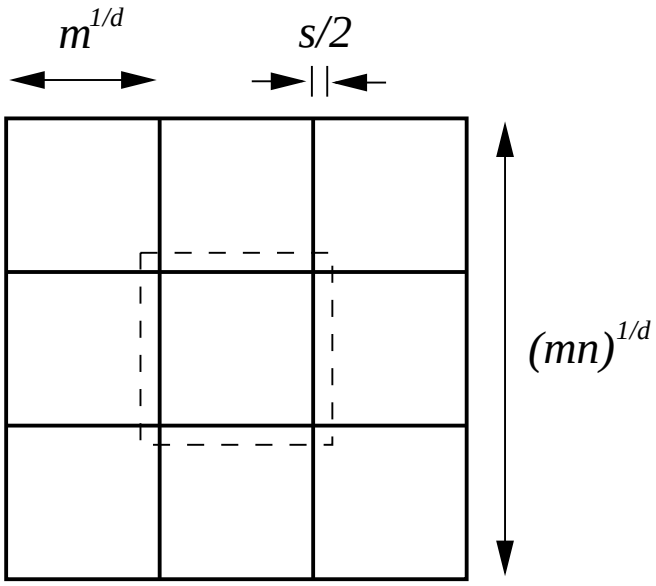

FIG. 6. Sketch of $n = 9$ cells (solid lines) used for subcritical renormalization on a square lattice of size $N = mn$. Along with the nine partitioning cells, one enlarged "supercell" (dashed lines) used in Ref. [43] to bound correlations for cluster sizes smaller than s (as described in the main text) is also shown.