ASYMPTOTIC DISTRIBUTION FOR THE COST OF LINEAR PROBING HASHING

SVANTE JANSON

Abstract. We study moments and asymptotic distributions of the construction cost, measured as the total displacement, for hash tables using linear probing. Four different methods are employed for different ranges of the parameters; together they yield a complete description. This extends earlier results by Flajolet, Poblete and Viola. The average cost of unsuccessful searches is considered too.

1. Introduction

Hashing with linear probing is a well-known algorithm that can be described as follows; here n and m are integers with $0 \leq n \leq m$. (For a thorough discussion, see Knuth [15, Section 6.4, in particular Algorithm 6.4.L].)

> n items $x_1, \dots, x_n$ are placed sequentially into a table with m cells $1, \dots, m$, using n integers $h_i \in \{1, \dots, m\}$, by inserting x_i into cell h_i if it is empty, and otherwise trying cells h_i+1, h_i+2, until an empty cell is found; all positions being interpreted modulo m.

In real applications, h_i is computed as $h(x_i)$ by some *hash function* h; in this paper, as in most theoretical analyses, it is assumed that the hash addresses h_i are random numbers, uniformly distributed on $\{1, \dots, m\}$ and independent. In other words, each of the m^n possible hash sequences $(h_i)_1^n$ has the same probability m^{-n}.

If item x_i is inserted into cell q_i, then its *displacement* $(q_i - h_i) \bmod m$, which is the number of unsuccessful probes when this item is inserted, is a measure of the cost of inserting it; it is also a measure of the cost of later finding the item in the table. The *total displacement* $D_{mn} := \sum_{i=1}^{n} (q_i - h_i) \bmod m$ is thus a measure of both the cost of constructing the table and of using it. (The average number of probes to find an element in the table is $D_{mn}/n + 1$.) Note that D_{mn} is an integer with $0 \leq D_{mn} \leq \binom{n}{2}$.

With our assumption that the numbers h_i be random, D_{mn} is a random variable, and the main purpose of the present paper is to give the asymptotic

---

Date: May 17, 2000; revised August 21, 2001.

1991 *Mathematics Subject Classification.* Primary: 68W40; Secondary: 60F05, 60J65.

Key words and phrases. Hashing, linear probing, parking problem, normal convergence, Poisson convergence, Brownian motion, Brownian excursion area, Airy distribution, Stein's method.

This is a preprint of an article accepted for publication in Random Structure & Algoritms

distribution of D_{mn} as $m, n \to \infty$. This has earlier been done by Flajolet, Poblete and Viola [9] for the two most important cases: *full tables* ($n = m$) (and *almost full tables*, $n = m - 1$), and *sparse tables* ($n/m \to a$, with $0 < a < 1$). They found that in the sparse case D_{mn} is asymptotically normal, with both variance and expectation growing like n (or m), while in the full case $n^{-3/2} D_{mn}$ has a non-normal limiting distribution, which equals the distribution of the area under a standard Brownian excursion. (This distribution had earlier been studied by, among others, Louchard [19, 20] and Takács [27]. It is, up to a factor $\sqrt{8}$, called the *Airy distribution* in [9].) We extend their results to other ranges of n as follows.

Theorem 1.1. *Suppose that $m \to \infty$.*

(i) *If $n/\sqrt{m} \to a$ for some a with $0 \le a < \infty$, then D_{mn} is asymptotically Poisson distributed: $D_{mn} \xrightarrow{\mathrm{d}} \mathrm{Po}(a^2/2)$.*

(ii) *If $n \gg \sqrt{m}$ and $m - n \gg \sqrt{m}$, then D_{mn} is asymptotically normal: $(D_{mn} - \mathbb{E}\, D_{mn})/(\operatorname{Var} D_{mn})^{1/2} \xrightarrow{\mathrm{d}} N(0,1)$.*

(iii) *If $(m - n)/\sqrt{m} \to a$ for some a with $0 \le a < \infty$, then $n^{-3/2} D_{mn} \xrightarrow{\mathrm{d}} W_a$, for some non-degenerate random variable W_a.*

In all cases the result holds with convergence of all moments.

Remark 1.2. By saying that $X_k \xrightarrow{\mathrm{d}} X$ with convergence of all moments, we mean that besides $X_k \xrightarrow{\mathrm{d}} X$, we also have $\mathbb{E}\, X_k^r \to \mathbb{E}\, X^r$ for each positive integer r. As is well-known, this holds if (and only if) $X_k \xrightarrow{\mathrm{d}} X$ and $\sup_k \mathbb{E}\, X_k^r < \infty$ for each $r \ge 1$. Moreover, it entails the convergence of all absolute moments $\mathbb{E}\, |X_k|^r$ (r positive real), of all central moments $\mathbb{E}(X_k - \mathbb{E}\, X_k)^r$ (r positive integer) and $\mathbb{E}\, |X_k - \mathbb{E}\, X_k|^r$ (r positive real), and of all semi-invariants $\kappa_r(X_k)$ to the corresponding quantities for X.

Remark 1.3. As in all similar situations, the three cases in Theorem 1.1 do not exhaust all possibilities, since n might oscillate between, for example, $m^{1/3}$ and $m - m^{1/3}$, but they effectively do so since every sequence (m_k, n_k) with $m_k \to \infty$ has a subsequence belonging to one of the cases.

Theorem 1.1 thus exhibits a "phase transition" at $m - n \asymp \sqrt{m}$, where we lose asymptotic normality. The reason is that less dense hash tables consist of many small blocks, each of which is negligible, but for $m - n \asymp \sqrt{m}$, the largest block is of order n and contributes significantly; see the proofs below and Remark 4.2.

The limit random variable W_0 can by [9] be described as the area under a Brownian excursion. We give a related formula for W_a in terms of a Brownian bridge (or a Brownian excursion) in Section 2, and explicit (but complicated) formulae for its moments in Section 3. However, there does not seem to be any simple expression for the distribution of W_a, and we do not know any simple relation with other distributions.

It is a consequence of Theorem 1.1 that the distribution of W_a approaches a normal distribution as $a \to \infty$. This is an instance of a simple general result on continuity of the limits in this type of situations, but since it apparently is not well-known, we give the details (together with some moment asymptotics) in Section 6.

The expectation of D_{mn} is $\frac{n}{2}\bigl(Q_0(m,n-1)-1\bigr)$ [9, 16], cf. [14], [15, Theorem 6.4.K]. (See [15] or [16] for the definition of Q_r.) A similar exact formula for the variance is given by Flajolet, Poblete and Viola [9, Theorem 4], see also [15, Exercise 6.4-68], [16] and (3.16)–(3.17) below.

It follows readily from the exact formula above and the bounds (6.4-43) in [15] that $\mathbb{E}\, D_{mn} \sim n^2/2(m-n)$ as $n \to \infty$, provided $m - n \gg \sqrt{m}$; this was found for fixed m/n already by Knuth [14]. For the variance, Flajolet, Poblete and Viola [9, Theorem 5] found in the case $n/m = \alpha \in (0,1)$ fixed the asymptotic formula (1.2) below (in a sharper form with a second-order term too). We can extend that to other ranges of m and n as follows. (Note that the cases (i) and (ii) overlap.)

Theorem 1.4. *Suppose that $n, m \to \infty$.*

(i) *If $n/m \to 0$, then $\mathbb{E}\, D_{mn} \sim \operatorname{Var} D_{mn} \sim n^2/2m$.*
(ii) *If $n \gg \sqrt{m}$ and $m - n \gg \sqrt{m}$, then, with $\alpha := n/m$,*

$$\mathbb{E}\, D_{mn} \sim \frac{\alpha}{2(1-\alpha)} n = \frac{n^2}{2(m-n)}, \tag{1.1}$$

$$\operatorname{Var} D_{mn} \sim \frac{6\alpha - 6\alpha^2 + 4\alpha^3 - \alpha^4}{12(1-\alpha)^4} n = \frac{6n^2m^3 - 6n^3m^2 + 4n^4m - n^5}{12(m-n)^4}. \tag{1.2}$$

In particular, if $n/m \to 1$ and $m - n \gg \sqrt{m}$, then

$$\operatorname{Var} D_{mn} \sim n^5/4(m-n)^4.$$

(iii) *If $(m-n)/\sqrt{m} \to a$ for some a with $0 \le a < \infty$, then $\mathbb{E}\, D_{mn} \sim n^{3/2}\, \mathbb{E}\, W_a$ and $\operatorname{Var} D_{mn} \sim n^3 \operatorname{Var} W_a$, where $\mathbb{E}\, W_a$ and $\operatorname{Var} W_a$ are given by Corollary 3.4.*

Remark 1.5. Alternatively, (1.2) can be shown by the method in [9]. It can be verified that the asymptotic expansion [9, (7)] for Q_0 is valid also if $\alpha = n/m$ is not constant, provided the error term $O(m^{-5})$ is changed to $O((1-\alpha)^{-11}m^{-5})$; (1.2) then follows by some algebra involving lots of cancellations. Our method has the advantage, however, of yielding the main term directly. On the other hand, the method in [9] yields any desired number of terms in an asymptotic expansion, while our method, in the present version, yields only the leading term.

A common variation of hashing problems is to consider *confined* hashing only, meaning that we consider only hash sequences that leave the last position empty (thus we assume $n < m$). (In particular, there is no wrapping around from m to 1; indeed, confined hashing can equivalently be described as hashing into $m-1$ cells, conditioned on never wrapping around.) Confined hashing is also known as the *parking problem* [18], [15, Exercise 6.4-29].

As is well-known, symmetry and the fact that the hashing table always has $m-n$ empty locations show that the number of confined hash sequences is

$$\frac{m-n}{m}m^n = (m-n)m^{n-1}, \tag{1.3}$$

and that the distribution of the total displacement is the same for the confined case as for the unrestricted case. Hence Theorem 1.1 and the other results in this paper are valid for confined hashing too. Moreover, when proving any result we can choose between the confined and unrestricted versions. This is very advantageous; it turns out that some of our arguments work for one version and some for the other.

There are also variations of the hashing algorithm above, such as "last-come-first-served" and "Robin Hood" [15, Answer 6.4-67], where the displacements of individual items may differ from the version above but the total displacement is the same; the results in this paper are thus valid for these versions too.

We will prove the three parts of Theorem 1.1 (in reverse order) by four different methods in the next four sections, giving two different proofs of part (iii). There are two reasons for giving both proofs: First, we find both interesting; secondly, they give different information about the limit W_a, see Theorems 2.2 and 3.3. The proofs will as a byproduct yield Theorem 1.4(i)(ii) too (Theorem 1.4(iii) is an immediate consequence of Theorem 1.1).

As mentioned above, the average cost of searching for an element in the table (after it has been constructed) is given by $D_{mn}/n+1$, and is thus asymptotically described by the results above. On the other hand, let U_{mn} be the average cost of an unsuccessful search, i.e. the average number of probes used until giving up when searching for an element *not* in the table, beginning at a random cell h; we average over h so U_{mn} becomes a function of the table, and thus a random variable. (This average is relevant in applications where a hash table is constructed once, and then used for many searches. The distribution of individual search costs will not be considered in this paper.) Note that U_{mn} is the same as the average number of probes needed to extend the table by one item.

The expectation of U_{mn} is $\mathbb{E}\, U_{mn} = \frac{1}{2}Q_1(m,n) + \frac{1}{2}$ [15, Theorem 6.4.K]. We give a corresponding exact formula for the variance in Theorem 7.3. (Higher moments could be obtained by the same method.)

For asymptotics, we have the following companion results to Theorems 1.1 and 1.4. The asymptotics for $\mathbb{E}\, U_{mn}$ in Theorem 1.7(i)(ii) follow easily from the exact formula above, using [15, (6.4-43)] for (ii). The other results are proved in Section 7.

Theorem 1.6. *Suppose that $m \to \infty$.*

(i) *If $n/\sqrt{m} \to a$ for some a with $0 \le a < \infty$, then $mU_{mn} - m - n \xrightarrow{\mathrm{d}} \operatorname{Po}(3a^2/2)$.*
(ii) *If $n \gg \sqrt{m}$ and $m-n \gg \sqrt{m}$, then U_{mn} is asymptotically normal: $(U_{mn} - \mathbb{E}\, U_{mn})/(\operatorname{Var} U_{mn})^{1/2} \xrightarrow{\mathrm{d}} N(0,1)$.*

(iii) *If* $(m-n)/\sqrt{m} \to a$ *for some* a *with* $0 \le a < \infty$, *then* $\frac{2}{m}U_{mn} \xrightarrow{\mathrm{d}} V_a$ *for some random variable* V_a, *which is non-degenerate for* $0 < a < \infty$ *while* $V_0 = 1$.

In all cases the result holds with convergence of all moments.

The normalizations in (i) and (iii) are partly explained by the fact, which is an easy consequence of (7.1) below, that mU_{mn} is an integer with $m+n \le mU_{mn} \le \binom{m}{2}$, and thus $\frac{2}{m}U_{mn} \le 1 + \frac{1}{m}$.

Theorem 1.7. *Suppose that* $n, m \to \infty$, *with* $n < m$.

(i) *If* $n/m \to 0$, *then* $\mathbb{E}\, U_{mn} = 1 + \frac{n}{m} + \frac{3n(n-1)}{2m^2} + o(\frac{n^2}{m^2})$ *and* $\operatorname{Var} U_{mn} \sim 3n^2/2m$.

(ii) *If* $n \gg \sqrt{m}$ *and* $m - n \gg \sqrt{m}$, *then, with* $\alpha := n/m$,

$$\mathbb{E}\, U_{mn} \sim \frac{1}{2} + \frac{1}{2(1-\alpha)^2} = \frac{2m^2 - 2mn + n^2}{2(m-n)^2},$$

$$\operatorname{Var} U_{mn} \sim \frac{3\alpha^2}{2(1-\alpha)^6} m^{-1} = \frac{3n^2m^3}{2(m-n)^6}.$$

In particular, if $n/m \to 1$ *and* $m - n \gg \sqrt{m}$, *then* $\mathbb{E}\, U_{mn} \sim n^2/2(m-n)^2$ *and* $\operatorname{Var} U_{mn} \sim 3n^5/2(m-n)^6$.

(iii) *If* $(m-n)/\sqrt{m} \to a$ *for some* a *with* $0 \le a < \infty$, *then* $\mathbb{E}\, U_{mn} \sim \frac{1}{2} n \,\mathbb{E}\, V_a$ *and* $\operatorname{Var} U_{mn} \sim \frac{1}{4} n^2 \operatorname{Var} V_a$, *where* $\mathbb{E}\, V_a$ *and* $\operatorname{Var} V_a$ *are given by Theorem 7.4.*

Finally, in Section 8 we discuss the joint distribution of D_{mn} and U_{mn}.

Acknowledgements. I thank Donald Knuth for drawing my attention to the study of hashing with linear probing, and Philippe Flajolet for helpful comments.

2. The dense case: Brownian limits

In this section we give our first proof of the limit theorem for D_{mn} when $(m-n)/\sqrt{m} \to a < \infty$. The convergence in distribution is an easy consequence of a limit theorem for the profile of hashing in terms of some stochastic processes related to Brownian motion [5, Theorem 4.1]; since that result is given in a technically more complicated context than used here, we sketch the argument in a slightly simpler version.

Let, for $i = 1, \dots, m$, X_i be the number of items x_k with hash address $h_k = i$, and let $S_i := \sum_{j=1}^{i} X_j$, $0 \le i \le m$. Thus $S_0 = 0$ and $S_m = \sum_1^m X_j = n$. Moreover, let H_i be the number of items that make an attempt to be inserted in cell i, whether they succeed or not. We call $(H_i)_{i=1}^m$ the *profile* of the hashing. Since the total displacement equals the number of unsuccessful probes and the total number of probes is $\sum_i H_i$, of which n are successful,

$$D_{mn} = \sum_{i=1}^{m} H_i - n. \tag{2.1}$$

It is convenient to extend the definition of X_i, S_i and H_i to all integers i, with $X_{i+m} = X_i$, $H_{i+m} = H_i$, $S_{i+m} = S_i + n$ for all i. (Thus $S_i = \sum_{j=1}^{i} X_j$ for all $i \geq 0$ and $S_i = -\sum_{i+1}^{0} X_j$ for $i < 0$; in any case $S_i = X_i + S_{i-1}$.) Then H_i can be computed as follows [5, Proposition 5.3], cf. [15, Exercise 6.4-32].

Lemma 2.1. *With X_i, S_i, H_i defined for all integers i as above,*

$$\begin{aligned} H_i &= \max_{j \leq i} \left(\sum_{k=j}^{i} X_k - (i-j) \right) = \max_{j \leq i} (S_i - S_{j-1} - i + j) \\ &= S_i - i - \min_{k<i} (S_k - k) + 1. \end{aligned}$$

Proof. For $i - m < j \leq i$, there are $\sum_{k=j}^{i} X_k$ items that first try one of the cells $\{j, \dots, i\}$, and at most $i-j$ of them can be accomodated in $\{j, \dots, i-1\}$, so at least $\sum_{k=j}^{i} X_k - (i-j)$ try cell i; hence, $H_i \geq \sum_{k=j}^{i} X_k - (i-j)$. The periodicity shows that this holds for $j \leq i - m$ too.

Conversely, if $j = j_0 + 1$, where j_0 is the largest integer less than i where there are no unsuccessful probes, it is easily seen that $H_l = \sum_{k=j}^{l} X_k - (l-j)$ for $j \leq l \leq i$; in particular $H_i = \sum_{k=j}^{i} X_k - (i-j)$. $\square$

Consider the random function $S_{\lfloor mt \rfloor} - nt$, $0 \leq t \leq 1$; note that it vanishes for both $t = 0$ and $t = 1$. This function equals $n(\beta_{mn}(t) - t)$, where β_{mn} is the empirical distribution function of $\{h_k/m\}_{k=1}^{n}$. Letting $U_1, \dots, U_n$ be independent random variables with a uniform distribution on $[0,1]$, we can take $h_k = \lceil mU_k \rceil$, and then $\beta_{mn}(t) = \beta'_n(\lfloor mt \rfloor/m)$, where β'_n is the empirical distribution function of $\{U_k\}_{k=1}^{n}$. Now, it is well-known [4, Theorem 16.4] that $\sqrt{n}(\beta'_n(t) - t) \xrightarrow{\mathrm{d}} b(t)$, where $b(t)$ is a standard Brownian bridge, and the convergence is in the Skorohod topology on $D[0,1]$. It follows that

$$\begin{aligned} \frac{1}{\sqrt{n}}(S_{\lfloor mt \rfloor} - nt) &= \sqrt{n}(\beta_{mn}(t) - t)) \\ &= \sqrt{n}\big(\beta'_n(\lfloor mt \rfloor/m) - \lfloor mt \rfloor/m + O(1/m)\big) \\ &\xrightarrow{\mathrm{d}} b(t). \end{aligned}$$

Multiplying by $\sqrt{n/m} \to 1$ and adding $(nt - \lfloor mt \rfloor)/\sqrt{m} \to -at$, we obtain

$$\frac{1}{\sqrt{m}}(S_{\lfloor mt \rfloor} - \lfloor mt \rfloor) \xrightarrow{\mathrm{d}} b(t) - at.$$

Hence, using Lemma 2.1 and the mapping theorem [4, Theorem 5.1], extending b periodically to a function on $(-\infty, \infty)$, we have in $D[0,1]$

$$\frac{1}{\sqrt{m}} H_{\lfloor mt \rfloor} \xrightarrow{\mathrm{d}} b(t) - at - \min_{s \leq t}\big(b(s) - as\big) = \max_{s \leq t}\big(b(t) - b(s) - a(t-s)\big). \quad (2.2)$$

Consequently, by the mapping theorem again,

$$\frac{1}{m^{3/2}} \sum_{i=1}^{m} H_i = \int_0^1 \frac{1}{\sqrt{m}} H_{\lfloor mt \rfloor}\, dt \xrightarrow{\mathrm{d}} \int_0^1 \max_{s \leq t}\big(b(t) - b(s) - a(t-s)\big)\, dt. \quad (2.3)$$

Together with (2.1) and $n/m \to 1$, this proves that $n^{-3/2}D_{mn}$ converges in distribution as asserted, with the following description of the limit distribution.

Theorem 2.2. *The limit W_a in Theorem 1.1(iii) can be constructed by*

$$W_a := \int_0^1 \max_{s\leq t}\bigl(b(t)-b(s)-a(t-s)\bigr)\,dt$$

for a Brownian bridge b on $[0,1]$, periodically extended to $(-\infty,\infty)$.

In order to show moment convergence, it suffices by Remark 1.2 to show that each moment $\mathbb{E}(D_{mn}/n^{3/2})^r$ is bounded, and since D_{mn} increases with n, it suffices to prove this for $n=m$. Moreover, by Lemma 2.1, $\max_i H_i \leq 2\max_i|S_i-i|+1$, and thus by (2.1),

$$\begin{aligned} m^{-3/2}D_{mm} &\leq 2m^{-1/2}\max_i|S_i-i| = 2m^{1/2}\max_i|\beta_{mm}(i/m)-i/m| \\ &= 2m^{1/2}\max_i|\beta'_m(i/m)-i/m| \leq 2m^{1/2}\max_t|\beta'_m(t)-t|, \end{aligned}$$

and all moments of the latter variable are bounded, for example by the (much stronger) Dvoretzky-Kiefer-Wolfowitz inequality [8], which completes the first proof of Theorem 1.1(iii). We omit the details, since we give another proof of moment convergence in the next section.

Remark 2.3. If $a=0$, then W_0 equals the integral of the stochastic process $\max_s(b(t)-b(s)) = b(t)-\min_s b(s)$, which by a theorem by Vervaat [28] has the same distribution as a standard Brownian excursion $e(t)$ up to a random shift. The shift does not affect the integral, and thus we can take $W_0=\int_0^1 e(t)$, the Brownian excursion area, as found by [9]. More generally, it follows from Vervaat's result that we can take

$$W_a := \int_0^1 \max_{0\leq s\leq t}\bigl(e(t)-e(s)-a(t-s)\bigr)\,dt$$

too [5]. (This can also be derived by arguing as above with confined hashing instead of the unconfined version, but the details become technically more complicated, cf. [5, 6, 7].) Furthermore, it follows from [5, Theorem 2.2] that W_a also can be defined as the integral of a reflecting Brownian bridge $|b|$ conditioned on having local time at 0 equal to a.

3. The dense case: moments

Our second proof of Theorem 1.1(iii) is based on expressions for generating functions given by Knuth [16] (see also [9]). We work with the confined version, and thus assume $n<m$; the results for $n=m$ follow from the case $n=m-1$, since the displacement of the last item is less than n and thus $D_{m,m-1}\leq D_{m,m} < D_{m,m-1}+m$.

Following [16], we let $F_{mn}(x)$ be the generating function for the total displacement in the confined version of the problem; thus

$$\mathbb{E}\,x^{D_{mn}} = F_{mn}(x)/F_{mn}(1), \tag{3.1}$$

where, see (1.3),

$$F_{mn}(1) = (m-n)m^{n-1}. \tag{3.2}$$

Next, using the bivariate generating function

$$F(x,z) := \sum_{n=0}^{\infty} F_{n+1,n}(x)z^n/n!, \tag{3.3}$$

Knuth [16, (1.5)] showed that

$$F_{mn}(x) = n![z^n]F(x,z)^{m-n}. \tag{3.4}$$

By (3.1) and (3.4) we have

$$\mathbb{E}\binom{D_{mn}}{k} = [w^k]\,\mathbb{E}(1+w)^{D_{mn}} = [w^k z^n]F(1+w,z)^{m-n}/[z^n]F(1,z)^{m-n}. \tag{3.5}$$

Knuth [16, (4.2)] further showed that

$$F(1+w,z) = \sum_{k=0}^{\infty} w^k W_k'(z), \tag{3.6}$$

where W_k is the exponential generating function for the number of connected labelled graphs with $k-1$ more edges than vertices, which by Wright [29] can be expressed in terms of the tree function

$$T(z) := \sum_{i=1}^{\infty} \frac{i^{i-1}z^i}{i!}. \tag{3.7}$$

In the notation of [13] we have $W_k = \widehat{C}_{k-1}$, where by [13, (8.13)], for $r \geq 1$,

$$\widehat{C}_r(z) = \sum_{d=0}^{3r+2} \hat{c}_{rd} \frac{T(z)^{3r+2-d}}{\bigl(1-T(z)\bigr)^{3r-d}}.$$

Expanding $T^{3r+2-d} = (1-(1-T))^{3r+2-d}$ by the binomial theorem, this yields

$$\widehat{C}_r(z) = \sum_{j=-2}^{3r} c^*_{rj}\bigl(1-T(z)\bigr)^{-j}$$

with the leading coefficient $c^*_{r,3r} = \hat{c}_{r0} = c_r$, where c_r is as in [13, §8]. Consequently, using the fact that $T'(z) = T(z)/z\bigl(1-T(z)\bigr)$, for $r \geq 1$,

$$\widehat{C}_r'(z) = \sum_{-2}^{3r} j c^*_{rj}\bigl(1-T(z)\bigr)^{-j-1}T'(z) = \frac{T(z)}{z}\sum_{-2}^{3r} j c^*_{rj}\bigl(1-T(z)\bigr)^{-j-2}. \tag{3.8}$$

For $r<1$ we instead have, by [13, §3], $\widehat{C}_{-1}(z)=U(z)=T(z)-\frac12T(z)^2$ and $\widehat{C}_0(z)=\widehat{V}(z)=\frac12\ln\bigl(1-T(z)\bigr)^{-1}-\frac12T(z)-\frac14T(z)^2$, which yield

$$\begin{aligned}\widehat{C}'_{-1}(z)&=T'(z)\bigl(1-T(z)\bigr)=\frac{T(z)}{z},\\ \widehat{C}'_0(z)&=\frac{T'(z)}{2\bigl(1-T(z)\bigr)}-\tfrac12T'(z)-\tfrac12T(z)T'(z)\\ &=\frac{T(z)}{z}\Bigl(\tfrac12\bigl(1-T(z)\bigr)^{-2}-\bigl(1-T(z)\bigr)^{-1}+\tfrac12\Bigr).\end{aligned}\tag{3.9}$$

Hence, for all $k\geq 0$,

$$W'_k(z)=\widehat{C}'_{k-1}(z)=\frac{T(z)}{z}f_k\bigl(T(z)\bigr),\tag{3.10}$$

where $f_k(t)$ is a polynomial in $(1-t)^{-1}$. Here $f_0(t)=1$, while for $k\geq1$, f_k has degree $3(k-1)+2=3k-1$ in $(1-t)^{-1}$; more precisely

$$f_k(t)=\omega_k(1-t)^{-(3k-1)}+\dots,\tag{3.11}$$

where the leading coefficient is given by $\omega_1=\frac12$ and

$$\omega_k=3(k-1)c^*_{k-1,3(k-1)}=3(k-1)c_{k-1},\qquad k\geq2.\tag{3.12}$$

(These are the same ω_k as in [9], as follows e.g. from (3.22) below.) For future use we note that $\omega_2=3c_1=5/8$; this and further numerical values are given in [9, Table 1], see also the table of $\hat{c}_{kd}$ in [13, §8]. We record also, see [16, (4.5)],

$$f_1(t)=\frac{t^2}{2(1-t)^2},\qquad\qquad f_2(t)=\frac{24t^3-11t^4+2t^5}{24(1-t)^5}.\tag{3.13}$$

Let $f(w,t):=\sum_0^\infty w^kf_k(t)$. Then (3.6) and (3.10) yield that [16, (4.4)]

$$F(1+w,z)=\frac{T(z)}{z}f(w,T(z)),\tag{3.14}$$

which using Lagrange inversion leads to, as shown by [16, (5.1)], cf. [9, (31)],

$$[z^n]F(1+w,z)^{m-n}=[t^n]e^{mt}(1-t)f(w,t)^{m-n}.\tag{3.15}$$

Consequently, for $k\geq1$, using $f_0=1$,

$$\begin{aligned}[w^kz^n]F(1+w,z)^{m-n}&=[w^kt^n]e^{mt}(1-t)f(w,t)^{m-n}\\ &=[t^n]e^{mt}(1-t)\sum_{j=1}^k\binom{m-n}{j}\sum_{\substack{k_1,\dots,k_j\geq1\\ \sum k_i=k}}\prod_{i=1}^j f_{k_i}(t).\end{aligned}$$

Moreover, by (3.15), or by (3.4) and (3.2),

$$[z^n]F(1,z)^{m-n}=[t^n]e^{mt}(1-t)=\frac{m^{n-1}}{n!}(m-n).$$

Hence, by (3.5),

$$\mathbb{E}\binom{D_{mn}}{k}=\frac{m^{1-n}n!}{m-n}\sum_{j=1}^{k}\binom{m-n}{j}\sum_{\substack{k_1,\dots,k_j\geq 1\\ \sum k_i=k}}[t^n]e^{mt}(1-t)\prod_{i=1}^{j}f_{k_i}(t), \tag{3.16}$$

where, as shown above, f_{k_i} is a polynomial in $(1-t)^{-1}$. Now, cf. [16, (5.3)],

$$[t^n]e^{mt}(1-t)^{-r-1}=m^nQ_r(m,n)/n! \tag{3.17}$$

where

$$Q_r(m,n)=\sum_{j=0}^{n}\binom{r+j}{j}\frac{n!}{m^j(n-j)!}. \tag{3.18}$$

The right hand side of (3.16) can thus be expressed as a linear combination of a number of different Q_r. (See [16, (5.4)–(5.5)] for the first two cases.) We need the following straightforward asymptotics for Q_r in our range.

Lemma 3.1. *If $r\geq 0$ is a fixed integer, and $n,m\to\infty$ with $(m-n)/\sqrt{m}\to a\geq 0$, then*

$$Q_r(m,n)\sim q_r(a)n^{(r+1)/2},$$

with

$$q_r(a):=\frac{1}{r!}\int_0^\infty x^r e^{-ax-x^2/2}dx. \tag{3.19}$$

Moreover, $Q_{-1}(m,n)=1$ and $Q_{-2}(m,n)=1-n/m=O(n^{-1/2})$.

Proof. Denote the terms in the sum in (3.18) by b_j. If $j\sim xn^{1/2}$ for some $x>0$, then

$$\begin{aligned}b_j&=\binom{j+r}{r}m^{-j}\frac{n!}{(n-j)!}=\frac{(j+O(1))^r}{r!}\Bigl(1-\frac{m-n}{m}\Bigr)^j\exp\Bigl(\sum_{i=0}^{j-1}\ln\Bigl(1-\frac{i}{n}\Bigr)\Bigr)\\&=\frac{j^r}{r!}\exp\Bigl(-j\frac{m-n}{m}-\frac{j^2}{2n}+o(1)\Bigr)\sim\frac{n^{r/2}x^r}{r!}e^{-ax-x^2/2}.\end{aligned}$$

Moreover, for all $j\leq n$,

$$\begin{aligned}b_j=\binom{j+r}{r}m^{-j}\frac{n!}{(n-j)!}&\leq(1+rj^r)\exp\Bigl(\sum_{i=0}^{j-1}\ln\Bigl(1-\frac{i}{n}\Bigr)\Bigr)\\&\leq(1+rj^r)\exp\Bigl(-\frac{j(j-1)}{2n}\Bigr)\end{aligned}$$

and dominated convergence yields

$$\begin{aligned}n^{-(r+1)/2}Q_r(m,n)=n^{-(r+1)/2}\sum_{j=0}^{n}b_j&=\int_0^{(n+1)n^{-1/2}}n^{-(r+1)/2}b_{\lfloor n^{1/2}x\rfloor}n^{1/2}\,dx\\&\to\int_0^\infty\frac{x^r}{r!}e^{-ax-x^2/2}\,dx=q_r(a).\end{aligned}$$

The formulae for Q_{-1} and Q_{-2} are immediate. ∎

Lemma 3.1 and (3.17) show that the leading terms in (3.16) come from the highest powers of $(1-t)^{-1}$. More precisely, since by (3.11) $\prod_1^j f_{k_i}$ has degree $\sum_i(3k_i-1)=3k-j$ with leading term $\prod_{i=1}^j \omega_{k_i}(1-t)^{j-3k}$, (3.16), (3.17) and Lemma 3.1 yield, for some a_{kjl},

$$\begin{aligned}
\mathbb{E}\binom{D_{mn}}{k} &= \frac{m}{m-n}\sum_{j=1}^{k}\binom{m-n}{j}\Bigg(\sum_{\substack{k_1,\dots,k_j\ge 1\\ \sum k_i=k}}\prod_{i=1}^{j}\omega_{k_i}\cdot Q_{3k-j-2}(m,n)\\
&\qquad\qquad + \sum_{l=0}^{3k-j-1} a_{kjl}Q_{l-2}(m,n)\Bigg)\\
&= m\sum_{j=1}^{k}\left(\frac{(m-n)^{j-1}}{j!}+O(n^{(j-2)/2})\right)\\
&\qquad\cdot\left(\sum_{\substack{k_1,\dots,k_j\ge 1\\ \sum k_i=k}}\prod_{i=1}^{j}\omega_{k_i}\cdot q_{3k-j-2}(a)n^{(3k-j-1)/2}+o(n^{(3k-j-1)/2})\right)\\
&= mn^{3k/2-1}\sum_{j=1}^{k}\left(\frac{a^{j-1}}{j!}+o(1)\right)\left(\sum_{\substack{k_1,\dots,k_j\ge 1\\ \sum k_i=k}}\prod_{i=1}^{j}\omega_{k_i}\cdot q_{3k-j-2}(a)+o(1)\right).
\end{aligned}$$

Thus, if we define, for $k\ge 1$,

$$\psi_k(a) := k!\sum_{j=1}^{k}\left(\sum_{\substack{k_1,\dots,k_j\ge 1\\ \sum k_i=k}}\prod_{i=1}^{j}\omega_{k_i}\right)\frac{a^{j-1}}{j!}q_{3k-j-2}(a), \tag{3.20}$$

we have shown

$$\mathbb{E}\binom{D_{mn}}{k} = n^{3k/2}\left(\tfrac{1}{k!}\psi_k(a)+o(1)\right), \qquad k\ge 1,$$

which implies

$$n^{-3k/2}\,\mathbb{E}\,D_{mn}^k \to \psi_k(a), \qquad k\ge 1. \tag{3.21}$$

We have thus shown that all moments of $n^{-3/2}D_{mn}$ converge. This gives a proof of Theorem 1.1(iii) by the method of moments, and shows that $\mathbb{E}\,W_a^k=\psi_k(a)$, provided we can show that the moments $\psi_k(a)$ determine a unique distribution. A sufficient condition for this is that the sum $\sum_k \frac{\lambda^k}{k!}\psi_k(a)$ converges for all real λ (the sum then equals $\mathbb{E}\,e^{\lambda W_a}-1$). We observe that D_{mn} and thus $\mathbb{E}\,D_{mn}^k$ are increasing in n for fixed m, and thus (3.21) implies that $\psi_k(a)$ is a decreasing function of a. In particular, $\psi_k(a)\le\psi_k(0)$, so it suffices to consider $a=0$.

Moreover, (3.20) yields, using the doubling formula for the gamma function,

$$\begin{aligned}\psi_k(0) = k!\,\omega_k q_{3k-3}(0) &= \frac{k!}{(3k-3)!}\omega_k \int_0^\infty x^{3k-3} e^{-x^2/2}\,dx \\ &= \frac{k!}{(3k-3)!}\omega_k \cdot 2^{3k/2-2}\Gamma(3k/2-1) \\ &= 2^{1-3k/2}\pi^{1/2} k!\,\omega_k/\Gamma((3k-1)/2), \end{aligned} \tag{3.22}$$

which by (3.12), the asymptotics $c_r \sim (3/2)^r (r-1)!/2\pi$ as $r \to \infty$ [13, (8.7)] and Stirling's formula easily implies $\sum_k \lambda^k \psi_k(0)/k! < \infty$.

Remark 3.2. The fact that $\mathbb{E}\, e^{\lambda W_a} < \infty$ for every real λ is perhaps more simply verified using the results of Section 2; Theorem 2.2 yields $0 \le W_a \le 2\max_t |b(t)|$, and it is well-known that $\mathbb{E} \exp(2\lambda \max_t |b(t)|) < \infty$, cf. e.g. [4, (11.39) or (11.40)].

The relation (3.22) shows further, since $\psi_k(0) = \mathbb{E}\, W_0^k > 0$, that $\omega_k > 0$ for all $k \ge 1$; hence $\psi_k(a) > 0$ for all $a \ge 0$.

We summarize the results obtained on W_a.

Theorem 3.3. *The limit random variables W_a have the moments $\mathbb{E}\, W_a^k = \psi_k(a)$, $k \ge 1$, with ψ_k defined in (3.20). In particular,*

$$\mathbb{E}\, W_a = \omega_1 q_0(a) = \tfrac{1}{2} q_0(a)$$

and

$$\mathbb{E}\, W_a^2 = 2\omega_2 q_3(a) + \omega_1^2 a q_2(a) = \tfrac{5}{4} q_3(a) + \tfrac{1}{4} a q_2(a).$$

Moreover, the moment generating function $\mathbb{E}\, e^{\lambda W_a}$ is finite for each λ, and thus the distribution of W_a is determined by the moments $\psi_k(a)$. □

The functions $q_k(a)$, and thus the moments $\mathbb{E}\, W_a^k = \psi_k(a)$, can be expressed in terms of the normal distribution function Φ. Indeed, by the change of variable $x + a = y$,

$$\begin{aligned} q_0(a) &= \int_0^\infty e^{-ax-x^2/2}\,dx = e^{a^2/2}\int_a^\infty e^{-y^2/2}\,dy = \sqrt{2\pi} e^{a^2/2}(1-\Phi(a)) \\ &= \sqrt{2\pi} e^{a^2/2}\Phi(-a). \end{aligned}$$

Moreover,

$$q_1(a) = \int_0^\infty (x+a) e^{-ax-x^2/2}\,dx - a q_0(a) = 1 - a q_0(a)$$

and, for $k \ge 2$, by integration by parts,

$$k q_k(a) = \int_0^\infty \frac{x^{k-1}}{(k-1)!}(x+a) e^{-ax-x^2/2}\,dx - a q_{k-1}(a) = q_{k-2}(a) - a q_{k-1}(a).$$

By induction, any q_k can thus recursively be expressed as $\alpha_k(a) + \beta_k(a)q_0(a)$, where α and β are polynomials of degree $k-1$ and k, respectively. For example,

$$\begin{aligned} q_2(a) &= \tfrac{1}{2}\big((1+a^2)q_0(a) - a\big), \\ q_3(a) &= \tfrac{1}{6}\big((2+a^2) - (3a+a^3)q_0(a)\big). \end{aligned}$$

Hence, the expressions for the first two moments of W_a in Theorem 3.3 can be rewritten:

Corollary 3.4. *For any $a \geq 0$,*

$$\begin{aligned} \mathbb{E}\, W_a &= \tfrac{1}{2}q_0(a) = \sqrt{\tfrac{\pi}{2}}e^{a^2/2}\Phi(-a), \\ \mathbb{E}\, W_a^2 &= \tfrac{5}{4}q_3(a) + \tfrac{1}{4}aq_2(a) = \tfrac{1}{12}\big(5 + a^2 - (6a+a^3)q_0(a)\big) \\ &= \tfrac{1}{12}\big(5 + a^2 - (6a+a^3)\sqrt{2\pi}e^{a^2/2}\Phi(-a)\big). \end{aligned}$$ □

Asymptotics as $a \to \infty$ are considered in Section 6.

4. The sparse case: normality

We exploit, as several other authors [6, 9, 16] the simple fact that a confined hash table with n items in m cells decomposes into $m - n$ blocks, each ending with an empty cell, where each block can be regarded as a separate almost full confined hash table. More precisely, a hash sequence $\{h_i\}$ giving a hash table with block lengths $\ell_1, \dots, \ell_N$, where $N = m - n$ and $\sum_i \ell_i = m$, can be constructed by first partitioning $\{1, \dots, n\}$ into subsets $\{A_j\}_{j=1}^N$ with $|A_j| = \ell_j - 1$, and then for each j choosing $(h_i)_{i\in A_j}$ that after a simple relabelling corresponds to a hash sequence yielding a confined hash table with $\ell_j - 1$ items and ℓ_j cells. (Note that we define the block lenghts to include the final, empty cell.)

Since, by (1.3), there are $\ell^{\ell-2}$ confined hash sequences for $\ell - 1$ items and ℓ cells, it follows that the number of confined hash sequences for n items in m cells yielding block lengths $\ell_1,\dots,\ell_N$ equals

$$\binom{n}{\ell_1 - 1, \dots, \ell_N - 1}\prod_{j=1}^N \ell_j^{\ell_j-2} = n!\prod_{j=1}^N \frac{\ell_j^{\ell_j-2}}{(\ell_j-1)!} = n!\prod_{j=1}^N \frac{\ell_j^{\ell_j-1}}{\ell_j!}.$$

Consequently, the probability that a random confined hash table has block lengths $\ell_1, \dots, \ell_N$ is proportional to $\prod_j \ell_j^{\ell_j-1}/\ell_j!$.

However, if λ is any real number with $0 < \lambda \leq e^{-1}$, so that $T(\lambda)$ defined by (3.7) is finite, and $X_1, \dots, X_N$ are independent random variables with the common Borel distribution

$$\mathbb{P}(X_j = \ell) = \frac{1}{T(\lambda)}\frac{\ell^{\ell-1}}{\ell!}\lambda^\ell, \qquad \ell = 1, 2, \dots, \tag{4.1}$$

then the conditional probability that $(X_1, \dots, X_N) = (\ell_1, \dots, \ell_N)$ given that $\sum_j X_j = m$ is also proportional to $\prod_j \ell_j^{\ell_j-1}/\ell_j!$. Consequently, the proportionality factors have to agree, and the sequence of block lengths in a random confined hash table has the same distribution as $(X_1, \dots, X_N)$ conditioned on

$\sum_j X_j = m$. Moreover, given the block lengths, the blocks can be regarded as independent almost full confined hash tables; in particular, the sums of displacements inside the blocks are distributed as the total displacements for independent almost full hash tables of sizes equal to the given block lengths, and we obtain the following result.

Lemma 4.1. *Suppose $0 \le n < m$ and let $N = m - n$. Let $0 < \lambda \le e^{-1}$ and let $(X_1, Y_1), \dots, (X_N, Y_N)$ be independent random vectors with a common distribution obtained by first selecting X_j according to (4.1) and then, if $X_j = \ell$, letting Y_j be distributed as the total displacement $D_{\ell,\ell-1}$. Then, for a random hash table with n items and m cells, the block lengths and the sums of displacements inside each block are distributed as $(X_1, Y_1), \dots, (X_N, Y_N)$ conditioned on $\sum_{j=1}^N X_j = m$. In particular, the distribution of the total displacement D_{mn} equals the conditional distribution of $\sum_{j=1}^N Y_j$ given $\sum_{j=1}^N X_j = m$.* □

Remark 4.2. Lemma 4.1 is closely related to the relation (3.4) for generating functions derived in [9, 16], and our proof partly repeats arguments there, but we use a more probabilistic formulation.

There is further a one-to-one correspondence between hash tables and rooted forests, see e.g. [15, Exercise 6.4-31] and [6], and the lemma is essentially the same as a result used by Pavlov [17, 21, 22] to study random rooted forests. In particular, the distribution of the length of the largest block is given by [21].

We will use Lemma 4.1 together with the following general asymptotic result for conditioned distributions, which is proved (in a slightly more general form) in [12]. (The method of proof is similar to the saddle point method analysis of a generating function in [9], but in more probabilistic terms. Related conditional limit theorems, proved by the same method, are given in, for example, [10, 11].)

Lemma 4.3. *Suppose that, for each k, $(X, Y) = (X(k), Y(k))$ is a pair of random variables such that X is integer valued, and that $N = N(k)$ and $m = m(k)$ are integers. Suppose further that for some γ and c (independent of k), with $0 < \gamma \le 2$ and $c > 0$, the following hold, where $\sigma_X^2 := \operatorname{Var} X$, $\sigma_Y^2 := \operatorname{Var} Y$ and all limits are taken as $k \to \infty$:*

(i) $\mathbb{E}\, X = m/N$.
(ii) $0 < \sigma_X^2 < \infty$.
(iii) *For every integer $r \ge 3$, $\mathbb{E}\,|X - \mathbb{E}\, X|^r = o(N^{r/2-1}\sigma_X^r)$.*
(iv) $\sigma_X^2 = O(N^{2/\gamma-1})$.
(v) *$\varphi_X(s) := \mathbb{E}\, e^{isX}$ satisfies $1 - |\varphi_X(s)| \ge c \min(|s|^\gamma, s^2\sigma_X^2)$ for $|s| \le \pi$.*
(vi) $0 < \sigma_Y^2 < \infty$.
(vii) *For every integer $r \ge 3$, $\mathbb{E}\,|Y - \mathbb{E}\, Y|^r = o(N^{r/2-1}\sigma_Y^r)$.*
(viii) *The correlation $\rho := \operatorname{Cov}(X, Y)/\sigma_X\sigma_Y$ satisfies* $\limsup |\rho| < 1$.

Let, for each k, (X_i, Y_i) be i.i.d. copies of (X, Y), and let $S_N := \sum_1^N X_i$, $T_N := \sum_1^N Y_i$ and $\tau^2 := \sigma_Y^2(1 - \rho^2) = \sigma_Y^2 - \operatorname{Cov}(X, Y)^2/\sigma_X^2$. Then, as $k \to \infty$, the conditional distribution of $(T_N - N\,\mathbb{E}\, Y)/N^{1/2}\tau$ given $S_N = m$ converges to a standard normal distribution. In other words, if $U = U_k$ is a random variable *whose distribution equals the conditional distribution of T_N given $S_N = m$, then*

$$\frac{U - N\,\mathbb{E}\,Y}{N^{1/2}\tau} \xrightarrow{\mathrm{d}} N(0,1). \tag{4.2}$$

Moreover, $\mathbb{E}\,U = N\,\mathbb{E}\,Y + o(N^{1/2}\tau)$ and $\operatorname{Var} U \sim N\tau^2$, and thus also

$$\frac{U - \mathbb{E}\,U}{(\operatorname{Var} U)^{1/2}} \xrightarrow{\mathrm{d}} N(0,1). \tag{4.3}$$

The limits (4.2) *and* (4.3) *hold with convergence of all moments.*

Remark 4.4. Since $\mathbb{E}\,|X - \mathbb{E}\,X|^r \le 2^r\,\mathbb{E}\,|X-a|^r$ for any real a and $r \ge 1$ (a consequence of Minkowski's inequality), it suffices in (iii) to estimate any $\mathbb{E}\,|X-a|^r$, for example $\mathbb{E}\,|X|^r$, and similarly in (vii).

Note further that (viii) is equivalent to $\tau^2 = \Theta(\sigma_Y^2)$, and that τ^2 is unchanged if Y is replaced by $Y + aX + b$ for any real constants a and b (which changes U by the constant $am + bN$ only).

It remains to show that the assumptions of Lemma 4.3 are satisfied with (X,Y) as in Lemma 4.1 for a suitable choice of λ. We begin with some estimates; we state them in greater generality than needed here (although we do not strive for maximal generality), partly in order to stress the properties of the random variables that really are important in our proof.

Lemma 4.5. *Let X be an integer valued random variable and let $p_j = \mathbb{P}(X = j)$. Suppose that $\eta > 0$ is such that there exists a j_0 with $p_{j_0} \ge \eta$ and $p_{j_0+1} \ge \eta$. Then $|\,\mathbb{E}\,e^{isX}| \le 1 - \eta s^2/5$ for $|s| \le \pi$.*

Proof. Let $\theta = \arg \mathbb{E}\,e^{isX}$. Thus, for $|s| \le \pi$,

$$\begin{aligned}
1 - |\,\mathbb{E}\,e^{isX}| &= 1 - \operatorname{Re}\mathbb{E}\,e^{isX - i\theta} = 1 - \operatorname{Re}\sum_j p_j e^{isj - i\theta} = \sum_j p_j\bigl(1 - \cos(js-\theta)\bigr)\\
&\ge \eta\bigl(1 - \cos(j_0 s - \theta) + 1 - \cos((j_0+1)s - \theta)\bigr)\\
&= 2\eta\bigl(1 - \cos\tfrac{s}{2}\cos((j_0+\tfrac12)s - \theta)\bigr) \ge 2\eta\bigl(1-\cos\tfrac{s}{2}\bigr)\\
&\ge 2\eta\frac{s^2}{\pi^2}.
\end{aligned}$$

□

Lemma 4.6. *Let $0 < \gamma < 1$, $\kappa > 0$ and $\lambda_0 > 0$, and let $a_0, a_1, \dots,$ be non-negative real numbers such that*

$$a_j \sim \kappa j^{-\gamma-1}\lambda_0^{-j} \qquad \text{as } j \to \infty. \tag{4.4}$$

Let, for $0 < \lambda \le \lambda_0$, X_λ be a random variable with the distribution

$$\mathbb{P}(X_\lambda = j) = a_j\lambda^j/F(\lambda),$$

where $F(\lambda) = \sum_{j=0}^\infty a_j\lambda^j$. Then $\mathbb{E}\,X_{\lambda_0} = \infty$, but if $\lambda < \lambda_0$, then $0 < \mathbb{E}\,X_\lambda^r < \infty$ for every $r > 0$. Asymptotically, if $r > \gamma$ is fixed, then as $\lambda \uparrow \lambda_0$, with $\kappa_0 = \kappa/F(\lambda_0)$,

$$\mathbb{E}\,X_\lambda^r \sim \kappa_0\Gamma(r-\gamma)(1-\lambda/\lambda_0)^{-(r-\gamma)}. \tag{4.5}$$

In particular, defining $\mu_\lambda := \mathbb{E}\,X_\lambda$ and $\sigma_\lambda^2 := \operatorname{Var} X_\lambda$,

$$\mu_\lambda \sim \kappa_0\Gamma(1-\gamma)(1-\lambda/\lambda_0)^{-(1-\gamma)} \tag{4.6}$$

and thus

$$\sigma_\lambda^2 \sim \mathbb{E}\, X_\lambda^2 \sim \kappa_0^{-1/(1-\gamma)}(1-\gamma)\Gamma(1-\gamma)^{-1/(1-\gamma)}\mu_\lambda^{(2-\gamma)/(1-\gamma)} \tag{4.7}$$

and more generally, for every $r > \gamma$,

$$\mathbb{E}\, X_\lambda^r \sim \kappa_0^{(1-r)/(1-\gamma)}\Gamma(r-\gamma)\Gamma(1-\gamma)^{-(r-\gamma)/(1-\gamma)}\mu_\lambda^{(r-\gamma)/(1-\gamma)}. \tag{4.8}$$

Moreover, there exists a positive constant c such that for $\lambda_0/2 \le \lambda \le \lambda_0$ and $|s| \le \pi$,

$$1 - |\,\mathbb{E}\, e^{isX_\lambda}| \ge c \min\bigl(|s|^\gamma, s^2\sigma_\lambda^2\bigr). \tag{4.9}$$

Proof. The assertions about existence of moments are immediate.

Replacing a_j by $a_j\lambda_0^j$ and λ by λ/λ_0, we may assume that $\lambda_0 = 1$. Further, let $\delta = -\ln\lambda$; note that $\delta \sim 1-\lambda$ as $\lambda \uparrow \lambda_0 = 1$. Then, by dominated convergence,

$$\begin{aligned}
\delta^{r-\gamma}\,\mathbb{E}\, X_\lambda^r &= \delta^{r-\gamma}F(\lambda)^{-1}\sum_{j=0}^\infty j^r a_j e^{-\delta j} \\
&= F(\lambda)^{-1}\int_0^\infty \delta^{r-\gamma}\lfloor x/\delta\rfloor^r a_{\lfloor x/\delta\rfloor} e^{-\delta\lfloor x/\delta\rfloor}\delta^{-1}\,dx \\
&\to F(\lambda_0)^{-1}\int_0^\infty \kappa x^{r-\gamma-1}e^{-x}\,dx = \kappa_0\Gamma(r-\gamma).
\end{aligned}$$

This proves (4.5) and, as a special case, (4.6); together these yield (4.8). It follows further that $(\mathbb{E}\, X_\lambda)^2/\,\mathbb{E}\, X_\lambda^2 \asymp (1-\lambda)^\gamma \to 0$ as $\lambda \uparrow 1$, whence $\sigma_\lambda^2 \sim \mathbb{E}\, X_\lambda^2$ and (4.7) holds.

To prove (4.9), let $\varphi_\lambda(s) = \mathbb{E}\exp(isX_\lambda)$. Let $j_0 \ge 1$ be such that $a_j > 0$ for $j \ge j_0$, and let $c_1 := \inf_{j\ge j_0} j^{\gamma+1}a_j > 0$, $s_0 := j_0^{-1}$, $\lambda_1 := \exp(-s_0)$. First, for any $\lambda \in [1/2, 1]$, we can apply Lemma 4.5 with $\eta = \min(a_{j_0}, a_{j_0+1})2^{-j_0-1}/F(1)$, which implies that for $1/2 \le \lambda \le \lambda_1$ and $|s| \le \pi$,

$$1 - |\varphi_\lambda(s)| \ge \tfrac15\eta s^2 \ge \tfrac15\eta(\mathbb{E}\, X_{\lambda_1}^2)^{-1}s^2\sigma_\lambda^2,$$

and for any $\lambda \ge 1/2$ and $s_0 \le s \le \pi$,

$$1 - |\varphi_\lambda(s)| \ge \tfrac15\eta s^2 \ge \tfrac15\eta s_0^{2-\gamma}s^\gamma;$$

in both cases verifying (4.9) for a suitably small $c > 0$. It remains to consider the case $\lambda_1 < \lambda \le 1$ and $|s| < s_0$; we may further assume $0 < s < s_0$ because $|\varphi_\lambda(-s)| = |\varphi_\lambda(s)|$ and the case $s = 0$ is trivial. Let $\theta = \arg\varphi_\lambda(s)$. Then

$$1 - |\varphi_\lambda(s)| = 1 - \operatorname{Re}\bigl(\varphi_\lambda(s)e^{-i\theta}\bigr) = \sum_{j=0}^\infty F(\lambda)^{-1}e^{-j\delta}a_j\operatorname{Re}(1 - e^{ijs-i\theta}). \tag{4.10}$$

Let $J = \min\bigl(\frac1s, \frac1\delta\bigr) \ge j_0$, $I_1 = [J, 2J]$ and $I_2 = [4J, 5J]$. The sets $\{e^{its-i\theta} : t \in I_k\}$, $k = 1, 2$, are two intervals of length $Js \le 1$ on the unit circle, separated by $2Js$ (note that $6Js \le 2\pi$); hence at least one of them is disjoint from $\{e^{iu} : |u| < Js\}$, which implies that for some choice of k (1 or 2) and every

$t \in I_k$, $\cos(ts - \theta) \le \cos(Js) \le 1 - \frac{1}{3}J^2s^2$. Consequently, (4.10) yields, for some $c_2, c_3 > 0$,

$$\begin{aligned}
1 - |\varphi_\lambda(s)| &\ge \sum_{j\in I_k} e^{-j\delta} a_j \bigl(1 - \cos(js - \theta)\bigr)/F(\lambda) \ge \sum_{j\in I_k} e^{-5} a_j \tfrac{1}{3} J^2 s^2 / F(\lambda_0) \\
&= c_2 J^2 s^2 \sum_{j\in I_k} a_j \ge c_2 J^2 s^2 c_1 \sum_{j\in I_k} j^{-\gamma-1} \ge c_1 c_2 J^2 s^2 \lfloor J \rfloor (5J)^{-\gamma-1} \\
&\ge c_3 s^2 J^{2-\gamma} = c_3 \min\bigl(s^\gamma, s^2 \delta^{\gamma-2}\bigr).
\end{aligned}$$

Since $\sigma_\lambda^2 \asymp \delta^{\gamma-2}$ by (4.5), (4.9) holds in this case too if $c > 0$ is small enough. $\square$

Proof of Theorem 1.1(ii). We change the notation slightly, and let, for $0 < \lambda < e^{-1}$, (X_λ, Y_λ) be a random vector with the distribution defined in Lemma 4.1 (there denoted (X_j, Y_j)). Thus X_λ has the Borel distribution (4.1), with probability generating function $\mathbb{E}\, z^{X_\lambda} = T(\lambda z)/T(\lambda)$, where T is the tree function (3.7).

It is a well-known fact (also for much more general exponential families of distributions) that $\lambda \mapsto \mathbb{E}\, X_\lambda$ is a continuous, strictly increasing function of $\lambda \in (0, e^{-1})$. [Sketch of proof: $\mathbb{E}\, X_\lambda = \lambda T'(\lambda)/T(\lambda)$ which shows continuity, and if $0 < \lambda < \lambda_1 < e^{-1}$ and $b = \lambda_1/\lambda > 1$, then $\mathbb{E}\, X_{\lambda_1} = \mathbb{E}\, X_\lambda b^{X_\lambda} / \mathbb{E}\, b^{X_\lambda} > \mathbb{E}\, X_\lambda$ by the FKG-inequality (calculate $\mathbb{E}(X' - X'')(b^{X'} - b^{X''}) > 0$ for two independent copies X' and X'' of X_λ).] Since further $\mathbb{E}\, X_\lambda \to 1$ as $\lambda \downarrow 0$ and $\mathbb{E}\, X_\lambda \to \infty$ as $\lambda \uparrow e^{-1}$, there exists for every $\mu > 1$ a unique $\lambda(\mu) \in (0, e^{-1})$ such that $\mathbb{E}\, X_{\lambda(\mu)} = \mu$, and the function $\mu \mapsto \lambda(\mu)$ is continuous.

Similarly, also higher moments $\mathbb{E}\, X_\lambda^r$ are continuous (and increasing) functions of λ.

For n and m with $0 < n < m$, we apply Lemma 4.3 with $N = m - n$ and $(X, Y) = (X_\lambda, Y_\lambda)$ for $\lambda = \lambda(m/N)$. Thus condition (i) holds by construction. (Actually, (4.32) below implies the explicit formula $\lambda = (n/m)e^{-n/m}$, but we do not need this.) Lemma 4.1 shows that D_{mn} has the same distribution as U, so we may take $U = D_{mn}$.

In order to verify the remaining conditions, we consider three subcases separately: $n/m \to 0$, $n/m \to a$ with $0 < a < 1$ (the case studied by [9]), and $n/m \to 1$. (It suffices to consider these three subcases, although they do not exhaust all possibilities, since every sequence (m_k, n_k) with $m_k \to \infty$ has a subsequence belonging to one of the subcases; cf. Remark 1.3.)

Case 1: $n/m \to 0$; $m/N \to 1$.
We verify the conditions of Lemma 4.3 with $\gamma = 2$.

In this case $\lambda = \lambda(m/N) \to 0$, and thus $T(\lambda) \sim \lambda$. We have

$$\begin{aligned}
\mathbb{P}(X = 1) &= \lambda/T(\lambda) \to 1, \\
\mathbb{P}(X = 2) &= \lambda^2/T(\lambda) \sim \lambda, \\
\mathbb{P}(X = 3) &= \tfrac{9}{6}\lambda^3/T(\lambda) \sim \tfrac{3}{2}\lambda^2.
\end{aligned}$$

Hence,

$$\frac{n}{N} = \frac{m}{N} - 1 = \mathbb{E}(X-1) \sim \mathbb{P}(X=2) \sim \lambda$$

and thus $\lambda \sim n/N \sim n/m$. Moreover,

$$\mathbb{E}\,|X-1|^r \sim \mathbb{P}(X=2) \sim \lambda \sim n/m$$

for every $r > 0$; in particular

$$\operatorname{Var} X = \operatorname{Var}(X-1) \sim \mathbb{E}(X-1)^2 \sim n/m,$$

which implies conditions (ii) and (iv) (with $\gamma = 2$), and further, for any $r > 2$, using $\lambda^{-1} \sim m/n = o(m)$,

$$\mathbb{E}\,|X-1|^r/\sigma_X^r \sim \lambda/\lambda^{r/2} = \lambda^{-(r/2-1)} = o(m^{r/2-1}) = o(N^{r/2-1}),$$

which yields (iii), cf. Remark 4.4. Since $\min\bigl(\mathbb{P}(X=1), \mathbb{P}(X=2)\bigr) \sim \lambda \sim \sigma_X^2$, Lemma 4.5 shows that (v) holds too.

For Y we have, from the definition, $Y = 0$ when $X \le 2$, and $\mathbb{P}(Y = 0 \mid X = 3) = 2/3$, $\mathbb{P}(Y = 1 \mid X = 3) = 1/3$; thus, for every $r > 0$,

$$\mathbb{E}\,Y^r = \tfrac{1}{3}\,\mathbb{P}(X=3) + O(\lambda^3) \sim \tfrac{1}{2}\lambda^2. \tag{4.11}$$

Hence, $\sigma_Y^2 \sim \frac{1}{2}\lambda^2 \sim \frac{1}{2}(n/m)^2$, and for every $r > 2$, now using $\lambda^{-1} \sim m/n = o(m^{1/2})$ (by the assumption $n \gg m^{1/2}$)

$$\mathbb{E}\,Y^r/\sigma_Y^r = O(\lambda^{-(r-2)}) = o(m^{r/2-1}) = o(N^{r/2-1}),$$

so (vi) and (vii) hold.

Finally, $\mathbb{E}(XY) \sim \frac{1}{3}\,\mathbb{P}(X=3) \cdot 3 \sim \frac{3}{2}\lambda^2$ and thus

$$\rho = \operatorname{Cov}(X,Y)/\sigma_X\sigma_Y = O(\lambda^2/\lambda^{1/2}\lambda) = O(\lambda^{1/2}),$$

so $\rho \to 0$ and (viii) holds.

Consequently, Lemma 4.3 applies and shows $(D_{mn} - \mathbb{E}\,D_{mn})/(\operatorname{Var} D_{mn})^{1/2} \xrightarrow{\mathrm{d}} N(0,1)$, with convergence of all moments. Note, for future use, that

$$\tau^2 \sim \sigma_Y^2 \sim \tfrac{1}{2}\lambda^2 \sim \frac{n^2}{2m^2}. \tag{4.12}$$

Case 2: $n/m \to a$, $0 < a < 1$; $m/N \to b := 1/(1-a)$.
Again we take $\gamma = 2$. In this case $\lambda = \lambda(m/N) \to \lambda(b)$, and thus the distribution of (X,Y) converges to $(X_{\lambda(b)}, Y_{\lambda(b)})$, together with all moments; in particular, $\sigma_X^2 \to \operatorname{Var}(X_{\lambda(b)}) > 0$. It is easily verified that all assumptions of Lemma 4.3 hold, cf. [12, Corollary 2.1]; note that (v) follows from Lemma 4.5 and that (viii) follows because the correlation coefficient $\rho(X_{\lambda(b)}, Y_{\lambda(b)})$ does not equal ± 1 since both $\{X_{\lambda(b)} = 3,\ Y_{\lambda(b)} = 0\}$ and $\{X_{\lambda(b)} = 3,\ Y_{\lambda(b)} = 1\}$ have positive probabilities. Thus the result follows from Lemma 4.3.

Case 3: $n/m \to 1$; $m/N \to \infty$.
In this case, $\lambda \to \lambda_0 = e^{-1}$ and we verify the conditions with $\gamma = 1/2$.

We are in the set-up of Lemma 4.6, with $a_j = j^{j-1}/j!$, $j \ge 1$, and $F(\lambda) = T(\lambda)$, the tree function in (3.7). By Stirling's formula, $a_j \sim (2\pi)^{-1/2} j^{-3/2} e^j$ as $j \to \infty$, so (4.4) holds with $\gamma = 1/2$, $\kappa = (2\pi)^{-1/2}$ and $\lambda_0 = e^{-1}$; we further have, as is well-known, $F(\lambda_0) = T(e^{-1}) = 1$, so $\kappa_0 = \kappa$. Hence, Lemma 4.6 applies, which by (4.9) yields (v). Moreover, it shows that for every $r > 1/2$,

$$\mathbb{E}\, X_\lambda^r \sim \frac{\Gamma(r-1/2)}{\sqrt{2\pi}}(1-e\lambda)^{1/2-r}. \tag{4.13}$$

In particular, cf. the exact formulae (4.25), (4.30) below,

$$\mu_\lambda = \mathbb{E}\, X_\lambda \sim 2^{-1/2}(1-e\lambda)^{-1/2}, \tag{4.14}$$
$$\sigma_\lambda^2 \sim \mathbb{E}\, X_\lambda^2 \sim 2^{-3/2}(1-e\lambda)^{-3/2} \sim \mu_\lambda^3 = (m/N)^3. \tag{4.15}$$

By assumption, $N^2 \gg n$, and thus $\mu_\lambda = m/N \sim n/N \ll N$, which yields $\sigma_\lambda^2 = O(\mu_\lambda^3) = O(N^3)$, i.e. (iv). Similarly, for $r > 2$,

$$\mathbb{E}\, X_\lambda^r/\sigma_\lambda^r = O(\mu_\lambda^{2r-1}/\mu_\lambda^{3r/2}) = O(\mu_\lambda^{r/2-1}) = o(N^{r/2-1}),$$

which verifies (iii).

Next, by the construction of Y_λ,

$$\mathbb{E}(Y_\lambda^r \mid X_\lambda = \ell) = \mathbb{E}\, D_{\ell,\ell-1}^r,$$

and by the already proved Theorem 1.1(iii), for every $r > 0$,

$$\ell^{-3r/2}\,\mathbb{E}\, D_{\ell,\ell-1}^r \to \mathbb{E}\, W_0^r \qquad \text{as } \ell \to \infty.$$

Hence, fixing r, for every $\varepsilon > 0$ there exists ℓ_ε such that

$$|\,\mathbb{E}\, D_{\ell,\ell-1}^r - \ell^{3r/2}\,\mathbb{E}\, W_0^r| < \varepsilon \ell^{3r/2} \qquad \text{for } \ell \ge \ell_\varepsilon;$$

letting C_ε be the maximum of the left hand side for $1 \le \ell < \ell_\varepsilon$, we see that for every ℓ

$$(\mathbb{E}\, W_0^r - \varepsilon)\ell^{3r/2} - C_\varepsilon \le \mathbb{E}(Y_\lambda^r \mid X_\lambda = \ell) = \mathbb{E}\, D_{\ell,\ell-1}^r \le (\mathbb{E}\, W_0^r + \varepsilon)\ell^{3r/2} + C_\varepsilon$$

and thus

$$(\mathbb{E}\, W_0^r - \varepsilon)X_\lambda^{3r/2} - C_\varepsilon \le \mathbb{E}(Y_\lambda^r \mid X_\lambda) \le (\mathbb{E}\, W_0^r + \varepsilon)X_\lambda^{3r/2} + C_\varepsilon, \tag{4.16}$$

which yields, by taking the expectation,

$$(\mathbb{E}\, W_0^r - \varepsilon)\,\mathbb{E}\, X_\lambda^{3r/2} - C_\varepsilon \le \mathbb{E}\, Y_\lambda^r \le (\mathbb{E}\, W_0^r + \varepsilon)\,\mathbb{E}\, X_\lambda^{3r/2} + C_\varepsilon.$$

Together with (4.8) this easily implies that for every $r > 1/3$, as $\lambda \to e^{-1}$,

$$\begin{aligned}\mathbb{E}\, Y_\lambda^r &\sim \mathbb{E}\, W_0^r\,\mathbb{E}\, X_\lambda^{3r/2} \sim \mathbb{E}\, W_0^r\,\kappa_0^{2-3r}\Gamma(3r/2-1/2)\Gamma(1/2)^{1-3r}\mu_\lambda^{3r-1}\\ &= 2^{3r/2-1}\pi^{-1/2}\Gamma(3r/2-1/2)\,\mathbb{E}\, W_0^r\,\mu_\lambda^{3r-1}.\end{aligned} \tag{4.17}$$

More generally, by first multiplying (4.16) by X_λ^s, it follows similarly that if $s, r \ge 0$ with $3r/2 + s > 1/2$, then

$$\begin{aligned}\mathbb{E}\, X_\lambda^s Y_\lambda^r &\sim \mathbb{E}\, W_0^r\,\mathbb{E}\, X_\lambda^{s+3r/2}\\ &\sim \kappa_0^{2-2s-3r}\Gamma(s+3r/2-1/2)\Gamma(1/2)^{1-2s-3r}\mu_\lambda^{2s+3r-1}\,\mathbb{E}\, W_0^r\\ &= 2^{s+3r/2-1}\pi^{-1/2}\Gamma(s+3r/2-1/2)\,\mathbb{E}\, W_0^r\mu_\lambda^{2s+3r-1}.\end{aligned} \tag{4.18}$$

In particular, using $\mathbb{E}\, W_0 = \sqrt{\pi/8}$ and $\mathbb{E}\, W_0^2 = 5/12$ [20, 9],

$$\mathbb{E}\, Y_\lambda \sim \sqrt{2/\pi}\, \mathbb{E}\, W_0 \mu_\lambda^2 = \tfrac{1}{2}(m/N)^2 \tag{4.19}$$

$$\sigma_Y^2 \sim \mathbb{E}\, Y_\lambda^2 \sim 4\pi^{-1/2}\Gamma(5/2)\, \mathbb{E}\, W_0^2 \mu_\lambda^5 = 3\, \mathbb{E}\, W_0^2 \mu_\lambda^5 = \tfrac{5}{4}(m/N)^5 \tag{4.20}$$

and, by (4.18),

$$\mathbb{E}\, X_\lambda Y_\lambda \sim 2^{3/2}\pi^{-1/2}\Gamma(2)\, \mathbb{E}\, W_0 \mu_\lambda^4 = \mu_\lambda^4 = (m/N)^4.$$

Thus, $\operatorname{Cov}(X_\lambda, Y_\lambda) \sim \mathbb{E}\, X_\lambda Y_\lambda \sim \mu_\lambda^4 = (m/N)^4$ and

$$\rho \sim \frac{\mathbb{E}\, X_\lambda Y_\lambda}{(\mathbb{E}\, X_\lambda^2\, \mathbb{E}\, Y_\lambda^2)^{1/2}} \sim \frac{\mu_\lambda^4}{(\mu_\lambda^3 \frac{5}{4} \mu_\lambda^5)^{1/2}} = \sqrt{\tfrac{4}{5}},$$

which shows (viii). Furthermore,

$$\tau^2 \sim \bigl(3\, \mathbb{E}\, W_0^2 - \tfrac{8}{\pi}(\mathbb{E}\, W_0)^2\bigr)\mu_\lambda^5 = \tfrac{1}{4}\Bigl(\frac{m}{N}\Bigr)^5. \tag{4.21}$$

Finally, for $r \geq 3$, by (4.20) and (4.17),

$$\mathbb{E}\, Y_\lambda^r / \sigma_Y^r = O(\mu_\lambda^{3r-1}/\mu_\lambda^{5r/2}) = O(\mu_\lambda^{r/2-1}) = o(N^{r/2-1}),$$

which verifies (vii), and again the result follows by Lemma 4.3. □

Proof of Theorem 1.4 (ii). In the case $n/m \to 0$, Lemma 4.3 and (4.11), (4.12) show that

$$\mathbb{E}\, D_{mn} = N\, \mathbb{E}\, Y + o(N^{1/2}\tau) \sim \frac{n^2}{2m},$$

$$\operatorname{Var} D_{mn} \sim N\tau^2 \sim \frac{n^2}{2m},$$

verifying Theorem 1.4(i) and (ii) when $m^{1/2} \ll n \ll m$.

Similarly, when $n/m \to 1$ and $m - n \gg m^{1/2}$, Lemma 4.3 and (4.19), (4.21) yield Theorem 1.4(ii) for this case.

In the case $\alpha = n/m \to a \in (0, 1)$, finally, it follows from Lemma 4.3 that $\mathbb{E}\, D_{mn}$ and $\operatorname{Var} D_{mn}$ are asymptotically proportional to N, and thus to n. In order to obtain explicit expressions, we argue as follows, using the generating functions explored in Section 3. (As stated in Section 1, these asymptotics were found by [14] and [9], respectively, directly from the exact formulae. Nevertheless, we find the alternative proof given here interesting.)

By the definition of Y_λ, (3.1), (4.1), (3.2) and (3.3),

$$\begin{aligned}\mathbb{E}\, w^{Y_\lambda} &= \sum_{\ell=1}^{\infty} \mathbb{E}\, w^{D_{\ell,\ell-1}}\, \mathbb{P}(X_\lambda = \ell) = \sum_{\ell=1}^{\infty} \frac{F_{\ell,\ell-1}(w)}{F_{\ell,\ell-1}(1)} \frac{\ell^{\ell-1}\lambda^\ell}{\ell!\, T(\lambda)} \\ &= \sum_{\ell=1}^{\infty} F_{\ell,\ell-1}(w) \frac{\lambda^\ell}{(\ell-1)!\, T(\lambda)} = \frac{\lambda}{T(\lambda)} F(w, \lambda)\end{aligned}$$

and thus, by (3.6) and (3.10), for $k = 0, 1, \dots,$

$$\mathbb{E} \binom{Y_\lambda}{k} = [w^k]\, \mathbb{E}(1+w)^{Y_\lambda} = \frac{\lambda}{T(\lambda)} [w^k] F(1+w, \lambda) = \frac{\lambda}{T(\lambda)} W_k'(\lambda) = f_k(T(\lambda)). \tag{4.22}$$

More generally, we similarly obtain, for $j = 0, 1, \dots,$

$$\mathbb{E}\big(w^{Y_\lambda} X_\lambda^j\big) = \sum_{\ell=1}^{\infty} F_{\ell,\ell-1}(w)\ell^j \frac{\lambda^\ell}{(\ell-1)!\,T(\lambda)} = \frac{1}{T(\lambda)}\Big(\lambda\frac{d}{d\lambda}\Big)^j \Big(\lambda F(w,\lambda)\Big)$$

and thus

$$\begin{aligned}\mathbb{E}\Big(\binom{Y_\lambda}{k} X_\lambda^j\Big) &= [w^k]\frac{1}{T(\lambda)}\Big(\lambda\frac{d}{d\lambda}\Big)^j \Big(\lambda F(1+w,\lambda)\Big)\\ &= \frac{1}{T(\lambda)}\Big(\lambda\frac{d}{d\lambda}\Big)^j \Big(T(\lambda) f_k(T(\lambda))\Big). \qquad (4.23)\end{aligned}$$

For any differentiable function h, we have

$$\lambda\frac{d}{d\lambda}\big(h(T(\lambda)\big) = \lambda T'(\lambda)h'(T(\lambda)) = \frac{T(\lambda)}{1-T(\lambda)}h'(T(\lambda)); \qquad (4.24)$$

in other words, $\lambda\frac{d}{d\lambda} = \frac{T}{1-T}\frac{d}{dT}$. Hence, (4.22), (4.23) and (3.13) yield by simple calculations, dropping the λ from the notation,

$$\mathbb{E}\,X = \frac{1}{T}\Big(\lambda\frac{d}{d\lambda}\Big)T = \frac{1}{T}\frac{T}{1-T}\frac{d}{dT}T = \frac{1}{1-T}, \qquad (4.25)$$

$$\mathbb{E}\,X^2 = \frac{1}{T}\Big(\frac{T}{1-T}\frac{d}{dT}\Big)^2 T = \frac{1}{(1-T)^3}, \qquad (4.26)$$

$$\mathbb{E}\,Y = f_1(T) = \frac{T^2}{2(1-T)^2}, \qquad (4.27)$$

$$\mathbb{E}\,Y^2 = 2\,\mathbb{E}\binom{Y}{2} + \mathbb{E}\,Y = 2f_2(T) + f_1(T) = \frac{6T^2+6T^3+7T^4-4T^5}{12(1-T)^5}, \qquad (4.28)$$

$$\mathbb{E}\,XY = \frac{1}{T}\frac{T}{1-T}\frac{d}{dT}\big(Tf_1(T)\big) = \frac{3T^2-T^3}{2(1-T)^4}, \qquad (4.29)$$

which by further straightforward calculations lead to

$$\operatorname{Var} X = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 = \frac{T}{(1-T)^3}, \qquad (4.30)$$

$$\tau^2 = \operatorname{Var} Y - \operatorname{Cov}(X,Y)^2/\operatorname{Var} X = \frac{6T^2-6T^3+4T^4-T^5}{12(1-T)^5}. \qquad (4.31)$$

The condition $\mathbb{E}\,X = m/N$ and (4.25) yield $1-T = N/m$ and thus

$$T = n/m = \alpha. \qquad (4.32)$$

Lemma 4.3 and (4.27), (4.31), (4.32) now yield (1.1) and (1.2). $\square$

5. The very sparse case: Poisson behaviour

Theorem 1.1(i) is much simpler than the other parts and is given mainly for completeness. It too can be shown using Lemma 4.1 (for example using Holst [11, Corollary 3.5]), but we prefer a direct approach, using a related occupancy problem.

Let D'_{mn} be the number of cells where at least two items make their first try, i.e. using the notation of Section 2, the number of j with $X_j \geq 2$. It is easily seen that if $X_j + X_{j+1} \leq 2$ for all j, then no item is displaced more than one step and $D_{mn} = D'_{mn}$. Consequently, using symmetry,

$$\mathbb{P}(D_{mn} \neq D'_{mn}) \leq m\,\mathbb{P}(X_1 + X_2 \geq 3) \leq mn^3\,\mathbb{P}(h_1, h_2, h_3 \in \{1,2\}) = 8\frac{n^3}{m^2} \to 0. \tag{5.1}$$

Moreover, it is easy to check by the method of moments or by Stein's method, see for example [2, Theorem 6.B], that $D'_{mn} \xrightarrow{\mathrm{d}} \mathrm{Po}(a^2/2)$. By (5.1), then $D_{mn} \xrightarrow{\mathrm{d}} \mathrm{Po}(a^2/2)$ too.

Remark 5.1. The argument shows more generally Poisson convergence in the form $\mathrm{d_{TV}}\big(D_{mn}, \mathrm{Po}(n^2/2m)\big) \to 0$, where $\mathrm{d_{TV}}$ denotes the total variation distance [2], even for $n^2/2m \to \infty$ as long as $n = o(m^{2/3})$.

Remark 5.2. Instead of approximating D_{mn} by D'_{mn}, we could just as well use the number of pairs (i,j), $i < j$ with $h_i = h_j$; this is a variable arising in birthday problems, and again it is easy to prove that it is asymptotically Poisson distributed, see e.g. [2, Theorem 5.G (with Γ the complete graph K_n)].

To show moment convergence, it suffices by Remark 1.2 to show that $\mathbb{E}\,D^r_{mn} = O(1)$ for each r. This can presumably be verified by a direct combinatorial analysis, but we argue instead as follows.

Suppose to the contrary that there is an integer $r \geq 1$ such that $\mathbb{E}\,D^r_{mn}$ is unbounded; then there is a sequence (m_k, n_k) with $n_k^2/m_k \to a^2$ and a sequence $\omega_k \to \infty$ such that $\mathbb{E}\,D^r_{m_k n_k} \geq \omega_k^{2r}$ for all k. We can further assume $\omega_k \ll \sqrt{m_k}$. Define $n'_k = \lfloor \omega_k m_k^{1/2} \rfloor$. Then $n'_k > n_k$ for large k, and thus $\mathbb{E}\,D^r_{m_k n'_k} \geq \mathbb{E}\,D^r_{m_k n_k} \geq \omega_k^{2r}$. On the other hand, Theorems 1.1(ii) and 1.4(ii) apply to $D_{m_k n'_k}$, and it follows from the moment convergence that

$$\mathbb{E}\,D^r_{m_k n'_k} \sim (\mathbb{E}\,D_{m_k n'_k})^r \sim \Big(\frac{(n'_k)^2}{2m_k}\Big)^r \sim 2^{-r}\omega_k^{2r}.$$

This yields the desired contradiction, proving $\mathbb{E}\,D^r_{mn} = O(1)$ and completing the proof of Theorem 1.1(i).

The moment estimates in Theorem 1.4(i) now follow for the case $n/m^{1/2} \to a > 0$. The case $n/m^{1/2} \to \infty$ was treated in Section 4, but it remains to consider the rather trivial case $n/m^{1/2} \to 0$, when $\mathbb{P}(D_{mn} \neq 0) \sim n^2/2m \to 0$. As remarked in the introduction, the exact formula for $\mathbb{E}\,D_{mn}$ easily yields $\mathbb{E}\,D_{mn} \sim n^2/2m$ in this case. We do not know any simple argument for the variance, but the exact formula for $\mathbb{E}\,D^2_{mn}$ in [9, Theorem 4] yields after straightforward (but tedious) calculations $\mathbb{E}\,D^2_{mn} \sim n^2/2m$ too, as required.

6. Asymptotics for the limits W_a

In this section we study the asymptotics of the distribution of the limit variables W_a as $a \to \infty$.

Theorem 6.1. *As $a \to \infty$, we have* $\mathbb{E}\, W_a \sim \frac{1}{2}a^{-1}$, $\operatorname{Var} W_a \sim \frac{1}{4}a^{-4}$ *and*

$$\frac{W_a - \mathbb{E}\, W_a}{(\operatorname{Var} W_a)^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \tag{6.1}$$

with convergence of all moments.

Proof. In this proof we use the notation $\widetilde{X} := (X - \mathbb{E}\, X)/(\operatorname{Var} X)^{1/2}$ for the standardization of a random variable X.

Theorem 1.1(iii) shows that, for any $a > 0$, $m^{-3/2} D_{m,m-\lfloor am^{1/2}\rfloor} \xrightarrow{\mathrm{d}} W_a$ with convergence of all moments, and thus also

$$\widetilde{D}_{m,m-\lfloor am^{1/2}\rfloor} = \bigl(m^{-3/2} D_{m,m-\lfloor am^{1/2}\rfloor}\bigr)^{\sim} \xrightarrow{\mathrm{d}} \widetilde{W}_a. \tag{6.2}$$

The space of all probability distributions on $\mathbb{R}$ is metrizable (see e.g. [4, Appendix III]); let d denote a metric on this space (for example the well-known Lévy metric, but any metric will do). If X and Y are random variables, we write $d(X,Y)$ for the distance between their distributions. Then (6.2) shows that for every $a > 0$, there is an integer $m(a)$ such that defining $n(a) := m(a) - \lfloor am(a)^{1/2}\rfloor$ we have

$$d(\widetilde{D}_{m(a),n(a)}, \widetilde{W}_a) < a^{-1}. \tag{6.3}$$

We may further assume $m(a) > 4a^2$, and thus $m(a) - n(a) \le am(a)^{1/2} \le \frac{1}{2}m(a)$.

Now let $a \to \infty$. Then $m(a) \to \infty$, $n(a) \ge \frac{1}{2}m(a)$ and $m(a) - n(a) \gg m(a)^{1/2}$, and thus by Theorem 1.1(ii)

$$d(\widetilde{D}_{m(a),n(a)}, N(0,1)) \to 0. \tag{6.4}$$

Combining (6.3) and (6.4) yields $d(\widetilde{W}_a, N(0,1)) \to 0$, which proves (6.1).

To prove moment convergence, we use the same argument, now taking $d(X,Y) := |\,\mathbb{E}\, X^r - \mathbb{E}\, Y^r|$ for a fixed integer r.

Finally, Theorem 1.4(iii) shows that if $m \to \infty$ and $n = m - \lfloor am^{1/2}\rfloor$, then

$$(m-n)n^{-2}\,\mathbb{E}\, D_{m,n} \to a\,\mathbb{E}\, W_a$$

and

$$(m-n)^4 n^{-5} \operatorname{Var} D_{m,n} \to a^4 \operatorname{Var} W_a,$$

which by a similar argument and Theorem 1.4(ii) yield $a\,\mathbb{E}\, W_a \to 1/2$ and $a^4 \operatorname{Var} W_a \to 1/4$ as $n \to \infty$. □

More precise estimates of the moments of W_a are easily obtained using the formulae in Section 3. Indeed, a Taylor expansion of $e^{-x^2/2}$ in the definition (3.19) yields

$$\begin{aligned} q_r(a) &= \frac{1}{r!}\int_0^\infty x^r e^{-ax}\bigl(1 - x^2/2 + O(x^4)\bigr)\,dx \\ &= a^{-r-1} - \frac{(r+1)(r+2)}{2}a^{-r-3} + O(a^{-r-5}). \end{aligned} \tag{6.5}$$

Consequently, Theorem 3.3 yields

$$\mathbb{E}\, W_a = \tfrac{1}{2} q_0(a) = \tfrac{1}{2} a^{-1} - \tfrac{1}{2} a^{-3} + O(a^{-5}), \tag{6.6}$$

$$\mathbb{E}\, W_a^2 = \tfrac{5}{4} q_3(a) + \tfrac{1}{4} a q_2(a) = \tfrac{1}{4} a^{-2} - \tfrac{1}{4} a^{-4} + O(a^{-6}), \tag{6.7}$$

and thus

$$\operatorname{Var} W_a = \mathbb{E}\, W_a^2 - (\mathbb{E}\, W_a)^2 = \tfrac{1}{4} a^{-4} + O(a^{-6}). \tag{6.8}$$

The same method yields further terms in (6.5)–(6.8), giving asymptotic expansions of $\mathbb{E}\, W_a$ and $\operatorname{Var} W_a$ in powers of a^{-1} up to an arbitrary degree, but we leave the details to the reader. The method yields asymptotics for higher moments too.

Note that by (6.6) and (6.8), the distributional limit (6.1) can be written

$$2a^2(W_a - 1/2a) \xrightarrow{\mathrm{d}} N(0,1), \qquad \text{as } a \to \infty.$$

7. Unsuccessful search

In an unsuccessful search, we start searching at a random cell h and probe successive cells until we reach an empty cell when we give up. (We assume throughout this section that $n < m$ so that there is at least one empty cell.) The number of probes used when starting in a block of length ℓ thus ranges from 1 to ℓ, and if the hash tables have block lengths $\ell_1, \dots, \ell_N$, with $N = m-n$ and $\sum_i \ell_i = m$, the average unsuccessful search time U_{mn} is given by

$$U_{mn} = \frac{1}{m} \sum_{j=1}^{N} \sum_{i=1}^{\ell_j} i = \frac{1}{m} \sum_{j=1}^{N} \binom{\ell_j + 1}{2} = \frac{1}{2m} \widehat{U}_{mn} + \frac{1}{2}, \tag{7.1}$$

where we for convenience define

$$\widehat{U}_{mn} := \sum_{j=1}^{N} \ell_j^2.$$

Note that for given m and n, $\widehat{U}_{mn}$, and thus U_{mn}, is largest when one block has length $n+1$ and the others length 1, and smallest when all block lengths are as equal as possible, i.e. when all are $\lfloor m/(m-n) \rfloor$ or $\lceil m/(m-n) \rceil$.

Brownian limits. First, we adapt the Brownian approach in Section 2, assuming $n < m$ and $(m-n)/\sqrt{n} \to a \geq 0$. The empty cells occur when $H_i = 0$, and thus the block lengths, normalized as $(\ell_i - 1)/m$, are the lengths of the excursions (i.e. the zero-free intervals) of the random function $H_{\lfloor mt \rfloor}$ or $m^{-1/2} H_{\lfloor mt \rfloor}$. By (2.2), the latter random function converges in distribution to $Y_a(t) := \max_{s \leq t} \bigl(b(t) - b(s) - a(t-s)\bigr)$, and it is reasonable to conjecture that the lengths of its excursions converge to the lengths of the excursions of Y_a.

(We consider the excursions in an interval $[t_0, t_0+1]$ with $Y_a(t_0) = 0$; equivalently, we consider $[0,1]$ but allow an excursion to wrap around from 1 to 0.) It follows from a result by Vervaat [28], see Remark 2.3, that these have the same distribution as the lengths of the excursions of $Z_a(t) := \max_{0 \leq s \leq t} \bigl(e(t) - e(s) - a(t-s)\bigr)$ in $[0,1]$.

However, the convergence of the lengths does *not* follow from the argument above alone, since taking the excursion lengths is not a continuous operation; nevertheless, it has been verified by Chassaing and Louchard [6]. More precisely, they show in [6] that if $(L_i)_{i=1}^\infty$ is the sequence consisting of the block lengths $\ell_1,\dots,\ell_N$ arranged in decreasing order, followed by infinitely many zeroes, and $(J_i)_{i=1}^\infty$ is the sequence of the excursion lengths of Z_a arranged in decreasing order, then $(L_i/m)_1^\infty \xrightarrow{\mathrm{d}} (J_i)_1^\infty$ as random elements of ℓ^1. Since $(x_i)\mapsto\sum x_i^2$ is a continuous functional on ℓ^1, this immediately yields

$$\widehat{U}_{mn}/m^2=\sum_1^\infty (L_i/m)^2 \xrightarrow{\mathrm{d}} \sum_1^\infty J_1^2,$$

which by (7.1) yields Theorem 1.6(iii) with the following description of the limit. (Moment convergence is immediate since $0\le \widehat{U}_{mn}/m^2\le 1$.)

Theorem 7.1. *The limit V_a can be constructed as the sum of the squares of the excursion lengths of the stochastic process*

$$Z_a(t):=\max_{0\le s\le t}\bigl(e(t)-e(s)-a(t-s)\bigr),\qquad 0\le t\le 1. \qquad \Lambda$$

As remarked above, Z_a can here be replaced by Y_a defined above. Moreover, the excursion lengths of Z_a or Y_a have several different, equivalent descriptions, which lead to the following alternative characterizations of V_a, see further [1, 3, 5, 6, 23, 24]. (We exclude the trivial case $a=0$ when $V_a=1$.)

Theorem 7.2. *Let $0<a<\infty$. The limit V_a can be constructed as any of the following random variables.*

(i) *The sum of the squares of the excursion lengths of a Brownian bridge on $[0,1]$ conditioned on having local time a at 1.*
(ii) *The sum of the squares of the excursion lengths of a Brownian motion on $[0,1]$ conditioned on having local time a at 1.*
(iii) *The sum of the squares of the jumps of a standard stable subordinator of index $1/2$ on $[0,a]$ conditioned on having value 1 at a.* (*Note that this value equals the sum of the jumps.*)
(iv) *The sum of the squares of a^2 times the jumps of a standard stable subordinator of index $1/2$ on $[0,1]$ conditioned on having value a^{-2} at 1.*
(v) *The sum $\sum x_i^2$ of the squares of the points in a Poisson process $\{x_i\}_1^\infty$ on $(0,\infty)$ with intensity $a/\sqrt{2\pi x^3}$, conditioned on $\sum x_i=a$.*
(vi) *The sum of the squares of the component sizes of $\mathbf{X}(-\log a)$, where $\mathbf{X}(t)$ denotes the standard additive coalescent* [1].
(vii) *Let $\xi_1,\xi_2,\dots$ be independent standard normal variables and define $S_k=\sum_1^k\xi_i^2$ (with $S_0=0$) and $R_k=\frac{a^2}{S_{k-1}+a^2}-\frac{a^2}{S_k+a^2}$; then define $V_a=\sum_1^\infty R_k^2$.*

Proof. The equivalence of the seven constructions is well-known, also on the level of random sequences of lengths, jumps, etc. More specifically, first it is well-known, cf. [25, §VI.2 and §XII.2], that the excursion lengths of a Brownian motion in $[0,1]$ are the jumps of the inverse $\tau_s := \inf\{t : T_t > s\}$ of the local time process T_t in the interval $0 \le s \ge T_1$, that τ_s is a stable subordinator of index $1/2$, and that the sizes of the jumps of τ_s for $0 \le s \le a$ are given by a Poisson process on $(0,\infty)$ with intensity $a/\sqrt{2\pi x^3}$. The equivalence of (i), (iii) and (v) now follows easily, cf. e.g. [24]. Moreover, a simple rescaling yields the equivalence of (iii) and (iv). By [24, Theorem 5.1], (i) and (ii) are equivalent. The equivalence of (iv), (vi) and (vii) follows by [1, Theorems 3, 4 and Corollary 5].

Finally, these constructions may be connected to Theorem 7.1 in several ways. First, [6] gives a direct proof that the normalized block lengths L_i/m, taken in order of arrival of the first item, converge to the the sequence (R_k) in (vii), which implies $\widehat{U}_{mn}/m^2 \xrightarrow{\mathrm{d}} \sum R_k^2$ and thus (vii). Secondly, the equivalence of (i) and Theorem 7.1 follows by [5]. Thirdly, by the equivalence between random hash tables and random forests mentioned in Remark 4.2, (vi) follows easily from the limit result [1, Proposition 2]. ☐

Moments. For the generating function approach in Section 3, we let $\widehat{F}_{mn}$ be the generating function for $\widehat{U}_{mn}$ in the confined version; thus

$$\mathbb{E}\, x^{\widehat{U}_{mn}} = \widehat{F}_{mn}(x)/\widehat{F}_{mn}(1), \tag{7.2}$$

where $\widehat{F}_{mn}(1) = F_{mn}(1) = (m-n)m^{n-1}$ by (3.2). In the case $m = n+1$, there is only a single block of length $n+1$, and thus $\widehat{U}_{mn} = (n+1)^2$ is non-random, so

$$\widehat{F}_{n+1,n}(x) = x^{(n+1)^2}\widehat{F}_{n+1,n}(1) = (n+1)^{n-1}x^{(n+1)^2}.$$

We define, as in (3.3),

$$\widehat{F}(x,z) = \sum_{n=0}^{\infty} \widehat{F}_{n+1,n}(x)\frac{z^n}{n!} = \sum_{n=0}^{\infty} \frac{(n+1)^{n-1}}{n!} x^{(n+1)^2} z^n = \sum_{m=1}^{\infty} \frac{m^{m-1}}{m!} x^{m^2} z^{m-1}, \tag{7.3}$$

and (3.4) holds with $\widehat{F}$. It is this time somewhat more convenient to study $\widehat{F}(e^w,z)$ instead of $\widehat{F}(1+w,z)$. Then, by (7.2) and (3.4), (3.5) is replaced by

$$\begin{aligned}\tfrac{1}{k!}\,\mathbb{E}\,\widehat{U}_{mn}^k &= [w^k]\,\mathbb{E}\, e^{w\widehat{U}_{mn}} = [w^k]\widehat{F}_{mn}(e^w)/\widehat{F}_{mn}(1)\\ &= [w^k z^n]\widehat{F}(e^w,z)^{m-n}/[z^n]\widehat{F}(1,z)^{m-n}.\end{aligned} \tag{7.4}$$

Moreover, we write $\widehat{F}(e^w,z) = \sum_0^\infty w^k \widehat{W}_k(z)$, where the power series $\widehat{W}_k$ are given by, cf. (7.3) and (3.7),

$$\begin{aligned}\widehat{W}_k &:= [w^k]\widehat{F}(e^w,z) = \sum_{m=1}^{\infty} \frac{m^{m-1}}{m!}[w^k]e^{m^2 w} z^{m-1} = \frac{1}{k!}\sum_{m=1}^{\infty}\frac{m^{m-1+2k}}{m!}z^{m-1}\\ &= \frac{1}{k!} z^{-1}\Big(z\frac{d}{dz}\Big)^{2k} T(z).\end{aligned} \tag{7.5}$$

By (4.24), for $j = 0, 1, \dots,$

$$\begin{aligned} z\frac{d}{dz}T(1-T)^{-j} &= \frac{T}{1-T}\big((1-T)^{-j} + jT(1-T)^{-j-1}\big) \\ &= jT(1-T)^{j-2} - (j-1)(1-T)^{-j-1}. \end{aligned}$$

It follows by induction that

$$\big(z\frac{d}{dz}\big)^k T(z) = T(z)g_k\big(T(z)\big), \tag{7.6}$$

where $g_0(t) = 1$ and, for $k \geq 1$, $g_k(t)$ is a polynomial in $(1-t)^{-1}$ of degree $2k-1$ with leading coefficient $(2k-3)!! = (2k-2)!/2^{k-1}(k-1)!$. For future use we record the first cases:

$$\begin{aligned} g_0(t) &= 1, \\ g_1(t) &= (1-t)^{-1}, \\ g_2(t) &= (1-t)^{-3}, \\ g_3(t) &= 3(1-t)^{-5} - 2(1-t)^{-4}, \\ g_4(t) &= 15(1-t)^{-7} - 20(1-t)^{-6} + 6(1-t)^{-5}. \end{aligned} \tag{7.7}$$

Consequently, (7.5) shows that we now have, instead of (3.10),

$$\widehat{W}_k(z) = \frac{T(z)}{z}\hat{f}_k\big(T(z)\big),$$

where $\hat{f}_k(t) = \frac{1}{k!}g_{2k}(t)$ is a polynomial in $(1-t)^{-1}$; if $k \geq 1$, then $\hat{f}_k$ has degree $4k-1$ and leading coefficient

$$\widehat{\omega}_k = \frac{(4k-3)!!}{k!} = \frac{(4k-2)!}{2^{2k-1}k!\,(2k-1)!}.$$

In particular, $\widehat{\omega}_1 = 1$, $\widehat{\omega}_2 = 15/2$ and, more precisely,

$$\begin{aligned} \hat{f}_1(t) &= g_2(t) = (1-t)^{-3}, \\ \hat{f}_2(t) &= \tfrac{1}{2}g_4(t) = \tfrac{15}{2}(1-t)^{-7} - 10(1-t)^{-6} + 3(1-t)^{-5}. \end{aligned} \tag{7.8}$$

Defining $\hat{f}(w,t) := \sum_0^\infty w^k \hat{f}_k(t)$, the arguments of Section 3 now yield (3.15) with $F(1+w,z)$ and $f(w,t)$ replaced by $\widehat{F}(e^w,z)$ and $\hat{f}(w,t)$, and then (3.16) with $\mathbb{E}\binom{D_{mn}}{k}$ and f_{k_i} replaced by $\frac{1}{k!}\mathbb{E}(\widehat{U}_{mn})^k$ and $\hat{f}_{k_i}$.

We pause to observe that (3.17) now yields explicit expressions for the moments of $\widehat{U}_{mn}$. In particular, for $k = 1$ and 2 we obtain, using (7.8),

$$\mathbb{E}\,\widehat{U}_{mn} = m^{1-n}n!\,[t^n]e^{mt}(1-t)\hat{f}_1(t) = m^{1-n}n!\,[t^n]e^{mt}(1-t)^{-2} = mQ_1(m,n)$$

and

$$\begin{aligned} \mathbb{E}\,\widehat{U}_{mn}^2 &= m^{1-n}n!\,[t^n]e^{mt}(1-t)\big(2\hat{f}_2(t) + (m-n-1)\hat{f}_1(t)^2\big) \\ &= 15mQ_5(m,n) + m(m-n-21)Q_4(m,n) + 6mQ_3(m,n). \end{aligned}$$

Returning to U_{mn} by (7.1), we obtain the following exact results; the expectation was found already by Knuth [14], [15, Theorem 6.4.K].

Theorem 7.3. *If* $0 \le n < m$, *then* $\mathbb{E}\, U_{mn} = \frac{1}{2}Q_1(m,n) + \frac{1}{2}$ *and*

$$\begin{aligned}\operatorname{Var} U_{mn} &= \frac{1}{4m^2}\operatorname{Var}\widehat{U}_{mn}\\ &= \frac{1}{4m}\bigl(15Q_5(m,n) + (m-n-21)Q_4(m,n) + 6Q_3(m,n) - mQ_1(m,n)^2\bigr).\end{aligned}$$

□

For asymptotics when $(m-n)/\sqrt{m} \to a \ge 0$, we use Lemma 3.1 and obtain in analogy with (3.21)

$$n^{-2k}\,\mathbb{E}\,\widehat{U}_{mn}^k \to \widehat{\psi}_k(a), \qquad k \ge 1, \tag{7.9}$$

where

$$\widehat{\psi}_k(a) := k! \sum_{j=1}^{k} \Biggl(\sum_{\substack{k_1,\dots,k_j \ge 1\\ \sum k_i = k}} \prod_{i=1}^{j} \widehat{\omega}_{k_i} \Biggr) \frac{a^{j-1}}{j!} q_{4k-j-2}(a). \tag{7.10}$$

Since $0 \le m^{-2}\widehat{U}_{mn} \le 1$, the moment convergence (7.9) implies convergence in distribution $m^{-2}\widehat{U}_{mn} \overset{\mathrm{d}}{\to} V_a$, for some V_a with $0 \le V_a \le 1$. This shows Theorem 1.6(iii) with the following characterization of the limit, as well as Theorem 1.7(iii).

Theorem 7.4. *The limit random variables* V_a *have the moments* $\mathbb{E}\, V_a^k = \widehat{\psi}_k(a)$, $k \ge 1$, *with* $\widehat{\psi}_k$ *defined in* (7.10). *In particular,*

$$\begin{aligned}\mathbb{E}\, V_a &= \widehat{\omega}_1 q_1(a) = q_1(a),\\ \mathbb{E}\, V_a^2 &= 2\widehat{\omega}_2 q_5(a) + \widehat{\omega}_1^2 a q_4(a) = 15 q_5(a) + a q_4(a).\end{aligned}$$

Moreover, $0 \le V_a \le 1$, *and thus the distribution of* V_a *is determined by the moments* $\widehat{\psi}_k(a)$. □

Again, the moments can be expressed in terms of the normal distribution function Φ, but we leave the details to the reader.

The normal case. We obtain immediately the following analogue and consequence of Lemma 4.1.

Lemma 7.5. *Suppose* $0 \le n < m$ *and let* $N = m - n$. *Let* $0 < \lambda \le e^{-1}$ *and let* $X_1, \dots, X_N$ *be independent random variables with the common distribution* (4.1). *Then the distribution of* $\widehat{U}_{mn}$ *equals the conditional distribution of* $\sum_{j=1}^N X_j^2$ *given* $\sum_{j=1}^N X_j = m$. □

In the cases $n/m \to a \in (0,1)$ and $n/m \to 1$, $m - n \gg m^{1/2}$, we apply Lemma 4.3 as before, still with $X = X_\lambda$ but now taking $Y = \hat{Y} := X^2$. The verification of the conditions is essentially as before; in the case $n/m \to 1$, and thus $\mu = m/N \to \infty$, we use that, by (4.13), (4.14) and (4.15),

$$\begin{aligned}
\mathbb{E}\,\hat{Y} &= \mathbb{E}\,X^2 \sim \mu^3,\\
\sigma_{\hat{Y}}^2 &\sim \mathbb{E}\,\hat{Y}^2 = \mathbb{E}\,X^4 \sim 15\mu^7,\\
\operatorname{Cov}(X,\hat{Y}) &\sim \mathbb{E}\,X\hat{Y} = \mathbb{E}\,X^3 \sim 3\mu^5,\\
\tau^2 &\sim 15\mu^7 - (3\mu^5)^2/\mu^3 = 6\mu^7,
\end{aligned}$$

and for any $r \ge 3$

$$\mathbb{E}\,\hat{Y}^r/\sigma_{\hat{Y}}^r = O(\mu^{4r-1}/\mu^{7r/2}) = O(\mu^{r/2-1}) = o(N^{r/2-1}).$$

This yields Theorem 1.6 in these cases.

In the case $n/m \to 0$, $n \gg m^{1/2}$, we cannot use Lemma 4.3 as stated with $Y = X^2$, since then $\rho \to 1$. Instead we take $Y = \hat{Y}' := (X-1)(X-2) = X^2 - 3X + 2$, which again vanishes for $X = 1$ or 2 yielding $\rho \to 0$; the conditions of Lemma 4.3 are easily verified. Note that if $\sum_1^j X_j = m$, then $\sum_1^N Y_j = \sum_1^N X_j^2 - 3m + 2N$, and thus this Y yields results for $\widehat{U}_{mn} - 3m + 2N = \widehat{U}_{mn} - m - 2n$, which is just as good.

For the moment estimates in Theorem 1.7(ii), we obtain from Lemma 4.3 in the case $n/m \to 1$, by the estimates above, $\mathbb{E}\,\widehat{U}_{mn} \sim N\mu^3$ and $\operatorname{Var}\widehat{U}_{mn} \sim 6N\mu^7$, which by (7.1) imply the corresponding estimates for U_{mn} in Theorem 1.7.

To treat also the other cases, we note that by (4.1) and (7.6), for any $\lambda \in (0,1)$,

$$\mathbb{E}\,X^k = \frac{1}{T(\lambda)}\sum_{\ell=1}^{\infty}\frac{\ell^{\ell-1+k}}{\ell!}\lambda^\ell = \frac{1}{T(\lambda)}\Bigl(\lambda\frac{d}{d\lambda}\Bigr)^k T(\lambda) = g_k\bigl(T(\lambda)\bigr); \tag{7.11}$$

in particular $\mathbb{E}\,X = g_1(T(\lambda)) = (1-T(\lambda))^{-1}$, and substituting $\mu = \mathbb{E}\,X = m/N$ for $(1-T(\lambda))^{-1}$ in (7.11), we obtain $\mathbb{E}\,X^k$ as a polynomial in μ. By (7.7), we have explicitly

$$\begin{aligned}
\mathbb{E}\,\hat{Y} &= \mathbb{E}\,X^2 = g_2\bigl(T(\lambda)\bigr) = \bigl(1-T(\lambda)\bigr)^{-3} = \mu^3,\\
\mathbb{E}\,\hat{Y}^2 &= \mathbb{E}\,X^4 = g_4\bigl(T(\lambda)\bigr) = 15\mu^7 - 20\mu^6 + 6\mu^5,\\
\mathbb{E}\,X\hat{Y} &= \mathbb{E}\,X^3 = g_3\bigl(T(\lambda)\bigr) = 3\mu^5 - 2\mu^4,
\end{aligned}$$

and thus

$$\begin{aligned}
\sigma_X^2 &= \mathbb{E}\,X^2 - \mu^2 = \mu^3 - \mu^2,\\
\sigma_{\hat{Y}}^2 &= \mathbb{E}\,\hat{Y}^2 - (\mathbb{E}\,\hat{Y})^2 = 15\mu^7 - 21\mu^6 + 6\mu^5,\\
\operatorname{Cov}(X,\hat{Y}) &= \mathbb{E}\,X\hat{Y} - \mathbb{E}\,X\,\mathbb{E}\,\hat{Y} = 3\mu^5 - 3\mu^4,\\
\tau^2 &= \sigma_{\hat{Y}}^2 - (\operatorname{Cov}(X,\hat{Y}))^2/\sigma_X^2 = 6\mu^7 - 12\mu^6 + 6\mu^5.
\end{aligned} \tag{7.12}$$

Consequently, for $\alpha \to a > 0$ and $m - n \gg m^{1/2}$, Lemma 4.3 yields

$$\mathbb{E}\,\widehat{U}_{mn} \sim N\,\mathbb{E}\,\hat{Y} = N\mu^3 = \frac{m^3}{(m-n)^2},$$

(as is more easily obtained directly from the exact formula $mQ_1(m,n)$) and

$$\operatorname{Var}\widehat{U}_{mn} \sim N\tau^2 = 6N(\mu-1)^2\mu^5 = 6N\Big(\frac{n}{N}\Big)^2\Big(\frac{m}{N}\Big)^5 = 6\frac{n^2m^5}{(m-n)^6}, \tag{7.13}$$

which yields the corresponding claims in Theorem 1.7 by (7.1).

In the case $n/m \to 0$, $n \gg m^{1/2}$, with $Y = (X-1)(X-2) = X^2 - 3X + 2$, we still have (7.12), cf. Remark 4.4, and thus (7.13).

Poisson limits. Let M_ℓ be the number of blocks of length ℓ. It is easily seen that if $X_j + X_{j+1} + X_{j+2} \le 2$ for all j, then all blocks have lengths at most 3 (i.e. they have at most 2 occupied cells), so $M_\ell = 0$ for $\ell \ge 4$; the constraints $\sum M_\ell = m - n$ and $\sum \ell M_\ell = m$ then yield $M_1 = m - 2n + M_3$ and $M_2 = n - 2M_3$, and thus, by (7.1),

$$mU_{mn} = M_1 + 3M_2 + 6M_3 = m + n + M_3.$$

Moreover, in this case, M_3 equals the number V of pairs of items that make their first try in the same cell or in adjacent ones, i.e. V equals the number of pairs (i,j), $i<j$, such that $|h_i - h_j| \le 1 \pmod m$, cf. Remark 5.2.

Assume now that $n/\sqrt{m} \to a \ge 0$. Arguing as in (5.1) we then find

$$\mathbb{P}(mU_{mn} - m - n \ne V) \le m\,\mathbb{P}(X_1 + X_2 + X_3 \ge 3) = O\Big(\frac{n^3}{m^2}\Big) \to 0.$$

Furthermore, it is easy to check by the method of moments or by Stein's method that $V \xrightarrow{\mathrm{d}} \operatorname{Po}(3a^2/2)$ (this can be regarded as a generalized birthday problem), and Theorem 1.6(i) follows. Moment convergence can be verified as in Section 5.

Asymptotics of V_a. The same proof as for Theorem 6.1 now yields the corresponding result for V_a.

Theorem 7.6. *As $a \to \infty$, we have $\mathbb{E}\,V_a \sim a^{-2}$, $\operatorname{Var} V_a \sim 6a^{-6}$ and*

$$\frac{V_a - \mathbb{E}\,V_a}{(\operatorname{Var} V_a)^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \tag{7.14}$$

with convergence of all moments. □

More refined moment asymtotics follow from Theorem 7.4 and (6.5); for example $\mathbb{E}\,V_a = a^{-2} - 3a^{-4} + O(a^{-6})$.

8. Joint limits

The methods in this paper easily yield joint convergence of D_{mn} and U_{mn} (after appropriate normalizations) in all cases. In the normal case, this leads to the following result. (We leave the other cases to the reader.)

Theorem 8.1. *If $n \gg \sqrt{m}$ and $m - n \gg \sqrt{m}$, then D_{mn} and U_{mn} are jointly asymptotically normal. Moreover, if $\alpha := n/m$, then their covariance and* *correlation have the asymptotics*

$$\operatorname{Cov}(D_{mn}, U_{mn}) \sim \frac{\alpha^2}{2(1-\alpha)^5} = \frac{n^2 m^3}{2(m-n)^5}, \tag{8.1}$$

$$\operatorname{Corr}(D_{mn}, U_{mn}) \sim (3 - 3\alpha + 2\alpha^2 - \tfrac{1}{2}\alpha^3)^{-1/2}. \tag{8.2}$$

In other words, if further $n/m \to a \in [0,1]$, then $(D_{mn} - \mathbb{E}\, D_{mn})/(\operatorname{Var} D_{mn})^{1/2}$ and $(U_{mn} - \mathbb{E}\, U_{mn})/(\operatorname{Var} U_{mn})^{1/2}$ converge jointly in distribution to a pair of normal variables with means 0, variances 1 and covariance

$$\rho = (3 - 3a + 2a^2 - \tfrac{1}{2}a^3)^{-1/2}. \tag{8.3}$$

Proof. Joint normal convergence follows easily from Lemma 4.3 by the Cramér–Wold device, see [12, Corollary 2.2].

For the asymptotic covariance, this yields, with Y as in Section 4,

$$\operatorname{Cov}(D_{mn}, \widehat{U}_{mn}) \sim N\big(\operatorname{Cov}(Y, X^2) - \operatorname{Cov}(Y, X)\operatorname{Cov}(X^2, X)/\operatorname{Var} X\big),$$

which yields (8.1) by straightforward calculations using (4.23) and (4.32) (most terms are already evaluated in Sections 4 and 7); we omit the details. Finally, (8.2) follows from (8.1) and the asymptotic variances given in Theorems 1.4 and 1.7. □

Remark 8.2. It is easily verified that the limiting correlation (or covariance) in (8.3) is an increasing function of a, which is $\sqrt{1/3}$ for $a = 0$ and $\sqrt{2/3}$ for $a = 1$.

References

[1] D.J. Aldous & J. Pitman, The standard additive coalescent. *Ann. Probab.* **26** (1998), 1703–1726.

[2] A.D. Barbour, L. Holst & S. Janson, *Poisson Approximation.* Oxford University Press, Oxford, 1992.

[3] J. Bertoin, A fragmentation process connected to Brownian motion. *Probab. Th. Rel. Fields* **117** (2000), 289–301.

[4] P. Billingsley, *Convergence of Probability Measures.* Wiley, New York, 1968.

[5] P. Chassaing & S. Janson, A Vervaat-like path transformation for the reflected Brownian bridge conditioned on its local time at 0. *Ann. Probab.*, to appear.

[6] P. Chassaing & G. Louchard, Phase transition for parking blocks, Brownian excursion and coalescence. *Rand. Struct. Alg.*, to appear.

[7] P. Chassaing & J.F. Marckert, Parking functions, empirical processes and the width of rooted labelled trees. *Electronic J. Combin.* **8** (2001), #R14.

[8] A. Dvoretzky, J. Kiefer & J. Wolfowitz, Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** (1956), 642–669.

[9] P. Flajolet, P. Poblete & A. Viola, On the analysis of linear probing hashing. *Algorithmica* **22** (1998), 490–515.

[10] L. Holst, Two conditional limit theorems with applications. *Ann. Statist.* **7** (1979), 551–557.

[11] L. Holst, Some conditional limit theorems in exponential families. *Ann. Probab.* **9** (1981), 818–830.

[12] S. Janson, Moment convergence in conditional limit theorems. *J. Appl. Probab.* **38** (2001), 421–437.

[13] S. Janson, D.E. Knuth, T. Łuczak & B. Pittel, The birth of the giant component. *Rand. Struct. Alg.* **4** (1993), 233–358.
[14] D.E. Knuth, Notes on "open" addressing. Unpublished notes, 1963. Available at `http://www.wits.ac.za/helmut/first.ps`
[15] D.E. Knuth, *The Art of Computer Programming. Vol. 3: Sorting and Searching.* 2nd ed., Addison-Wesley, Reading, Mass., 1998.
[16] D.E. Knuth, Linear probing and graphs. *Algorithmica* **22** (1998), 561–568.
[17] V.F. Kolchin, *Random Mappings.* Nauka, Moscow, 1984 (Russian). English transl.: Optimization Software, New York, 1986.
[18] A.G. Konheim & B. Weiss, An occupancy discipline and applications. *SIAM J. Appl. Math.* **14** (1966), 1266–1274.
[19] G. Louchard, Kac's formula, Lévy's local time and Brownian excursion. *J. Appl. Probab.* 21 (1984), 479–499.
[20] G. Louchard, The Brownian excursion area: a numerical analysis. *Comput. Math. Appl.* 10 (1984), 413–417. Erratum: *Comput. Math. Appl. Part A* 12 (1986), 375.
[21] Yu. L. Pavlov, The asymptotic distribution of maximum tree size in a random forest. *Teor. Verojatnost. i Primenen.* **22** (1977), no. 3, 523–533 (Russian). English transl.: *Th. Probab. Appl.* **22** (1977), no. 3, 509–520.
[22] Yu. L. Pavlov, *Random forests.* Karelian Centre Russian Acad. Sci., Petrozavodsk, 1996 (Russian). English transl.: VSP, Zeist, The Netherlands, 2000.
[23] J. Pitman, Coalescent random forests. *J. Combin. Theory A* **85** (1999), 165–193.
[24] J. Pitman & M. Yor, Arcsine laws and interval partitions derived from a stable subordinator. *Proc. London Math. Soc.* (3) **65** (1992), 326–356.
[25] D. Revuz & M. Yor, *Continuous Martingales and Brownian Motion.* 3^{rd} edition, Springer, Berlin, 1999.
[26] J. Spencer, Enumerating graphs and Brownian motion. *Commun. Pure Appl. Math.* **50** (1997), 291–294.
[27] L. Takács, A Bernoulli excursion and its various applications. *Adv. in Appl. Probab.* 23 (1991), 557–585.
[28] W. Vervaat, A relation between Brownian bridge and Brownian excursion. *Ann. Probab.* **7** (1979), 143–149.
[29] E.M. Wright, The number of connected sparsely edged graphs. *J. Graph Th.* **1** (1977), 317–330.

Department of Mathematics, Uppsala University, PO Box 480, S-751 06 Uppsala, Sweden

E-mail address: `svante.janson@math.uu.se`