The Annals of Probability
2000, Vol. 28, No. 3, 1162–1194

SIZE OF THE LARGEST CLUSTER UNDER ZERO-RANGE INVARIANT MEASURES[1]

BY INTAE JEON, PETER MARCH AND BORIS PITTEL

Ohio State University

We study the finite zero-range process with occupancy-dependent rate function $g(\cdot)$. Under the invariant measure, which can be written explicitly in terms of g, particles are distributed over sites and we regard all particles at a fixed site as a cluster. In the density one case, that is, equal numbers of particles and sites, we determine asymptotically the size of the largest cluster, as the number of particles tends to infinity, and determine its dependence on the rate function.

0. Introduction. An ordered partition of m with n parts is any nonnegative integer solution η of the equation

$$\eta_1 + \eta_2 + \cdots + \eta_n = m.$$

One may think of the distribution of m balls over n urns as typical of a variety of problems in probability theory leading to random ordered partitions, or random allocations, as they are sometimes called. We are interested in the asymptotic behavior as $n, m \to \infty$ of ordered partitions when they are chosen at random according to the probability distribution

$$\nu(\eta) = \mathscr{Z}^{-1} \prod_{i=1}^{n} g!(\eta_i)^{-1}. \tag{0.1}$$

Here $g(k)$ is a given positive function, $g!(k) = g(k)g(k-1)\cdots g(1)$, with $g!(0) = 1$, is the generalized factorial function and $\mathscr{Z}$ is the appropriate normalizing constant.

Here is one way distributions like this arise naturally. Consider n Markovian particles with transition matrix $P_{i,j}$ moving around on finite number m of sites, subject to the following interaction. If there are k particles at a given site then the waiting time for the (chronologically) first of these particles to move is exponential of rate $g(k)$ independent of the disposition of all other particles. This gives an informal description of the finite zero-range Markov process $Z(t) = (Z_1(t), Z_2(t), \dots, Z_n(t))$, whose state space is the set of ordered partitions. If $P_{i,j}$ is symmetric and irreducible then, as Spitzer [42] discovered, the corresponding zero-range process has ν as its unique invariant probability measure. (See also [18, 27, 7]). Thus we can consider ν as the stationary distribution of the n-tuple $Z = (Z_1, Z_2, \dots, Z_n)$, Z_j being the random occupancy of site j.

Received February 1999; revised January 2000.
[1]Supported in part by NSF Grant DMS-98-03410.
AMS 1991 *subject classifications*. Primary 60K35; secondary 82C22.
Key words and phrases. Zero-range process, equilibrium measure, cluster size, random partition, local limit theorem.

Let us pause to make two convenient normalizations. First, ν is invariant under a change of scale $g \to \lambda g$, so we may and shall assume that $g(1) = 1$. Second, for simplicity we will consider the case $n = m$, so the density of particles is one. The results and their proofs for the more general case $m = \gamma n$, the density γ being fixed, are quite similar.

Notice that if $g(k) = k$ then the particles are independent so we can use this case as a benchmark. Accordingly, when $g(k)$ is superlinear, particles tend to leave highly occupied sites much more quickly than in the independent case so that one expects ν to be concentrated on configurations having parts of comparable size. However, if $g(k)$ is sublinear then particles tend to linger at highly occupied sites so that one expects ν to be concentrated on lumpy configurations in which most sites are sparsely occupied and just a few sites contain most of the particles. Our goal is to verify this picture by describing the limiting behavior of the occupancy vector Z under the stationary distribution ν, by focusing on $Z_n^* = \max\{Z_1, Z_2, \dots, Z_n\}$, the population size of the most occupied site. Thinking of all particles at a given site as a cluster, we are interested, then, in the asymptotic behavior of the largest cluster.

The starting point for our study was the one-parameter family of rate functions

$$g(k) = k^{-\alpha}, \qquad -\infty < \alpha < \infty.$$

Evidently $\alpha = -1$ corresponds to the benchmark case of independent particles and if $\alpha = 0$ then ν reduces to the uniform distribution over ordered partitions. Our family of rate functions contains these two canonical examples as well as the superlinear case, $\alpha < -1$, the sublinear case, $\alpha > -1$ and the interesting situation $\alpha > 0$, in which the rate function tends to zero.

THEOREM 1. *Let* $g(k) = k^{-\alpha}$.

(a) *If* $\alpha > 1$ *then* $n - Z_n^*$ *converges to* 0 *in probability*.
(b) *If* $\alpha = 1$ *then* $n - Z_n^*$ *converges weakly to a Poisson distribution of parameter* 1.
(c) *If* $0 < \alpha < 1$ *then* $(n - Z_n^*)/n^{1-\alpha}$ *converges to* 1 *in probability*.
(d) *If* $\alpha = 0$ *then* $Z_n^*/\log n$ *converges to* $\log 2$ *in probability*.
(e) *If* $\alpha < 0$ *then* $Z_n^* \log\log n/\log n$ *converges to* $-1/\alpha$ *in probability*.

This result demonstrates two striking transitions. The first one, occurring at $\alpha = 1$, we call the condensation transition, since if $\alpha > 1$ then all particles condense into a single giant cluster. The giant cluster sheds a finite number of particles when $\alpha = 1$ and for $0 < \alpha < 1$ the number of "outcasts" is of the order $n^{1-\alpha}$. The second transition, which we call the gelation transition, occurs at $\alpha = 0$ since $Z_n^*/n \to 0$ or 1 accordingly as $\alpha \le 0$ or $\alpha > 0$. This situation is reminiscent of the emergence of a giant component in random graphs; see [5]. This phenomenon can also be viewed as conceptually close to the sol–gel transition in colloid physics. There, under certain conditions, a cluster of size proportional to the system size (the gel) emerges from a suspension of sticky particles (the sol). See [43] or [17]. An earlier, less complete version of Theorem 1 has been proved by one of the authors in his thesis [19].

Our interest in the zero-range process was kindled by the desire to find new, tractable situations in which particles coagulate or aggregate together in such a way that one cluster comes to have a fixed fraction of all available particles. There is a large scientific literature on phenomena of this type, (e.g., [8, 12, 15, 16, 17, 26, 39, 41, 42, 44, 45], among others. See the survey paper [1].) There is an accompanying mathematical literature centering mainly on the Smoluchowski, or coagulation–fragmentation equations (e.g., [1, 2, 10, 19, 20, 28, 29, 30]) and on models based on random graphs, (e.g., [35, 36, 37, 6]).

While the zero-range process has been studied quite extensively at least for increasing rate functions, the case of decreasing g, and especially g tending to zero, has not received as much attention. This could be because in these cases the infinite particle system may be difficult or impossible to construct. Nevertheless, the sequence of finite systems and the corresponding sequence of invariant measures is quite explicit and reveals some rather interesting features. There is another, combinatorial, attraction as well. Because ν has a product structure, it assigns equal probability to configurations which are permutations of one another, which means ν descends to a probability measure on partitions of n.

Emboldened by the case $g(k) = k^{-\alpha}$, we turned to the general problem of identifying classes of rates $g(\cdot)$ for which the largest cluster contains either almost all n particles, or $o(n)$ particles. We also wanted to determine an intermediate class of rates for which the largest cluster is asymptotic to cn, with $c \in (0, 1)$. As Theorem 1 demonstrates, this cannot happen for $g(k) = k^{-\alpha}$, $(-\infty < \alpha < \infty)$.

To make formulations of the results we obtained more comprehensible and coherent, we need to touch briefly on one of the tools we use. Colloquially it is known as the conditioning device, and it has been used for analysis of random partitions by [40, 25, 24, 22, 13, 3, 33, 34], among others. In fact, the approach has its roots in statistical mechanics (cf. [23]).

Consider a one-parameter family of i.i.d. integer valued random variables $X_1, X_2, \dots, X_n$ with common distribution

$$P\{X = k\} = G^{-1}(x)\frac{x^k}{g!(k)}, \qquad G(x) = \sum_{k'} \frac{x^{k'}}{g!(k')}, \tag{0.2}$$

where, depending on the rate function g, either $k, k' \in \mathbb{N}$ or $k, k' \leq n$. We call these the *untruncated* and *truncated* cases, respectively. It is an easy calculation that, *independent* of x,

$$(Z_1, Z_2, \dots, Z_n) \overset{\mathscr{D}}{\equiv} (X_1, X_2, \dots, X_n | X_1 + X_2 + \cdots + X_n = n).$$

In other words, the occupancy numbers $Z_1, \dots, Z_n$ are distributed like $X_1, \dots, X_n$, conditioned to have a fixed sum. How to best choose x? Intuitively, the larger $P\{\sum_j X_j = n\}$ is, the less "intrusive" is the conditioning; thus to choose x to maximize the probability in question is quite natural. Presumably, this choice makes the expected value of X as close as possible to 1, the particle density.

Here is how it works out in a simple case. When $g(k)$ increases (e.g., $g(k) = k^{-\alpha}$, $\alpha \le 0$), the series $G(x)$ has a positive radius of convergence R and, in addition,

$$\lim_{x\uparrow R} \frac{xG'(x)}{G(x)} = \infty, \qquad G(x) \stackrel{\text{def}}{=} \sum_{k\ge 0} x^k/g!(k).$$

A standard argument shows that $EX = xG'(x)/G(x)$ increases with x. Thus the *untruncated* X, with x chosen to make $EX = 1$, is a reasonable choice in this case. Since $x < R$, the distribution of X has an exponential tail and a standard local limit theorem shows that $\Pr\{\sum_j X_j = n\}$ is of order $n^{-1/2}$. Thus the probability of the conditioning event is under control and we are led to a result for "large rates."

THEOREM 2. *Suppose $R > 0$ and $\lim_{x\uparrow R} xG'(x)/G(x) > 1$. Then there exists a unique $\bar{x}$ such that $EX = 1$. Let $m = m(n)$ be such that*

$$\lim_{n\to\infty} nP\{X \ge m\} = \lambda \in (0, \infty).$$

Then $m(n) \to \infty$, $m(n) = O(\log n)$ and $Z_n^*/m(n) \to 1$, in probability.

The precise role played by λ is laid out below in Theorem 2.1. Needless to say, the asymptotic behavior of $m(n)$ is, in the main, independent of λ. Theorem 2 applies to $g(k) = k^{-\alpha}$, $\alpha \le 0$, and we recover parts (d) and (e) of Theorem 1, by finding that

$$m(n) = \begin{cases} (1 + o(1)) \log_2 n, & \text{if } \alpha = 0, \\ \dfrac{\log n}{-\alpha \log\log n}(1 + o(1)), & \text{if } \alpha < 0. \end{cases}$$

The last relation comes from

$$\sum_{k\ge m} \frac{\bar{x}^k}{(k!)^{-\alpha}} = (1 + o(1)) \frac{\bar{x}^m}{(m!)^{-\alpha}}, \qquad m \to \infty,$$

and the Stirling formula for $m!$.

An interesting situation arises if $\lim_{x\uparrow R} xG'(x)/G(x) \le 1$. Then $R < \infty$ and the series for G and G' both converge at $x = R$, but $kR^k/g!(k)$ and $R^k/g!(k)$ approach 0 slower than any exponential. Choosing $x = R$ in (0.2), we see that X cannot have an exponential tail. Now there are many ways this can occur. In this paper, for simplicity, we confine ourselves to the case of a polynomial tail, with $P\{X = j\}$ being of order $j^{-\beta}$. We consider the case

$$g(j) = R\left(1 + \frac{\beta}{j} + O\big(j^{-(1+\delta)}\big)\right), \qquad \delta > 0.$$

Since $EX \le 1 < \infty$, we must have $\beta > 2$. Again, we must estimate $P\{X_1 + X_2 + \cdots + X_n = n\}$ but standard local limit theorems no longer apply. Substantial new effort is required and the extra condition $\beta > 3$ allows us to bring an estimate of Nagaev to bear on the problem. We are led to a result for “moderate rates.” Here g has the specific form above, $\beta > 3$ and we choose $x = R$.

THEOREM 3. (a) *If* $EX = 1$, *then*, *in probability*, $Z_n^*/n^{(\beta-1)^{-1}}$ *is bounded from both zero and infinity*. (b) *If* $EX < 1$, *then* Z_n^*/n *converges to* $1 - EX$ *in probability*. *In addition*, *the second largest cluster is* $o(n^{1/2}\log^2 n)$ *in probability*.

Theorems 2 and 3 demonstrate a double jump phenomenon, analogous to that discovered by Erdős and Rényi [11] in 1960 in their pioneering study of evolving random graphs. Namely, depending on the value $\lim_{x\uparrow R} xG'/G$, the largest cluster size is either of order n, or $n^{1/(\beta-1)}$, or $\log n$. Similar behavior was found in random graph models of polymerization (c.f. [35, 36, 37]).

The attentive reader has certainly noticed by now that had we considered the more general case of γn particles, we would have had to compare the above limit with γ, the density of particles, instead of 1 to decide which one of the three modes holds. The following picture seems to emerge here. For any density γ, choose the parameter x so that the expectation of X is as close as possible to γ. If x is inside the radius of convergence of G then the maximum occupancy over all sites behaves roughly like the maximum of i.i.d. random variables X_j, with expectation γ. Otherwise, the gelation transition occurs when this best possible expectation falls below γ, and then one of the remaining sites becomes a dumping ground for the huge number of particles “rejected” by the rest of the sites.

Finally, consider the case $R = 0$ which occurs when the rate function tends to zero. A special case is $g(k) = k^{-\alpha}$ $(\alpha > 0)$, which featured in Theorem 1. We need to exercise some control over the way g vanishes. Specifically, we assume that $g(k)$ admits a smooth interpolation $g(x)$, $x \in [1,\infty)$, such that:

1. $g(x) \downarrow 0$, as $x \uparrow \infty$.
2. $r(x) := -d(\log g(x))/d\log x$, the logarithmic rate of $1/g(x)$ is bounded and nonincreasing on $[1,\infty)$.
3. The logarithmic rate $s(x) = d\log r(x)/d\log x$ of $r(x)$ tends to zero as $x \to \infty$.
4. $\lim_{x\to\infty}[r(x)\log^{1/2} x - \log\log x] = \infty$.

NOTES. (i) Condition 2 is similar to the condition of the representation theorem for slowly varying functions. (See [4], Theorem 1.3.1.)

(ii) Given $r(x)$, we can write

$$g(x) = g(1)\exp\left\{-\int_1^x \frac{r(t)}{t}\,dt\right\}.$$

(iii) For example, if $g(k) = k^{-\alpha}$, then $r(x) \equiv \alpha$ and the conditions are obviously met. We emphasize that conditions 1–3 can perhaps be somewhat relaxed, at the price of more complicated proofs and weaker results. However, the last condition, which says basically that $r(x)$ cannot go to zero too fast, seems much harder to nudge. It is easy to see that condition (4) excludes rates that vanish slower than $\exp(-c\,\log^{1/2} x)$, such as $g(k) = a\log^{-b} k$, $b > 0$.

A common feature in existing studies of random partitions is that truncation is not needed for the conditioning device to work. However if $g(k) \downarrow 0$, the series $\sum_{k\geq 0} x^k/g!(k)$ converges for $x = 0$ only. So, breaking with tradition, we restrict the sum to the range $\{0, 1, \dots, n\}$ and let $x = x_n \to 0$ at a properly chosen rate. As before, we need good asymptotic estimates of $P\{\sum_j X_j = n\}$, especially from below.

No such limit theorems are available in case of the vanishing rates, and so a fairly large portion of our effort goes into the proof that the probability in question is at least $\exp(-cng(n)r(n))$. For $g(k) = k^{-\alpha}$, $0 < \alpha < 1$, the bound becomes $\exp(-c'n^{\beta})$, $\beta \in (0, 1)$, which is much smaller than for increasing rates. Considering how much smaller the probability is here, it came as a surprise to us that the conditioning device, with a twist, was indispensable for vanishing rates as well.

Here is our main result for vanishing rates. It illustrates the condensation transition identified in Theorem 1.

THEOREM 4. *Under conditions* 1–4, $(n - Z_n^*)/ng(n)$ *converges to* 1 *in probability. Thus, with high probability, the giant cluster absorbs all but* $ng(n)(1 + o(1))$ *particles. Furthermore*:

(a) *If* $r(\infty) = 1$, *then* $ng(n) \leq 1$, *and* $n - Z_n^*$ *is asymptotically Poisson with parameter* $ng(n)$; *more precisely, for fixed* k *and* $n \to \infty$:

$$P\{Z_n^* = n - k\} = (1 + o(1))\,\exp(-ng(n))\frac{(ng(n))^k}{k!}.$$

(b) *If* $r(\infty) > 1$, *then* $n - Z_n^*$ *converges to* 0 *in probability.*

Parts (a), (b) and (c) of Theorem 1 follow directly from Theorem 4. To be sure, it was this earlier result that led us to a conjecture "$Z_n^* \approx n - ng(n)$ for $g(k) \downarrow 0$ sufficiently fast." Structurally, the proof of Theorem 4 is close to the original argument, the main new question being quantification of "sufficiently fast." The answer turned out to be condition 4. It describes, however incompletely, a boundary in "rate space," separating classes of rate functions under which random ordered partitions behave rather differently. For further discussion of the condensation transition, see [21].

The rest of the paper is organized as follows. In Section 1 we prove, for completeness, Spitzer's formula (0, 1) for the stationary distribution ν and show, for the monotone rate $g(\cdot)$, that $\nu(\eta)$ is monotone with respect to the "dominance order" on unordered partitions. It Section 2 we prove Theorems 2 and 3, and in Section 3 we prove Theorem 4. We conclude, in Section 4, with a list of open problems and questions.

1. Preliminaries. For any positive integers m and n, let Ω_n^m denote the set of ordered partitions of m with n parts, that is the set of nonnegative integer solutions of

$$\eta_1 + \eta_2 + \cdots + \eta_n = m.$$

Let $g!(k) = g(k)g(k-1)\cdots g(1)$ with the convention that $g!(0) = g(0) = 1$. We include for the reader's convenience a proof of the fact that ν is the unique invariant measure of the finite zero-range process on Ω_n^m [42].

LEMMA 1.1 (Spitzer [42]). *Let $P_{i,j}$ be a symmetric, irreducible transition matrix and let g be a positive rate function. For any $\eta \in \Omega_n^m$, let*

$$\mu(\eta) = \prod_{i=1}^{n} g!(\eta(i))^{-1}$$

and set

$$\nu(\eta) = \tfrac{1}{\mathscr{Z}}\mu(\eta) \quad \textit{where} \quad \mathscr{Z} = \sum_{\eta \in \Omega_n^m} \mu(\eta).$$

Then ν is the reversible invariant measure for the finite zero-range process corresponding to $P_{i,j}$ and g.

PROOF. The zero-range process is generated by the operator

$$L_n f(\eta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\big[f(\eta^{i,j}) - f(\eta)\big]P_{i,j}\,g(\eta_i).$$

Here, if $\eta(i) = 0$ then $\eta^{i,j} = \eta$ and if $\eta(i) > 0$ then

$$\eta^{i,j}(k) = \eta(k) + \delta_{j,k} - \delta_{i,k},$$

where δ is the Kronecker symbol. It suffices to show that the detailed balance condition holds. Indeed,

$$\begin{aligned}
\frac{g(\eta(i))P_{ij}\nu(\eta)}{g(\eta^{i,j}(j))P_{ji}\nu(\eta^{i,j})} &= \frac{g(\eta(i))\prod_{k=1}^{n} 1/g!(\eta(k))}{g(\eta(j)+1)\prod_{k=1}^{n} 1/g!(\eta^{i,j}(k))}\\
&= \frac{g(\eta(i))g!(\eta(i)-1)g!(\eta(j)+1)}{g(\eta(j)+1)g!(\eta(i))g!(\eta(j))}\\
&= 1. \qquad \square
\end{aligned}$$

Let $S(n)$ be the permutation group of n letters. For any $\sigma_n \in S(n)$, and for any $\eta = (\eta(1),\dots,\eta(n)) \in \Omega_n^m$, let $\sigma_n\eta = (\eta(\sigma_n(1)), \eta(\sigma_n(2)),\dots,\eta(\sigma_n(n)))$.

LEMMA 1.2. *For any $\eta \in \Omega_n^m$ and $\sigma \in S(n)$, let ν be the invariant measure corresponding to g. Then $\nu(\eta) = \nu(\sigma\eta)$.*

For the proof, evidently $\nu(\sigma\eta)$ and $\nu(\eta)$ differ only in the order of multiplication of their factors.

On certain subsets of the state space we can find the configuration of maximum probability. Let

$$A_k = \left\{\eta \in \Omega_n^m \colon \max_{1\le i\le n} \eta(i) \ge k\right\};$$

so that $\Omega_n^m \setminus A_{k+1}$ is the set of configurations in which no site has more than k particles.

LEMMA 1.3. *Let $m = lk + r$, $0 \le r < k$, and let $\eta_* = (k, k, \dots, k, r, 0, 0, \dots, 0)$, where the k's are repeated l times. Let g be a rate function with corresponding invariant measure ν on Ω_n^m. Then $\eta_* \in \Omega_n^m \setminus A_{k+1}$ and for any $\eta \in \Omega_n^m \setminus A_{k+1}$:*

(a) *$\nu(\eta_*) \ge \nu(\eta)$, if g is decreasing.*
(b) *$\nu(\eta_*) \le \nu(\eta)$, if g is increasing.*

PROOF. Let us suppose g is decreasing and prove (a); the proof of item (b) is similar. For the purpose of comparing probabilities, we may suppose by Lemma 2 that the entries of any configuration are ordered in decreasing order:

$$\eta(1) \ge \eta(2) \ge \cdots \ge \eta(n).$$

First, let us compare probabilities of η and $\eta^{i,j}$ where $i > j$ so that a particle from a less occupied site has been moved to a site of higher occupancy. Then

$$\begin{aligned}
\frac{\nu(\eta)}{\nu(\eta^{i,j})} &= \frac{g!(\eta^{i,j}(1))\cdots g!(\eta^{i,j}(n))}{g!(\eta(1))\cdots g!(\eta(n))} \\
&= \frac{g!(\eta^{i,j}(i))g!(\eta^{i,j}(j))}{g!(\eta(i))g!(\eta(j))} \\
&= \frac{g!(\eta(i)-1)g!(\eta(j)+1)}{g!(\eta(i))g!(\eta(j))} \\
&= \frac{g!(\eta(i)+1)}{g(\eta(i))} \\
&\le 1,
\end{aligned}$$

where the last inequality is true since g is decreasing. Second, for any $\eta_* \in \Omega_n^m \setminus A_{k+1}$, there exists a sequence of transpositions of the form

$$\eta \to \eta_1 \to \eta_2 \to \cdots \to \eta_s = \eta_*,$$

where for each $1 \le r \le s-1$, $\eta_{r+1} = \eta_r^{i,j}$ for some $i > j$. By the comparison above,

$$\nu(\eta_*) = \nu(\eta_s) \ge \nu(\eta). \qquad \square$$

Let us consider the equation

$$x_1 + x_2 + \cdots + x_n = m.$$

It is well known that the number of nonnegative solutions is $\binom{n+m-1}{n-1}$ and the number of positive solutions is $\binom{m-1}{n-1}$; [38]. The following lemma records these facts in useful form.

LEMMA 1.4. *Let* $\widetilde{\Omega}_n^m = \{\eta \in \Omega_n^m \colon \eta(i) \geq 1$ *for all* $i\}$. *Then*:

(a) $|\widetilde{\Omega}_n^m| = \binom{m-1}{n-1}$.
(b) $|\Omega_n^m| = \binom{n+m-1}{n-1} = \binom{n+m-1}{m}$.

From now on, to make things a bit simpler, we will consider only the case $n = m$, and we will drop m from the notation. Thus $\Omega_n^m = \Omega_n$, etc.

2. Large and moderate rates. We prove Theorems 2.1 and 2.2, which are more detailed versions of Theorems 2 and 3 in the introduction. Consider the case when R, the radius of convergence of $G(x) = \sum_{j\geq 0} x^j/g!(j)$, is positive. Assume first that

$$\lim_{x\uparrow R} \rho(x) > 1, \qquad \rho(x) \stackrel{\text{def}}{=} \frac{xG'(x)}{G(x)}.$$

Since by Cauchy's inequality,

$$\rho'(x) = \frac{1}{xG^2(x)}\left(\left(\sum_j x^j j^2/g!(j)\right)\left(\sum_{j'} x^{j'}/g!(j')\right) - \left(\sum_j jx^j/g!(j)\right)^2\right) > 0,$$

there exists a unique $\bar{x} < R$ such that $\rho(\bar{x}) = 1$. Introduce the random variable X with distribution

$$p_j = P\{X = j\} = G(\bar{x})^{-1}\frac{\bar{x}^j}{g!(j)}, \qquad j \geq 0.$$

Since $\bar{x} < R$, there exists $c \in (0, 1)$ such that $p_{j+1}/p_j \leq c$ for all large enough j's. Consequently $p_j = O(c^j)$ as $j \to \infty$, that is, the tail of X is exponentially thin, and also $\sum_{j\geq\mu} p_j$ is of order p_μ exactly, as $\mu \to \infty$. In particular, $EX^k < \infty$ for every $k \in \mathbb{N}$, and, of course, $EX = 1$ by the definition of $\bar{x}$.

THEOREM 2.1. *Suppose* $m = m(n)$ *is such that*

$$\lim_{n\to\infty} nP\{X \geq m\} = \lambda \in (0, \infty).$$

Let $\mathscr{C}_n$ *stand for the total number of clusters of size at least* m. *Then* $m(n) \to \infty$, $m(n) = O(\log n)$, $\mathscr{C}_n$ *converges in distribution to a Poisson with parameter* λ, *and* $Z_n^*/m(n)$ *converges to* 1 *in probability*.

PROOF. We begin by observing that $m = O(\log n)$, since $p_j = O(c^j)$. Introduce $X_n^* = \max_{1 \le j \le n} X_j$, and let $X_1^{(m)}, \dots, X_n^{(m)}$ be the independent copies of $X^{(m)}$, which is X conditioned on $\{X < m\}$. In addition, let $q(m) = P\{X \ge m\}$.

According to ([25], Lemma 2, page 10) we have

$$P\{Z_n^* < m\} = P\{X_n^* < m\} \frac{P\left\{\sum_{j=1}^n X_j^{(m)} = n\right\}}{P\left\{\sum_{j=1}^n X_j = n\right\}}. \tag{2.1}$$

By the local limit theorem ([9], [14]),

$$P\left\{\sum_{j=1}^n X_j = \nu\right\} = \frac{1+o(1)}{\sqrt{2\pi\sigma^2(X)n}} \exp\left\{-\frac{(\nu - n\mathbf{E}X)^2}{2n\sigma^2(X)}\right\},$$

uniformly for $|\nu - nEX| = O(n^{1/2})$. A careful study of the proof reveals that the analogous formula holds for $X_1^{(m)}, \dots, X_n^{(m)}$ uniformly for $m > 0$. This is so because the effect of conditioning on the event $\{X \le m\}$ can be estimated uniformly using remainder terms in expansions of $E\big(\exp(iuX^{(m)})\big)$, notably $E(|u| \cdot |X^{(m)}|^3 \wedge |X^{(m)}|^2)$; see [9]. Furthermore, $\sigma^2(X^{(m)}) \to \sigma^2(X)$, as $m \to \infty$, and

$$\begin{aligned} n - nEX^{(m)} &= n\frac{E\{X\colon X \ge m\} - P\{X \ge m\}}{P\{X < m\}} \\ &\le 2nE\{X\colon X \ge m\} \\ &= O(nq(m)m) \\ &= O(\log n). \end{aligned}$$

So $n - nEX^{(m)} = o(n^{1/2})$, and we have

$$\begin{aligned} P\left\{\sum_{j=1}^n X_j^{(m)} = n\right\} &= \frac{1+o(1)}{\sqrt{2\pi n\sigma^2(X^{(m)})}} \exp\left\{-\frac{(n - nEX^{(m)})^2}{2n\sigma^2(X^{(m)})}\right\} \\ &= \frac{1+o(1)}{\sqrt{2\pi n\sigma^2(X)}}. \end{aligned}$$

Therefore [see (2.1)], by the definition of $m = m(n)$,

$$\begin{aligned} P\{Z_n^* < m\} &= (1+o(1))P\{X_n^* < m\} \\ &= (1+o(1))(1 - P\{X \ge m\})^n \\ &\to e^{-\lambda}. \end{aligned} \tag{2.2}$$

Thus $P\{\mathscr{C}_n = 0\} \to e^{-\lambda}$, and it remains to show that

$$P\{\mathscr{C}_n = k\} \to e^{-\lambda}\lambda^k/k!$$

for every fixed $k \ge 1$. To this end we notice first that $nq(m_1) \to 0$ for $m_1 = [(1+\varepsilon)m]$ and a fixed $\varepsilon > 0$. So $P\{Z_n^* < m_1\} \to 1$ as $n \to \infty$; that is,

$P\{\mathscr{C}_n = \widetilde{\mathscr{C}}_n\} \to 1$, where $\widetilde{\mathscr{C}}_n$ is the total number of clusters with size between m and $m_1 - 1$. Thus, by [5] Chapter 1, we need to show only that the binomial moments of $\widetilde{\mathscr{C}}_n$ converge to those of the Poisson of parameter λ.

Now $\widetilde{\mathscr{C}}_n = \sum_{j=1}^n I_j$, with I_j being the indicator of the event $\{Z_j \in [m, m_1)\}$. Hence, as in (2.1),

$$E\binom{\widetilde{c}_n}{k} = \binom{n}{k} \frac{\sum_{m \le l_1, \dots, l_k \le m_1 - 1} \prod_{t=1}^k p_{l_1} \cdot P\{\sum_{j=k+1}^n X_j = n - \sum_{s=1}^k l_s\}}{P\{\sum_{j=1}^n X_j = n\}}.$$

To see why this formula holds, note that $\binom{n}{k}$ is the number of ways to select k sites from m while the numerator is the probability that the X_j's associated with the selected sites have values in $[m, m_1 - 1]$, and that $\sum_{j=1}^n X_j = n$. Here, since none of l_j, $(j \le k)$, exceeds m_1,

$$n - \sum_{s=1}^k l_s = n - k + O(km_1) = n - k + O(\log n).$$

So, uniformly for $l_1, \dots, l_k < m_1$, the probability in the numerator is asymptotic to the probability in the denominator. Therefore

$$\begin{aligned} E\binom{\widetilde{\mathscr{C}}_n}{k} &= (1 + o(1)) \frac{\left(n \sum_{l=m}^{m_1 - 1} p_l\right)^k}{k!} \\ &= (1 + o(1)) \frac{(nq_m)^k}{k!} \\ &\to \frac{\lambda^k}{k!}. \end{aligned} \tag{2.3}$$

Hence $\widetilde{\mathscr{C}}_n$ converges in distribution to a Poisson (λ). It remains to show that $P\{Z_n^* \ge (1 - \varepsilon)m(n)\} \to 1$ for every $\varepsilon \in (0, 1)$. Fix a positive integer a. Using the properties of the distribution $\{p_j\}$, it is easy to show that

$$b_1 c^{-a} \lambda \le \liminf_{n \to \infty} nq(m(n) - a) \le \liminf_{n \to \infty} nq(m(n) - a) \le b_2 c^{-a} \lambda;$$

here b_1, b_2 do not depend on a. Let $\widehat{\mathscr{C}}_n$ denote the number of clusters of size $m(n) - a$ at least. For every subsequence $\{n_l\}$ such that there exists

$$\lambda' := \lim_{n \to \infty} n_l q(m(n_l)) - a),$$

we see that $\lambda' \in [b_1 c^{-a} \lambda, b_2 c^{-a} \lambda]$ and that $\widehat{\mathscr{C}}_{n_l}$ is asymptotically Poisson (λ'). So

$$\begin{aligned} \lim_{n_l \to \infty} P\{Z_{n_l}^* \ge m(n_l) - a\} &= \lim_{n_l \to \infty} P\{\widehat{\mathscr{C}}_{n_l} > 0\} \\ &= 1 - e^{-\lambda'} \\ &\ge 1 - \exp(-b_1 c^{-a} \lambda). \end{aligned}$$

Therefore, for every fixed $a < 0$,

$$\liminf_{n \to \infty} P\{Z_n^* \ge m(n) - a\} \ge 1 - \exp(-b_1 c^{-a} \lambda).$$

Since $m(n) - a > (1-\varepsilon)m(n)$ for large enough n, we get then

$$\liminf_{n\to\infty} P\{Z_n^* \geq (1-\varepsilon)m(n)\} \geq 1 - \exp(-b_1 c^{-a}\lambda),$$

and letting $a \to \infty$ completes the proof. □

Now suppose that $\lim_{x\uparrow R} \rho(x) \leq 1$. Then necessarily $R < \infty$, and the series for $G(x)$ and $G'(x)$ both converge at $x = R$. Define the random variable X by

$$p_j = P\{X = j\} \stackrel{\text{def}}{=} (G(R))^{-1} \frac{R^j}{g!(j)}, \qquad j \geq 0.$$

We know that $\rho = \rho(R) \leq 1$ and of course, $\rho = EX$. Observe also that, unlike in the case of unbounded rates, the tail of X cannot be exponentially thin, since then R would not be the radius of convergence.

We consider the family of rate functions

$$g(j) = a\left(1 + \frac{\beta}{j} + O\big(j^{-(1+\delta)}\big)\right), \qquad a, \delta > 0.$$

Evidently $R = a$, and both $G(R)$ and $G'(R) < \infty$ if $\beta > 2$, in which case p_j is of order $j^{-\beta}$ exactly.

THEOREM 2.2. *Let g be as above and suppose $\beta > 3$.*

(a) *If $\rho < 1$ then Z_n^*/n converges to $1-\rho$ in probability and, furthermore, the size of the second largest cluster is $o(n^{1/2}\log^2 n)$ in probability.*

(b) *Suppose $\rho = 1$ and let $\lambda \in (0,\infty)$ be given. Then there exists $m = m(n,\lambda)$ such that*

$$\lim_{n\to\infty} nP\{X \geq m\} = \lambda$$

and $m(n,\lambda)$ is exactly of order $n^{(\beta-1)^{-1}}$. Let $\mathscr{C}_n$ denote the number of clusters of size at least m. Then $\mathscr{C}_n$ converges in distribution to a Poisson with parameter λ. Consequently, in probability, $Z_n^*/n^{(\beta-1)^{-1}}$ is bounded away from both 0 and ∞, hence Z_n^* is of order $n^{(\beta-1)^{-1}}$ exactly.

PROOF. (a) Pick $0 < \varepsilon < \min\{\rho^{-1}(1-\rho), 1\}$ and introduce $\nu_\pm = n - (1\mp\varepsilon)(n-1)\rho$, so that

$$\nu_-/n \to (1-\rho) - \varepsilon\rho, \qquad \nu_+/n \to (1-\rho) + \varepsilon\rho.$$

The proof of (a) is broken up into a number of steps which combine to show that $\nu_- \leq Z_n^* \leq \nu_+$ with probability tending to 1. First, we show that $P\{X_1 + \cdots + X_n\}$ vanishes at most like a power of n and then, using this fact, that $P\{Z_n^* \geq \nu_+\}$ tends to zero. Next, we show that for a properly chosen σ, $P\{[n^\sigma] \leq Z_n^* \leq \nu_-\}$ and $P\{Z_n^* \geq [n^\sigma]\}$ vanish as well. There is a twist. By allowing ε in the definition of ν_- to depend correctly on n we can tease out of the proof an estimate of the size of the second largest cluster.

STEP 1. Introducing $\widetilde{X} = \max_{1\le k\le n-1} X_k$, we have

$$\begin{aligned}P\left\{\sum_{j=1}^{n} X_j = n\right\} &\ge nP\left\{\sum_{j=1}^{n} X_j = n,\, X_n > \widetilde{X}\right\}\\ &\ge n \sum_{|(m/(n-1)\rho)-1|\le\varepsilon} P\left\{\sum_{j=1}^{n-1} X_j = m,\, \widetilde{X} < n-m,\, X_n = n-m\right\}\\ &\ge cnP\{X=n\} \sum_{|(m/(n-1)\rho)-1|\le\varepsilon} P\left\{\sum_{j=1}^{n-1} X_j = m,\, \widetilde{X} \le \nu_-\right\},\end{aligned}$$

since $P\{X = n-m\} = O(P\{X = n\})$ uniformly for the range of m in question. Consequently,

$$\begin{aligned}P\left\{\sum_{j=1}^{n} X_j = n\right\} &\ge cnP\{X=n\}(P\{X\le\nu_-\})^{n-1}\\ &\quad\times \sum_{|(m/(n-1)\rho)-1|\le\varepsilon} P\left\{\sum_{j=1}^{n-1} X_j^{(\nu_-)} = m\right\},\end{aligned}\tag{2.4}$$

where $X_1^{(\nu_-)}, \dots, X_{n-1}^{(\nu_-)}$ are independent copies of $X^{(\nu_-)}$, which is X conditioned on the event $\{X \le \nu_-\}$. Observe that, for a fixed $t \in \mathbb{R}$,

$$E\exp\bigl(itX^{(\nu_-)}/(n-1)\bigr) = E\exp(itX/(n-1))\bigl(1+O(n^{-(\beta-1)})\bigr)$$

so that

$$E\,\exp\left(\frac{it}{n-1}\sum_{j=1}^{n-1} X_j^{(\nu_-)}\right) = \bigl(1+O(n^{-(\beta-2)})\bigr)E\,\exp\left(\frac{it}{n-1}\sum_{j=1}^{n-1} X_j\right),$$

and the last expression converges to $e^{it\rho}$. Hence,

$$(n-1)^{-1}\sum_{j=1}^{n-1} X_j^{(\nu_-)} \to \rho$$

in probability, and the last probability in (2.4) tends to 1. In addition, we see that

$$(P\{X\le\nu_-\})^{n-1} = (1-P\{X>\nu_-\})^{n-1} \to 1,$$

because $P\{X > \nu_-\} = O(\nu_-^{1-\beta}) = o(n^{-1})$. Thus we have proved that

$$P\left\{\sum_{j=1}^{n} X_j = n\right\} \ge cnP\{X=n\} \ge c'n^{1-\beta}.\tag{2.5}$$

STEP 2. Clearly $h(\theta) := Ee^{-\theta X}$ exists and is continuously differentiable for all $\theta > 0$. So, using the Chernoff-type inequality [i.e., the Markov inequality

$P\{Y > y\} \le y^{-1}EY$ for $Y = \exp(-\theta\sum_{j=1}^{n-1} X_j)$ and $y = \exp(-\theta(n-\nu_+))\big]$, we have

$$P\left\{\sum_{j=1}^{n} X_j = n,\, X_n^* \ge \nu_+\right\} \le nP\left\{\sum_{j=1}^{n-1} X_j \le n-\nu_+\right\}$$
$$\le n\exp\big[(n-1)(\log h(\theta) + \theta\rho(1-\varepsilon))\big], \qquad \theta > 0.$$

Since $h'(0) = -EX = -\rho$, using the bound for $\theta > 0$ chosen sufficiently small, we see that the above probability is exponentially small. Consequently, by (2.5),

$$P\{Z_n^* \ge \nu_+\} = \frac{P\{\sum_{j=1}^{n} X_j = n,\, X_n^* \ge \nu_+\}}{P\{\sum_{j=1}^{n} X_j = n\}} \to 0, \tag{2.6}$$

as $n \to \infty$.

STEP 3. Pick $\sigma \in ((\beta+3)/2\beta, 1)$ and define $l = [n^\sigma]$. Also, let $a > 0$ be fixed, set $\varepsilon = an^{-1/2}\log n$ and let $\nu_- = n - (1+\varepsilon)(n-1)\rho$. Then

$$P\left\{\sum_{j=1}^{n} X_j = n,\, l \le X_n^* \le \nu_-\right\} \le c'nP\{X = l\}P\left\{\sum_{j=1}^{n-1} X_j \ge n-\nu_-\right\}.$$

Now,

$$n - \nu_- = (n-1)\rho + x, \qquad x := (n-1)\varepsilon\rho,$$

and we have

$$\frac{x}{\log x} \ge (2a\rho + o(1))\sqrt{n} \ge \sqrt{n},$$

provided $a > 1/2\rho$. Using a theorem due to A. V. Nagaev ([31], Chapter 8, Section 4) we obtain

$$P\left\{\sum_{j=1}^{n-1} X_j \ge n-\nu_-\right\} = O(nP\{X \ge x\}).$$

Therefore,

$$P\left\{\sum_{j=1}^{n} X_j = n,\, l \le X_n^* \le \nu_-\right\} = O\big(n^{-(\beta\sigma+(\beta-1)/2)+2}\big)$$
$$= o\left(P\left\{\sum_{j=1}^{n} X_j = n\right\}\right),$$

since $P\{\sum_{j=1}^{n} X_j = n\}$ is of order $n^{-(\beta-1)}$ at least, and $\sigma > (\beta+3)/2\beta$. Hence $Z_n^* \notin [l, \nu_-]$ with probability approaching 1.

STEP 4. Next,

$$P\left\{\sum_{j=1}^{n} X_j = n,\, X_n^* \le l\right\} \le P\left\{\sum_{j=1}^{n} X_j^{(l)} = n\right\}$$
$$\le P\left\{\sum_{j=1}^{n} X_j^{(l)} \ge n\right\};$$

here $X^{(l)}$ is X conditioned on $\{X \le l\}$. The last probability is bounded above by

$$\exp\Bigl(n\bigl(\log E\exp\bigl(\theta X^{(l)}\bigr) - \theta\bigr)\Bigr) \qquad \forall\, \theta > 0.$$

(The moment generating function exists since the range of $X^{(l)}$ is bounded from above.) For θ of order $n^{-\sigma'}$, where $\sigma' \in (\sigma, 1)$, we have

$$\begin{aligned}\exp(\theta X^{(l)}) &= 1 + \theta X^{(l)} + O\bigl((\theta X^{(l)})^2\bigr)\\ &= 1 + \bigl(1 + O\bigl(n^{-(\sigma'-\sigma)}\bigr)\bigr)\theta X^{(l)}.\end{aligned}$$

Hence

$$E\exp(\theta X^{(l)}) = 1 + \bigl(1 + O(n^{-(\sigma'-\sigma)})\bigr)\theta E X^{(l)},$$

where $EX^{(l)} \to \rho < 1$. Therefore the above probability is of order $\exp(-e^* n\theta) = \exp(-c^* n^{1-\sigma'})$, which is subexponentially small, $o(P\{\sum_{j=1}^{n} X_j = n\})$ that is. Therefore with probability approaching 1, $Z_N^* \ge l$.

Summarizing Steps 1 through 4, we conclude that $Z_n^* \in [\nu_-, \nu_+]$ with probability approaching 1.

STEP 5. Let Z_n^{**} denote the size of the second largest cluster. Fix $\delta > 0$. On the event $\{Z_n^* \ge \nu_-, Z_n^{**} \ge \delta n^{1/2}\log^2 n\}$ the total size of all clusters, other than the two largest ones, is at most $(n-2)\widetilde{\rho}$, where $\widetilde{\rho} = \rho - \delta n^{-1/2}(\log^2 n)/2$. As in Step 2,

$$P\left\{\sum_{j=1}^{n-2} X_j \le (n-2)\widetilde{\rho}\right\} = (O\exp(-n^{1/2})),$$

and we obtain

$$P\bigl\{Z_n^* \ge \nu_-, Z_n^{**} \ge \delta n^{1/2}\log^2 n\bigr\} \to 0.$$

Since $P\{Z_n^* < \nu_-\} \to 0$ as well, we see

$$Z_n^{**} = o_p\bigl(n^{1/2}\log^2 n\bigr).$$

(b) The proof of this part mimics, to a large extent, the proof of Theorem 2.1. We begin with introducing $m = m(n)$ such that $\lim_{n\to\infty} nP\{X \ge m\} = \lambda$, $\lambda > 0$ being fixed. Since $P\{X \ge \mu\}$ is asymptotic, within a constant factor, to $\mu^{1-\beta}$ as $\mu \to \infty$, $m = m(n)$ is asymptotic to $(n/\lambda)^{(\beta-1)^{-1}}$. Then we select $m_1 = n^a$, where $a \in \bigl((\beta-1)^{-1}, 1\bigr)$. Clearly $nP\{X \ge m_1\} = O(n^{1-a(\beta-1)}) \to 0$, so that $\mathscr{E}_n$, the number of clusters of size at least m, coincides, with probability approaching 1, with $\widetilde{\mathscr{C}}_n$, the number of clusters of size from the interval $[m, m_1 - 1]$. Since $m_1 = o(n)$, the computation of $E\binom{\bar{\mathscr{C}}_n}{k}$ proceeds exactly as that for $\widetilde{\mathscr{C}}_n$ in the proof of Theorem 2.1. So $\widetilde{\mathscr{C}}_n$ is asymptotically Poisson with parameter λ, and then so is $\mathscr{C}_n$. In particular,

$$P\{Z_n^* < m(n)\} = P\{\mathscr{C}_n = 0\} \to e^{-\lambda}, \qquad n \to \infty.$$

Since $m(n) \sim c(n/\lambda)^{(\beta-1)^{-1}}$, we can translate the last result into

$$\lim_{n\to\infty} P\left\{Z_n^* < xn^{(\beta-1)^{-1}}\right\} = \exp\left[-\left(\frac{c}{x}\right)^{\beta-1}\right].$$

Consequently, with high probability, Z_n^* is of order $n^{(\beta-1)^{-1}}$ exactly. $\square$

3. Vanishing rates. Consider, finally, the case $g(k) \downarrow 0$. Concerning g we assume:

1. $g(x)$ is defined for $x \in [0, \infty)$, $g(x) \equiv 1$ for $x \in [0, 1]$, $g(x) \downarrow 0$ as $x \uparrow \infty$, $g(x)$ twice continuously differentiable for $x > 1$.
2. $r(x) = -xg'(x)/g(x)$ is bounded and decreasing on $[1, \infty)$.
3. $s(x) = xr'(x)/r(x)$ tends to 0 as $x \to \infty$.
4. $\lim_{x\to\infty}[r(x)\log^{1/2} x - \log\log x] = \infty$.

REMARKS. Before getting down to brass tacks, let's make a few remarks about these four conditions.

(a) Let $f(x) = 1/g(x)$. Then $r(x)$ and $s(x)$ are the derivatives of $\log f(x)$ and $\log r(x)$ with respect to $\log x$. Evidently,

$$f(x) = \exp\left\{\int_1^x \frac{r(t)}{t}\,dt\right\}$$

and so condition 2 implies that $f(x) = O(x^c)$, for some $c > 0$. Thus, $g(x) \geq c'x^{-c}$. It is just as easy to see that

$$g(x) = x^{-r(\infty)+o(1)},$$

so that, roughly speaking, if $r(\infty) > 0$ then g behaves like a power function.

(b) It follows from 4 that $\log f(x) \gg \log^{1/2} x$, whence $r(x)\log f(x) \to \infty$. Consequently, we see that $g(x) = o(r(x))$.

(c) We will need the *average* logarithmic rate of $f(x)$, defined by

$$\bar{r}(x) = \frac{1}{x}\int_1^x r(y)\,dy.$$

For $x > 1$, we have $\bar{r}(x) = r(x) + o(r(x))$. To see it, note that

$$\begin{aligned}\bar{r}(x) &= \frac{1}{x}\left[xr(x) - r(1) - \int_1^x yr'(y)\,dy\right]\\ &= r(x) - \frac{r(1)}{x} - \bar{r}(x)\frac{\int_1^x r(y)s(y)\,dy}{\int_1^x r(y)\,dy}\\ &= r(x)\left(1 - \frac{r(1)}{xr(x)}\right) + o(\bar{r}(x)),\end{aligned}$$

since by 3 $\lim_{x\to\infty} s(x) = 0$ and, by 4, $\int_0^\infty r(y)\,dy = \infty$. Therefore, again using 4,

$$r(x) = \bar{r}(x) + o(\bar{r}(x)) \Rightarrow \bar{r}(x) = r(x) + o(r(x)). \tag{3.1}$$

As for condition 4, we have believed that it would have been sufficient to require only that $\lim_{x\to\infty} r(x)\log^{1/2} x = \infty$. However, thanks to the thoughtful questions made by a referee, we have realized that we need 4, which is a bit stronger.

Now, the centerpiece of Theorem 4 as stated in the introduction is the following.

THEOREM 3.1. *Under the assumptions above*, $(n - Z_n^*)/ng(n)$ *converges to* 1 *in probability*.

The proof of Theorem 3.1 and the rest of Theorem 4 consists of a step-by-step identification of the parts of $\{0, 1, 2, \dots, n\}$ that, taken together, form the set of *unlikely* values of Z_n^*. Needless to say, this set is going to be

$$\{j: |(n-j)/ng(n) - 1| > \varepsilon\}, \qquad \varepsilon > 0.$$

First we show that Z_n^* is extremely unlikely to be small compared to n, and the magnitude of Z_n^* is measured intrinsically, by the value of $g(Z_n^*)$, the current disintegration rate of the largest cluster.

LEMMA 3.1. *If* $\beta_n \in (0, 1)$ *is such that* $(1 - \beta_n)\log f(n) \to \infty$, *then*

$$P\{g(Z_n^*) \ge g(n)^{\beta_n}\} = \exp(-(1-\beta_n)n\log f(n) + O(n)).$$

PROOF. Let k be the largest integer such that $g(k) \ge g(n)^{\beta_n}$, that is $f(k) \le f(n)^{\beta_n}$. Introduce the configuration

$$\eta_1 = (k, k, \dots, k, r, 0, 0, \dots, 0),$$

where k is repeated l times, and let $\eta_2 = (n, 0, 0, \dots, 0)$. By Lemma 1.3,

$$P\{Z = \eta_1\} \ge P\{Z = \eta\} \quad \text{for all } \eta \in \{\eta: Z^* \le k\}.$$

Thus,

$$\begin{aligned} P\{g(Z_n^*) \geq g(n)^{\beta_n}\} &= P\{Z_n^* \leq k\} \\ &\leq P\{Z = \eta_1\}|\{Z^* \leq k\}| \\ &\leq \frac{P\{Z = \eta_1\}}{P\{Z = \eta_2\}}|\Omega_n| \\ &\leq \frac{g!(n)}{(g!(k))^{l+1}} 4^n, \end{aligned}$$

since, by Lemma 1.4, $|\Omega_n| = \binom{2n-1}{n}$. For an integer m,

$$\begin{aligned} g!(m) &= \exp\left[-\sum_{j=1}^{m} \log f(j)\right] \\ &= \exp\left[-m \log f(m) + \int_0^m r(y)\,dy + O(\log f(m))\right] \\ &= \exp[-m \log f(m) + O(m)], \end{aligned}$$

since by Remark 1, $\log f(m) = O(\log m)$. Therefore,

$$\begin{aligned} \frac{g!(n)}{(g!(k))^{l+1}} &= \frac{\exp[-n \log f(n) + O(n)]}{(\exp[-k \log f(k) + O(k)])^{l+1}} \\ &= \exp[-n \log f(n) + n \log f(k) + O(n)] \\ &= \exp[-(1 - \beta_n) n \log f(n) + O(n)]. \end{aligned}$$

□

Our next, much bigger, step is to show that it is also unlikely for Z_n^* to have any value between $\max\{j\colon g(j) \geq g(n)^{\beta_n}\}$ and $n - cng(n)$, if c is sufficiently large. For this purpose we turn again to the conditioning device. Since now the radius of convergence of $\sum_j x^j/g!(j)$ is zero, we are forced to consider $X \leq n$ and $x = x_n \to 0$, the latter being necessary to avoid a hopelessly small value of $P\{\sum_i X_i = n\}$. By analogy with the proofs in Section 2, it would seem natural to choose x_n so as to make EX close to 1, the particle density. Since $P\{X = n\}/P\{X = 0\} = x^n/g!(n)$, the requirement $EX \sim 1$ leads to x_n being of order $g(n)$. However, with this choice, the resulting variance of X is unbounded which, essentially, means that we will have to bound $P\{\sum_i X_i = n\}$ from below without being able to apply a local limit theorem. If that is the case, we may just as well forget about $EX \sim 1$ and pick an x_n of order $g(n)$ which does allow us to find a working bound for that probability in a reasonably direct way.

To that end, we define x_n as the root of the equation

$$\frac{x^n}{g!(n)} = g(n)^3. \tag{3.2}$$

Thus

$$x_n = (g(n))^{3/n} (g!(n))^{1/n}.$$

Since $-\log g(n) = \log f(n) = O(\log n)$, the first factor is $1 + O(n^{-1}\log n)$. Besides, from the proof of Lemma 3.1, we have

$$\begin{aligned}(g!(n))^{1/n} &= \exp\biggl(-\log f(n) + \frac{1}{n}\int_0^n r(y)\,dy + O(n^{-1}\log n)\biggr)\\ &= (1 + O(n^{-1}\log n))g(n)e^{\bar{r}(n)}.\end{aligned}$$

Thus

$$x_n = (1 + O(n^{-1}\log n))g(n)e^{\bar{r}(n)}. \tag{3.3}$$

So x_n is indeed of order $g(n)$.

To proceed, we need to have a close look at the sequence $\{q_t\} = \{x^t/g!(t)\}$ which when normalized will give the probability distribution of X. Calculating as before, but with more precision, we have for every $1 \le t \le n$,

$$\begin{aligned}\frac{(g!(n))^{t/n}}{g!(t)} &= \exp\biggl(-\frac{t}{n}\sum_{j=0}^{n}\log f(j) + \sum_{j=0}^{t}\log f(j)\biggr)\\ &= \exp\biggl(-t\log f(n) + \frac{t}{n}\int_0^n r(y)\,dy + t\log f(t)\\ &\qquad\qquad - \int_0^t r(y)\,dy + O(\log n)\biggr)\\ &= \exp\Bigl(-t\bigl[\log\bigl(e^{-\bar{r}(n)}f(n)\bigr) - \log\bigl(e^{-\bar{r}(t)}f(t)\bigr)\bigr] + O(\log n)\Bigr).\end{aligned}$$

Now an easy computation shows that, for $x \in [1, n]$,

$$x\frac{(d/dx)\bigl(e^{-\bar{r}(x)}f(x)\bigr)}{e^{-\bar{r}(x)}f(x)} = \bar{r}(x) \ge \frac{x-1}{x}r(x) \ge \frac{x-1}{x}r(n),$$

since $r(x)$ decreases. Therefore,

$$\begin{aligned}\frac{(g!(n))^{t/n}}{g!(t)} &\le \exp\bigl(-tr(n)\bigl(\log(ne^{1/n}) - \log(te^{1/t})\bigr) + O(\log n)\bigr)\\ &= \exp\bigl(-tr(n)(\log n - \log t) + O(\log n)\bigr),\end{aligned}$$

so that

$$q_t = \frac{x^t}{g!(t)} \le n^c\left(\frac{t}{n}\right)^{tr(n)}. \tag{3.4}$$

Another inequality we will need is

$$r(x_2)\log\frac{x_2}{x_1} \le \log\frac{f(x_2)}{f(x_1)} \le r(x_1)\log\frac{x_2}{x_1}, \qquad 1 \le x_1 \le x_2, \tag{3.5}$$

which follows from the definition and monotonicity of $r(x)$ on $[1, \infty)$.

Finally, to avoid cumbersome expressions in our inequalities we will use below the shorthand

$$A_n \overset{p}{=} B_n, \quad A_n \overset{p}{\leq} B_n, \quad A_n \overset{p}{\geq} B_n$$

for, respectively,

$$n^{d_1} B_n \leq A_n \leq n^{d_2} B_n, \quad n^d A_n \leq B_n \quad \text{and} \quad A_n \geq n^d B_n,$$

where d, d_i are suitable constants of any sign. (p stands for "polynomially.")

LEMMA 3.2. *Let $M_n = \sum_{t=0}^{n} q_t$, so that $p_t = P\{X = t\} = q_t/M_n$. Then*

$$M_n = 1 + x_n + O\big(g(n)^2\big)$$

and

$$M_n^n \overset{p}{=} \exp\Big[ne^{\bar{r}(n)} g(n)(1 + O(g(n)))\Big].$$

PROOF. Let $t_n = [\varepsilon_n n]$ and $\tau_n = [(1-\varepsilon_n)n]$, and $\varepsilon_n = c_1 \log n/(nr(n))$, with $c_1 > 0$ being fixed. Using the unimodality of q_t and the bound (3.4), together with $g(n) \overset{p}{\geq} 1$, and $r(n) \gg \log^{-1/2} n$, we see that

$$\sum_{t_n \leq t \leq \tau_n} q_t = O(n \max(q_{t_n}, q_{\tau_n})) = O(nq_{\tau_n}) = O(n^{c-c_1+1}) = O(g^2(n)),$$

where the last estimate holds if c_1 is chosen sufficiently large. [The reason we are content with the obviously weak bound $O(g^2(n))$ will become clear shortly.] It remains to consider the tail values of t.

Let $t \leq t_n$. The ratio $q_t/q_{t-1} = x_n/g(t)$ increases for $t \geq 1$, and by (3.3), (3.5) and (3.1),

$$\begin{aligned}\frac{x_n}{g(t_n)} &= \exp\left(\bar{r}(n) - \log\frac{f(n)}{f(t_n)} + O(n^{-1}\log n)\right)\\ &\leq \exp\left(\bar{r}(n) - (1/2 + o(1))\bar{r}(n)\log\frac{n}{t_n} + O(n^{-1}\log n)\right)\\ &\leq e^{-r(n)}.\end{aligned} \tag{3.6}$$

the last inequality being rather crude, but sufficient. Indeed, since $g(n) = O(r(n))$ [even $o(r(n))$],

$$\sum_{t=3}^{t_n} q_t \leq \frac{q_3}{1 - e^{-r(n)}} = O\left(\frac{g(n)^3}{r(n)}\right) = O\big(g(n)^2\big)$$

and

$$\sum_{t \leq t_n} q_t = \sum_{t=0}^{2} q_t + O(g(n)^2) = 1 + x_n + O(g(n)^2). \tag{3.7}$$

Let's turn to $t \ge \tau_n$. The ratio $x_n/g(t)$ attains its *minimum* at τ_n, and analogously to (3.6),

$$\frac{x_n}{g(\tau_n)} \ge \exp\left(\bar{r}(n) - 2\bar{r}(n)\log\frac{n}{\tau_n}\right) \ge \exp(r(n)/2). \tag{3.8}$$

According to (3.2), $q_n = g(n)^3$. So, as in the case of small t's,

$$\sum_{t=\tau_n}^{n} q_t = O\left(\frac{q_n}{r(n)}\right) = O\left(\frac{g(n)^3}{r(n)}\right) = O(g(n)^2).$$

Combining (3.7), the last relation, and (3.3), completes the proof for the estimate of M_n. The estimate for M_n^n follows from

$$M_n^n = (1 + x_n + O(g(n)^2))^n = \exp\bigl[n\log(1 + x_n + O(g(n)^2))\bigr],$$

(3.3) and the fact that $\log(1+z) = z + O(z^2)$ as $z \to 0$. $\square$

Now we can estimate the probability of the conditioning event $\{\sum_i X_i = n\}$ from below.

LEMMA 3.3. *We have*

$$P\left\{\sum_i X_i = n\right\} \overset{p}{\ge} \exp\Bigl[-(e^{r(n)} - 1)ng(n)(1 + o(1))\Bigr]. \tag{3.9}$$

PROOF. Given k, $0 \le k < n$, let

$$n_k = (n-k, 1, 1, \dots, 1, 0, 0, \dots, 0).$$

where the 1's are repeated k times and the 0's $n-k-1$ times. Let

$$B_k = \{\eta \in \Omega_n \colon \eta = \sigma_n \eta_k \text{ for some } \sigma_n \in S(n)\}.$$

Evidently, for $k < n-1$, there are exactly $n\binom{n-1}{k}$ configurations in B_k. Therefore,

$$\begin{aligned} P\left\{\sum_i X_i = n\right\} &\ge \max_{0\le k<n-1} P\{X_1 = \eta_k(1), \dots, X_n = \eta_k(n)\} \cdot |B_k| \\ &= \max_{0\le k<n-1} n\binom{n-1}{k}\frac{q_1^k q_{n-k}}{M_n^n} \\ &\overset{p}{\ge} \max_{0\le k<n-1} \binom{n}{k}\frac{q_1^k q_{n-k}}{M_n^n} \\ &= \frac{x_n^n}{M_n^n}\max_{0\le k<n-1}\binom{n}{k}\frac{1}{g!(n-k)}. \end{aligned} \tag{3.10}$$

Here, by (3.3) and Lemma 3.2,

$$\begin{aligned} \frac{x_n^n}{M_n^n} &\overset{p}{=} \exp\bigl(n\log(g(n)e^{\bar{r}(n)}) - ne^{\bar{r}(n)}g(n) + O(ng(n)^2)\bigr) \\ &= \exp\bigl(n\log g(n) - n\bar{r}(n) - ne^{\bar{r}(n)}g(n) + O(ng(n)^2)\bigr). \end{aligned} \tag{3.11}$$

Next, using the Stirling formula $\nu! \sim (2\pi\nu)^{1/2}(\nu/e)^\nu$, $(\nu \to \infty)$, and its corollary $\nu! \le c(\nu+1)^{1/2}(\nu/e)^\nu$, $(\forall\, \nu \ge 0)$, we obtain, again neglecting polynomial factors, that

$$\frac{\binom{n}{k}}{g!(n-k)} \overset{p}{\geq} \exp\biggl(k \log \frac{n}{k} + (n-k) \log \frac{n}{n-k} - (n-k) \log f(n-k) - \int_0^{n-k} r(y)\, dy\biggr).$$

The exponent attains its maximum at $k = \kappa$ such that

$$\log \frac{n}{\kappa} - \log \frac{n}{n-\kappa} - \log f(n-\kappa) = 0, \tag{3.12}$$

and it is easy to show that κ is asymptotic to $n/f(n) = ng(n)$. More precisely

$$\kappa = ng(n)(1 - (r(n) - 1)g(n) + O(g(n)^2)). \tag{3.13}$$

By (3.12), the maximum value equals

$$n \log \frac{n}{\kappa} - \int_0^{n-\kappa} r(y)\, dy.$$

Here, by (3.13),

$$n \log \frac{n}{\kappa} = -n \log g(n) - n(r(n) - 1)g(n) + O(ng(n)^2)$$

and

$$\begin{aligned} -\int_0^{n-\kappa} r(y)\, dy &= -\int_0^n r(y)\, dy + \int_{n-\kappa}^n r(y)\, dy \\ &= -n\bar{r}(n) + nr(n)g(n) + O(nr(n)g(n)^2). \end{aligned}$$

The estimate for the integral from $n - \kappa$ to n comes from condition 3 and it implies that, uniformly in $y \in [n - \kappa, n]$.

$$r(y) = r(n)(1 + O(\kappa/n)) = r(n)(1 + O(g(n)).$$

So the maximum is

$$-n \log g(n) + ng(n) - n\bar{r}(n) + O(ng(n)^2) \tag{3.14}$$

at least. Combining this with (3.10) and (3.11) completes the proof. □

We are in a position now to prove a preliminary upper bound for the probability that Z_n^* assumes a "midrange" value.

LEMMA 3.4. (a) *If*

$$\alpha_n = \min\left\{\frac{1}{2}; \left(\frac{r(n)}{\log f(n)}\right)^{1/2}\right\},$$

then

$$\lim_{n\to\infty} \alpha_n \log f(n) = \infty, \qquad \limsup_{n\to\infty} \frac{\alpha_n}{r(n)} = 0. \tag{3.15}$$

(b) *Set* $\beta_n = 1 - \alpha_n$ *and* $k = \max\{j\colon g(j) \geq g(n)^{\beta_n}\}$. *If* $\delta_n = o(f(n))$ *then for every* $\rho \in (0.1)$,

$$P\{k < Z_n^* \leq n - \delta_n g(n)n\}$$
$$\overset{p}{\leq} \exp\Big[(e^{r(n)} - 1 - \delta_n r(n))ng(n) + O(nr(n)(\delta_n g(n))^2)\Big] + \exp(-n^\rho).$$

The following corollary is a direct consequence.

COROLLARY 3.5. *If* $\lim_{n\to\infty} nr(n)g(n)/(\log n) = \infty$ *and* $\delta > \lim_{n\to\infty} \big(e^{r(n)} - 1\big)/r(n)$, *then*

$$\lim_{n\to\infty} P\big\{Z_n^* > n - \delta ng(n)\big\} = 1.$$

PROOF OF LEMMA 3.4. Part (a) follows from $\lim_{n\to\infty} r(n)\log f(n) = \infty$, implied by condition 4. Turning to (b), the probability is bounded above by

$$\frac{P\{k < \max X_i \leq n - \delta_n ng(n)\}}{P\{\sum_i X_i = n\}} \leq \frac{nP\{k < X_1 \leq n - \delta_n ng(n)\}}{P\{\sum_i X_i = n\}}.$$

The numerator of the last fraction is of order

$$n \max\big\{P\{X_1 = k\}, P\{X_1 = [n - \delta_n ng(n)]\}\big\}.$$

By (3.4), the second probability is of order

$$n^c \exp\left(-t\bar{r}(n)\log\frac{n}{t}\right)\bigg|_{t=n-\delta_n ng(n)} = n^c \exp\big(-\delta_n ng(n)\bar{r}(n)(1 + O(\delta_n g(n)))\big).$$

Therefore, using Lemma 3.3,

$$\frac{nP\{X_1 = [n - \delta_n ng(n)]\}}{P\{\sum_i X_i = n\}}$$
$$\overset{p}{\leq} \exp\Big(\big(e^{r(n)} - 1 - \delta_n \bar{r}(n)\big)ng(n) + O(nr(n)(\delta_n g(n))^2)\Big).$$

Further, since $f(k+1) \geq f(n)^{\beta_n}$, we have [see (3.5) and the second relation in (3.15)]

$$\begin{aligned} k+1 &\geq n\exp\left(-\alpha_n(1+o(1))\frac{\log f(n)}{r(n)}\right) \\ &\geq n\exp(-o(\log f(n))) \\ &\to \infty. \end{aligned}$$

So, combining this with $f(k) \le f(n)^{\beta_n}$, (3.3), (3.5) and the first relation in (3.15), we have

$$
\begin{aligned}
P\{X_1 = k\} &\stackrel{p}{=} \frac{x_n^k}{g!(k)} \\
&= \exp(-k(\log f(n) - \log f(k)) + O(k)) \\
&\le \exp(-k\alpha_n \log f(n) + O(k)) \\
&\le \exp\Bigl(-0.5 n r(n)\lambda(n) e^{-\lambda(n)}\Bigr),
\end{aligned}
$$

where

$$\lambda(n) := \frac{\alpha_n \log f(n)}{r(n)}.$$

Here, by (3.15), both $\lambda(n) \to \infty$ and $\lambda(n) r(n) \to \infty$ and

$$\lambda(n) \le \frac{\log f(n)}{\sqrt{r(n)\log f(n)}} = o(\log n).$$

So, for every $a < 1$ and sufficiently large n,

$$n r(n)\lambda(n) e^{-\lambda(n)} \ge n^a$$

and

$$
\begin{aligned}
\frac{n r(n)\lambda(n) e^{-\lambda(n)}}{n r(n) g(n)} &\ge \frac{e^{-\lambda(n)}}{g(n)} \\
&= \exp\biggl[\log f(n)\biggl(1 - \frac{\alpha_n}{r(n)}\biggr)\biggr] \ge \log^{1/2} n.
\end{aligned}
$$

Thus, given $\rho < 1$, there exists a number $n(\rho)$ such that

$$\frac{n P\{X_1 = k\}}{P\{\sum_{t=1}^n X_t = n\}} \le e^{-n^\rho}.$$

for all $n \ge n(\rho)$. □

We use the last estimate to show next that, for $r(\infty) \ge 1$, the largest cluster contains, with high probability, all but finitely many particles,

LEMMA 3.6. *If $r(\infty) \ge 1$, then the sequence $\{n - Z_n^*\}$ is tight; that is, $n - Z_n^*$ is bounded in probability.*

PROOF. Integrating the differential inequality

$$\frac{x f'(x)}{f(x)} = r(x) \ge 1, \qquad f(1) = 1,$$

we get $f(x) \geq x$ and $g(x) \leq x^{-1}$, $(x \geq 1)$. Therefore, by Lemma 3.2, $M_n = 1 + O(n^{-1})$ and $M_n^n = O(1)$. Hence

$$\begin{aligned} P\{X_1 + \cdots + X_n = n\} &\geq nP\{X_1 = n, X_j = 0, \forall j \geq 2\} \\ &= n\frac{x_n^n/g!(n)}{M_n^n} \qquad (3.16) \\ &\geq cng(n)^3, \qquad c \in (0, \infty). \end{aligned}$$

To proceed, notice that $\lim_{x\to\infty} r(x) \log f(x) = \infty$. So we can use Lemma 3.4 with $\delta_n = c_0 \log n/(ng(n))$, selecting c_0 large enough to outpower the polynomial factor implicit in the statement and also Lemma 3.1 to obtain

$$\begin{aligned} P\{Z_n^* \leq n - c_0 \log n\} &= P\{Z_n^* \leq k\} + P\{k < Z_n^* \leq n - \delta_n g(n)n\} \\ &\leq o(1) + \exp(-(c_0/2) \log n) \\ &= o(1). \end{aligned}$$

Furthermore, for $t > n - c_0 \log n$, the ratio q_t/q_{t-1} is at least $\exp(r(n)/2) \geq e^{1/2}$; see (3.8). Therefore, given an integer N,

$$\begin{aligned} P\{n - Z_n^* \geq N\} &= P\{Z_n^* \leq n - N\} \\ &= o(1) + P\{n - c_0 \log n < Z_n^* \leq n - N\} \\ &\qquad [\text{using } (3.16)] \\ &\leq o(1) + \frac{n}{cng(n)^3} P\{n - c_0 \log n < X_1 \leq n - N\} \\ &\leq o(1) + \frac{1}{cg(n)^3} \cdot q_n \frac{e^{-N/2}}{1 - e^{-1/2}} \\ &= o(1) + O(e^{-N/2}), \end{aligned}$$

which completes the proof. □

LEMMA 3.7. *Suppose* $r(\infty) > 1$. *Then* $P\{Z_n^* = n\} \to 1$.

PROOF. By Lemma 3.6, it suffices to show that for every fixed $k \geq 1$,

$$\lim_{n\to\infty} \frac{P\{Z_n^* = n - k\}}{P\{Z_n^* = n\}} = 0.$$

The denominator equals $\mathcal{Z}^{-1}(n/g!(n))$, and the numerator is bounded above by

$$P\{\mathbf{Z} = (n - k, k, 0, \ldots, 0)\}n|\Omega_{n-1}^k| \leq \mathcal{Z}^{-1}\big(g!(n - k)g!(k)\big)^{-1}(2n)^{k+1},$$

from Lemmas (1.3) and (1.4). Here, since k is fixed,

$$\begin{aligned} \frac{g!(n)}{g!(n - k)} &\leq \exp\big((n - k) \log f(n - k) - n \log f(n) + O(1)\big), \\ &\leq \exp\big(-k \log f(n) + O(1)\big). \end{aligned}$$

According to the proof of Lemma 3.6, $f(n) \geq n^{r(\infty)}$, so the ratio of the probabilities in question is $O(n^k/n^{kr(\infty)}) = o(1)$. $\square$

LEMMA 3.8. *Suppose* $r(\infty) = 1$. *Then* $n - Z_n^*$ *is asymptotically Poisson with parameter* $ng(n) \leq 1$; *that is, for every fixed* $k \geq 0$,

$$P\{Z_n^* = n - k\} = (1 + o(1)) \exp(-ng(n)) \frac{(ng(n))^k}{k!}.$$

PROOF. Introduce C_n the total number of clusters, that is, the total number of occupied sites. Clearly $C_n \leq n - Z_n^* + 1$ and the equality means that all the clusters, besides the largest one, are of size 1. For a fixed k,

$$\begin{aligned} P\{Z_n^* = n - k, C_n = k + 1\} &= \mathscr{Z}^{-1} n \binom{n-1}{k} \frac{x_n^n}{g!(n-k)} \\ &= (1 + o(1)) \frac{n x_n^n}{\mathscr{Z} g!(n)}, \frac{(ng(n))^k}{k!}, \end{aligned}$$

while, for $l \leq k$,

$$\begin{aligned} P\{Z_n^* = n - k, C_n = l\} &= O\left(n \binom{n-1}{l-1} \frac{x_n^n}{\mathscr{Z} g!(n-k)}\right) \\ &= O\Big(n^{-(k+1-l)} P\{Z_n^* = n - k, C_n = k + 1\}\Big). \end{aligned}$$

Therefore, for every fixed k,

$$P\{Z_n^* = n - k\} = (1 + o(1)) \frac{n x_n^n}{\mathscr{Z} g!(n)} \frac{(ng(n))^k}{k!}$$

and, since $n - Z_n^*$ is tight,

$$1 = (1 + o(1)) \frac{n x_n^n}{\mathscr{Z} g!(n)} \sum_{k \geq 0} \frac{(ng(n))^k}{k!},$$

that is,

$$P\{Z_n^* = n - k\} = (1 + o(1)) \exp(-ng(n)) \frac{(ng(n))^k}{k!}. \qquad \square$$

So we have proved parts (a) and (b) of Theorem 4. The main assertion of the theorem follows from a combination of the next (and last) two lemmas.

LEMMA 3.9. *Let* $r(\infty) < 1$ *and suppose* $\lim_{n\to\infty}[r(n)\log^{1/2} n - \log\log n] = \infty$. *Introduce the quantities*

$$\gamma_n^* = \frac{1}{2} \min\left\{1, \left(\frac{r(n)}{1 - r(n)}\right)^2\right\}, \qquad \gamma_n = \gamma_n^* - r(n) \log^{-1/2} n$$

and let $\varepsilon_n = (ng(n))^{-\gamma_n}$. *Then* (a) $\gamma_n = (1+o(1))\gamma_n^*$ *and* $\lim_{n\to\infty}\varepsilon_n = 0$. *Furthermore, let* $k_n = (1-\varepsilon_n)ng(n)$. *Then* (b)

$$\lim_{n\to\infty} P\{Z_n^* > n - k_n\} = 0,$$

or, more precisely, for every $0 < c < 1/2$ *there exists* $n(c)$ *such that if* $n > n(c)$ *then*

$$P\{Z_n^* > n - k_n\} \le \exp\bigl(-c(ng(n))^{1-2\gamma_n}\bigr).$$

PROOF. Part (a) follows from the observation that $g(n) = n^{-r(\infty)+o(1)}$ and the conditions on $r(\infty)$ and $r(n)$.

(b) In a nutshell, we partition the event $\{Z_n^* > n - k_n\}$ into the subevents $\{Z_n = j, C_n = l\}$ and determine a pair (j, l) that maximizes the contribution to the overall probability bound.

To start, notice that for all j,

$$\bigl|\{Z_n^* = j, C_n = l\}\bigr| \le l\binom{n}{l}\binom{n-j-1}{l-2}, \qquad 1 \le l \le n-j+1,$$

with equality if $j > n/2$. Also notice that, from Lemma 1.3, the allocation $(j, n-j-l+2, \overbrace{1,\dots,1}^{l-2}, 0,\dots,0)$ is a *maximal* element of $\{Z_n^* = j, C_n = l\}$ for large j. So, denoting $\mu = \mathscr{Z}\nu$, $(\mu(\Omega_n) = \mathscr{Z})$, we have

$$\begin{aligned}\mu\{Z_n^* > n-k_n\} \le \sum_{\substack{j>n-k_n\\ 1\le l\le n-j+1}} l\binom{n}{l}\binom{n-j-1}{l-2}\\ \times\{g!(j)g!(n-j-l+2)\}^{-1}.\end{aligned}\tag{3.17}$$

We'll use Stirling's formula for the usual factorials and the approximation

$$\frac{1}{g!(m)} = \exp(h(m) + O(\log m)),$$

where

$$h(m) := m\log f(m) - \int_0^m r(z)\,dz.$$

Now, switching variables to $k = n - j + 1$, and *enlarging* the range of l a bit, we transform the estimate in (3.17) to

$$\mu\{Z_n^* > n-k_n\} \overset{p}{\le} \sum_{0\le l\le k\le k_n} \exp\bigl(H_n(k,l)\bigr),\tag{3.18}$$

where

$$\begin{aligned}H_n(x,y) := y\log\frac{n}{y} + (n-y)\log\frac{n}{n-y} + y\log\frac{x}{y} + (x-y)\log\frac{x}{x-y}\\ + h(n-x) + h(x-y).\end{aligned}$$

Here we use the convention $0\log c/0 := 0$.

Our task is to characterize a point (x_n, y_n) such that

$$\max\{H_n(x, y)\colon 0 \le y \le x \le k_n\} = H_n(x_n, y_n).$$

From (3.18), for $y < x$,

$$\begin{aligned}\frac{\partial H_n(x, y)}{\partial x} &= \log \frac{xf(x-y)}{(x-y)f(n-x)},\\ \frac{\partial H_n(x, y)}{\partial y} &= \log \frac{(x-y)(n-y)}{y^2 f(x-y)}.\end{aligned} \tag{3.19}$$

These relations owe their relative simplicity to the product type form of the original summands in (3.17). We notice that, for $y < x$ in question, the sum of the derivatives is

$$\begin{aligned}\log \frac{x(n-y)}{y^2 f(n-x)} &\ge O(k_n/n) + \log \frac{n}{k_n f(n)}\\ &\ge O(g(n)) + \log(1-\varepsilon_n)^{-1}\\ &> 0.\end{aligned}$$

since $\varepsilon_n \gg g(n)$, which follows from a stronger inequality $\varepsilon_n^2 \gg g(n)$ proved below. Therefore, $H_n(x, y)$ is strictly increasing with x along every straight line $x - y \equiv \text{const}$, and

$$\max\{H_n(x, y)\colon 0 \le y \le x \le k_n\} = \max\{H_n(k_n, y)\colon y \le k_n\},$$

so, in particular, $x_n = k_n$. Introduce

$$\sigma_n = \max\left\{0, \frac{1-2r(n)}{(1-r(n))^2}\right\} + r(n)\log^{-1/2} n.$$

So $\sigma_n > 0$, and it follows from

$$1 - \sigma_n = \min\left\{1, \left(\frac{r(n)}{1-r(n)}\right)^2\right\} + r(n)\log^{-1/2} n,$$

and $r^2(n)\log n \to \infty$ that $(1-\sigma_n)\log n \to \infty$. Therefore, $\sigma_n < 1$ for n large enough. Then, for $y \le k_n - (ng(n))^{\sigma_n}$, using $g(n) \le n^{-r(n)}$ we have

$$\begin{aligned}\frac{\partial H_n(k_n, y)}{\partial y} &\ge O(1) + \log\left(\frac{n(ng(n))^{\sigma_n(1-r(n))}}{(ng(n))^2}\right)\\ &= O(1) + \log\left(\frac{n}{(ng(n))^{2-\sigma_n(1-r(n))}}\right)\\ &\ge O(1) + (2r(n) - 1 + \sigma_n(1-r(n))^2)\log n\\ &\to \infty.\end{aligned}$$

by definition of σ_n and $r(n)\log^{1/2} n \to \infty$. So, whatever y_n is, it must satisfy

$$k_n - (ng(n))^{\sigma_n} \le y_n \le k_n.$$

It remains to sharply estimate $H_n(k_n, y_n)$. The dominating summands are

$$y_n \log \frac{n}{y_n} = k_n \log \frac{n}{k_n} + O\big((ng(n))^{\sigma_n} \log n\big),$$

$$(n - y_n) \log \frac{n}{n - y_n} = y_n + O\big(y_n^2/n\big) = k_n + O\big((ng(n))^{\sigma_n}\big),$$

$$h(n - k_n) = -(n - k_n) \log g(n) - n\bar{r}(n) + O\big((ng(n))^{\sigma_n}\big),$$

by definition of σ_n, and condition (3). The sum of the three remaining terms is $O((ng(n))^{\sigma_n} \log n)$. Thus,

$$\begin{aligned} H_n(k_n, y_n) = {} & k_n \log \frac{n}{k_n} + k_n - (n - k_n) \log g(n) - n\bar{r}(n) \\ & + O\big((ng(n))^{\sigma_n} \log n\big). \end{aligned} \tag{3.20}$$

Now, while proving Lemma 3.3, we actually demonstrated [see (3.14)] that

$$\mu\{\Omega_n\} \geq \exp(-n \log g(n) + ng(n) - n\bar{r}(n) + O(ng(n)^2)). \tag{3.21}$$

These last two relations, combined with (3.18) and some simple algebra, lead to

$$\begin{aligned} P\{Z_n^* > n - k_n\} &= \frac{\mu\{Z_n^* > n - k_n\}}{\mu\{\Omega_n\}} \\ &\overset{p}{\leq} \exp(-\mathscr{H}(\varepsilon_n) ng(n) + O((ng(n))^{\sigma_n} \log n + ng(n)^2)), \end{aligned} \tag{3.22}$$

where

$$\mathscr{H}(z) := z - (1 - z) \log(1 - z)^{-1}, \qquad z \in (0, 1).$$

Notice that $\mathscr{H}(z) \sim z^2/2$, $(z \to 0)$. Let us prove that $-\mathscr{H}(\varepsilon_n) ng(n)$ dominates the remainder terms. Recalling the definitions of ε_n, σ_n and noting $g(n) = n^{-r(\infty)+o(1)}$, we see that

$$\begin{aligned} \frac{ng(n)\varepsilon_n^2}{(ng(n))^{\sigma_n} \log n} &= \frac{(ng(n))^{1 - 2\gamma_n - \sigma_n}}{\log n} \\ &= \frac{\exp\big[(\log n)(1 - r(\infty) + o(1))(1 - 2\gamma_n - \sigma_n)\big]}{\log n}, \end{aligned}$$

where (miraculously?)

$$1 - 2\gamma_n - \sigma_n = r(n) \log^{-1/2} n.$$

So the ratio in question approaches infinity as fast as $\exp(r(n) \log^{1/2} n - \log\log n)$. Consequently, $ng(n)\varepsilon_n^2 \gg \log n$, which will be needed to neutralize a polynomial factor implicit in (3.23). Likewise, by $g(n) \leq n^{-r(n)}$,

$$\frac{ng(n)\varepsilon_n^2}{ng(n)^2} = \frac{n^{-2\gamma_n}}{g(n)^{1+2\gamma_n}} \geq \exp\big[(r(n) - 2\gamma_n(1 - r(n)) \log n\big]$$

and

$$r(n) - 2\gamma_n(1 - r(n)) = \max\left\{2r(n) - 1, \frac{(1 - 2r(n))r(n)}{1 - r(n)}\right\}$$
$$+ 2r(n)(1 - r(n)) \log^{-1/2} n$$
$$\geq 2r(n)(1 - r(n)) \log^{-1/2} n.$$

So the second ratio approaches infinity even faster, as $\exp(2r(n)\log^{1/2} n)$. We conclude that

$$P\{Z_n^* > n - k_n\} \leq \exp(-c(ng(n))^{1-2\gamma_n}) \qquad \forall\, c < 1/2,$$

if $n \geq n(c)$. This completes the proof. $\square$

LEMMA 3.10. *Under the conditions and in the notation of Lemma* 3.9,

$$\lim_{n\to\infty} P\{Z_n^* \leq n - \kappa_n\} = 0,$$

where $\kappa_n := (1 + \varepsilon_n)ng(n)$, *and more explicitly*, *the probability in question has the same upper bound as the one in Lemma* 3.9.

PROOF. By Corollary 3.5, it suffices to show that

$$\lim_{n\to\infty} P\big\{n - \hat{\kappa}_n \leq Z_n^* \leq n - \kappa_n\big\} = 0,$$

where $\hat{\kappa}_n := \delta n g(n)$, and $\delta > (e^{r(\infty)} - 1)/(r(\infty))$; $((e^0 - 1)/0 = 1$, by definition). Just as in the previous proof,

$$\mu\big\{n - \hat{\kappa}_n \leq Z_n^* \leq n - \kappa_n\big\} \overset{p}{\leq} \sum_{\substack{\kappa_n \leq k < \hat{\kappa}_n \\ 0 \leq l \leq k}} \exp\big(H_n(k, l)\big).$$

Given $x \in [\kappa_n, \hat{\kappa}_n]$, and $x - y \geq (ng(n))^{\sigma_n}$, we can estimate [see (3.19)]

$$\frac{\partial H_n(x, y)}{\partial y} \geq O(1) + \log \frac{(ng(n))^{\sigma_n(1-r(n))} n}{(ng(n))^2}$$
$$\geq O(1) + \big(\sigma_n(1 - r(n))^2 + 2r(n) - 1\big) \log n$$
$$\to \infty.$$

Hence, for fixed x, $H_n(x, y)$ attains its maximum at a point $y(x)$ such that $x - y(x) \leq (ng(n))^{\sigma_n} = o(x)$. Therefore, analogously to (3.20),

$$\max\big\{H_n(x, y)\colon y \in [0, x]\big\} = H_n(x) + O\big((ng(n))^{\sigma_n} \log n\big), \tag{3.23}$$

where

$$H_n(x) := x \log \frac{n}{x} + x - (n - x) \log g(n) - n\bar{r}(n).$$

Now, $H_n(x)$ is convex and attains its *absolute* maximum at $x_0 = ng(n) < \kappa_n = (1+\varepsilon_n)ng(n)$. Thus,

$$\max\{H_n(x)\colon x \in [\kappa_n, \hat{\kappa}_n]\} = H_n(\kappa_n)$$

and, using also (3.21) and (3.23), we arrive at

$$\begin{aligned} P\{n - \hat{\kappa}_n \le Z_n^* \le n - \kappa_n\} &\overset{p}{\le} \exp(-\mathscr{H}(-\varepsilon_n)ng(n) \\ &\qquad + O\big((ng(n))^{\sigma_n} + ng(n)^2\big)\big) \\ &\le \exp\big(-c(ng(n))^{1-2\gamma_n}\big) \qquad \forall\, c < 1/2, \end{aligned}$$

if $n \ge n(c)$. $\square$

Thus, under the condition $\lim_{x\to\infty}(r(x)\log^{1/2} x - \log\log x) = \infty$, the total number of particles in the remaining less populous sites has a distribution sharply concentrated around $ng(n)$, with random fluctuations of order $\varepsilon_n ng(n)$, at most. If our derivation is any indication, and we think it is, then $r(\infty) = 1/2$ appears to be a certain threshold value for the magnitude of the random fluctuations. For $r := r(\infty) < 1/2$, their order is roughly $(ng(n))^{1-(r/(1-r))^2/2}$, and it is $(ng(n))^{1/2}$ for $r > 1/2$.

4. Open problems and questions.

1. We conjecture that under the conditions of Lemmas 3.9 and 3.10, $n - Z_n^*$ is asymptotically Gaussian with mean $ng(n)$ and standard deviation $o(ng(n))$ (cf. [32]).
2. What happens if $g(x) \downarrow 0$, but $r(x)\log^{1/2} x \not\to \infty$? Is it possible that in this case $n - Z_n^*$ is not sharply concentrated around $ng(n)$? Are there the rates in this class for which the limiting distribution has more than one pronounced peak?
3. We have proved that, for $\rho := \lim_{x\uparrow R} xG'(x)/G(x) < 1$, $Z_n^*/n \to 1-\rho$ in probability, if the tail of X corresponding to $x = R$ is polynomially thin. We conjecture that the conclusion is true for other tails as well, such as $p_j \sim \exp\big(-n^{1-\beta}\big)$, $\beta \in (0,1)$. The latter arises for example when $g(k) \sim R\big(1+k^{-\beta}\big)$. The underlying difficulty is a lack of a local limit theorem for "very large" deviations when the common distribution of the independent summands does not satisfy Cramér's condition.

Acknowledgment. Our paper was read very carefully by two masterful referees who provided us with specific advice and two sets of numerous critical comments. We are very much indebted to them for their thoughtful criticism.

REFERENCES

[1] ALDOUS, D. J. (1999). Deterministic and stochastic models for coalescence (aggregation, coagulation): a review of the mean-field theory for probabilists. *Bernoulli* **5** 3–48.

[2] ALDOUS, D. J. (1998). Emergence of the giant component in special Marcus–Lushnikov process. *Random Structures Algorithms* **12** 179–196.

[3] ARRATIA, R. and TAVARÉ, S. (1994). Independent process approximations for random combinatorial structures. *Adv. Math.* **104** 90–154.
[4] BINGHAM, N. H., GOLDIE, C. M. and TEUGELS, J. L. (1987). *Regular Variation*. Cambridge Univ. Press.
[5] BOLLOBÁS, B. (1985). *Random Graphs*. Academic Press, New York.
[6] BOLLOBÁS, B., GRIMMETT, G. and JANSON, S. (1996). The random cluster model on the complete graph. *Probab. Theory Related Fields* **104** 283–317.
[7] DE MASI, A. and PRESUTTI, E. (1991). Mathematical methods for hydrodynamic limits. *Lecture Notes in Math.* **1501** 33–51. Springer, New York.
[8] DRAKE, R. L. (1972). A general mathematical survey of the coagulation equation. In *Topics in Current Aerosol Research* (G. M. Hidy and J. R. Brock, eds.) **3** 51–119. Pergamon Press, Oxford.
[9] DURRETT, R. (1991). *Probability: Theory and Examples*. Wadsworth & Brooks/Cole, Belmont, CA.
[10] DURRETT, R., GRANOVSKY, B. and GUERON, S. (1997). The equilibrium behavior of reversible coagulation fragmentation processes. Preprint.
[11] ERDŐS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* **5** 17–61.
[12] FLORY, P. J. (1941). Molecular size distribution in three-dimensional polymers III. Tetrafunctional branching units. *J. Amer. Chemical Soc.* **63** 3096–3100.
[13] FRISTEDT, B. (1993). The structure of random partitions of large integers. *Trans. Amer. Math. Soc.* **337** 703–735.
[14] GNEDENKO, B. V. and KOLMOGOROV, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Reading, MA.
[15] GONZALEZ, A. E. (1993). Universality of colloid aggregation in the reaction limit: The computer simulations. *Phys. Rev. Lett.* **71** 2248–2251.
[16] GUERON, S. and LEVIN, S. A. (1995). The Dynamics of Group Formation. *Math. Biosci.* **128** 243–264.
[17] HENDRICKS, E., ERNST, M. and ZIFF, R. (1983). Coagulation equations with gelation. *J. Statist. Phys.* **31** 519–563.
[18] HOLLEY, R. (1970). A class of interactions in an infinite particle system. *Adv. Math.* **5** 291–309.
[19] JEON, I. (1996). Gelation phenomena. Ph.D. dissertation, Ohio State Univ.
[20] JEON, I. (1998). Existence of gelling solutions for coagulation fragmentation equations. *Comm. Math. Phys.* **194** 541–567.
[21] JEON, I. and MARCH, P. (2000). Condensation transition for zero range invariant measures. In *Stochastic Models. Proceedings of the International Conference on Stochastic Models in Honor of Professor Donald A. Dawson* (Luis G. Gorostiza, B. Gail Ivanoff, eds.) 233-244.
[22] KARP, R. M. and SIPSER, M. (1981). Maximum matchings in sparse random graphs. In *Proceedings of the Twenty-Second Annual IEEE Symposium on Foundations of Computing* 364–375. IEEE, Nashville.
[23] KHINCHIN, A. I. (1949). *Mathematical Foundations of Statistical Mechanics*. Dover, New York.
[24] KOLCHIN, V. F. (1986). Random mappings. Optimization Software, Inc., Publication Division, New York.
[25] KOLCHIN, V. F., SEVAST'YANOV, B. A. and CHISTYAKOV, V. P. (1978). *Random Allocations*. Winston, Washington, D.C.
[26] LEYVRAZ, F. and TSCHUDI, H. (1982). Critical kinetics near gelation. *J. Phys. A* **15** 1951–1964.
[27] LIGGETT, T. M. (1985). *Interacting Particle Systems*. Springer, New York.
[28] MCLEOD, J. B. (1962). On an infinite set of non-linear differential equations I, II. *Quart. J. Math.* Oxford *Ser.* (*2*) **13** 119–128, 193–205.
[29] NORRIS, JAMES R. (1999). Smoluchowski's coagulation equation: Uniqueness, non-uniqueness and a hydrodynamic limit for the stochastic coalescent. *Ann. Appl. Probab.* **9** 78–109.
[30] NORRIS, J. R. (2000). Cluster coagulation. *Comm. Math Phys.* **209** 407–435.

[31] PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, New York.
[32] PITTEL, B. (1990). On tree census and the giant component in sparse random graphs. *Random Structures Algorithms* **1** 311–342.
[33] PITTEL, B. (1997). On a likely shape of the random Ferrers diagram. *Adv. in Appl. Math.* **18** 432–488.
[34] PITTEL, B. (1997). Random set partitions: asymptotics of subset counts. *J. Combin. Theory Ser. A* **79** 326–359.
[35] PITTEL, B., WOYCZYNSKI, W. A. and MANN, J. A. (1989). Random tree-type partitions as a model for acyclic polymerization: Gaussian behavior of the subcritical sol phase. In *Random Graphs* (J. Jaworski, M. Karonski and A. Rucinski, eds.) **87** 1–40. Wiley, New York.
[36] PITTEL, B. (1990). Random tree-type partitions as a model for acyclic polymerization: Holtsmark (3/2-stable) distribution of the supercritical gel. *Ann. Probab.* **18** (1990) 319–341. [Correction (1992) **20** 1105–1106.]
[37] PITTEL, B. and WOYCZYNSKI, W. A. (1990). A graph-valued Markov process as rings-allowed polymerization model: subcritical behavior. *SIAM J. Appl. Math.* **50** 1200–1220.
[38] RIORDAN, J. (1968). *Combinatorial Identities*. Wiley, New York.
[39] SCOTT, W. T. (1967). Poisson statistics in distributions of coalescing droplets. *J. Atmospheric Sci.* **24** 221–225.
[40] SHEPP, L. A. and LLOYD, S. P. (1966). Ordered cycle lengths in a random permutation. *Trans. Amer. Math. Soc.* **12** 340–357.
[41] SILK, J. (1980). *Star Formation*. Geneva Observatory, Sauverny, Switzerland.
[42] SPITZER, F. (1970). Interaction of Markov processes. *Adv. in Math.* **5** 246–290.
[43] STOCKMAVER, W. H. (1943). Theory of molecular size distribution and gel formation in branched chain polymers. *Chem. Phys.* **11** 45–55.
[44] VAN DONGEN, P. G. J. and ERNST, M. H. (1983). Pre- and post-gel size distributions in (ir)reversible polymerization. *J. Phys. A* **16** L327–L332.
[45] ZIFF, R., HENDRICKS, E. and ERNST, M. (1982). Critical properties for gelation: A kinetic approach. *Phys. Rev. Lett.* **49** 593–595.

INTAE JEON
DEPARTMENT OF MATHEMATICS
CATHOLIC UNIVERSITY OF KOREA
43-1, YOKKOK 2-DONG, WONMI-GU
PUCHON CITY, KYONGGI-DO, 420–743
KOREA

P. MARCH
B. PITTEL
DEPARTMENT OF MATHEMATICS
OHIO STATE UNIVERSITY
100 MATHEMATICS BUILDING
231 WEST 18TH AVENUE
COLUMBUS, OHIO 43210-1174