Combinatorics, Probability and Computing (2017) 00, 1–29.
DOI: 10.1017/S0963548301004989 Printed in the United Kingdom

Asymmetric Rényi Problem

M. DRMOTA[1]† and A. MAGNER[2]‡ and W. SZPANKOWSKI§

[1] Institute for Discrete Mathematics and Geometry, TU Wien, Vienna, Austria, A-1040 Wien, Wiedner Hauptstr. 8–10

[2] Coordinated Science Lab, UIUC, Champaign, IL 61820, USA

[3] Department of Computer Science, Purdue University, IN 47907, USA

michael.drmota@tuwien.ac.at, anmagner@illinois.edu, and szpan@purdue.edu

Received 31 October 2016

In 1960 Rényi in his Michigan State University lectures asked for the number of random queries necessary to recover a hidden bijective labeling of n distinct objects. In each query one selects a random subset of labels and asks, which objects have these labels? We consider here an asymmetric version of the problem in which in every query an object is chosen with probability $p > 1/2$ and we ignore "inconclusive" queries. We study the number of queries needed to recover the labeling in its entirety (H_n), to recover at least one element (F_n), and to recover a randomly chosen element (D_n). This problem exhibits several remarkable behaviors: D_n converges in probability but not almost surely; H_n and F_n exhibit phase transitions with respect to p in the second term. We prove that for $p > 1/2$ with high probability (whp) we need $H_n = \log_{1/p} n + \frac{1}{2}\log_{p/(1-p)} \log n + o(\log\log n)$ queries to recover the entire bijection. This should be compared to its symmetric ($p = 1/2$) counterpart established by Pittel and Rubin, who proved that in this case one requires $H_n = \log_2 n + \sqrt{2\log_2 n} + o(\sqrt{\log n})$ queries. As a bonus, our analysis implies novel results for random PATRICIA tries, as it turns out that the problem is probabilistically equivalent to the analysis of the height, fillup level, typical depth, and profile of a PATRICIA trie built from n independent binary sequences generated by a biased(p) memoryless source.

1. Introduction

In his lectures in the summer of 1960 at Michigan State University, Alfred Rényi discussed several problems related to random sets [19]. Among them there was a problem regarding

† Research partially supported by the Austrian Science Foundation FWF Grant No. F50-02.
‡ Research supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370.
§ Research partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by the NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01.

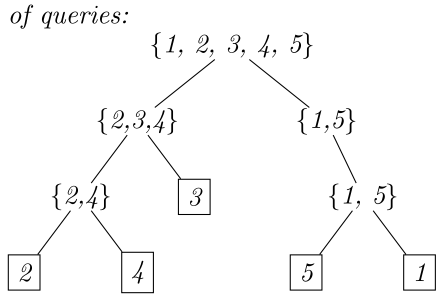
recovering a labeling of a set X of n distinct objects by asking random subset questions of the form "which objects correspond to the labels in the (random) set B?" For a given method of randomly selecting queries, Rényi's original problem asks for the typical behavior of the number of queries necessary to recover the hidden labeling.

Formally, the unknown labeling of the set X is a bijection ϕ from X to a set A of labels (necessarily with equal cardinality n), and a query takes the form of a subset $B \subseteq A$. The response to a query B is $\phi^{-1}(B) \subseteq X$.

Our contribution in this paper is a precise analysis of several parameters of Rényi's problem for a particular natural probabilistic model on the query sequence. In order to formulate this model precisely, it is convenient to first state a view of the process that elucidates its tree-like structure. In particular, a sequence of queries corresponds to a refinement of partitions of the set of objects, where two objects are in different partition elements if they have been distinguished by some sequence of queries. More precisely, the refinement works as follows: before any questions are asked, we have a trivial partition $\mathfrak{P}_0 = X$ consisting of a single class (all objects). Inductively, if $\mathfrak{P}_{j-1}$ corresponds to the partition induced by the first $j-1$ queries, then $\mathfrak{P}_j$ is constructed from $\mathfrak{P}_{j-1}$ by splitting each element of $\mathfrak{P}_{j-1}$ into at most two disjoint subsets: those objects that are contained in the preimage of the jth query set B_j and those that are not. The hidden labeling is recovered precisely when the partition of X consists only of singleton elements. An instance of this process may be viewed as a rooted binary tree (which we call the *partition refinement tree*) in which the jth level, for $j \geq 0$, corresponds to the partition resulting from j queries; a node in a given level corresponds to an element of the partition associated with that level. A right child corresponds to a subset of a parent partition element that is included in the subsequent query, and a left child corresponds to a subset that is not included. See Example 1 for an illustration.

Example 1 (Demonstration of partition refinement). *Consider an instance of the problem where* $X = [5] = \{1, ..., 5\}$, *with labels* (d, e, a, c, b) *respectively (so* $A = \{a, b, c, d, e\}$*). Consider the following sequence of queries:*

1 $B_1 = \{b, d\} \mapsto \{1, 5\}$
2 $B_2 = \{a, b, d\} \mapsto \{1, 3, 5\}$,
3 $B_3 = \{a, c, d\} \mapsto \{1, 3, 4\}$,



Each level $j \geq 0$ *of the tree depicts the partition* $\mathfrak{P}_j$, *where a right child node corresponds to the subset of objects in the parent set which are contained in the response to the* j*th query. Singletons are only explicitly depicted in the first level in which they appear. We can determine the labels of all objects using the tree and the sequence of queries: for example, to determine the label of the object* 3, *we traverse the tree until we reach the leaf* *corresponding to* 3*. This indicates that the label corresponding to* 3 *is in the singleton set*

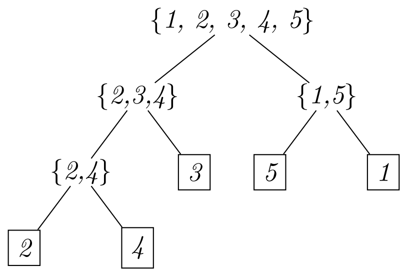
$$\neg B_1 \cap B_2 = \{a, c, e\} \cap \{a, b, d\} = \{a\}.$$

Note that leaves of the tree always correspond to singleton sets.

In this work we consider a version of the problem in which, in every query, each label is included independently with probability $p > 1/2$ (the *asymmetric case*) and we *ignore inconclusive queries*. In particular, if a candidate query fails to non trivially split some element of the previous partition, we modify the query by deciding again independently whether or not to include each label of that partition element with probability p. We perform this modification until the resulting query splits every element of the previous partition non trivially. See Example 2.

Example 2 (Ignoring inconclusive queries). *Continuing Example 1, the query B_2 fails to split the partition element $\{1, 5\}$, so it is an example of an inconclusive query and would be modified in our model to, say, $B_2' = \phi(\{1, 3\})$. The resulting refinement of partitions is depicted as a tree here. Note that the tree now does not contain non-branching paths and that B_2 is ignored in the final query sequence.*

1 $B_1 = \{b, d\} \mapsto \{1, 5\}$
2 $B_2' = \{a, d\} \mapsto \{1, 3\}$
3 $B_3 = \{a, c, d\} \mapsto \{1, 3, 4\}$.



We study three parameters of this random process: H_n, the number of such queries needed to recover the entire labeling; F_n, the number needed before at least one element is recovered; and D_n, the number needed to recover an element selected uniformly at random. Our objective is to present precise probabilistic estimates of these parameters and to study the distributional behavior of D_n.

The symmetric version (i.e., $p = 1/2$) of the problem (with a variation) was discussed by Pittel and Rubin in [18], where they analyzed the typical value of H_n. In their model, a query is constructed by deciding whether or not to include each label from A independently with probability $p = 1/2$. To make the problem interesting, they added a constraint similar to ours: namely, a query is, as in our model, admissible if and only if it splits every nontrivial element of the current partition. In contrast with our model, however, Pittel and Rubin completely discard inconclusive queries (rather than modifying their inconclusive subsets as we do). Despite this difference, the model considered in [18] is probabilistically equivalent to ours for the symmetric case. Our primary contribution is the analysis of the problem in the asymmetric case ($p > 1/2$), but our methods of proof allow us to recover the results of Pittel and Rubin.

The question asked by Rényi brings some surprises. For the symmetric model ($p = 1/2$)

Pittel and Rubin [18] were able to prove that the number of necessary queries is with high probability (whp) (see Theorem 2.1)

$$H_n = \log_2 n + \sqrt{2 \log_2 n} + o(\sqrt{\log n}). \tag{1.1}$$

In this paper, we develop a different method that could be used to re-establish this result and prove that for $p > 1/2$ the number of queries grows whp as

$$H_n = \log_{1/p} n + \frac{1}{2} \log_{p/q} \log n + o(\log \log n), \tag{1.2}$$

where $q := 1-p$. Note a phase transition in the second term. We show that another phase transition, also in the second term, occurs in the asymptotics for F_n (see Theorem 2.2):

$$F_n = \begin{cases} \log_{1/q} n - \log_{1/q} \log \log n + o(\log \log \log n) & p > q \\ \log_2 n - \log_2 \log n + o(\log \log n) & p = q = 1/2. \end{cases} \tag{1.3}$$

We also state in Theorem 2.3 some interesting probabilistic behaviors of D_n. We have $D_n/\log n \to 1/h(p)$ (in probability) where $h(p) := -p \log p - q \log q$, but we do not have almost sure convergence.

We establish these results in a novel way by considering first the *external profile* $B_{n,k}$, whose analysis was, until recently, an open problem of its own (the second and third authors gave a precise analysis of the external profile in an important range of parameters in [13, 14], but the present paper requires really nontrivial extensions). The external profile at level k is the number of bijection elements revealed by the kth query (one may also define the *internal* profile at level k as the number of non-singleton elements of the partition immediately after the kth query). Its study is motivated by the fact that many other parameters, including all of those that we mention here, can be written in terms of it. Indeed, $\Pr[D_n = k] = \mathbb{E}[B_{n,k}]/n$, $H_n = \max\{k : \ B_{n,k} > 0\}$, and $F_n = \min\{k : \ B_{n,k} > 0\} - 1$.

We now discuss our new results concerning the probabilistic behavior of the external profile. We establish in [14, 13] precise asymptotic expressions for the expected value and variance of $B_{n,k}$ in the *central range*, that is, with $k \sim \alpha \log n$, where, for any fixed $\epsilon > 0$, $\alpha \in (1/\log(1/q) + \epsilon, 1/\log(1/p) - \epsilon)$ (the left and right endpoints of this interval are associated with F_n and H_n, respectively). Specifically, it was shown that both the mean and the variance are of the same (explicit) polynomial order of growth (with respect to n). More precisely, expected value and variance grow for $k \sim \alpha \log n$ as

$$H(\rho(\alpha), \log_{p/q}(p^k n)) \ \frac{n^{\beta(\alpha)}}{\sqrt{C \log n}}$$

where $\beta(\alpha) \leq 1$ and $\rho(\alpha)$ are complicated functions of α, C is an explicit constant, and $H(\rho, x)$ is a function that is periodic in x. The oscillations come from infinitely many regularly spaced saddle points that we observe when inverting the Mellin transform of the Poisson generating function of $\mathbb{E}[B_{n,k}]$. Finally, in [14] we prove a central limit theorem; that is, $(B_{n,k} - \mathbb{E}[B_{n,k}])/\sqrt{\mathrm{Var}[B_{n,k}]} \to \mathcal{N}(0,1)$ where $\mathcal{N}(0,1)$ represents the standard normal distribution.

In order to establish the most interesting results claimed in the present paper for H_n and F_n, the analysis sketched above does not suffice: we need to estimate the mean and the variance of the external profile *beyond* the range $\alpha \in (1/\log(1/q)+\epsilon, 1/\log(1/p)-\epsilon)$; in particular, for F_n and H_n we need expansions at the left and right side, respectively, of this range.

Having described most of our main results, we mention an important equivalence pointed out by Pittel and Rubin [18]. They observed that their version of the Rényi process resembles the construction of a digital tree known as a PATRICIA trie[1] [12, 21]. In fact, the authors of [18] show that H_n is probabilistically equivalent to the height (longest path) of a PATRICIA trie built from n binary strings generated independently by a memoryless source with bias $p = 1/2$ (that is, with a "1" generated with probability p; this is often called the *Bernoulli model with bias p*); the equivalence is true more generally, for $p \geq 1/2$. It is easy to see that F_n is equivalent to the fillup level (depth of the deepest full level), D_n to the typical depth (depth of a randomly chosen leaf), and $B_{n,k}$ to the external profile of the tree (the number of leaves at level k; the internal profile at level k is similarly defined as the number of non-leaf nodes at that level). We spell out this equivalence in the following simple claim.

Lemma 1.1 (Equivalence of parameters of the Rényi problem with those of PATRICIA tries). *Any parameter (in particular, H_n, F_n, D_n, and $B_{n,k}$) of the Rényi process with bias p that is a function of the partition refinement tree is equal in distribution to the same function of a random PATRICIA trie generated by n independent infinite binary strings from a memoryless source with bias $p \geq 1/2$.*

Proof. In a nutshell, we couple a random PATRICIA trie and the sequence of queries from the Rényi process by constructing both from the same sequence of binary strings from a memoryless source. We do this in such a way that the resulting PATRICIA trie and the partition refinement tree are isomorphic with probability 1, so that parameters defined in terms of either tree structure are equal in distribution.

More precisely, we start with n independent infinite binary strings $S_1, \ldots, S_n$ generated according to a memoryless source with bias p, where each string corresponds, in a way to be made precise below, to a unique element of the set of labels (for simplicity, we assume that $A = [n]$, and S_j is associated to the object j, for $j \in [n]$; intuitively, S_j encodes the decision, for each query, of whether or not to include j). These induce a PATRICIA trie T, and our goal is to show that we can simulate a Rényi process using these strings, such that the corresponding tree T_R is isomorphic to T as a rooted plane– oriented tree (see Example 2). The basic idea is as follows: we maintain for each string S_j an index k_j, initially set to 1. Whenever the Rényi process demands that we make a decision about whether or not to include label j in a query, we include it if and only if $S_{j,k_j} = 1$, and then increment k_j by 1.

Clearly, this scheme induces the correct distribution on queries. Furthermore, the re-

[1] We recall that a trie is a binary digital tree, where data that are represented by binary strings are stored at leaves of the tree according to finite prefixes of the corresponding binary strings in a minimal way such that all appearing prefixes are different. A PATRICIA trie is a trie in which non-branching paths are *compressed*; that is, there are no unary paths.

sulting partition refinement tree (ignoring inconclusive queries) is easily seen to be isomorphic to T. Since the trees are isomorphic, the parameters of interest are equal in each case. □

Thus, our results on these parameters for the Rényi problem directly lead to novel results on PATRICIA tries, and vice versa. In addition to their use as data structures, PATRICIA tries also arise as combinatorial structures which capture the behavior of various processes of interest in computer science and information theory (e.g., in leader election processes without trivial splits [10] and in the solution to Rényi's problem which we study here [18, 2]).

Similarly, the version of the Rényi problem that allows inconclusive queries corresponds to results on tries built on n binary strings from a memoryless source. We thus discuss them in the literature survey below.

Now we briefly review known facts about PATRICIA tries and other digital trees when built over n independent strings generated by a memoryless source. Profiles of tries in both the asymmetric and symmetric cases were studied extensively in [16]. The expected profiles of digital search trees in both cases were analyzed in [6], and the variance for the asymmetric case was treated in [11]. Some aspects of trie and PATRICIA trie profiles (in particular, the concentration of their distributions) were studied using probabilistic methods in [4, 3]. The depth in PATRICIA for the symmetric model was analyzed in [2, 12] while for the asymmetric model in [20]. The leading asymptotics for the PATRICIA height for the symmetric Bernoulli model was first analyzed by Pittel [17] (see also [21] for suffix trees). The two-term expression for the height of PATRICIA for the symmetric model was first presented in [18] as discussed above (see also [2]). Finally, in [13, 14], the second two authors of the present paper presented a precise analysis of the external profile (including its mean, variance, and limiting distribution) in the asymmetric case, for the range in which the profile grows polynomially. The present work relies on this previous analysis, but the analyses for H_n and F_n involve a significant extension, since they rely on precise asymptotics for the external profile outside this central range.

Regarding methodology, the basic framework (which we use here) for analysis of digital tree recurrences by applying the Poisson transform to derive a functional equation, converting this to an algebraic equation using the Mellin transform, and then inverting using the saddle point method/singularity analysis followed by depoissonization, was worked out in [6] and followed in [16]. While this basic chain is common, the challenges of applying it vary dramatically between the different digital trees, and this is the case here. As we discuss later (see (2.5) and the surrounding text), this variation starts with the quite different forms of the Poisson functional equations, which lead to unique analytic challenges.

The plan for the paper is as follows. In the next section we formulate more precisely our problem and present our main results regarding $B_{n,k}$, H_n, F_n, and D_n, along with sketches of the derivations. Complete proofs for H_n (and a roadmap for the proof for F_n) are provided in Section 3. Finally Section 4 provides some background on the depoissonization step.

2. Main Results

In this section, we formulate precisely Rényi's problem and present our main results. Our goal is to provide precise asymptotics for three natural parameters of the Rényi problem on n objects with each label in a given query being included with probability $p \geq 1/2$: the number F_n of queries needed to identify at least a single element of the bijection, the number H_n needed to recover the bijection in its entirety, and the number D_n needed to recover an element of the bijection chosen uniformly at random from the n objects. If one wishes to determine the label for a particular object, these quantities correspond to the best, worst, and average case performance, respectively, of the random subset strategy proposed by Rényi.

We recall that we can express F_n, H_n, and D_n in terms of the *profile* $B_{n,k}$:

$$F_n = \min\{k : \ B_{n,k} > 0\} - 1, \ H_n = \max\{k : \ B_{n,k} > 0\}, \ \Pr[D_n = k] = \frac{\mathbb{E}[B_{n,k}]}{n}. \quad (2.1)$$

Using the first and second moment methods, we can then obtain upper and lower bounds on H_n and F_n in terms of the moments of $B_{n,k}$:

$$\Pr[H_n > k] \leq \sum_{j>k} \mathbb{E}[B_{n,j}], \qquad \Pr[H_n < k] \leq \frac{\mathrm{Var}[B_{n,k}]}{\mathbb{E}[B_{n,k}]^2}, \quad (2.2)$$

and

$$\Pr[F_n > k] \leq \frac{\mathrm{Var}[B_{n,k}]}{\mathbb{E}[B_{n,k}]^2}, \quad \Pr[F_n < k] \leq \mathbb{E}[B_{n,k}]. \quad (2.3)$$

The analysis of the distribution of D_n reduces simply to that of $\mathbb{E}[B_{n,k}]$, as in (2.1).

Having reduced the analyses of F_n, H_n, and D_n to that of the moments of $B_{n,k}$, we now explain our approach to the latter analysis, starting in Section 2.1 with a review of the work done in [13]. We will then show in Section 2.2 how the present paper requires extensions far beyond [13, 14] to give new results on the quantities of interest in the Rényi problem.

2.1. Basic facts for the analysis of $B_{n,k}$

Here we recall some facts, worked out in detail in [13], which will form the starting point of the analysis in the present paper. In order to derive our main results, we need proper asymptotic information about $\mathbb{E}[B_{n,k}]$ and $\mathrm{Var}[B_{n,k}]$ at the boundaries of this region.

We start by deriving a recurrence for the average profile, which we denote by $\mu_{n,k} := \mathbb{E}[B_{n,k}]$. It satisfies

$$\mu_{n,k} = (p^n + q^n)\mu_{n,k} + \sum_{j=1}^{n-1} \binom{n}{j} p^j q^{n-j} (\mu_{j,k-1} + \mu_{n-j,k-1}) \quad (2.4)$$

for $n \geq 2$ and $k \geq 1$, with some initial/boundary conditions; most importantly, $\mu_{n,k} = 0$ for $k \geq n$ and any n. Moreover, $\mu_{n,k} \leq n$ for all n and k owing to the elimination of inconclusive queries. This recurrence arises from conditioning on the number j of objects that are included in the first query. If $1 \leq j \leq n-1$ objects are included, then the conditional expectation is a sum of contributions from those objects that are included and those that aren't. If, on the other hand, all objects are included or all are excluded from the first potential query (which happens with probability $p^n + q^n$), then the partition element splitting constraint on the queries applies, the potential query is ignored as *inconclusive*, and the contribution is $\mu_{n,k}$.

The tools that we use to solve this recurrence (for details see [13, 14]) are similar to those of the analyses for digital trees [21] such as tries and digital search trees (though the analytical details differ significantly). We first derive a functional equation for the Poisson transform $\tilde{G}_k(z) = \sum_{m\geq 0} \mu_{m,k} \frac{z^m}{m!} e^{-z}$ of $\mu_{n,k}$, which gives

$$\tilde{G}_k(z) = \tilde{G}_{k-1}(pz) + \tilde{G}_{k-1}(qz) + e^{-pz}(\tilde{G}_k - \tilde{G}_{k-1})(qz) + e^{-qz}(\tilde{G}_k - \tilde{G}_{k-1})(pz).$$

This we write as

$$\tilde{G}_k(z) = \tilde{G}_{k-1}(pz) + \tilde{G}_{k-1}(qz) + \tilde{W}_{k,G}(z), \tag{2.5}$$

and at this point the goal is to determine asymptotics for $\tilde{G}_k(z)$ as $z \to \infty$ in a cone around the positive real axis. When solving (2.5), $\tilde{W}_{k,G}(z)$ significantly complicates the analysis because it has no closed-form Mellin transform (see below). Finally, depoissonization [21] will allow us to directly transfer the asymptotic expansion for $\tilde{G}_k(z)$ back to one for $\mu_{n,k}$ since $\mu_{n,k}$ is well approximated by $\tilde{G}_k(n)$.

To convert (2.5) to an equation that is more easy to handle, we use the *Mellin transform* [7], which, for a function $f : \mathbb{R} \to \mathbb{R}$ is given by

$$f^*(s) = \int_0^\infty z^{s-1} f(z)\, \mathrm{d}z.$$

Using the Mellin transform identities and defining $T(s) = p^{-s} + q^{-s}$, we end up with an expression for the Mellin transform $G_k^*(s)$ of $\tilde{G}_k(z)$ of the form

$$G_k^*(s) = \Gamma(s+1) A_k(s) (p^{-s} + q^{-s})^k = \Gamma(s+1) A_k(s) T(s)^k,$$

where $A_k(s)$ is an infinite series arising from the contributions coming from the function $\tilde{W}_{k,G}(z)$, and the fundamental strip of $\tilde{G}_k(z)$ (as sketched below) contains $(-k-1, \infty)$. It involves unknown $\mu_{m,j} - \mu_{m,j-1}$ for various m and j (see [13, 15]), that is:

$$A_k(s) = \sum_{j=0}^{k} T(s)^{-j} \sum_{m\geq j} T(-m)(\mu_{m,j} - \mu_{m,j-1}) \frac{\Gamma(m+s)}{\Gamma(s+1)\Gamma(m+1)}. \tag{2.6}$$

Locating and characterizing the singularities of $G_k^*(s)$ then becomes important. In [14] it is shown that for any k, $A_k(s)$ is entire, with zeros at $s \in \mathbb{Z} \cap [-k, -1]$, so that $G_k^*(s)$ is meromorphic, with possible simple poles at the negative integers less than $-k$. The fundamental strip of $\tilde{G}_k(z)$ then contains $(-k-1, \infty)$.

We then must asymptotically invert the Mellin transform to recover $\tilde{G}_k(z)$. The Mellin inversion formula for $G_k^*(s)$ is given by

$$\tilde{G}_k(z) = \frac{1}{2\pi i} \int_{\rho - i\infty}^{\rho + i\infty} z^{-s} G_k^*(s)\, \mathrm{d}s = \frac{1}{2\pi i} \int_{\rho - i\infty}^{\rho + i\infty} z^{-s} \Gamma(s+1) A_k(s) T(s)^k\, \mathrm{d}s, \tag{2.7}$$

where ρ is any real number inside the fundamental strip associated with $\tilde{G}_k(z)$.

2.2. Main results via extension of the analysis of $B_{n,k}$

Having explained the relevant functional equations and the integral representation (2.7) for $\tilde{G}_k(z)$, we now move on to describe the main results of this paper. For Theorem 2.1 and 2.2 we start with a sketch of the derivation whereas the proof of Theorem 2.3 is given immediately, The complete proof of Theorem 2.1 and a roadmap for Theorem 2.2, both for the case $p > q$, is given in Section 3.

2.2.1. Result on H_n Our first aim is to derive two-term expansions for the typical values of H_n and F_n. To do this for, e.g., H_n, we define, for $p \geq q$,

$$k_* = \log_{1/p} n + \psi_*(n),$$

where $\psi_*(n) = o(\log n)$ is a function to be determined. We also define

$$\psi_L(n) = (1-\epsilon)\psi_*(n) \qquad k_L = \log_{1/p} n + \psi_L(n) \tag{2.8}$$

$$\psi_U(n) = (1+\epsilon)\psi_*(n) \qquad k_U = \log_{1/p} n + \psi_U(n), \tag{2.9}$$

for arbitrarily small $\epsilon > 0$. We require that $\psi_*(n)$ be such that

$$\mathbb{E}[B_{n,k_L}] \to \infty, \qquad \mathbb{E}[B_{n,k_U}] \to 0, \tag{2.10}$$

and a proper upper bound for $\mathrm{Var}[B_{n,k_L}]$ (see Lemma 3.4). However, in order to make the following pre-analysis more transparent we will not dwell on the variance.

To determine a candidate for $\psi_*(n)$, we start with the inverse Mellin integral representation for $\tilde{G}_{k_*}(n)$:

$$\tilde{G}_{k_*}(n) = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} J_{k_*}(n,s)\,\mathrm{d}s, \tag{2.11}$$

where we define

$$\begin{aligned} J_k(n,s) &= n^{-s}T(s)^k\Gamma(s+1)A_k(s) \\ &= \sum_{j=0}^{k} n^{-s}T(s)^{k-j}\sum_{m\geq j} T(-m)(\mu_{m,j}-\mu_{m,j-1})\frac{\Gamma(m+s)}{\Gamma(m+1)}. \end{aligned} \tag{2.12}$$

Note that by depoissonization (see Section 4) we have

$$\mu_{n,k_*} = \tilde{G}_{k_*}(n) - \frac{n}{2}\tilde{G}''_{k_*}(n) + O(n^{-1+\epsilon}).$$

Indeed, because of the exponential decay of $A_k(s)\Gamma(s+1)$ along vertical lines, the entire integral is at most of the same order as the integrand on the real axis (we justify this more carefully in Section 3.1). Furthermore, since the second derivative has an additional factor $s(s+1)n^{-2}$ in the integrand we will get a similar bound for $\frac{n}{2}\tilde{G}''_{k_*}(n)$ which is just ρ^2/n times the corresponding bound for $\tilde{G}_{k_*}(n)$ and, thus, negligible in comparison to $\tilde{G}_{k_*}(n)$.

In this proof roadmap we focus on estimating the integrand $J_{k_*}(n,\rho)$, $\rho \in \mathbb{R}$, as precisely as possible. Using Lemma 3.1, we find (see (3.7) in Section 3.1) that the jth term in the representation (2.12) of $J_{k_*}(n,\rho)$ is of order

$$n^{-\rho}T(\rho)^{k_*-j}p^{j^2/2+O(j\log j)}, \tag{2.13}$$

where $\rho < 0$ and $T(\rho) = p^{-\rho} + q^{-\rho}$. Hence, by setting $j_0 = -\log_{1/p} T(\rho)$ we have

$$J_{k_*}(n,\rho) = O\left(n^{-\rho}T(\rho)^{k_*}p^{-j_0^2/2+O(j_0\log j_0)}\right). \tag{2.14}$$

Next we have to choose $\rho \in \mathbb{R}_-$ that minimizes this upper bound. Here we distinguish between the symmetric case $p = q = 1/2$ and the case $p > q$.

In the symmetric case we have $T(\rho) = 2^{\rho+1}$ and $j_0 = -\rho - 1$ and, thus,

$$J_{k_*}(n,\rho) = O\left(n^{-\rho}2^{(\rho+1)(\log_2 n+\psi_*(n))+\rho^2/2+O(|\rho|\log|\rho|)}\right).$$

Consequently by disregarding the error term $O(|\rho|\log|\rho|)$ the optimal choice of ρ is $\rho = -\psi_*(n)$ which gives the upper bound

$$J_{k_*}(n,\rho) = O\left(2^{\log_2 n-\psi_*(n)^2/2+O(|\psi_*(n)|\log|\psi_*(n)|)}\right).$$

Hence, the threshold for this upper bound is $\psi_*(n) = \sqrt{2\log_2 n}$. In particular it also follows that

$$J_{k_U}(n,\rho) = O\left(2^{-(2\epsilon+\epsilon^2)\log_2 n+O(\sqrt{\log n}\log\log n)}\right),$$

where $k_U = \log_{1/p} n + (1+\epsilon)\sqrt{2\log_2 n}$. We also note that we get the same bound if $\rho = -\psi_*(n) + O(1)$.

In the case $p > q$ we have to be slightly more careful. Nevertheless we can start with the upper bound (2.14) and obtain

$$J_{k_*}(n,\rho) = O\left(p^{(\rho-\log_{1/p}T(\rho))\log_{1/p}n-\psi_*(n)\log_{1/p}T(\rho)-(\log_{1/p}T(\rho))^2/2+O(j_0\log j_0)}\right).$$

¿From the representation $T(\rho) = p^{-\rho}(1 + (p/q)^\rho)$ we obtain

$$\log_{1/p} T(\rho) = \rho + \frac{(p/q)^\rho}{\log(1/p)} + O\left((p/q)^{2\rho}\right).$$

It is clear that we have to choose $\rho < 0$ that tends to $-\infty$ if $n \to \infty$. Hence, $\log_{1/p} T(\rho) = \rho + o(1)$ and consequently a proper choice for ρ is the solution of the equation

$$\frac{\partial}{\partial\rho}\left(-\frac{(p/q)^\rho}{\log(1/p)}\log_{1/p} n - \psi_*(n)\rho - \frac{\rho^2}{2}\right) = \frac{(p/q)^\rho\log(p/q)}{\log(1/p)}\log_{1/p} n - \psi_*(n) - \rho = 0.$$

Actually this gives $\rho < -\psi_*(n)$ and, thus,

$$\rho = -\log_{p/q}\log n + O(\log\log\log n).$$

With this choice the upper bound for $J_{k_*}(n,\rho)$ writes as

$$J_{k_*}(n,\rho) = O\left(p^{(\psi_*(n)+\rho)/\log(p/q)-\psi_*(n)\rho-\frac{\rho^2}{2}+O(j_0\log j_0)}\right) = O\left(p^{-\psi_*(n)\rho-\frac{\rho^2}{2}+O(j_0\log j_0)}\right).$$

This implies that the threshold for this upper bound is given by

$$\psi_*(n) = -\frac{\rho}{2} = \frac{1}{2}\log_{p/q}\log n + O(\log\log\log n).$$

In particular, if we replace $\psi_*(n)$ by $\psi_U(n) = \frac{1}{2}(1+\epsilon)\log_{p/q}\log n$ we obtain

$$J_{k_U}(n,\rho) = O\left(p^{\epsilon(\log_{p/q}\log n)^2/2+O(\log\log n\log\log\log n}\right) \tag{2.15}$$

and for $\psi_L(n) = (1-\epsilon)\,\frac{1}{2}\log_{p/q}\log n$,

$$J_{k_L}(n,\rho) = O\left(p^{-\epsilon(\log_{p/q}\log n)^2/2+O(\log\log n\log\log\log n)}\right). \tag{2.16}$$

The above pre-analysis suggests asymptotic estimates for $\tilde{G}_k(n)$ and, thus, by depoissonization estimates for $\mu_{n,k}$, which imply a two-term expansion for H_n. The complete proof of this result is given in Section 3.1. In summary, we formulate below our first main result.

Theorem 2.1 (Asymptotics for H_n). *With high probability,*

$$H_n = \begin{cases} \log_{1/p} n + \frac{1}{2}\log_{p/q}\log n + o(\log\log n) & p > q \\ \log_2 n + \sqrt{2\log_2 n} + o(\sqrt{\log n}) & p = q \end{cases}$$

for large n.

2.2.2. Result on F_n We take a similar approach for the derivation of F_n, with some differences. We set

$$k_* = \log_{1/q} n + \phi_*(n)$$

with

$$\phi_L(n) = (1+\epsilon)\phi_*(n), \qquad \phi_U(n) = (1-\epsilon)\phi_*(n),$$

and k_L and k_U, respectively, defined with ϕ_L (respectively, ϕ_U) in place of ϕ_*. The derivation of an estimate for the jth term of $J_{k_*}(n,\rho)$, $\rho \in \mathbb{R}$, is similar to that in Section 2.2.1, except now the asymptotics of $\Gamma(\rho+1)$ play a role (this is reflected in the proof, where $\Gamma(\rho+1)$ determines the location of the saddle points of the integrand). We find that the jth term is at most $q^{\lambda_j(n,\rho)}$, where

$$\lambda_j(n,\rho) = \rho(j-\phi_*(n)) + (j-\phi_*(n)-\log_{1/q} n)\log_{1/q}(1+(q/p)^\rho) - \rho\log_{1/q}\rho + O(\rho) \tag{2.17}$$

Optimizing over j gives $j = 0$. The behavior with respect to ρ depends on whether or not $p = q$, because $\log_{1/q}(1+(q/p)^\rho) = 1$ when $p = q$ and is dependent on ρ otherwise. Taking this into account and minimizing over all ρ gives an optimal value of

$$\rho = \begin{cases} 2^{-\phi_*(n)-1/\log 2} & p = q = 1/2, \\ \log_{p/q}\log n & p > 1/2. \end{cases}$$

Note that this corresponds to the real part of the saddle points in the proof. Plugging this into (2.17), setting the resulting expression equal to 0, and solving for $\phi_*(n)$ gives

$$\phi_*(n) = \begin{cases} -\log_2\log n + O(1) & p = q = 1/2 \\ -\log_{1/q}\log\log n & p > 1/2. \end{cases}$$

This heuristic derivation suggests that the following theorem holds. More details are given in Section 3.2.

Theorem 2.2 (Asymptotics for F_n). *With high probability,*

$$F_n = \begin{cases} \log_{1/q} n - \log_{1/q} \log\log n + o(\log\log\log n) & p > q \\ \log_2 n - \log_2 \log n + o(\log\log n) & p = q \end{cases}$$

for large n.

2.3. Result on D_n

We move to our results concerning D_n. To state them, we first need to observe that there is a natural way to define the sequence $\{D_n\}_{n\geq 0}$ on a single probability space, so that we may ask whether or not D_n, properly normalized, converges almost surely, and to what limiting value. This common space is defined by appealing to the correspondence between the sequence of Rényi problem queries and the growth of a random PATRICIA trie. For each $n \geq 0$, we define a tree T_n which is a PATRICIA trie constructed on n strings (equivalently, a terminating sequence of Rényi queries recovering a bijection between two sets of n elements): T_0 is an empty tree, and T_{n+1} is constructed from T_n by generating an independent string of i.i.d. Bernoulli(p) random variables and inserting this string into T_n. Then, for each n, D_n is the depth of a leaf chosen uniformly at random (and independent of everything else) from the leaves of T_n.

With this construction in mind, we have the following result about the convergence of D_n. Its proof combines known facts about the profile with the new ones proved here, as well as a proof technique that was used before in, e.g., [17].

Theorem 2.3 (Asymptotics of D_n). *For $p > 1/2$, the normalized depth $D_n/\log n$ converges in probability to $1/h(p)$ where $h(p) := -p\log p - q\log q$ is the entropy of a* Bernoulli(p) *random variable, but not almost surely. In fact,*

$$\liminf_{n\to\infty} D_n/\log n = 1/\log(1/q), \qquad \limsup_{n\to\infty} D_n/\log n = 1/\log(1/p) \tag{2.18}$$

almost surely.

Proof. The fact that $D_n/\log n$ converges in probability to $1/h(p)$ follows directly from the central limit theorem for D_n given in [21].

Next we show that (2.18) holds. Clearly $F_n \leq D_n \leq H_n$. Now let us consider the following sequences of events: A_n is the event that $D_n = F_n + 1$, and A'_n is the event that $D_n = H_n$. We note that all elements of the sequences are independent, and $\Pr[A_n] \geq 1/n$, $\Pr[A'_n] \geq 1/n$. This implies that $\sum_{n=1}^{\infty} \Pr[A_n] = \sum_{n=1}^{\infty} \Pr[A'_n] = \infty$, so that the Borel-Cantelli lemma tells us that both A_n and A'_n occur infinitely often almost surely.

In the next step we show that, almost surely, $F_n/\log n \to 1/\log(1/q)$ and $H_n/\log n \to 1/\log(1/p)$. Then (2.18) is proved. We cannot apply the Borel-Cantelli lemmas directly, because the relevant sums do not converge. Instead, we apply a trick which was used in [17]. We observe that both (F_n) and (H_n) are non-decreasing sequences. Next, we show that, on some appropriately chosen subsequence, both of these sequences, when divided by $\log n$, converge almost surely to their respective limits. Combining this with the observed monotonicity yields the claimed almost sure convergence, and, hence, the equalities in (2.18).

We illustrate this idea more precisely for H_n. By our analysis above, we know that

$$\Pr[|H_n/\log n - 1/\log(1/p)| > \epsilon] = O(e^{-\Theta(\log\log n)^2}).$$

Then we fix t, and we define $n_{r,t} = 2^{t^2 2^{2r}}$. On this subsequence, by the probability bound just stated, we can apply the Borel-Cantelli lemma to conclude that $H_{n_{r,t}}/\log(n_{r,t}) \to 1/\log(1/p) \cdot (t+1)^2/t^2$ almost surely. Moreover, for every n, we can choose r such that $n_{r,t} \le n \le n_{r,t+1}$. Then

$$H_n/\log n \le H_{n_{r,t+1}}/\log n_{r,t},$$

which implies

$$\limsup_{n\to\infty} \frac{H_n}{\log n} \le \limsup_{r\to\infty} \frac{H_{n_{r,t+1}}}{\log n_{r,t+1}} \frac{\log n_{r,t+1}}{\log n_{r,t}} = \frac{1}{\log(1/p)} \cdot \frac{(t+1)^2}{t^2}.$$

Taking $t \to \infty$, this becomes $1/\log(1/p)$, as desired. The argument for the lim inf is similar, and this establishes the almost sure convergence of H_n. The derivation is entirely similar for F_n. □

3. Proof of Theorems 2.1 and 2.2

We give a detailed proof of Theorem 2.1 and indicate the main lines of the proof of Theorem 2.2. We also concentrate just on the case $p > q$. The proof of the symmetric case can be done by the same techniques (properly adapted) but it just reproves the result by Pittel and Rubin [18].

3.1. Proof of Theorem 2.1

3.1.1. A-Priori Bounds for $\mu_{n,k}$ For the analysis of the profile around the height level, we need precise information about $\mu_{n,k}$ with $n \to \infty$ when k close to n. This is captured in the following lemma, which first appeared in a similar form in [15].

We consider $\mu_{n,k}$ where k is close to n, so we set $k = n - \ell$ and represent it as

$$\mu_{n,k} = \mu_{n,n-\ell} = n! C_*(p) p^{(n-\ell)(n-\ell+1)^2/2} q^{n-\ell} \xi_\ell(n),$$

where

$$C_*(p) = \prod_{j=2}^{\infty} (1 - p^j - q^j)^{-1} \cdot (1 + (q/p)^{j-2}),$$

$\xi_1(1) = 1/C_*(p)$ and for $n > \ell \ge 1$

$$\xi_\ell(n)(1 - p^n - q^n) = \sum_{J=1}^{\ell} \frac{\xi_{\ell+1-J}(n-J)}{J!} q^{-1} p^{\ell-n} (p^{n-J} q^J + p^J q^{n-J}). \tag{3.1}$$

Note that $\xi_\ell(n) = 0$ for $n \le \ell$.

Lemma 3.1 (Asymptotics for $\mu_{n,k}$, $k \to \infty$ and n near k). *Let $p \ge q$.*
(i) Precise estimate: *For every fixed $\ell \ge 1$ and $n \to \infty$*

$$\mu_{n,n-\ell} \sim n! C_*(p) p^{(n-\ell)^2/2 + (n-\ell)/2} q^{n-\ell} \xi_\ell,$$

where the sequence ξ_ℓ, $\ell \ge 1$ satisfies the recurrence

$$\xi_\ell = q^{-1} p^\ell \sum_{J=1}^{\ell} \frac{\xi_{\ell+1-J}}{J!} (q/p)^J \tag{3.2}$$

with $\xi_1 = 1$. Furthermore we have (for some positive constant C)

$$|\xi_{\ell+1-J}(n-J) - \xi_{\ell+1-J}| \le C(p^{n-\ell-1} + (q/p)^{n-\ell-1})/(\ell-J)!, \tag{3.3}$$

(ii) Upper bound: *We have $\xi_\ell(n) \le C_1/(\ell-1)!$ for some constant C_1 and, thus, for $1 \le k < n$ (and some constant C)*

$$\mu_{n,k} \le C \frac{n!}{(n-k-1)!} p^{(k^2+k)/2} q^k. \tag{3.4}$$

Proof. From the recurrence (3.1) it follows easily that for each $\ell \ge 1$ the limit $\xi_\ell = \lim_{n\to\infty} \xi_\ell(n)$ exists, and in particular for $\ell = 1$ we have $\xi_1 = 1$. Clearly this limits satisfy the recurrence (3.2).

Next we show by induction a uniform upper bound of the form $\xi_\ell(n) \le C_1/(\ell-1)!$ The induction step for $n > \ell > \ell_1$ runs as follows (where C_1 and ℓ_1 is appropriately chosen such that the upper bound is true for $\ell \le \ell_1$ and that $2/(q\ell_1(1-p^{\ell_1}-q^{\ell_1}) \le 1)$:

$$\begin{aligned}
\xi_\ell(n) &\le \frac{C_1}{1-p^n-q^n} \left(\sum_{J=1}^{\ell} \frac{p^{\ell-J} q^{J-1}}{J!(\ell-J)!} + \sum_{J=1}^{\ell} \frac{p^{\ell+J-n} q^{n-J-1}}{J!(\ell-J)!} \right) \\
&\le \frac{C_1}{\ell!(1-p^n-q^n)} \left(\frac{1}{q} \sum_{J=0}^{\ell} \binom{\ell}{J} p^{\ell-J} q^J + \frac{(q/p)^{n-\ell}}{q} \sum_{J=0}^{\ell+1} \binom{\ell}{J} p^J q^{\ell-J} \right) \\
&\le \frac{C_1}{(\ell-1)!} \frac{1}{\ell_1(1-p^{\ell_1}-q^{\ell_1})} \frac{2}{q} \le \frac{C_1}{(\ell-1)!}.
\end{aligned}$$

In a similar way we obtain the approximation estimate (3.3). We leave the details to the reader. □

3.1.2. Upper bound on H_n Now we set

$$k = k_U = \log_{1/p} n + \psi_U(n) = \log_{1/p} n + \frac{1}{2}(1+\epsilon)\log_{p/q} \log n. \tag{3.5}$$

just as in (2.9). We will first estimate the value of $J_k(n,s)$ (which is defined in (2.12)) for $s = \rho = -2\psi_U(n) + O(1) \in \mathbb{Z}^- - 1/2$ (as hinted at in Section 2).

Lemma 3.2. *Suppose that $p > q$, that $\epsilon > 0$, that k_U is given by (3.5), and that $\rho' = \lfloor \rho \rfloor + \frac{1}{2}$, where $\rho = -\log_{p/q} \log n + O(\log\log\log n)$ is the solution of the equation*

$$\frac{(p/q)^\rho \log(p/q)}{\log(1/p)} \log_{1/p} n + \psi_U(n) + \rho = 0.$$

Then we have for $k \ge k_U$

$$J_k(n,\rho') = O\left(T(\rho')^{k-k_U} p^{\epsilon(\log_{p/q} \log n)^2/2 + O(\log\log n \cdot \log\log\log n} \right). \tag{3.6}$$

Proof. First we observe that the assumption $\rho' \in \mathbb{Z}^- - 1/2$ assures that $|\Gamma(m+\rho')/\Gamma(m+1)| \le 1$ for all $m \ge 0$. Next by (3.4) of Lemma 3.1 we have $\mu_{m,j} = O\left(m^{j+1} p^{j^2/2+O(j)}\right)$ which implies that

$$\sum_{m \ge j} T(-m)\mu_{m,j} = O\left(p^{j^2/2+O(j\log j)}\right).$$

Hence, the jth term in the representation (2.12) of $J_k(n,\rho')$ can be estimated by

$$\left| n^{-\rho'} T(\rho')^{k-j} \sum_{m\ge j} T(-m)(\mu_{m,j} - \mu_{m,j-1}) \frac{\Gamma(m+\rho')}{\Gamma(m+1)} \right| \tag{3.7}$$
$$\le n^{-\rho'} T(\rho')^{k-j} \sum_{m \ge j} T(-m)(\mu_{m,j} + \mu_{m,j-1}) = O\left(n^{-\rho'} T(\rho')^{k-j} p^{j^2/2+O(j\log j)}\right).$$

Thus, we have shown (2.14) which implies (3.6) for $k = k_U$ (see (2.15). However, it is easy to extend it to larger k. Actually we get uniformly for $k \ge k_U$

$$J_k(n,\rho') = O\left(T(\rho')^{k-k_U} p^{\epsilon(\log_{p/q}\log n)^2/2+O(\log\log n \log\log\log n}\right)$$

for large n. □

Our next goal is to evaluate the integral (2.11) and to obtain a bound for $\mu_{n,k}$.

Lemma 3.3. *Suppose that $p > q$, that $\epsilon > 0$, and that k_U and ρ' are given as in Lemma 3.2. Then we have (for some $\delta > 0$)*

$$\mu_{n,k} = O\left(T(\rho')^{k-k_U} p^{\epsilon(\log_{p/q}\log n)^2/2+O((\log\log n)^{1-\delta})}\right) + O(n^{-1+\epsilon}) \tag{3.8}$$

uniformly for $k \ge k_U$.

Proof. Letting $\mathcal{C}$ denote the vertical line $\Re(s) = \rho'$, we evaluate the integral (2.11) by splitting it into an inner region $\mathcal{C}^I$ and outer tails $\mathcal{C}^O$:

$$\mathcal{C}^I = \{\rho' + it : \ |t| \le e^{(\log\log n)^{2-\delta}}\}, \quad \mathcal{C}^O = \{\rho' + it : \ |t| > e^{(\log\log n)^{2-\delta}}\},$$

where $0 < \delta < 1$ is some fixed real number. The inner region we evaluate by showing that it is of the same order as the integrand on the real axis, and the outer tails are shown to be negligible by the exponential decay of the Γ function.

It is easily checked that $|n^{-s}T(s)^{k-j}\Gamma(m+s)| \le n^{-\rho'}T(\rho')^{k-j}|\Gamma(m+\rho')|$. Thus,

$$|J_k(n,s)| \le T(\rho')^{k-k_U} \sum_{j=0}^{k} n^{-\rho'} T(\rho')^{k_U-j} \sum_{m\ge j} T(-m)|\mu_{m,j} - \mu_{m,j-1}| \frac{|\Gamma(m+\rho')|}{\Gamma(m+1)},$$

which can be upper bounded as in the proof of Lemma 3.2. Multiplying by the length of the contour, we get

$$\left| \int_{\mathcal{C}^I} J_k(n,s)\, \mathrm{d}s \right| = O\left(T(\rho')^{k-k_U} p^{\epsilon(\log_{p/q}\log n)^2/2+O((\log\log n)^{2-\delta})}\right).$$

We use the following standard bound on the Γ function: for $s = \rho' + it$, provided that

$|\mathrm{Arg}(s)|$ is less than and bounded away from π and $|s|$ is sufficiently large, we have

$$|\Gamma(s)| \le C|t|^{\rho'-1/2}e^{-\pi|t|/2}.$$

This is applicable on $\mathcal{C}^O$, and we again use the fact that $|T(s)| \le T(\rho')$ and $|\mu_{m,j} - \mu_{m,j-1}| \le m$, which yields an upper bound of the form

$$\begin{aligned}
\left|\sum_{m\ge j} T(-m)(\mu_{m,j}-\mu_{m,j-1})\frac{\Gamma(m+s)}{\Gamma(m+1)}\right| &= O\left(\sum_{m\ge j} T(-m)m\frac{|t|^{m+\rho'-1/2}e^{-\pi|t|/2})}{\Gamma(m+1)}\right)\\
&= O\left(p|t|^{\rho'+1/2}e^{-\pi|t|/2}e^{p|t|}\right),
\end{aligned}$$

where we have used the inequality

$$|t|^{\rho'-1/2}e^{-\pi|t|/2}\sum_{m\ge j}\frac{m(p|t|)^m}{m!} \le p|t|^{\rho'+1/2}e^{-\pi|t|/2}e^{p|t|} = e^{-\Theta(|t|)}.$$

Note that $-\pi/2 + p < 0$, so that we are left with $e^{-\Theta(|t|)+(\rho'+1/2)\log|t|}$. By our choice of $|t|$, this is simply $e^{-\Theta(|t|)}$, uniformly in j.

Furthermore, since $T(\rho') < 1$ we have

$$\sum_{j=0}^{k} n^{-\rho'}T(\rho')^{k-j} = O(n^{-\rho'}) = O\left(e^{\log n\log\log n}\right).$$

Hence, integrating this on $\mathcal{C}^O$ gives

$$\begin{aligned}
\left|\int_{\mathcal{C}^O} J_k(n,s)\,\mathrm{d}s\right| &= O\left(T(\rho')^{k-k_U}e^{\log n\log\log n}e^{-\Theta(e^{(\log\log n)^{2-\delta}})}\right)\\
&= O\left(T(\rho')^{k-k_U}e^{-\Theta(e^{(\log\log n)^{2-\delta}})}\right).
\end{aligned}$$

Adding these together gives

$$\begin{aligned}
\tilde{G}_k(n) &\le \left|\int_{\mathcal{C}^I} J_{k_U}(n,s)\,\mathrm{d}s + \int_{\mathcal{C}^O} J_{k_U}(n,s)\,\mathrm{d}s\right|\\
&= O\left(T(\rho')^{k-k_U}p^{\epsilon(\log_{p/q}\log n)^2/2+O((\log\log n)^{2-\delta})}\right).
\end{aligned}$$

Similarly we get a bound for $\tilde{G''}_k(n)$:

$$\tilde{G''}_k(n) = O\left(\rho'^2T(\rho')^{k-k_U}p^{\epsilon(\log_{p/q}\log n)^2/2+O((\log\log n)^{2-\delta})}\right).$$

Hence by depoissonization (see (4.5) from Section 4) we get

$$\mu_{n,k} = O\left(T(\rho')^{k-k_U}p^{\epsilon(\log_{p/q}\log n)^2/2+O((\log\log n)^{2-\delta})}\right) + O(n^{-1+\epsilon})$$

as needed. □

Our original goal was to bound the tail $\Pr[H_n > k_U]$ by the following sum which we split into two part:

$$\Pr[H_n > k_U] \le \sum_{k \ge k_U} \mu_{n,k} = \sum_{k=k_U}^{\lceil (\log n)^2 \rceil} \mu_{n,k} + \sum_{k=\lceil (\log n)^2 \rceil + 1}^{n} \mu_{n,k}.$$

The initial part can be bounded using (3.8), and the final part we handle using (3.4) in Lemma 3.1. Indeed, since $T(\rho') < 1$ the first sum can be bounded by

$$\sum_{k=k_U}^{\lceil (\log n)^2 \rceil} \mu_{n,k} \le e^{-\Theta(\epsilon \log\log n)^2}.$$

The second sum is at most

$$\sum_{k=\lceil (\log n)^2 \rceil + 1}^{\infty} \mu_{n,k} = \sum_{k=\lceil (\log n)^2 \rceil + 1}^{n} \mu_{n,k} \le n e^{-\Theta(\log n)^4} = e^{-\Theta(\log n)^4}.$$

Adding these upper bounds together shows that $\Pr[H_n > k_U] = e^{-\Theta(\epsilon(\log\log n)^2)} \to 0$, as desired.

3.1.3. Upper bound on the variance of the profile We consider now the case

$$k = k_L = \log_{1/p} n + \psi_L(n) = \log_{1/p} n + \psi(n), \qquad \psi(n) = \frac{1}{2}(1-\epsilon)\log_{p/q}\log n. \tag{3.9}$$

and start with an upper bound for the variance of the profile $\mathrm{Var}[B_{n,k}]$.

Lemma 3.4. *Suppose that $p > q$, that $\epsilon > 0$, and that k_L is given by (3.9). Then we have*

$$\mathrm{Var}[B_{n,k}] \quad = O\left(p^{-\epsilon(\log_{p/q}\log n)^2/2 + O((\log\log n)^{2-\delta})}\right). \tag{3.10}$$

Proof. The proof technique here is the same as for the proof of the upper bound on $\mu_{n,k}$. Our goal is to upper bound the expression

$$\tilde{V}_k(n) = \sum_{n \ge 0} \mathbb{E}[B_{n,k}^2] \frac{n^n}{n!} e^{-n} - \tilde{G}_k(n)^2 = \frac{1}{2\pi i} \int_{\rho' - i\infty}^{\rho' + i\infty} J_k^{(V)}(n,s)\, \mathrm{d}s,$$

where

$$J_k^{(V)}(n,s) = n^{-s} T(s)^k \Gamma(s+1) B_k(s),$$

and

$$B_k(s) = 1 - (s+1)2^{-(s+2)} + \sum_{j=1}^{k} T(s)^{-j} \frac{W_{j,V}^*(s)}{\Gamma(s+1)},$$

with

$$W_{j,V}^*(s) \quad = \quad \sum_{m \ge j} \frac{\Gamma(m+s)}{m!} [\ T(-m)(c_{m,j} - c_{m,j-1} + \mu_{m,j} - \mu_{m,j-1})$$

$$+T(s)2^{-(s+m)}\sum_{\ell=0}^{m}\mu_{\ell,j-1}\mu_{m-\ell,j-1}$$

$$\left. +2\sum_{\ell=0}^{m}\mu_{\ell,j-1}\mu_{m-\ell,j-1}p^{\ell}q^{m-\ell} - 2^{-(m+s)}\sum_{\ell=0}^{m}\mu_{\ell,j}\mu_{m-\ell,j}\right].$$

As above we need a bound on the moments of $B_{m,j}$ for m sufficiently close to j: for $\mu_{m,j} = \mathbb{E}[B_{m,j}]$, this is (3.4) in Lemma 3.1. It turns out that $c_{m,j} = \mathbb{E}[B_{m,j}(B_{m,j}-1)]$ satisfies a similar recurrence as $\mu_{m,j}$ (see [14]) and also similar inequality: for $j \to \infty$ and $m \geq j$,

$$c_{m,j} \leq \frac{m!}{(m-j-1)!}p^{j^2/2+O(j)}.$$

The proof is by induction and follows along the same lines as that of the upper bound in Lemma 3.1. Using this, we can upper bound the inverse Mellin integral as in the upper bound for $\tilde{G}_k(n)$.

In particular it follows that

$$\tilde{V}_{k_L}(n) = O\left(p^{-\epsilon(\log_{p/q}\log n)^2/2+O((\log\log n)^{2-\delta})}\right)$$

and similarly we have

$$\tilde{V}''_{k_L}(n) = O\left(\rho'^2 n^{-2} p^{-\epsilon(\log_{p/q}\log n)^2+O((\log\log n)^{2-\delta})}\right),$$

where $\rho' = -\log_{p/q}\log n + O(\log\log\log n)$. With the help of depoissonization, see (4.6), we thus obtain (3.10). □

3.1.4. Lower bound on H_n The most difficult part of the proof of Theorem 2.1 is to prove a lower bound for the expected profile.

Lemma 3.5. *Suppose that $p > q$, that $\epsilon > 0$, and that k_L is given by (3.9). Then we have*

$$\mu_{n,k_L} = \Omega\left(p^{-\epsilon(\log_{p/q}\log n)^2/2+O(\log\log n\log\log\log n}\right). \tag{3.11}$$

By combining Lemma 3.4 and Lemma 3.5 it immediately follows that

$$\Pr[H_n < k_L] \leq \frac{\mathrm{Var}[B_{n,k_L}]}{\mu^2_{n,k_L}} \to 0$$

which proves the lower bound on H_n.

The plan to prove Lemma 3.5 is as follows: we evaluate the inverse Mellin integral exactly by a residue computation. This results in a nested summation, which we simplify using the binomial theorem and the series of the exponential function. From this representation we will then detect several terms that contribute to the leading term in the asymptotic expansion.

Lemma 3.6. *Suppose that $\rho < 0$ but not an integer. Then we have*

$$\tilde{G}_k(n) = \sum_{j=0}^{k} \sum_{m \geq j} \kappa_{m,j} \left(\mu_{m,j} - \mu_{m,j-1}\right), \tag{3.12}$$

where

$$\kappa_{m,j} = \frac{T(-m)n^m}{m!} \sum_{\ell=(-\lceil m+\rho \rceil+1) \ \vee \ 0}^{\infty} \frac{(-n)^\ell}{\ell!} T(-m-\ell)^{k-j} \tag{3.13}$$

and $x \ \vee \ y$ denotes the maximum of x and y.

Proof. By shifting the line of integration and collecting residues we have

$$\frac{1}{2\pi i} \int_{\rho - i\infty}^{\rho + i\infty} n^{-s} T(s)^{k-j} \Gamma(m+s)\, ds = \sum_{l \geq \max\{0, -m-\rho\}} \frac{n^{m+\ell}(-1)^\ell}{\ell!} T(-\ell - m)^{k-j}.$$

Hence the lemma follows. □

We now choose ρ as $\rho = -j^* - 1$ and set $j_0 = \lceil j^* \rceil$, where j^* is the root of the equation

$$(q/p)^{j*}(k_L - j^*) = \frac{\log(1/p)}{\log(p/q)}(j^* - \psi_L(n)), \tag{3.14}$$

where $\psi_L(n) = \frac{1}{2}(1-\epsilon)\log_{p/q} \log n$.

In particular it follows that

$$\overline{r}_0 := (q/p)^{j_0}(k_L - j_0) \leq \overline{r}_1 := \frac{\log(1/p)}{\log(p/q)}(j_0 - \psi_L(n)), \tag{3.15}$$

If $j > j_0$ and $m \geq j$ then we certainly have $(-\lceil m+\rho \rceil + 1) \ \vee \ 0 = 0$, whereas for $j = j_0$ we have $(-\lceil j_0 + \rho \rceil + 1) \ \vee \ 0 = 1$.

Asymptotically we have $j^* = \log_{p/q} \log n - \log_{p/q} \log\log n + O(1)$. Hence we also have $j_0 = \log_{p/q} \log n - \log_{p/q} \log\log n + O(1)$ and $\rho = -\log_{p/q} \log n - \log_{p/q} \log\log n + O(1)$. We also want to mention that j^* can be considered as a function of ϵ. By implicit differentiation it follows that this function has bounded derivative. Thus j^* is almost constant and, thus, ρ is really constant for sufficiently small ϵ.

In what follows we will encounter several different asymptotic behaviors. In particular we will show that

$$\begin{aligned} \tilde{G}_k(n) &= D(p)C_*(p)p^{j_0(j_0+1)/2}q^{j_0-1}n^{j_0}p^{j_0(k-j_0)}e^{\overline{r}_0}\Phi\left(\frac{\overline{r}_1 - \overline{r}_0}{\sqrt{\overline{r}_0}}\right) \\ &+ O\left(p^{j_0(j_0+1)/2}q^{j_0-1}n^{j_0}p^{j_0(k-j_0)}\frac{\overline{r}_0^{\overline{r}_1}}{\Gamma(\overline{r}_1+1)}\right), \end{aligned} \tag{3.16}$$

where

$$D(p) = \sum_{L,M \geq 0} \xi_{L+1} \frac{(-1)^M}{M!} p^{((L+M)^2+L-M)/2} q^{-L-M}. \tag{3.17}$$

Note that $\overline{r}_0 \leq \overline{r}_1$ which implies that $\Phi\left(\frac{\overline{r}_1 - \overline{r}_0}{\sqrt{\overline{r}_0}}\right) \geq \frac{1}{2}$. Furthermore $\overline{r}_0^{\overline{r}_1}/\Gamma(\overline{r}_1+1) =$

$O(e^{\overline{r}_0}/\sqrt{\overline{r}_0})$. Thus, the first term is the asymptotically leading one and since $\tilde{G}_k(n) > 0$ it also follows that $D(p) \ge 0$. Unfortunately it seems that $D(p)$ is identically zero (which is conjectural relation as we verified by extensive symbolic and numerical calculations). If $D(p)$ would be positive then the proof of the lower bound would be finished. However we have to deal also with the case $D(p) = 0$. For this purpose we have to deal with the second order term more precisely. It follows that

$$\tilde{G}_k(n) \ge C(p) p^{j_0(j_0+1)/2} q^{j_0-1} n^{j_0} p^{j_0(k-j_0)} \frac{\overline{r}_0^{\overline{r}_1}}{\Gamma(\overline{r}_1+1)} \tag{3.18}$$

for some constant $C(p) > 0$. This is a quite involved calculation. Thus, in order to demonstrate the method we will concentrate only on few instances. Note also that this lower bound implies (3.11) since by definition $\overline{r}_1 < (p/q)\,\overline{r}_0$ so that $E := e^{-\overline{r}_0} \overline{r}_0^{\overline{r}_1}/\Gamma(\overline{r}_1+1) = e^{\Omega(\log\log n)}$. $E := e^{-\overline{r}_0} \overline{r}_0^{\overline{r}_1}/\Gamma(\overline{r}_1+1) = e^{\Omega(\log\log n)}$.

Before we start with the core part of the proof we mention that we will have also error terms that are smaller by a factor p^{j_0} or $(q/p)^{j_0}$ compared to the asymptotic leading term. However, it is easy to check that $p^{j_0} = o(E)$ and $(q/p)^{j_0} = o(E)$ for $\frac{1}{2} < p < 1$ so that we can safely neglect those error terms.

Note that for $j \le j_0$ and $j \le m \le j_0$ we have

$$\kappa_{m,j} = \frac{T(-m)n^m}{m!} \sum_{r=0}^{k-j} \binom{k-j}{r} p^{m(k-j-r)} q^{mr} \left(e^{-np^{k-j-r}q^r} - \sum_{\ell \le j_0 - m} \frac{(-n)^\ell}{\ell!} (p^{k-j-r}q^r)^\ell \right)$$

and otherwise

$$\kappa_{m,j} = \frac{T(-m)n^m}{m!} \sum_{r=0}^{k-j_0} \binom{k-j}{r} p^{m(k-j-r)} q^{mr)} e^{-np^{k-j-r}q^r}.$$

In view of the above discussion we can thus replace the term $T(-m)$ (in $\kappa_{m,j}$) by p^m; the resulting sum will be denoted by $\overline{\kappa}_{m,j}$. And we can also replace $\mu_{m,j} - \mu_{m,j-1}$ by

$$\overline{\nu}_{m,j} := -C_*(p) m! p^{j(j-1)/2} q^{j-1} \xi_{m-j+1}.$$

It is an easy exercise (by using the methods from below) to show that

$$\begin{aligned}\tilde{G}_k(n) &= C_*(p) \sum_{j=0}^{k} \sum_{m \ge j} \overline{\kappa}_{m,j} \xi_{m-j+1} m! p^{j(j-1)/2} q^{j-1} \\ &+ O\left(n^{j_0} T(-j^0)^{k-j_0} p^{j_0(j_0+1)/2} q^{j_0} \left(p^{j_0} + (q/p)^{j_0}\right)\right).\end{aligned}$$

In order to analyze the sum representation (3.12) we split it into several parts:

$$T_1 := \sum_{j > j_0} \sum_{m \ge j} \overline{\kappa}_{m,j} \overline{\nu}_{m,j}, \quad T_2 := \sum_{j \le j_0} \sum_{m > j_0} \overline{\kappa}_{m,j} \overline{\nu}_{m,j}, \quad T_3 := \sum_{j \le j_0} \sum_{m=j}^{j_0} \overline{\kappa}_{m,j} \overline{\nu}_{m,j}.$$

The most interesting part is third term that we will discuss first. Note that the exponential function $e^{-np^{k-j-r}q^r} = e^{-(q/p)^{r-r_1(j)}}$ behaves completely different for $r \le r_1(j)$ and for $r > r_1(j)$ where $r_1(j) = (j - \psi_(n)) \frac{\log(1/p)}{\log(p/q)}$. Hence it is convenient to split T_3 into three parts $T_{30} + T_{31} + T_{32}$, where the T_{30} and T_{31} correspond to the terms with $r \le r_1(j)$ and

T_{32} for those with $r > r_1(j)$. T_{30} involves the exponential function $e^{-np^{k-j-r}q^r}$ whereas T_{31} takes care of the polynomial sum $\sum_{\ell\le j_0-m}\frac{(-n)^\ell}{\ell!}(p^{k-j-r}q^r)^\ell$.

The most interesting term is T_{31}. However, it is more convenient to start with the other terms T_{30}, T_{32}, T_1, and T_2:

Lemma 3.7. *We have*

$$-T_{30} \le C_1(p)p^{j_0(j_0+1)/2}q^{j_0}n^{j_0}p^{j_0(k-j_0}\frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)},$$

where $C_1(p)$ is positive and depends on p.

Proof. We first note that all summands of T_{30} are negative. So $-T_{30}$ has just positive terms. We then use the upper bound $\xi_\ell \le C/(\ell-1)!$ and thus, led to consider the following sum which can be again upper bounded be estimating the m-sum with the help of the exponential function $e^{np^{k-j-r+1}q^r}$:

$$\begin{aligned} C\sum_{j\le j_0} p^{j(j-1)/2}q^j \sum_{m=j}^{j_0}\frac{p^m n^m}{(m-j)!}\sum_{r\le r_1(j)}\binom{k-j}{r}p^{m(k-j-r)}q^{mr}e^{-np^{k-j-r}q^r} \\ \le C\sum_{j\le j_0}p^{j(j+1)/2}q^j n^j\sum_{r\le r_1(j)}\binom{k-j}{r}p^{j(k-j-r)}q^{jr}e^{-np^{k-j-r}q^r(1-p)}. \end{aligned}$$

Let us start with the term related to $j = j_0$:

$$p^{j_0(j_0+1)/2}q^{j_0-1}n^{j_0}\sum_{r\le\overline{r}_1}\binom{k-j_0}{r}p^{j_0(k-j_0-r)}q^{j_0 r}e^{-np^{(k-j_0-r)}q^{r+1}}.$$

Recall that $\overline{r}_0 < \overline{r}_1$. Thus, the sum $r \le \overline{r}_1$ covers those r around $\overline{r}_0$ for which the binomial part $\binom{k-j_0}{r}p^{j_0(k-j_0-r)}q^{j_0 r}$ is maximal. However, the exponential part $e^{-np^{(k-j_0-r)}q^{r+1}} = e^{-q(q/p)^{r-\overline{r}_1}}$ is small enough so that the whole r-sum is dominated by the last summand:

$$\sum_{r\le\overline{r}_1}\binom{k-j_0}{r}p^{j_0(k-j_0-r)}q^{j_0 r}e^{-np^{(k-j_0-r)}q^r} \le C_1'(p)p^{j_0(k-j_0)}\frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)}.$$

If $j < j_0$ then we have

$$\sum_{r\le r_1(j)}\binom{k-j}{r}p^{j(k-j-r)}q^{jr}e^{-np^{k-j-r}q^{r+1}} \le C_1''(p)p^{j(k-j)}\frac{(k-j)^{r_1(j)}}{\Gamma(r_1(j)+1)}\left(\frac{q}{p}\right)^{jr_1(j)}.$$

The situation is here even easier since the summands are monotonically increasing in r (if $r \le r_1(j)$). Now observe that

$$\frac{(k-j)^{r_1(j)}}{\Gamma(r_1(j)+1)}\left(\frac{q}{p}\right)^{j_0 r_1(j)} = \frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)}e^{O(j-j_0)}.$$

Hence we obtain (up to a constant) the upper bound

$$\frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)}\sum_{j<j_0}p^{j(j+1)/2}q^j n^j e^{O(j-j_0)}\left(\frac{q}{p}\right)^{(j-j_0)r_1(j)}p^{j(k-j)}$$

$$= p^{j_0(j_0+1)/2} q^{j_0} n^{j_0} p^{j_0(k-j_0)} \frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)} \sum_{j<j_0} p^{(j-j_0)^2/2+O(j-j_0)} q^{j-j_0}.$$

Note that we have used the relation $np^{k-j} = (q/p)^{-r_1(j)}$. This proves the lemma. ☐

Lemma 3.8. *We have*

$$|T_{32}| \le C_2(p) p^{j_0(j_0+1)/2} q^{j_0} n^{j_0} p^{j_0(k-j_0} \frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)},$$

where $C_2(p)$ is positive and depends on p.

Proof. By using the inequalities $\xi_\ell \le C/(\ell-1)!$ and $|e^{-x} - \sum_{\ell\le L} x^\ell/\ell!| \le x^{L+1}/(L+1)!$, $0 \le x \le 1$, that we apply for $x = np^{k-j-r}q^r = (q/p)^{r-r_1(j)} \le 1$ (for $r > r_1(j)$) it is sufficient to consider the sum

$$\sum_{j\le j_0} p^{j(j-1)/2} q^j \sum_{m=j}^{j_0} \frac{p^m n^m}{(m-j)!} \sum_{r>r_1(j)} \binom{k-j}{r} p^{m(k-j-r)} q^{mr} \frac{(np^{k-j-r}q^r)^{j_0-m+1}}{(j_0-m+1)!}$$

$$= n^{j_0+1} \sum_{j\le j_0} p^{j(j-1)/2} q^j \sum_{m=j}^{j_0} \frac{p^m}{(m-j)!(j_0-m+1)!} \sum_{r>r_1(j)} \binom{k-j}{r} p^{(j_0+1)(k-j-r)} q^{(j_0+1)r}.$$

Clearly, the sum $\sum_{m=j}^{j_0} p^m/((m-j)!(j_0-m+1)!)$ is bounded from the above by $p^j(1+p)^{j_0-j+1}/(j_0-j+1)!$. Furthermore we use a usual tail estimate of a binomial sum to get

$$\begin{aligned}\sum_{r>r_1(j)} \binom{k-j}{r} p^{(j_0+1)(k-j-r)} q^{(j_0+1)r} &\le C_2' p^{(j_0+1)(k-j)} \frac{(k-j)^{r_1(j)}}{\Gamma(r_1(j)+1)} \left(\frac{q}{p}\right)^{(j_0+1)r_1(j)} \\ &\le C_2' \frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)} e^{O(j-j_0)} p^{(j_0+1)(k-j)} \left(\frac{q}{p}\right)^{r_1(j)}.\end{aligned}$$

Hence we obtain (up to a constant) the upper bound

$$\begin{aligned}&\frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)} n^{j_0} \sum_{j\le j_0} p^{j(j+1)/2} q^j \frac{e^{O(j-j_0)}}{(j_0-j+1)!} p^{(j_0-j)j_0} \\ &= p^{j_0(j_0+1)/2} q^{j_0} n^{j_0} p^{j_0(k-j_0)} \frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)} \sum_{j\le j_0} \frac{p^{(j-j_0)^2/2+O(j-j_0)} q^{j-j_0}}{(j_0-j+1)!},\end{aligned}$$

which completes the proof of the lemma. Note that we have again used the the relation $np^{k-j} = (q/p)^{-r_1(j)}$. ☐

Lemma 3.9. *We have*

$$-T_1 \le C_3(p) p^{j_0(j_0+1)/2} q^{j_0} n^{j_0} p^{j_0(k-j_0)} \frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)},$$

where $C_3(p)$ is positive and depends on p.

Proof. We proceed similarly to the proof of Lemma 3.7. We first use the inequality

$\xi_\ell \le C/(\ell-1)!$ and estimate the m-sum from the above by the exponential function $e^{np^{k-j-r+1}q^r}$. Hence, $-T_1$ is upper bounded (up to a constant) by

$$C\sum_{j>j_0} p^{j(j+1)/2}q^j n^j \sum_{r=0}^{k-j}\binom{k-j}{r}p^{j(k-j-r)}q^{jr}e^{-np^{k-j-r}q^r(1-p)}.$$

Again similarly to the proof of Lemma 3.7 we can estimate the r-sum by

$$\sum_{r=0}^{k-j}\binom{k-j}{r}p^{j(k-j-r)}q^{jr}e^{-np^{k-j-r}q^{r+1}} \le C_3'(p)p^{j(k-j)}\frac{(k-j)^{r_1(j)}}{\Gamma(r_1(j)+1)}\left(\frac{q}{p}\right)^{jr_1(j)}.$$

Actually for $r \le r_1(j)$ the same argument as in Lemma 3.7 applies whereas for $r > r_1(j)$ we get an upper bound by usual tail estimates of a binomial sum (as in Lemma 3.8).

Summing up we are then (almost) in the same situation as in Lemma 3.7. The only difference is that $j > j_0$ instead of $j \le j_0$. Thus we can proceed as above and obtain the proposed upper bound. □

Lemma 3.10. *We have*

$$-T_2 \le C_4(p)p^{j_0(j_0+1)/2}q^{j_0}n^{j_0}p^{j_0(k-j_0)}\frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)},$$

where $C_4(p)$ is positive and depends on p.

Proof. Here we have to argue slightly differently although we start again by estimating ξ_ℓ by $C/(\ell-1)!$. In particular the m-sum is now estimated by

$$\begin{aligned}\sum_{m>j_0}\frac{p^m n^m}{(m-j)!}p^{m(k-j-r)}q^{mr} &\le \sum_{m>j_0}\frac{p^m n^m}{(m-j_0)!}p^{m(k-j-r)}q^{mr}\\ &= p^{j_0}n^{j_0}p^{j_0(k-j-r)}q^{j_0 r}\left(e^{np^{k-j-r+1}q^r}-1\right)\end{aligned}$$

which leads to the sum

$$p^{j_0}n^{j_0}\sum_{j\le j_0}p^{j(j-1)/2}q^j\sum_{r=0}^{k-j}\binom{k-j}{r}p^{j_0(k-j-r)}q^{j_0 r}\left(e^{-np^{k-j-r}q^{r+1}}-e^{-np^{k-j-r}q^r}\right).$$

Again we distinguish between $r \le r_1(j)$ and $r > r_1(j)$. In the first case we proceed as in Lemma 3.7, whereas in the second case we have

$$e^{-np^{k-j-r}q^{r+1}}-e^{-np^{k-j-r}q^r} \le np^{k-j-r+1}q^r$$

so that we are can proceed as in Lemma 3.8. □

Finally we deal with T_{31} for which we will show later that it is positive and supersedes the other terms (although it is of the same order of magnitude).

First of all we regroup the summation by setting $m = j_0 - M$, $j = j_0 - M - L$, and $\ell = M - K$ which gives

$$T_{31} = C_*(p)p^{j_0(j_0+1)/2}q^{j_0-1}n^{j_0}p^{j_0(k-j_0)}\sum_{K\ge 0}\left(\frac{q}{p}\right)^{K\overline{r}_1}$$

$$\times \sum_{L\ge 0,\, M\ge K} \xi_{L+1} \frac{(-1)^{M-K}}{(M-K)!} p^{((L+M)^2+L-M)/2-K(L+M)} q^{-L-M}$$
$$\times \sum_{r\le r_1(j_0-M-L)} \binom{k-j_0+M+L}{r} \left(\frac{q}{p}\right)^{(j_0-K)r}.$$

We single out the case $K=0$ (and consider only the sum over K, M, r) which we write as

$$D(p)C_*(p) \sum_{r\le \overline{r}_1} \binom{k-j_0+L+M}{r} \left(\frac{q}{p}\right)^{j_0 r} + S_0,$$

where $D(p)$ is given by (3.17) and

$$\begin{aligned} S_0 \;&:=\; -C_*(p) \sum_{L,M\ge 0} \xi_{L+1} \frac{(-1)^M}{M!} p^{((L+M)^2+L-M)/2} q^{-L-M} \\ &\qquad \times \sum_{r_1(j_0-M-L)<r\le \overline{r}_1} \binom{k-j_0+L+M}{r} \left(\frac{q}{p}\right)^{j_0 r}. \end{aligned}$$

Note that

$$\sum_{r\le \overline{r}_1} \binom{k-j_0+L+M}{r} \left(\frac{q}{p}\right)^{j_0 r} = e^{\overline{r}_0} \Phi\left(\frac{\overline{r}_1-\overline{r}_0}{\sqrt{\overline{r}_0}}\right)\left(1+O\left(\frac{\log\log n}{\log n}(L+M)\right)\right),$$

where Φ denotes the distribution function of the normal distribution. Thus, we have proved the asymptotic relation (3.16) since the following property is easy to establish:

Lemma 3.11. *We have*

$$\sum_{K\ge 0} |S_K| \le C_5(p) p^{j_0(j_0+1)/2} q^{j_0} n^{j_0} p^{j_0(k-j_0)} \frac{\overline{r}_0^{\,\overline{r}_1}}{\Gamma(\overline{r}_1+1)},$$

where S_K, $K>0$, is defined by

$$\begin{aligned} S_K \;&:=\; C_*(p)\left(\frac{q}{p}\right)^{K\overline{r}_1} \sum_{L\ge 0,\, M\ge K} \xi_{L+1} \frac{(-1)^{M-K}}{(M-K)!} p^{((L+M)^2+L-M)/2-K(L+M)} q^{-L-M} \\ &\qquad \times \sum_{r\le r_1(j_0-M-L)} \binom{k-j_0+M+L}{r} \left(\frac{q}{p}\right)^{(j_0-K)r}. \end{aligned}$$

Proof. We just give a sketch of the proof. In all appearing r-sums the last summand is dominating which gives rise to the factor $\overline{r}_0^{\,\overline{r}_1}/\Gamma(\overline{r}_1+1)$. The remaining calculations are then quite close to the proofs of the previous Lemmas 3.7–3.10. □

As mentioned above we have to handle in particular the case $D(p)=0$, too. This means that we to be need more precise Lemmas 3.7–3.11. In principle we just have to make the calculations more accurate and more importantly we have to observe finally that all second order terms sum up to a positive contribution (3.18). It is not immediately clear that we end up with a positive contribution since T_{30}, T_1, and T_2 are certainly negative. In order to simplify (and shorten) the presentation we just present here a special case. The general case is just notationally much more involved but runs along the same lines. We suppose that $\overline{r}_1$ is an integer. Furthermore let us assume that $\log(1/p)/\log(p/q)$ is also an integer with implies that $r_j(j)$ is an integer, too, for all j. In this case we have

$$\sum_{r\le r_1(j_0-M-L)} \binom{k-j_0+M+L}{r} \left(\frac{q}{p}\right)^{(j_0-K)r} \sim \frac{\overline{r}_0^{\,\overline{r}_1}}{\overline{r}_1!} \left(\frac{\overline{r}_1}{\overline{r}_0}\right)^{(L+M)\frac{\log(1/p)}{\log(p/q)}} \frac{\left(\frac{q}{p}\right)^{-Kr_1(j_0-L-M)}}{1-\frac{\overline{r}_1}{\overline{r}_0}\left(\frac{q}{p}\right)^K}.$$

Note that $\overline{r}_0 \le \overline{r}_1 < \frac{p}{q}\overline{r}_0$. So everything is well defined. Furthermore we have

$$\left(\frac{q}{p}\right)^{(\overline{r}_1-r_j(j_0-L-M))K} = p^{(L+M)K}.$$

Thus, by summing up S_K, $K \ge 1$, we get

$$\sum_{K\ge 1} S_K \sim \frac{\overline{r}_0^{\,\overline{r}_1}}{\overline{r}_1!} C_*(p) \sum_{L\ge 0,\, M\ge 1} \xi_{L+1} p^{((L+M)^2+L-M)/2} q^{-L-M} \left(\frac{\overline{r}_1}{\overline{r}_0}\right)^{(L+M)\frac{\log(1/p)}{\log(p/q)}} \times \sum_{K=1}^{M} \frac{(-1)^{M-K}}{(M-K)!} \frac{1}{1-\frac{\overline{r}_1}{\overline{r}_0}\left(\frac{q}{p}\right)^K}.$$

The alternating sum

$$\sum_{K=1}^{M} \frac{(-1)^{M-K}}{(M-K)!} \frac{1}{1-\frac{\overline{r}_1}{\overline{r}_0}\left(\frac{q}{p}\right)^K}$$

is certainly negative for $M = 2$. But for $M \to \infty$ it converges to e^{-1}. Thus we (finally) obtain a positive contribution.

Similarly we can handle the other terms. For example, for the negative term T_{30} we can replace the constant $C_1(p)$ by

$$C_*(p) \sum_{J\ge 0} p^{J(J+1)/2} q^{-J} \sum_{R\ge 0} \left(\frac{q}{p}\right)^{JR} e^{-\left(\frac{q}{p}\right)^{-R}} \sum_{L=0}^{R} \xi_{L+1} p^L \left(\frac{q}{p}\right)^{-LR}.$$

If we compute all these terms we finally end up with a representation of the form (3.18) with a positive constant $C(p)$. The following table gives a sample evaluation of these constants (for the case $\overline{r}_0 = \overline{r}_1$).

p	0.5001	0.6	0.7	0.8	0.9	0.95
$C(p)/C_*(p)$	3633.132	1.283	16.055	1561.066	1.020×10^{12}	2.667×10^{39}

These sample computations indicate that the second order term provides a positive contribution as proposed. For a complete analysis we have to go through Lemmas 3.7–

3.11 in much more detail (we have to take care of several rounding effects if $\overline{r}_1$ and $\overline{r}_1(j)$ are not integers) and then make extensive numerical computations. Since we can bound derivatives with respect to p it is sufficient to check positivity on a proper finite grid plus an asymptotic analysis for $p \to 1$. So this finally proves the lower bound (3.18).

3.2. Proof of Theorem 2.2

The analysis of F_n runs along the same lines as for H_n. As already mentioned we will give only a roadmap of the proof since it actually much easier than that of H_n.

3.2.1. Lower bound on F_n The lower bound on F_n can be proven in two different ways. We can either use the inverse Mellin transform integral for $\tilde{G}_k(n)$

$$k = k_L = \log_{1/q} \log n - (1+\epsilon) \log_{1/q} \log\log n$$

evaluated at $\rho = \log_{p/q} \log n$. This leads to $\Pr[F_n < k] \le \mu_{n,k} \to 0$.

Alternatively we can use the correspondence between the Rényi process and the random PATRICIA trie construction, along with the relationship between PATRICIA tries and standard tries. Because of the path compression step in the construction of a PATRICIA trie from a trie, the fillup level for a PATRICIA trie is always greater than or equal to the fillup level for the associated trie. Furthermore, it is known (see [16]) that the fillup level in random tries for $p > 1/2$ is, with high probability,

$$\log_{1/q} n - \log_{1/q} \log\log n + o(\log\log\log n).$$

Thus, with high probability, this is also a lower bound for the F_n that we study.

3.2.2. Upper bound on F_n The upper bound proof for F_n follows along similar lines to the lower bound for H_n (though there are fewer complications). We set

$$k = k_U = \log_{1/q} n - (1-\epsilon) \log_{1/q} \log\log n,$$

and our goal is to show that $\mathrm{Var}[B_{n,k}] = o(\mathbb{E}[B_{n,k}]^2)$. First we get an upper bound for $\mathrm{Var}[B_{n,k}]$ in the same way as in the case of H_n (via inverse Mellin transform and Depoissonization) of the form

$$\mathrm{Var}[B_{n,k}] = O\left(q^{-\epsilon \log_{p/q} \log n \cdot \log 1/q \log\log n(1+o(1))}\right).$$

In order to obtain a corresponding lower bound for $\mu_{n,k} = \mathbb{E}[B_{n,k}]$ we use again the explicit representation

$$\tilde{G}_k(n) = \sum_{j=0}^{k} \sum_{m \ge j} \kappa_{m,j} (\mu_{m,j} - \mu_{m,j-1}), \tag{3.19}$$

where

$$\begin{aligned} \kappa_{m,j} &= \frac{T(-m)n^m}{m!} \sum_{\ell=0}^{\infty} \frac{(-n)^\ell}{\ell!} T(-m-\ell)^{k-j} \\ &= \frac{T(-m)}{m!} \sum_{r=0}^{k-j} \binom{k-j}{r} (np^r q^{k-j-r})^m \exp(-np^r q^{k-j-r}). \end{aligned} \tag{3.20}$$

We note that, because $\rho > 0$, there are no contributions from poles, so that the ℓ-sum begins with 0, in contrast to (3.13) which leads to the simplified form (3.20).

The heuristic derivation suggests that the main contribution to (3.19) comes from the terms $j = O(1)$ and $m = \rho \cdot p/q + O(1)$. In this range, the difference $\mu_{m,j} - \mu_{m,j-1}$ is estimable by the following lemma from [15] (see part (i) of Theorem 2.2) of that paper).

Lemma 3.12 (Precise asymptotics for $\mu_{m,j}$ when $j = O(1)$ and $m \to \infty$). *For $p > q$, $m \to \infty$, and $j = O(1)$, we have*

$$\mu_{m,j} \sim mq^j(1 - q^j)^{m-1}.$$

Note, in particular, that $\mu_{m,j} - \mu_{m,j-1}$ is strictly positive in this range. Applying this lemma, some algebra is required to show that the contribution of the (m, j)th term, with $m = \rho \cdot p/q + O(1)$ and $j = O(1)$, is

$$q^{-\epsilon \log_{p/q} \log n \cdot \log_{1/q} \log\log n(1+o(1))}. \tag{3.21}$$

To complete the necessary lower bound on the entire sum (3.19), we consider also the following sums:

$$\sum_{j=0}^{j'} \sum_{m=j}^{m'} \kappa_{m,j}(\mu_{m,j} - \mu_{m,j-1}) \quad \text{and} \quad \sum_{j>j'} \sum_{m \geq j} \kappa_{m,j}(\mu_{m,j} - \mu_{m,j-1}), \tag{3.22}$$

where j' and m' are sufficiently large fixed positive numbers. We note that the terms that are not covered by any of these sums may be disregarded, since by Lemma 3.12 they are non-negative.

It may be shown that both sums are smaller than the dominant term (3.21) by a factor of $e^{-\Theta(\rho)}$, both by upper bounding terms in absolute value and using the trivial bound $|\mu_{m,j} - \mu_{m,j-1}| \leq 2m$.

We thus arrive at

$$\mu_{n,k} \geq q^{-\epsilon \log_{p/q} \log n \cdot \log_{1/q} \log\log n(1+o(1))}. \tag{3.23}$$

Since this tends to ∞ with n, combining this with the upper bound for the variance yields the desired upper bound on $\Pr[F_n > k]$, which establishes the upper bound on F_n.

4. Depoissonization

4.1. Analytic Depoissonization

The Poisson transform $\tilde{G}(z)$ of a sequence g_n is defined by $\tilde{G}(z) = \sum_{n \geq 0} g_n \frac{z^n}{n!} e^{-z}$. If the sequence g_n is *smoothly enough* then we usually have $g_n \sim \tilde{G}(n)$ (as $n \to \infty$) which we call *Depoissonization.* In [9] a theory for *Analytic Depoissonization* is developed. For example, the basic theorem (Theorem 1) says that if

$$|\tilde{G}(z)| \leq B|z|^\beta \tag{4.1}$$

for $|z| > R$ and $|\arg(z)| \leq \theta$ (for some $B > 0$, $R > 0$, and $0 < \theta < \pi/2$) and

$$|\tilde{G}(z)e^z| \leq Ae^{\alpha|z|} \tag{4.2}$$

for $|z| > R$ and $\theta < |\arg(z)| \le \pi$ (for some $A > 0$ and $\alpha < 1$) then

$$g_n = \tilde{G}(n) + O(n^{\beta-1}). \tag{4.3}$$

Actually this expansion can be more precise by taking into account derivatives of $\tilde{G}(z)$. For example, we have

$$g_n = \tilde{G}(n) - \frac{n}{2}\tilde{G}''(n) + O(n^{\beta-2}). \tag{4.4}$$

In [14, Lemmas 1 and 18] it is show that $\tilde{G}_k(z) = \sum_{n\ge0}\mu_{n,k}\frac{z^n}{n!}e^{-z}$ satisfies (4.1) with $\beta = 1+\epsilon$ for any $\epsilon > 0$ and (4.2) for some $\alpha < 1$ uniformly for all $k \ge 0$. Thus, it follows uniformly for all $k \ge 0$

$$\mu_{n,k} = \tilde{G}_k(n) - \frac{n}{2}\tilde{G}_k''(n) + O(n^{\epsilon-1}). \tag{4.5}$$

The estimate (4.3) is not sufficient for our purposes (it only works if $\mu_{n,k}$ grows $\mu_{n,k}$ at least polynomially as in the *central range*). For the boundary region, where $k \sim \log_{1/p} n$ or $k \sim \log_{1/q} n$ we have to use (4.5) which means that we have to deal with derivatives of $\tilde{G}_k(z)$, too.

4.2. Poisson Variance

Next we discuss how the variance of a random variable can be handled with the help of the Poisson transform. First we assume that $\tilde{G}(z)$ is the Poisson transform of the expected values $\mu_n = \mathbb{E}[X_n]$ or a sequence of random variables. Furthermore we set

$$\tilde{V}(z) = \sum_{n\ge0}\mathbb{E}[X_n^2]\frac{z^n}{n!}e^{-z} - \tilde{G}(z)^2$$

which we denote the Poisson variance. This is not the Poisson transform of the variance. However, since we usually have $\mathbb{E}[X_n^2] \sim V(n) + G(n)^2$ and $\mathbb{E}[X_n] \sim G(n)$ it is expected that $\mathrm{Var}[X_n] \sim V(n)$. Actually this can be made precise with the help of (4.4). Suppose that $\tilde{G}(z)$ and $\tilde{V}(z)$ satisfy the property (4.1) and that $\tilde{G}(z)$ and $\tilde{V}(z) + \tilde{G}(z)^2$ the property (4.2). Then it follows that

$$\mathbb{E}[X_n] = \tilde{G}(n) - \frac{n}{2}\tilde{G}''(n) + O(n^{\beta-2})$$

and

$$\mathbb{E}[X_n^2] = \tilde{V}(n) + \tilde{G}(n)^2 - \frac{n}{2}\tilde{V}''(n) - n(\tilde{G}'(n))^2 - n\tilde{G}(n)\tilde{G}''(n) + O(n^{\beta-2})$$

from which it follows that

$$\begin{aligned}\mathrm{Var}[X_n] &= \tilde{V}(n) - \tfrac{n}{2}\tilde{V}''(n) - n(\tilde{G}'(n))^2 + \tfrac{1}{4}n^2(\tilde{G}''(n))^2 \\ &\quad + O(n^{2\beta-4}) + O(n^{\beta-2}\tilde{G}(n)) + O(n^{\beta}\tilde{G}''(n))\end{aligned} \tag{4.6}$$

In particular in our case we know that the Poisson transform $\tilde{G}_k(z)$ (of the sequence $\mu_{n,k} = \mathbb{E}[B_{n,k}]$) and the corresponding Poisson variance $\tilde{V}_k(z)$ satisfy the assumptions for $\beta = 1+\epsilon$ (for every fixed $\epsilon > 0$), see [14]. Thus we also obtain (4.6) in the present context.

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964.

[2] Luc Devroye. A note on the probabilistic analysis of PATRICIA trees. *Random Structures and Algorithms*, 3(2):203–214, March 1992.

[3] Luc Devroye. Laws of large numbers and tail inequalities for random tries and PATRICIA trees. *Journal of Computational and Applied Mathematics*, 142:27–37, 2002.

[4] Luc Devroye. Universal asymptotics for random tries and PATRICIA trees. *Algorithmica*, 42(1):11–29, 2005.

[5] Michael Drmota, Abram Magner, and Wojciech Szpankowski. Asymmetric Rényi Problem and PATRICIA Tries. Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms, Kraków, Poland, 4–8 July 2016.

[6] Michael Drmota and Wojciech Szpankowski. The expected profile of digital search trees. *Journal of Combinatorial Theory, Series A*, 118(7):1939–1965, October 2011.

[7] Philippe Flajolet, Xavier Gourdon, and Philippe Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[8] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, Cambridge, UK, 2009.

[9] Philippe Jacquet and Wojciech Szpankowski. Analytical depoissonization and its applications. *Theoretical Computer Science*, 201(1-2):1–62, July 1998.

[10] Svante Janson and Wojciech Szpankowski. Analysis of an asymmetric leader election algorithm. *Electronic Journal of Combinatorics*, 4:1–16, 1996.

[11] Ramin Kazemi and Mohammad Vahidi-Asl. The variance of the profile in digital search trees. *Discrete Mathematics and Theoretical Computer Science*, 13(3):21–38, 2011.

[12] Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2nd ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998.

[13] Abram Magner. *Profiles of PATRICIA Tries*. PhD thesis, Purdue University, December 2015.

[14] Abram Magner and Wojciech Szpankowski. *Profiles of PATRICIA Tries*. manuscript. https://www.cs.purdue.edu/homes/spa/papers/patricia2015.pdf

[15] Abram Magner, Charles Knessl, and Wojciech Szpankowski. Expected external profile of PATRICIA tries. *Proceedings of the Eleventh Workshop on Analytic Algorithmics and Combinatorics*, pages 16–24, 2014.

[16] Gahyun Park, Hsien-Kuei Hwang, Pierre Nicodème, and Wojciech Szpankowski. Profiles of tries. *SIAM Journal on Computing*, 38(5):1821–1880, 2009.

[17] Boris Pittel. Asymptotic growth of a class of random trees. *Annals of Probability*, 18:414–427, 1985.

[18] Boris Pittel and Herman Rubin. How many random questions are needed to identify n distinct objects? *Journal of Combinatorial Theory, Series A*, 55(2):292–312, 1990.

[19] Alfred Rényi. On random subsets of a finite set. *Mathematica*, 3:355–362, 1961.

[20] Wojciech Szpankowski. PATRICIA tries again revisited. *Journal of the ACM*, 37(4):691–711, October 1990.

[21] Wojciech Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, Inc., New York, NY, USA, 2001.