Profile of Tries

Gahyun Park[1], Hsien-Kuei Hwang[2], Pierre Nicodème[3], and Wojciech Szpankowski[4]

Abstract

Tries (from re*trie*val) are one of the most popular data structures on words. They are pertinent to (internal) structure of stored words and several splitting procedures used in diverse contexts. The *profile* of a trie is a parameter that represents the number of nodes (either internal or external) with the same distance to the root. It is a function of the number of strings stored in a trie *and* the distance from the root. Several, if not all, trie parameters such as height, size, depth, shortest path, and fill-up level can be uniformly analyzed through the (external and internal) profiles. Although profiles represent one of the most fundamental parameters of tries, they have been hardly studied in the past. The analysis of profiles is surprisingly arduous but once it is carried out it reveals unusually intriguing and interesting behavior. We present a detailed study of the distribution of the profiles in a trie built over strings generated by a memoryless source. We first derive recurrences satisfied by the expected profiles and solve them asymptotically for all possible ranges of the distance from the root. It appears that profiles of tries exhibit several fascinating phenomena. When moving from the root to the leaves of a trie, the growth of the expected profile varies. Near the root, the external profile is exponentially small (with the number of strings stored), then it decays in a logarithmic rate until it abruptly starts growing, first logarithmically and then polynomially; it then tends polynomially to zero again. Furthermore, the expected profiles of asymmetric tries are oscillating in a range where profiles grow polynomially, while symmetric tries are non-oscillating, in contrast to most shape parameters of random tries studied previously. Such a periodic behavior for asymmetric tries implies that the depth satisfies a central limit theorem, but not a local limit theorem of the usual form. Also the widest levels contain a linear number of nodes in symmetric tries, differing from the order $n/\sqrt{\log n}$ for asymmetric tries, n being the size of the trees. Finally, it is observed that profiles satisfy central limit theorems when the variance goes unbounded while near the height they are distributed according to Poisson laws. As a consequence of these results we find typical behaviors of the height, shortest path, fill-up level, and the depth. These results are derived here by methods of analytic algorithmics such as generating functions, Mellin transform, Poissonization and de-Poissonization, the saddle-point method, singularity analysis and uniform asymptotic analysis.

Key Words: Digital trees, tries, profile, depth, height, shortest path, fill-up level, analytic Poissonization, Mellin transform, saddle-point method, singularity analysis.

[1]Department of Computer Sciences, Purdue University, 250 N. University Street, West Lafayette, Indiana, 47907-2066, USA, gpark@cs.purdue.edu.

[2]Institute of Statistical Science, Academia Sinica, 11529 Taipei, Taiwan, hkhwang@stat.sinica.edu.tw. This work was partially supported by a grant from the National Science Council of Taiwan.

[3]Laboratory LIX, École polytechnique, 91128 Palaiseau Cedex, France, nicodeme@lix.polytechnique.fr.

[4]Department of Computer Sciences, Purdue University, 250 N. University Street, West Lafayette, Indiana, 47907-2066, USA, spa@cs.purdue.edu. This research was sponsored by NSF Grants CCR-0208709, CCF-0513636, and DMS-0503742, AFOSR Grant FA8655-04-1-3074, and NIH Grant R01 GM068959-01.

1 Introduction

Tries are prototype data structures useful for many indexing and retrieval purposes. They were first proposed by de la Briandais [9] in the late 1950's for information processing; Fredkin [28] suggested the current name as it being part of re*trie*val. Tries are multiway trees whose nodes are vectors of characters or digits. Due to their simplicity and efficiency, tries found widespread use in diverse applications ranging from document taxonomy to IP addresses lookup, from data compression to dynamic hashing, from partial-match queries to speech recognition, from leader election algorithms to distributed hashing tables (see [30, 51, 55, 82]). In this paper, we are concerned with probabilistic properties of the profiles of tries, where the *profile* of a tree is the sequence of numbers each counting the number of nodes with the same distance to the root. We discover several new phenomena in the profiles of tries built over strings generated by a random memoryless source, and develop asymptotic tools to describe them.

Structure and usefulness of tries. Tries are natural choice of data structures when the input records involve a notion of alphabets or digits. They are often used to store such data so that future retrieval can be made efficient. Given a sequence of n words over the alphabet $\{a_1, \ldots, a_m\}$, $m \geq 2$, we can construct a trie as follows. If $n = 0$, then the trie is empty. If $n = 1$, then a single (external) node holding the word is allocated. If $n \geq 1$, then the trie consists of a root (internal) node directing words to the m subtrees according to the first alphabet of each word, and words directed to the same subtree are themselves tries (see [51, 55, 82] for more details). For simplicity, we deal only with binary tries in this paper. Unlike other search trees such as digital search trees and binary search trees where records or keys are stored at the internal nodes, the *internal nodes* in tries are branching nodes used merely to direct records to each subtries, records being all stored in *external nodes* that are leaves of such tries. A trie has more internal nodes than external nodes (fixed to be n throughout this paper), differing from almost all other search trees. In Figure 1 we plot a binary trie of 5 strings.

The simple organizing procedure used to construct tries and the general efficiency they achieve make tries one of the most popular digital search trees. Since their invention, tries have found frequent use in many computer science applications. For example, tries are widely used in algorithms for automatically correcting words in texts (see [53]) and in algorithms for taxonomies and toolkits of regular language (see the Ph. D. Thesis [83]); they are also used to represent the event history in datarace detection for multi-threaded object-oriented programs (see [6]); another example is the internet IP addresses lookup problem (see [62, 77]), where the search time for the IP address problem is directly related to the distribution of the fill-up level (see below for a more precise definition) and other trie parameters. For applications to other problems in searching, sorting, dynamic hashing, coding, polynomial factorization, Lempel-Ziv compression schemes, and molecular biology, see [30, 82].

The structure of tries also have a close connection to several splitting procedures using coin-flipping; these include algorithms for resolving collisions in multi-access (or broadcast) communication models and algorithms for loser selection or leader election, etc.; see [45]. Thus most shape parameters in tries have direct interpretations in terms of other related objects.

Random tries under the Bernoulli model. Throughout the paper, we write $B_{n,k}$ to denote the number of external nodes (leaves) at distance k from the root; the number of internal nodes at distance k from the root is denoted by $I_{n,k}$. For simplicity, we will refer to $B_{n,k}$ as the *external profile* and $I_{n,k}$ the *internal profile*. Figure 1 shows a trie and its profiles.

In this paper we study the profiles of a trie built over n binary strings generated by a memoryless source. More precisely, we assume that the input is a sequence of n independent and identically distributed random

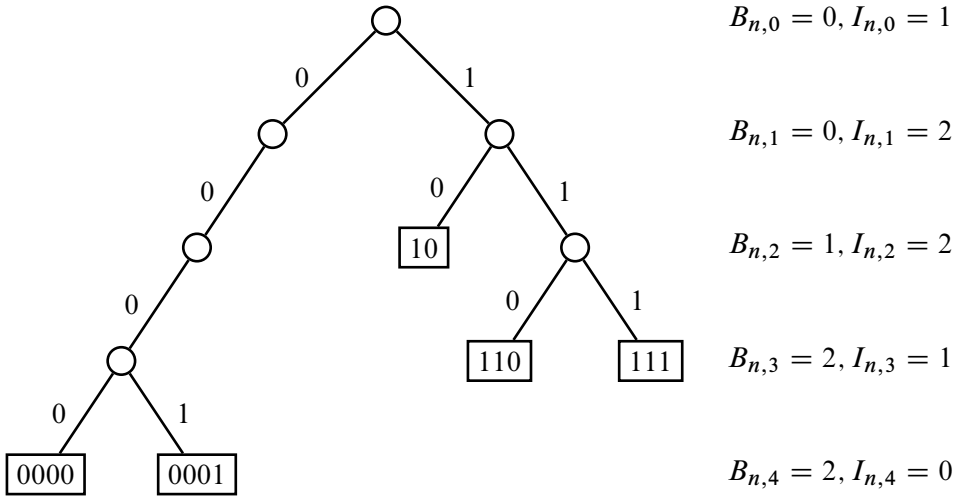

Figure 1: *A trie of $n = 5$ records and its profiles: the circles represent internal nodes and rectangles holding the records are external nodes.*

variables, each being composed of an infinite sequence of Bernoulli random variables with mean p, where $0 < p < 1$ is the probability of a "1" and $q := 1 - p$ is the probability of a "0". The corresponding trie constructed from these n bit-strings is called a *random trie*. This simple model may seem too idealized for practical purposes, however, the typical behaviors under such a model often hold under more general models such as Markovian or dynamical sources, although the technicalities are usually more involved; see for example [8, 12, 15, 36].

The motivation of studying the profiles is multifold. First, they are fine shape measures closely connected to many other cost measures on tries; some of them are indicated below. Second, they are also asymptotically close to the profiles of suffix trees, which in turn have a direct combinatorial interpretation in terms of words; see [37, 61, 81, 82] for more information and another interpretation in terms of urn models. Third, not only the analytic problems are mathematically challenging, but the diverse new phenomena they exhibit are highly interesting and unusual. Fourth, our findings imply several new results on other shape parameters (see Section 8). Finally, most properties of random tries have also a prototype character and are expected to hold for other varieties of digital search trees (and under more general random models), although the proofs are generally more complicated.

Major cost measures on random tries. Due to the usefulness of tries, many cost measures, discussed below, on random tries have been studied in the literature since the early 1970's, and most of these measures can be expressed and analyzed through the profiles studied in this paper:

- *depth*: the distance from the root to a randomly selected node; its distribution is given by the expected external profile divided by n; see [10, 12, 13, 21, 34, 37, 43, 54, 67, 74, 78, 79];
- *total path length*: the sum of distances between nodes and the root, or equivalently, $\sum_j jI_{n,j}$; see [8, 11, 44, 60, 59, 73, 74, 75, 78];
- *size*: the total number of internal nodes, or $\sum_j I_{n,j}$; see [8, 35, 37, 46, 51, 59, 69, 70, 73, 74, 75];
- *height*: the length of the longest path from the root, or $\max\{j : B_{n,j} > 0\}$; see [8, 11, 12, 13, 14, 23, 27, 34, 66, 67, 80];

- *shortest path*: the length of the shortest path from the root to an external node, or $\min\{j : B_{n,j} > 0\}$; see [66, 67];
- *fill-up (or saturation) level*: the largest full level, or $\max\{j : I_{n,j} = 2^j\}$, where the *levels* of a tree denote the sets of nodes with the same distance to the root; see [50];
- *Horton-Strahler number and stack-size*: certain notions of heights related to the traversal of tries; see [4, 17, 56, 57, 58];
- *distance of two randomly chosen nodes*; see [1, 7];
- *pattern occurrences in tries (including page usage or b-tries)*; see [23, 43, 46, 60, 74, 79];
- *one-sided height (or leader election or loser selection)*; see [22, 39, 68, 84, 85].

The reader is referred to the book [82] and the papers [15, 38, 74] for a systematic treatment of several of these quantities.

The general analytic context. The major difference between most previous study and the current paper is that we are dealing with asymptotics of bivariate recurrence, in contrast to univariate recurrences (with or without maximization or minimization) addressed in the literature.

To be more precise, we observe that by assumption of the model, the probability generating function $P_{n,k}(y) := \mathbb{E}(y^{B_{n,k}})$ of the external profile satisfies the recurrence

$$P_{n,k}(y) = \sum_{0\le j\le n} \binom{n}{j} p^j q^{n-j} P_{j,k-1}(y) P_{n-j,k-1}(y) \qquad (n \ge 2; k \ge 1), \tag{1}$$

with the initial conditions $P_{n,k}(y) = 1 + \delta_{n,1}\delta_{k,0}(y-1)$ when either $n \le 1$ and $k \ge 0$ or $k = 0$ and $n \ge 0$, where $\delta_{a,b}$ is the Kronecker symbol. Observe that this recurrence depends on two parameters n and k which makes the analysis quite challenging, as we will demonstrate in this paper. The probability generating functions of the internal profile satisfy the same recurrence (1) but with different initial conditions; see Section 6.

From (1), the moments of $B_{n,k}$ and $I_{n,k}$ (centered or not) are seen to satisfy a recurrence of the form

$$x_{n,k} = a_{n,k} + \sum_{0\le j\le n} \binom{n}{j} p^j q^{n-j} \left(x_{j,k-1} + x_{n-j,k-1}\right),$$

with suitable initial conditions, where $a_{n,k}$ are known (either explicitly or inductively). A standard approach is to consider the Poisson generating function $\tilde{f}_k(z) := e^{-z}\sum_n x_{n,k} z^n/n!$, which in turn satisfies the functional equation

$$\tilde{f}_k(z) = \tilde{g}_k(z) + \tilde{f}_{k-1}(pz) + \tilde{f}_{k-1}(qz),$$

with a suitable $\tilde{g}_k(z)$. This equation can be solved explicitly by a simple iteration argument and asymptotically by using the Mellin transform (see [24, 82]). The final step is to invert from the asymptotics of the Poisson generating function $\tilde{f}_k(z)$ to recover the asymptotics of $x_{n,k}$. This last step is guided by the *Poisson heuristic*, which roughly states that

$$\textit{if a sequence } \{x_n\}_n \textit{ is ``smooth enough,'' then } x_n \sim e^{-n}\sum_{j\ge 0} x_j n^j/j! \tag{2}$$

where $x_n \sim y_n$ if $\lim_{n\to\infty} x_n/y_n = 1$. Such a Poisson heuristic appeared in diverse contexts under different forms such as Borel summability and Tauberian theorems; it dated back to at least Ramanujan's Notebooks;

see the book by Berndt [3, pp. 57–66] for more details. It is known as *analytic de-Poissonization*, when justified by complex analysis and the saddle-point method, and was the subject of intensive analysis, resulting in a robust solution presented in [38].

By means of the Poisson heuristic (2), we expect that $\mu_{n,k} \sim e^{-n} \sum_{j\geq 0} \mu_{j,k} n^j/j!$. However, as we will see, such a heuristic holds in our case when $q^{2k}n \to 0$ but fails otherwise. The reason is that $\mu_{n,k}$ is too small in this range. Also it should be mentioned that the asymptotic analysis of the above functional equation is in general more intricate because we have an additional parameter k to be taken into account and we need uniformity for our asymptotic approximations in k (varying with n) and in z (in some region in the complex plane) in order to invert the results to obtain $x_{n,k}$ by suitable complex analysis.

Known results for profiles. As far as probabilistic properties of the profiles of random tries are concerned, very little is known in the literature. Since the distribution of the depth D_n in random tries is given by $\mathbb{P}(D_n = k) = \mu_{n,k}/n$, where $\mu_{n,k} := \mathbb{E}(B_{n,k})$, the asymptotics of the expected profile $\mu_{n,k}$ for $n \to \infty$ and varying $k = k(n)$ can be regarded as local limit theorems for D_n. Although many papers addressed the limiting behaviors of the depth, none dealt with the local limit theorem of D_n and the asymptotics of $\mu_{n,k}$ for varying k. We will see in the last section that our result implies an unusual type of local limit theorem for D_n. However, it should be mentioned that the central limit theorem for the depth was developed in [13, 35, 36].

On the other hand, Pittel [67] showed that the distribution of the number of pairs of input-strings having a common prefix of length at least k is asymptotically Poisson when k is close to the height. Devroye [14] showed that

$$\text{if} \quad \frac{\mathbb{E}(B_{n,k})}{\sqrt{n}} \to \infty \quad \text{then} \quad \frac{B_{n,k}}{\mathbb{E}(B_{n,k})} \to 1 \quad \text{in probability;}$$
$$\text{if} \quad \mathbb{E}(I_{n,k}) \to \infty \quad \text{then} \quad \frac{I_{n,k}}{\mathbb{E}(I_{n,k})} \to 1 \quad \text{in probability,}$$

under very general assumptions on the underlying models; see also [15] for further refinements. These represent known results concerning profiles. We will see that convergence in probability in both cases holds as long as the variance tends to infinity.

Sketch of the major phenomena. In the next section we present an in-depth discussion of our results. Here, we briefly summarize our main findings. We focus mostly on the profiles of asymmetric tries (when $p \neq q$) since the symmetric tries (when $p = q = 1/2$) are comparatively easier. We will first derive asymptotic approximations to the average external profile $\mu_{n,k}$ for all ranges of k.

Our results show *inter alia* that for $k \leq (1-\varepsilon)\log n/\log(1/q)$ the average profile $\mu_{n,k}$ is exponentially small, where $\varepsilon > 0$ is small. When k increases and lies in the range $(\log n - \log\log\log n + O(1))/\log(1/q)$, then $\mu_{n,k}$ decays to zero logarithmically until $k > k^*$ for a specific threshold k^* in this range beyond which $\mu_{n,k}$ suddenly grows unbounded in a logarithmic rate. The rate becomes polynomial $\Theta(n^\upsilon)$ for some $0 < \upsilon \leq 1$ when

$$\frac{1}{\log(1/q)}(1+\varepsilon)\log n \leq k \leq \frac{2}{\log(1/(p^2+q^2))}(1-\varepsilon)\log n.$$

Surprisingly enough, for this range of k an oscillating factor emerges in the expected profile behavior, that is, $\mathbb{E}(B_{n,k}) \approx G(\log_{p/q} p^k n)n^\upsilon/\sqrt{\log n}$, where G is a bounded periodic function. Such a behavior is a consequence of an infinite number of saddle-points appearing in the integrand of the associated Mellin integral transform. This was first observed by Nicodème [61]. For larger values of k, these oscillations disappear since the behavior of the expected profile is dominated by a polar singularity.

Analogous results also hold for the internal profile. Also we prove that the variances of both profiles are asymptotically of the same order as their expected values. This suggests a central limit theorem for both external and internal profiles for a wide range of k. We show that this is indeed true; furthermore, we also show that for k near the height the limiting distribution of the profiles becomes Poisson. Some of these results were already anticipated in [64] and they constitute the Ph.D. thesis of the first author [65].

Profiles of digital and non-digital log-trees. In passing, we observe that most random trees in the discrete probability literature fall into two major categories according to their expected height being of order $\sqrt{n}$ (referred to as *square-root trees* for brevity) or of order $\log n$ (referred to as *log trees*), where n is the tree size. While most random square-root trees were introduced in combinatorics and probability, the majority of log trees arise from data structures and computer algorithms.

We can further classify log trees into "digital type" and "non-digital type" log trees, according to the nature of construction (or search) of the tree. Profiles of non-digital type search trees of logarithmic height for which *binary search trees* are representative have received much recent attention, and are showed to exhibit several interesting phenomena such as bimodality of the variance, and multifaceted behaviors of the limiting distributions; see [5, 19, 20, 29, 32] for more information. In contrast, profiles of digital type search trees were much less addressed and most properties remain unknown; see [14, 15, 67] for tries and [2, 40] for digital search trees. We will show that the limiting behaviors of the profiles are very different from those of non-digital search trees. In particular, while in no range will the normalized profiles in random binary search trees lead to asymptotic normality (in the sense of convergence in distribution), profiles of random tries, when properly centered and normalized, all converge to the standard normal law when the variance goes unbounded in the limit. As is often the case for proving asymptotic normality, we need more precise asymptotic approximation to the variance, rendering our analysis more complicated.

Organization of the paper. The paper is organized as follows. In the next section, we (rather informally) present a more detailed summary of our main findings. This section is to help the reader to comprehend the richness of our results in their fullness but without resorting to rather abstruse mathematical formulations. Sections 3–8 are devoted to precise formulations of our results. This paper contains two major parts: The first part, Section 3, develops the asymptotic tools we need for deriving the diverse asymptotic approximations to the expected external profile $\mu_{n,k}$. Most proofs of the second part (Sections 4–8) are then sketched because they extend the same methods of proof as in the first part. Except for Sections 7 and 8, we assume $p \neq q$ throughout this paper. Among these sections, Section 4 derives asymptotics of the variance of $B_{n,k}$, the corresponding results of convergence in distribution being given in Section 5. The internal profiles are addressed in Section 6 and results for symmetric tries are given in Section 7. Consequences of our findings are discussed in Section 8 where we establish typical behaviors of the height, the width, the shortest path, the fill-up level, and the right-profile, as well as a rather atypical local limit theorem for the depth.

2 Summary of main results

In this section we discuss informally our main results. We focus here on describing the major phenomena arising in the analysis of profiles rather than presenting the precise and complicated results to which we devote all the remaining sections of this paper.

Crucial to our analysis of the profiles is the asymptotics of the expected profiles. Not only are the results fundamental and highly interesting, but also the analytic methods we used are of certain generality.

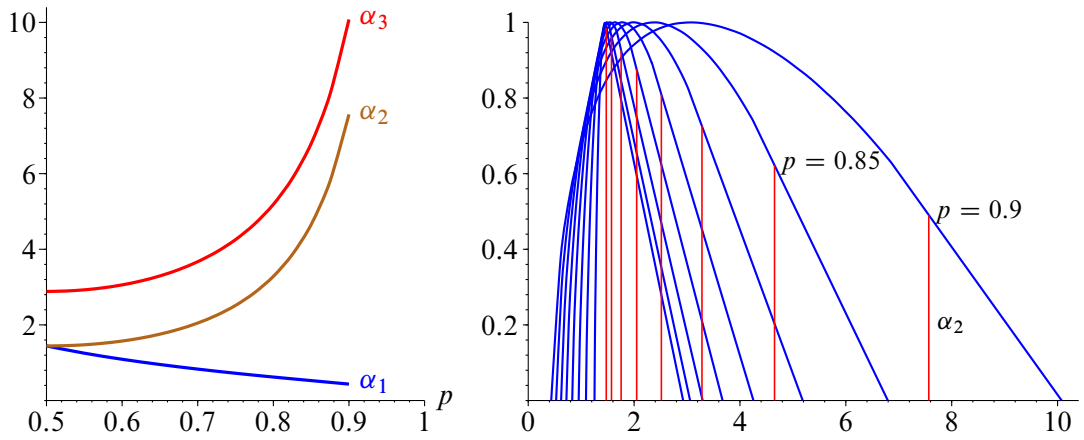

Figure 2: *Left: A plot of α_1, α_2, and α_3 (defined in (5)) as functions of p. Right: The (non-zero) limiting order of $\log \mu_{n,k}/\log n$ plotted against $\alpha = \lim_n k/\log n$ for $p = 0.55, 0.6, \ldots, 0.9$ (the spans of the curves increase as p grows). The vertical lines represent the positions of α_2 (to the right of which the curves are straight lines); see (4).*

From (1), we see that the expected external profile $\mu_{n,k} := \mathbb{E}(B_{n,k})$ satisfies the following recurrence

$$\mu_{n,k} = \sum_{0\le j\le n} \binom{n}{j} p^j q^{n-j} (\mu_{j,k-1} + \mu_{n-j,k-1}), \tag{3}$$

for $n \ge 2$ and $k \ge 1$ with the initial values $\mu_{n,0} = 0$ for all $n \ne 1$ and 1 for $n = 1$. Furthermore, $\mu_{0,k} = 0, k \ge 0$ and $\mu_{1,k} = 0$ for $k \ge 1$ and equal to 1 when $k = 0$.

The polynomial growth of $\mu_{n,k}$. In Section 3, we solve asymptotically (3) for various ranges of k when $p \ne q$; a crude description of the asymptotics of $\mu_{n,k}$ is as follows.

$$\frac{\log \mu_{n,k}}{\log n} \to \begin{cases} 0, & \text{if } \alpha \le \alpha_1; \\ -\rho + \alpha \log(p^{-\rho} + q^{-\rho}), & \text{if } \alpha_1 \le \alpha \le \alpha_2; \\ 2 + \alpha \log(p^2 + q^2), & \text{if } \alpha_2 \le \alpha \le \alpha_3; \\ 0, & \text{if } \alpha \ge \alpha_3, \end{cases} \tag{4}$$

where

$$\alpha_1 := \frac{1}{\log(1/q)}, \quad \alpha_2 := \frac{p^2+q^2}{p^2\log(1/p) + q^2\log(1/q)}, \quad \text{and} \quad \alpha_3 := \frac{2}{\log(1/(p^2+q^2))} \tag{5}$$

are delimiters of $\alpha := \lim_n k/\log n$ ($k = k(n)$), and

$$\rho := \frac{1}{\log(p/q)} \log\left(\frac{1-\alpha\log(1/p)}{\alpha\log(1/q)-1}\right).$$

Note that $\alpha_1 \le \alpha_2$; see Figure 2. The limiting estimate (4) gives a rough picture of $\mu_{n,k}$ as follows: $\mu_{n,k}$ is of polynomial growth rate when $\alpha_1 + \varepsilon \le \alpha \le \alpha_3 - \varepsilon$, and is smaller than any polynomial powers when $0 \le \alpha \le \alpha_1 - \varepsilon$ and $\alpha \ge \alpha_3 + \varepsilon$. Near the two boundaries α_1 and α_3, the behaviors of $\mu_{n,k}$ will undergo phase-changes from being sub-polynomial to being polynomial or the other way around.

More refined asymptotics. To derive more precise asymptotics of $\mu_{n,k}$ than the phase transitions (4) of the polynomial order of $\mu_{n,k}$, we divide all possible values of k into four overlapping ranges.

(I) *Elementary range*: $1 \le k \le \alpha_1(\log n - \log\log\log n + O(1))$;

(II) *Saddle-point range*: $\alpha_1(\log n - \log\log\log n + K_n) \le k \le \alpha_2(\log n - K_n\sqrt{\log n})$;

(III) *Gaussian transitional range*: $k = \alpha_2 \log n + o((\log n)^{2/3})$;

(IV) *Polar singularity range*: $k \ge \alpha_2 \log n + K_n\sqrt{\log n}$,

where, throughout this paper, $K_n \ge 1$ represents a (generic) sequence tending to infinity.

More precisely, in Theorem 1 we prove that for k lying in range (I) the expected external profile $\mu_{n,k}$ decays first exponentially fast (asymptotic to $q^k n(1-q^k)^{n-1}$). Then, when k is around $\alpha_1(\log n - \log\log\log n + \log(p/q - 1) + m\log(p/q))$ for some integer $m \ge 0$,

$$\mu_{n,k} \sim \frac{k^m}{m!} p^m q^{k-m} n e^{-np^m q^{k-m}},$$

which is of order

$$\mu_{n,k} = O\left(\frac{\log\log n}{\log^{\xi-m} n}\right),$$

for some ξ. Thus, for $m < \xi$ the expected external profile decays only logarithmically, but for $m \ge \xi$ it increases logarithmically.

The behavior of $\mu_{n,k}$ in range (II) is described in Theorem 2. The situation becomes highly nontrivial and interesting. More precisely, for $\alpha_1(1+\varepsilon)\log n \le k \le \alpha_2(1-\varepsilon)\log n$, we find that

$$\mu_{n,k} \sim G_1\left(\rho; \log_{p/q} p^k n\right) \frac{p^\rho q^\rho (p^{-\rho} + q^{-\rho})}{\sqrt{2\pi\alpha_{n,k}}\,\log(p/q)} \cdot \frac{n^{\upsilon_1}}{\sqrt{\log n}},$$

where ($\alpha_{n,k} := k/\log n$)

$$\upsilon_1 = -\rho + \alpha_{n,k}\log(p^{-\rho} + q^{-\rho}),$$
$$\rho = -\frac{1}{\log(p/q)} \log\left(\frac{-1 - \alpha_{n,k}\log q}{1 + \alpha_{n,k}\log p}\right),$$

and $G_1(\rho; x)$ is a periodic function. We plot in Figures 3 and 4 the periodic parts of $G_1(-1, x)$ for a few values of p and ρ, respectively. These oscillations are consequences of an infinite number of saddle-points appearing in the integrand of the associated Mellin transform of the expected profile.

Finally, in Theorem 3 we prove that for k in range (IV)

$$\mu_{n,k} \sim \frac{2pq}{p^2 + q^2} n^{\upsilon_2},$$

where $\upsilon_2 = 2 + \alpha_{n,k}\log(p^2+q^2)$, and the periodic function disappear. In this region, the asymptotic behavior of the expected profile is dictated by the expected number of pairs (of input-strings) having common prefixes of length at least k. This property is analytically reflected by a polar singularity in the associated Mellin transform. Asymptotics of $\mu_{n,k}$ in range (III) for $k = \alpha_2 \log n + o(\log^{2/3} n)$ is presented in Theorem 4. In this transitional range, the saddle-point coalesces with the polar singularity, so we use the Gaussian integral to describe the behavior of $\mu_{n,k}$.

In summary, our results roughly state that $\mu_{n,k} \to 0$ when $1 \le k \le k^*$ for some k^* close to $\alpha_1(\log n - \log\log\log n + O(1))$, then $\mu_{n,k}$ tends *abruptly* to infinity at a logarithmic rate when $k > k^*$. Such an

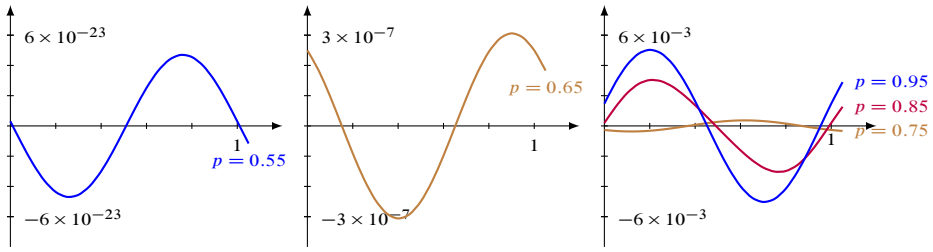

Figure 3: *The fluctuating part of the periodic function* $G_1(-1;x)$ *for* $p = 0.55, 0.65, \ldots, 0.95$ *and for* x *in the unit interval; its amplitude tends to zero when* $p \to 0.5^+$.

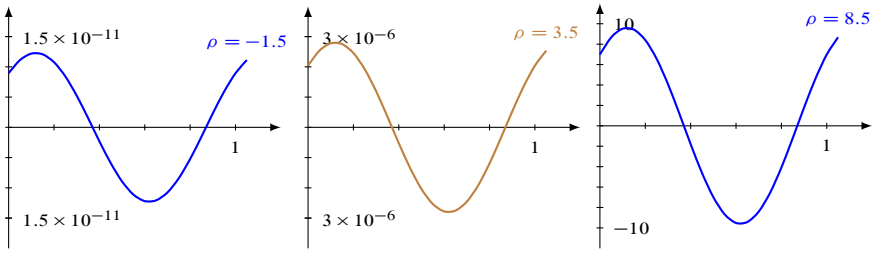

Figure 4: *The fluctuating part of the periodic function* $G_1(\rho;x)$ *for* $\rho \in \{-1.5, 3.5, 8.5\}$ *and* $x \in [0,1]$. *The amplitude increases as* ρ *grows.*

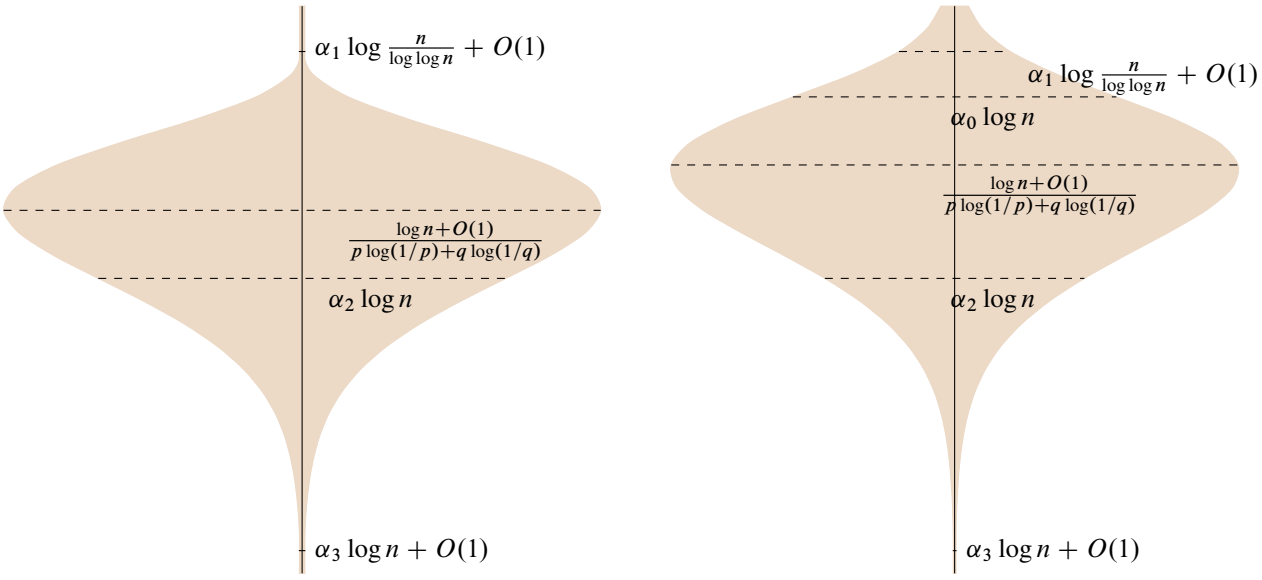

Figure 5: *The silhouettes of the expected external (left) and internal (right) profiles of an asymmetric trie* ($p = 0.75$). *Note that the right subtrees of the asymmetric trie has more nodes than their left siblings since* $p > 1/2$. *Also the first few levels contain almost no external nodes but almost full of internal nodes.*

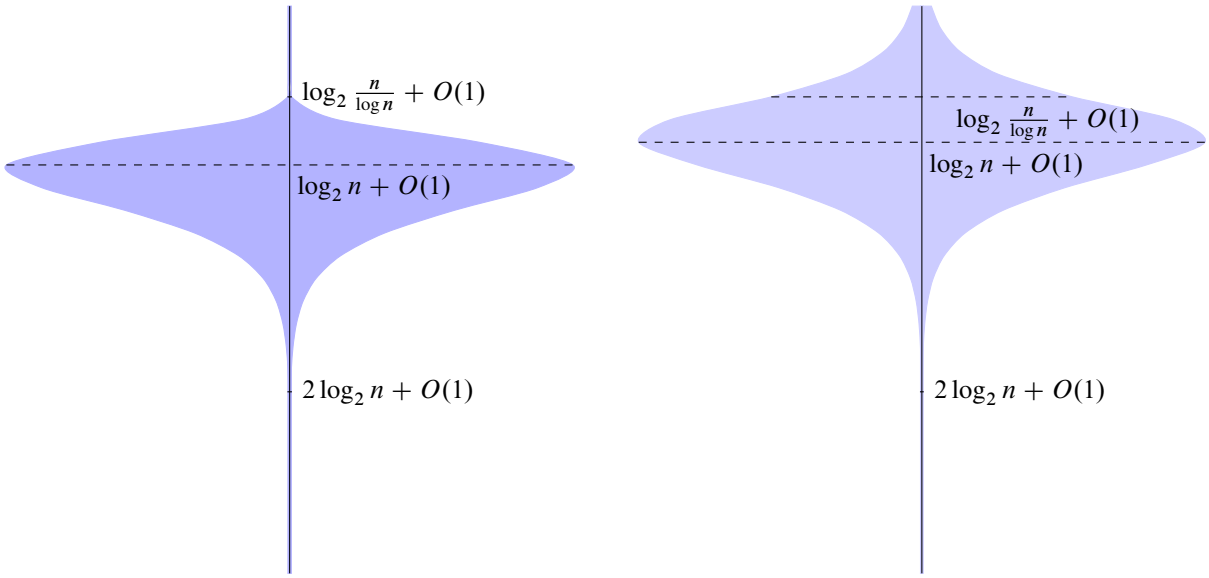

Figure 6: *The silhouettes of the expected external and internal profiles of a symmetric; compare Figure 5.*

abrupt change has already been observed before in the literature for the shortest path and the fill-up level (see [50, 67]), but not much is known for $\mu_{n,k}$ beyond that. Then we show that $\mu_{n,k}$ grows polynomially when k lies in the range $\alpha_1(1+\varepsilon)\log n \le k \le \alpha_3(1-\varepsilon)\log n$, reaching the peak where it is of order $n/\sqrt{\log n}$; it decays in a slower rate afterwards until it tends to zero again when $k \ge \alpha_3(\log n + K_n)$. A salient feature here is the presence of an oscillating function in the asymptotic approximation when $p \neq q$[1]. In Figure 5, a plot of the rough silhouettes of $\mu_{n,k}$ is presented.

Asymptotics of the expected internal profile. The expected value of the internal profile $\mathbb{E}(I_{n,k})$ is discussed in Section 6. In particular, the expected internal profile is asymptotically equivalent to 2^k for $k \le \alpha_0(\log n - K_n\sqrt{\log n})$, where $\alpha_0 := 2/(\log(1/p) + \log(1/q))$. When $k \ge \alpha_2(\log n + K_n\sqrt{\log n})$, then $\mathbb{E}(I_{n,k}) \sim (p^2+q^2)\mathbb{E}(B_{n,k})/pq$. Between these two ranges, it is again the infinite number of saddle-points that yield the dominant asymptotic approximation. Unlike $\mu_{n,k}$, an additional phase transition appears in the asymptotics of the $\mathbb{E}(I_{n,k})$ when $k = \alpha_0 \log n + O(\sqrt{\log n})$, reflecting the structural change of the internal nodes from being asymptotically full to being of the same order as the number of external nodes. The silhouettes of the expected internal profiles for a symmetric trie and an asymmetric ($p = 0.75$) trie are presented in Figure 6.

Variance and limiting distributions. In Section 4 we deal with the variance of the profile. In particular, in Theorem 7 we derive asymptotic approximations to the variance of the profile, which asymptotically turns out to be of the same order as the expected value for all ranges of $k \ge 1$, namely, $\mathbb{V}(B_{n,k}) = \Theta(\mathbb{E}(B_{n,k}))$. In fact, we show that $\mathbb{V}(B_{n,k}) \sim \mathbb{E}(B_{n,k})$ in range (I), for range (IV) $\mathbb{V}(B_{n,k}) \sim 2\mathbb{E}(B_{n,k})$, while in range (II) (polynomial growth) the variance and the expected profile differ only by the oscillating functions. The variance of the internal profile behaves almost identically to the variance of the external profile; roughly, $\mathbb{V}(I_{n,k}) = \Theta(\mathbb{V}(B_{n,k}))$ for all k. The methods used to derive these results are the same as the ones used in

[1]The expected values of many shape characteristics of random tries often exhibit the asymptotic pattern: $\sim F(\log_c n)n$ if $\log p/\log q$ is rational for some periodic function F and constant c expressible in terms of p, and $\sim Cn$ if $\log p/\log q$ is irrational; see [38, 74, 82]

Section 3.

We then prove, in Section 5, that both internal and external profiles, after proper normalization, are asymptotically normally distributed if and only if the variance tends to infinity (see Theorems 8 and 9). The limiting distribution is Poisson when the variance remains bounded away from zero and infinity. In particular, we will prove that when $\mathbb{V}(B_{n,k}) = \Theta(1)$, then

$$\mathbb{P}\left(B_{n,k} = 2m\right) = \frac{\lambda_0^m}{m!}\, e^{-\lambda_0} + o(1) \quad \text{and} \quad \mathbb{P}\left(B_{n,k} = 2m+1\right) = o(1),$$

where $\lambda_0 := pqn^2(p^2+q^2)^{k-1}$, while for $\mathbb{V}(I_{n,k}) = \Theta(1)$, we find

$$\mathbb{P}(I_{n,k} = m) = \frac{\lambda_1^m}{m!}\, e^{-\lambda_1} + o(1) \qquad (m = 0, 1, \dots),$$

where $\lambda_1 := n^2(p^2+q^2)^k/2$. These results hold for both symmetric and asymmetric tries, but the ranges where the variances become unbounded are different.

Symmetric tries. For the symmetric case, we have $\alpha_1 = \alpha_2 = 1/\log 2$. This means that the two ranges separated by α_2 coalesce into one for symmetric tries. The analysis then becomes simpler as shown in Section 7. An interesting property is that unlike asymmetric tries, the fattest levels of profiles of symmetric tries contain a linear number of nodes. The global picture of a random symmetric trie is roughly as follows ($\alpha_1 = 1/\log 2$):

- When $1 \le k \le \alpha_1(\log n - \log\log n + O((\log n)^{-1}))$, each level is almost full of internal nodes ($I_{n,k} \approx 2^k$), the number of external nodes tending to zero; in particular, the variances of both profiles tend to zero.
- When $\alpha_1(\log n - \log\log n + K_n/\log n) \le k \le 2\alpha_1(\log n - K_n)$, where K_n is any sequence tending to infinity, the variances of both profiles tend to infinity, and we prove the asymptotic normality of both profiles.
- When $k = 2\alpha_1(\log n + O(1))$, both profiles are asymptotically Poisson distributed, but $B_{n,k}$ assumes only even values.
- When $k \ge \alpha_1(\log n + K_n)$, then nodes appear very unlikely.

The last Section 8 describes some consequences of our main results. In particular, we point out a rather unusual form of the local limit theorem for the depth due to the oscillating factor in the expected profile. Then we apply our results to re-derive typical behavior for the height, shortest path and the fill-up level. Also the width and right-profile (counting only right branches and neglecting the left ones) are briefly discussed.

This completes the summary of our main results. Precise formulations and proofs are presented in the next five sections. Enjoy the reading!

3 Expected external profile

We derive asymptotic approximations to the expected external profile $\mu_{n,k}$ in this section, starting from a few useful expressions for $\mu_{n,k}$.

Notation. Throughout this paper, $p \in [1/2, 1)$ is fixed and $q = 1-p$. Let $k = k(n)$ and $\alpha := \lim_n k/\log n$, whenever the limit exists. The constants α_1, α_2, and α_3 are defined in (5). For convenience, we also write

$$L_n := \log n, \quad LL_n := \log\log n, \quad LLL_n := \log\log\log n.$$

The generic symbol ε is always used to represent a suitably small constant whose value may vary from one occurrence to another, and K_n denotes any sequence tending to infinity. The symbol $f(n) = \Theta(g(n))$ means that there are positive constants C and C' such that $C|g(n)| \le |f(n)| \le C'|g(n)|$.

3.1 Exact expressions and integral representations

Let $M_k(z) := \sum_{n\ge 0} \mu_{n,k} z^n/n!$ denote the exponential generating function of $\mu_{n,k}$ and $\tilde{M}_k(z) := e^{-z} M_k(z)$ be the Poisson generating function.

Lemma 1. *The Poisson generating function $\tilde{M}_k(z)$ satisfies the integral representation*

$$\tilde{M}_k(z) = \frac{1}{2\pi i} \int_{(\rho)} z^{-s} \Gamma(s+1) g(s) \left(p^{-s} + q^{-s}\right)^k \mathrm{d}s, \tag{6}$$

for $k \ge 1$ and $\Re(z) > 0$, where Γ denotes the Gamma function, $g(s) := 1 - 1/(p^{-s} + q^{-s})$ and the integration path $\int_{(\rho)}$ stands for the integral (upwards) along the vertical line with real part equal to ρ. The integral with $\rho > -2$ is absolutely convergent for $\Re(z) > 0$.

Proof. By taking derivative with respective to y on both sides of (1) and then substituting $y = 1$, we see that $\mu_{n,k}$ satisfies the recurrence (3) with the initial conditions $\mu_{n,k} = \delta_{n,1}\delta_{k,0}$ when either $n \le 1$ and $k \ge 0$ or $k = 0$ and $n \ge 0$. Note that

$$\mu_{n,1} = n\left(pq^{n-1} + qp^{n-1}\right) \qquad (n \ge 2).$$

It follows that

$$M_k(z) = e^{qz} M_{k-1}(pz) + e^{pz} M_{k-1}(qz) \qquad (k \ge 2),$$

with $M_1(z) = z(pe^{qz} + qe^{pz} - 1)$. Thus $\tilde{M}_k(z)$ satisfies

$$\tilde{M}_k(z) = \tilde{M}_{k-1}(pz) + \tilde{M}_{k-1}(qz). \tag{7}$$

Iterating this equation leads to

$$\tilde{M}_k(z) = \sum_{0\le j<k} \binom{k-1}{j} \tilde{M}_1(p^j q^{k-1-j} z), \tag{8}$$

from which (6) follows since the Mellin transform

$$M_1^*(s) = \int_0^\infty \tilde{M}_1(z) z^{s-1} ds$$

of $\tilde{M}_1(z) = pze^{-pz} + qze^{-qz} - ze^{-z}$ equals

$$M_1^*(s) = \Gamma(s+1)(p^{-s} + q^{-s} - 1),$$

where $\Re(s) > -2$, and the Mellin transform of $M_1(p^j q^{k-1-j} z)$ is

$$p^{-sj} q^{(-s(k-1-j)} M_1^*(s)$$

(see [24, 82]).

To justify the absolute convergence of the integral, we apply the Stirling formula for the Gamma function (with complex parameter)

$$\Gamma(s+1) = \sqrt{2\pi s}\left(\frac{s}{e}\right)^s \left(1 + O\left(|s|^{-1}\right)\right),$$

uniformly as $|s| \to \infty$ and $|\arg s| \le \pi - \varepsilon$, which implies that

$$|\Gamma(\rho + it)| = \Theta(|t|^{\rho-1/2} e^{-\pi|t|/2}), \tag{9}$$

uniformly for $|t| \to \infty$ and $\rho = o(|t|^{2/3})$.

The integrand in (6) is analytic for $\Re(s) > -2$ and bounded above by

$$z^{-\rho-it}\Gamma(\rho+1+it)g(\rho+it)\left(p^{-\rho-it} + q^{-\rho-it}\right)^k = O\left(|z|^{-\rho}|t|^{\rho+1/2}e^{-\pi|t|/2+\arg(z)t}(p^{-\rho}+q^{-\rho})^k\right),$$

for large $|t|$. This completes the proof of the lemma. □

Corollary 1. *The expected external profile $\mu_{n,k}$ satisfies, for $n, k \ge 1$,*

$$\mu_{n,k} = \sum_{0\le j\le k}\binom{k}{j} p^j q^{k-j} n\left(1 - p^j q^{k-j}\right)^{n-1} - \sum_{0\le j<k}\binom{k-1}{j} p^j q^{k-1-j} n\left(1 - p^j q^{k-1-j}\right)^{n-1}, \tag{10}$$

and the integral representation

$$\mu_{n,k} = \frac{1}{2\pi i}\int_{(\rho)} \frac{\Gamma(n+1)\Gamma(s+1)}{\Gamma(n+1+s)} g(s)\left(p^{-s}+q^{-s}\right)^k \mathrm{d}s \qquad (\rho > -2). \tag{11}$$

Proof. By definition and (8)

$$M_k(z) = \sum_{0\le j<k}\binom{k-1}{j} p^j q^{k-1-j} z\left(pe^{(1-p^{j+1}q^{k-1-j})z} + qe^{(1-p^j q^{k-j})z} - e^{(1-p^j q^{k-1-j})z}\right).$$

Thus (the symbol $[z^n]f(z)$ denoting the coefficient of z^n in the Taylor expansion of $f(z)$)

$$\begin{aligned}\mu_{n,k} &= n![z^n]M_k(z)\\ &= \sum_{0\le j<k}\binom{k-1}{j}\left(p^{j+1}q^{k-1-j}n\left(1-p^{j+1}q^{k-1-j}\right)^{n-1} + p^j q^{k-j} n\left(1-p^j q^{k-j}\right)^{n-1}\right)\\ &\quad - \sum_{0\le j<k}\binom{k-1}{j} p^j q^{k-1-j} n\left(1-p^j q^{k-1-j}\right)^{n-1}.\end{aligned}$$

Rearranging the indices of the first sum, we obtain (10).

On the other hand, it also follows from (7) that, denoting by $\tilde{\mu}_{n,k} := n![z^n]\tilde{M}_k(z)$,

$$\tilde{\mu}_{n,k} = (p^n + q^n)\tilde{\mu}_{n,k-1} \qquad (k \ge 2),$$

with

$$\tilde{\mu}_{n,1} = \sum_{2\le j\le n}\binom{n}{j}(-1)^{n-j}\mu_{j,1} = (-1)^n n\left(1 - p^n - q^n\right) \qquad (n \ge 1).$$

Iterating this recurrence yields

$$\tilde{\mu}_{n,k} = (-1)^n n \left(1 - p^n - q^n\right) \left(p^n + q^n\right)^{k-1} \qquad (n, k \geq 1).$$

By definition, we have for $n \geq 2$

$$\mu_{n,k} = \sum_{0 \leq j \leq n} \binom{n}{j} \tilde{\mu}_{j,k} = \sum_{2 \leq j \leq n} \binom{n}{j} (-1)^j j \left(1 - p^j - q^j\right) \left(p^j + q^j\right)^{k-1}. \tag{12}$$

The last sum falls under the so called Rice integral representation for finite differences (see [26, 82]) from which we conclude

$$\mu_{n,k} = \frac{1}{2\pi i} \int_{(\rho)} \frac{\Gamma(n+1)\Gamma(-s)}{\Gamma(n+1-s)} s \left(1 - p^s - q^s\right) \left(p^s + q^s\right)^{k-1} \mathrm{d}s.$$

This gives (11). Absolute convergence of the integral in (11) when $\Re(s) > -2$ is justified as above. Note that $g(-1) = 0$. □

Remarks. The integral representation (11) follows formally from interchanging the Cauchy and Mellin integrals as shown below

$$\begin{aligned} \mu_{n,k} &= \frac{n!}{2\pi i} \int z^{-n-1} e^z \tilde{M}_k(z) \mathrm{d}z \\ &= \frac{n!}{2\pi i} \int z^{-n-1} e^z \left(\frac{1}{2\pi i} \int z^{-s} \Gamma(s+1) g(s) \left(p^{-s} + q^{-s}\right)^k \mathrm{d}s \right) \mathrm{d}z \\ &= \frac{n!}{2\pi i} \int \Gamma(s+1) g(s) \left(p^{-s} + q^{-s}\right)^k \left(\frac{1}{2\pi i} \int z^{-n-1-s} e^z \mathrm{d}z \right) \mathrm{d}s \\ &= \frac{n!}{2\pi i} \int \frac{\Gamma(s+1)}{\Gamma(n+s+1)} g(s) \left(p^{-s} + q^{-s}\right)^k \mathrm{d}s. \end{aligned} \tag{13}$$

Although all steps here can be justified by analytic properties of the functions involved (which is essentially the estimates needed by the saddle-point method), the way we proved (11), based solely on finite differences, does not rely on any analytic properties.

Note that since the Mellin transform of $x(1-x)^{n-1}$, $x \in (0,1)$, equals $\Gamma(n)\Gamma(s+1)/\Gamma(n+1+s)$, the exact expression (10) also follows from (11) by expanding $(p^{-s} + q^{-s})^k$ and then integrating term by term. For numerical purposes, the expression (10) is preferable to (12), especially when k is not too large.

On the other hand, the closed-form expression (10) can also be proved directly by either a direct combinatorial argument (see [67] for similar details) or an urn model argument (see [61]).

3.2 Road map of the proof through de-Poissonization

From the preceding analysis, we have two different integral representations at choice: the Rice integral (11) and the Cauchy integral (13). The approach via Rice integral (11) is simpler than that via Cauchy and Mellin integrals (13), but the latter can be easily amended for computing the variance and limiting distribution as will be evident from Sections 4 and 5. It is for this reason that we use here the route via Cauchy and Mellin integrals.

By the Poisson heuristic (2), we anticipate the asymptotic equivalence $\mu_{n,k} \sim \tilde{M}_k(n)$. We will see that this holds when $q^{2k} n \to 0$ but requires suitable modification when $q^{2k} n \not\to 0$.

A simple analytic de-Poissonization result. Define a sequence of (Charlier) polynomials $\tau_\ell(n)$ by

$$\tau_\ell(n) := n![z^n]e^z(z-n)^\ell = \ell![z^\ell](1+z)^n e^{-nz} \qquad (\ell = 0, 1, \dots).$$

Then $\tau_0(n) = 1$, $\tau_1(n) = 0$, $\tau_2(n) = -n$, $\tau_3(n) = 2n$, and $\tau_4(n) = 3n^2 - 6n$. Note that $\tau_\ell(n)$ is a polynomial in n of degree $\lfloor \ell/2 \rfloor$.

Proposition 1. *Let $f(z) := \sum_n a_n z^n/n!$ be an entire function and $\tilde{f}(z) := e^{-z} f(z)$ be the Poisson transform of f, where a_n is a given sequence. Write $z = re^{i\theta}$. If*

$$|f(z)| \le f(r)e^{-cr\theta^2} \tag{14}$$

holds uniformly for $r \ge 0$ and $|\theta| \le \pi$, where $f(r) \ge 0$, and

$$\tilde{f}^{(\ell)}(ne^{i\theta}) = O\left(\delta(n)^\ell \tilde{f}(n)\right) \qquad (\ell = 0, 1, \dots), \tag{15}$$

uniformly for $|\theta| \le \theta_1$, where $\theta_1 \ge n^{-1/2+\varepsilon}$ and $\delta(n) = o(n^{-1/2})$, then for any $\ell_0 \ge 2$

$$a_n = \sum_{0 \le \ell < \ell_0} \frac{\tilde{f}^{(\ell)}(n)}{\ell!} \tau_\ell(n) + O\left(n^{\ell_0/2}\delta(n)^{\ell_0} \tilde{f}(n)\right). \tag{16}$$

Proof. By the Cauchy formula and the condition (14), we need only to assess the integral

$$a_n = \frac{n!}{2\pi i} \int_{\substack{|z|=n \\ |\arg(z)| \le \theta_0}} z^{-n-1} e^z \tilde{f}(z) \mathrm{d}z + O\left(n! n^{-n} f(n) \int_{\theta_0}^{\infty} e^{-cn\theta^2} \mathrm{d}\theta\right) \tag{17}$$

where $\theta_0 = n^{-2/5}$. By Stirling's formula, we see that the O-term in (17) is bounded above by

$$O\left(n^{1/2} \tilde{f}(n) n^{-1/2} e^{-cn^{1/5}}\right) = O\left(e^{-cn^{1/5}} \tilde{f}(n)\right), \tag{18}$$

which is negligible in comparison to the main term $\tilde{f}(n)$. It remains to evaluate the first term in (17). To that purpose, we expand $\tilde{f}(z)$ at $z = n$ and then integrate term by term, the error term introduced being of the form

$$\frac{n!}{2\pi i (\ell_0 - 1)!} \int_{\substack{|z|=n \\ |\arg(z)| \le n^{-2/5}}} z^{-n-1} e^z (z-n)^{\ell_0} \int_0^1 (1-t)^{\ell_0 - 1} \tilde{f}^{(\ell_0)}(n + (z-n)t) \mathrm{d}t \mathrm{d}z,$$

for any $\ell_0 \ge 1$, which is easily seen, by (15), to be bounded above by the O-term in (16); see [31] for similar details. Since $\delta(n) = o(n^{-1/2})$, this proves the asymptotic nature of (16). □

Remark. In particular, we have

$$\begin{aligned} a_n &= \tilde{f}(n) + O(n\delta^2(n)\tilde{f}(n)), \\ a_n &= \tilde{f}(n) - \frac{n}{2}\tilde{f}''(n) + O\left(n^2\delta^4(n)\tilde{f}(n)\right), \end{aligned} \tag{19}$$

for large n.

The theorem indicates that, when the *regularity condition* (14) and the *smoothness condition* (15) both hold for $\tilde{M}_k(z)$, the asymptotics of $\mu_{n,k}$ is reduced to that of its Poisson generating function $\tilde{M}_k(z)$ for large z near the real axis. Our effort in this section is mostly devoted to finding the uniform bounds for justifying the de-Poissonization result (16), which holds for $\mu_{n,k}$ when $q^{2k}n \to 0$. Note that although the condition (14) may seem too strong for our purposes, it can be checked rather systematically in the cases studied in this paper; see [38] for weaker conditions.

On the other hand, we show that when (16) fails (which is the case when $q^{2k}n \not\to 0$), the same proof given above through the Cauchy integral (17) can be appropriately amended because (18) also holds in this case. Thus when deriving our asymptotic estimates for $\mu_{n,k}$, we will either follow the de-Poissonization route through Proposition 1 or evaluate the integral (13) directly using (17).

3.3 Range (I): An elementary analysis

We show in this section that when $1 \le k \le \alpha_1(L_n - LLL_n + O(1))$, the asymptotics of $\mu_{n,k}$ are dictated by one or two terms in the first sum of (10). Although asymptotics of $\mu_{n,k}$ in this range can be easily derived by (10) using only elementary arguments, we will use a lengthier analytic approach based on Cauchy's integral representation since this approach is later on readily amended for the asymptotics of the variance. Define

$$k_m := \alpha_1 \left(L_n - LLL_n + \log\left(\frac{p}{q} - 1\right) + m \log \frac{p}{q} \right) \qquad (m \ge 0), \tag{20}$$
$$S_{n,k,j} := \binom{k}{j} p^j q^{k-j} n \left(1 - p^j q^{k-j}\right)^{n-1} \qquad (0 \le j \le k).$$

For convenience, define $k_{-1} = 0$.

Our first result says that $\mu_{n,k}$ is asymptotic to $S_{n,k,m}$ when $k_{m-1} < k < k_m$ except when k is close to the boundaries, where the corresponding neighboring term (either $S_{n,k,m-1}$ or $S_{n,k,m+1}$) is of the same order.

Theorem 1 (Asymptotics of $\mu_{n,k}$ in Range (I))**.** *Assume $m \ge 0$. If*

$$k_{m-1} + \frac{\alpha_1 K_n}{LL_n} \le k \le k_m - \frac{\alpha_1 K_n}{LL_n}, \tag{21}$$

then

$$\mu_{n,k} = S_{n,k,m} \left(1 + O((m+1)e^{-K_n})\right). \tag{22}$$

If $k = k_m + \alpha_1 x / LL_n$, where $x = o(\sqrt{LL_n})$, then

$$\mu_{n,k} = S_{n,k,m} \left(1 + \frac{p\alpha_1 e^x}{q(m+1)}\right) \left(1 + O\left(x^2 LL_n^{-1} + (m+1) L_n^{-(1-q/p)}\right)\right). \tag{23}$$

Remark. Since $\log(p/q) < 1$ for $p \in (1/2, e/(e+1))$, the interval (21) may contain no integer in it.

By Theorem 1, the proofs of the following special cases are straightforward.

Corollary 2. *If $k \ge 1$ and $q^k n \to \infty$, then*

$$\mu_{n,k} \sim q^k n (1 - q^k)^{n-1};$$

if $q^{2k} n \to 0$ and $k \le \alpha_1(L_n - LLL_n + K_n)$, then

$$S_{n,k,m} \sim \frac{k^m}{m!} p^m q^{k-m} n e^{-p^m q^{k-m} n} \qquad (m \ge 0).$$

On the other hand, the estimate

$$\mu_{n,k} = \Theta(S_{n,k,m}) \tag{24}$$

holds uniformly for $k_{m-1} \le k \le k_m$, $m \ge 0$.

The proof of Theorem 1 is based on evaluating the Cauchy integral (13) along the circle $|z| = n$. By the same arguments used in the proof of Proposition 1 (see (18)). Observe that

$$\mu_{n,k} = \frac{n!}{2\pi i} \int_{\substack{|z|=n \\ |\arg(z)| \le \theta_0}} z^{-n-1} e^z \tilde{M}_k(z) \mathrm{d}z + O\left(e^{-cn^{1/5}} \tilde{M}_k(n)\right), \tag{25}$$

where the O-term is justified by applying the following estimate for $M_k(z)$.

Lemma 2. *Uniformly for $r \geq 0$ and $|\theta| \leq \pi$*

$$|M_k(re^{i\theta})| \leq M_k(r)e^{-cr\theta^2} \qquad (r > 0; |\theta| \leq \pi), \tag{26}$$

for all $k = k(n) \geq 1$ and some constant $c > 0$.

The proof of (26) follows directly from the next proposition in view of (8) and $[z^n]M_1(z) \geq 0$.

Proposition 2. *Let $f(z)$ be an entire function and $z = re^{i\theta}$, where $r \geq 0$ and $|\theta| \leq \pi$. If*

$$|e^z f(z)| \leq e^r f(r) \qquad (r \geq 0; |\theta| \leq \pi), \tag{27}$$

where $f(r) \geq 0$, then the sum $f_k(z) := \sum_{0\leq j\leq k} \binom{k}{j} f(p^j q^{k-j} z)$ satisfies

$$|e^z f_k(z)| \leq e^r f_k(r) e^{-cr\theta^2}, \tag{28}$$

uniformly for $k \geq 0$, $r \geq 0$ and $|\theta| \leq \pi$, where $c > 0$ is independent of z and k.

Proof. By (27) and the elementary inequality

$$1 - \cos\theta \geq \frac{2}{\pi^2}\theta^2 \qquad (|\theta| \leq \pi), \tag{29}$$

we obtain

$$\begin{aligned}
|e^z f_k(z)| &\leq \sum_{0\leq j\leq k} \binom{k}{j} e^{(1-p^j q^{k-j})r\cos\theta} e^{p^j q^{k-j} r} f(p^j q^{k-j} r) \\
&\leq \sum_{0\leq j\leq k} \binom{k}{j} e^{(1-p^j q^{k-j})r(1-2\theta^2/\pi^2)} e^{p^j q^{k-j} r} f(p^j q^{k-j} r) \\
&\leq e^{-2r\theta^2(1-p^k)/\pi^2} e^r f_k(r).
\end{aligned}$$

This proves (28) with, say $c = 2(1-p)/\pi^2$. □

Proof of (22) in Theorem 1. We next evaluate $\tilde{M}_k(z)$ more precisely in the following lemma whose proof is presented in Appendix A.

Let

$$S_{k,m}(z) := \binom{k-1}{m} p^m q^{k-m} z e^{-p^m q^{k-m} z}.$$

Lemma 3. (i) ($m = 0$) *If $1 \leq k \leq k_0 - \alpha_1 K_n / LL_n$, then*

$$\tilde{M}_k(z) = q^k z e^{-q^k z} \left(1 + O(e^{-K_n})\right), \tag{30}$$

uniformly for $|z| = n$ and $\arg(z) = o(LL_n^{-1/2})$.
(ii) ($m \geq 1$) *If*

$$k = \alpha_1 \left(L_n - LLL_n + \log(p/q - 1) + m\log(p/q) - \eta\right), \tag{31}$$

where $m \geq 1$ and

$$\frac{K_n}{LL_n} \leq \eta \leq \log(p/q) - \frac{K_n}{LL_n},$$

then

$$\tilde{M}_k(z) = S_{k,m}(z)\left(1 + O(me^{-K_n})\right), \tag{32}$$

uniformly for $|z| = n$ and $\arg(z) = o(LL_n^{-1/2})$.

Using the above lemma, we now prove Theorem 1. It remains to evaluate the integral in (25). We first consider the case $m = 0$. By substituting (30) into the integral in (25), and by completing the arc $|\arg(z)| \le \theta_0$ to a full circle, we see that

$$\begin{aligned}\frac{n!}{2\pi i}\int_{\substack{|z|=n\\ |\arg(z)|\le\theta_0}} z^{-n-1}e^z\tilde{M}_k(z)\mathrm{d}z &= \frac{q^k n!}{2\pi i}\int_{\substack{|z|=n\\ |\arg(z)|\le\theta_0}} z^{-n}e^{(1-q^k)z}\mathrm{d}z + O(E_1)\\ &= q^k n![z^{n-1}]e^{(1-q^k)z} + O(E_2) + O(E_1),\end{aligned}$$

where

$$\begin{aligned}E_1 &:= e^{-K_n}n!n^{-n}q^{kn}\int_{-\theta_0}^{\theta_0} e^{(1-q^k)n\cos\theta}\mathrm{d}\theta,\\ E_2 &:= q^k n!n^{1-n}\int_{\theta_0}^{\pi} e^{(1-q^k)n\cos\theta}\mathrm{d}\theta.\end{aligned}$$

By the inequality (29), we have

$$\begin{aligned}E_1 &= O\left(e^{-K_n}n^{1/2}q^k n e^{-q^k n}\int_{-\infty}^{\infty} e^{-2n(1-q^k)\theta^2/\pi^2}\mathrm{d}\theta\right)\\ &= O\left(e^{-K_n}q^k n e^{-q^k n}\right).\end{aligned}$$

Similarly,

$$E_2 = O\left(q^k n e^{-q^k n}n^{-1/10}e^{-2n^{1/5}/\pi^2}\right).$$

This completes the proof of (22) when $m = 0$. For $m \ge 1$, we proceed in a similar manner but using part(ii) of Lemma 3. This proves Theorem 1.

Proof of (23) in Theorem 1. We now consider the remaining gaps when k is of the form (31) with $\eta = x/LL_n$, where $x = o(\sqrt{LL_n})$. In this case, the same analysis as above shows that both terms $S_{k,m}(z)$ and $S_{k,m+1}(z)$ are asymptotically close, so that

$$\tilde{M}_k(z) = \left(S_{k,m}(z) + S_{k,m+1}(z)\right)\left(1 + O(E_3)\right), \tag{33}$$

where the error E_3 introduced is bounded above by

$$\begin{aligned}E_3 &= O\left(\sum_{0\le j<m}\left|\frac{S_{k,j}(z)}{S_{k,m}(z)}\right| + \sum_{m+2\le j\le k}\left|\frac{S_{k,j}(z)}{S_{k,m}(z)}\right|\right)\\ &= O\left((m+1)L_n^{-(1-qe^{\eta}\cos\theta/p)}\right) + O\left(m!\sum_{j\ge 2}\frac{(p\alpha_1/q)^j}{(j+m)!}L_n^{j-\frac{(p/q)^j-1}{p/q-1}e^{\eta}\cos\theta}\right)\\ &= O\left((m+1)L_n^{-(1-q/p)} + (m+1)^{-1}L_n^{-(p/q-1)}\right)\\ &= O\left((m+1)L_n^{-(1-q/p)}\right),\end{aligned}$$

since $1 - q/p \le p/q - 1$, where we used the inequality

$$\frac{t^j - 1}{t-1} \ge \frac{t+1}{2}\,j \qquad (t > 1;\, j \ge 2),$$

and $\theta = o(LL_n^{-1/2})$. Thus the same analysis as above gives

$$\mu_{n,k} = \frac{k^m}{m!}p^m q^{k-m}ne^{-p^m q^{k-m}n}\left(1 + \frac{pL_n^{1-e^{\eta}}}{q(m+1)\log(1/q)}\right)\left(1 + O\left((m+1)L_n^{-(1-q/p)}\right)\right),$$

which implies (23). □

3.4 Range (II): A saddle-point analysis

We now assume that

$$\alpha_1 \left(L_n - LLL_n + K_n\right) \le k \le \alpha_2 (L_n - K_n \sqrt{L_n}), \tag{34}$$

and proceed by the saddle-point method (see [82, 86]) to derive the following main result of this subsection.

Theorem 2 (Asymptotics of $\mu_{n,k}$ in Range (II))**.** *If k satisfies (34), then*

$$\mu_{n,k} = G_1\left(\rho; \log_{p/q} p^k n\right) \frac{n^{-\rho}\,(p^{-\rho}+q^{-\rho})^k}{\sqrt{2\pi\beta_2(\rho)k}} \left(1 + O\left(\frac{1}{k(p/q)^\rho} + \frac{1}{k(\rho+2)^2}\right)\right), \tag{35}$$

where $\rho = \rho(n,k) > -2$ is chosen to satisfy the saddle-point equation

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}\rho}\left(\rho^\rho e^{-\rho} n^{-\rho}(p^{-\rho}+q^{-\rho})^k\right) = 0, & \text{if } \rho \ge 1; \\ \dfrac{\mathrm{d}}{\mathrm{d}\rho}\left(n^{-\rho}(p^{-\rho}+q^{-\rho})^k\right) = 0, & \text{if } \rho \le 1, \end{cases} \tag{36}$$

and

$$\begin{aligned} \beta_2(\rho) &:= \frac{p^{-\rho}q^{-\rho}\log(p/q)^2}{(p^{-\rho}+q^{-\rho})^2}, \\ G_1(\rho;x) &= \sum_{j\in\mathbb{Z}} g(\rho + i t_j)\Gamma(\rho+1+it_j)e^{-2j\pi i x} \qquad (t_j := 2j\pi/\log(p/q)) \end{aligned} \tag{37}$$

where $g(s) = 1 - 1/(p^{-s}+q^{-s})$, and $G_1(\rho,x)$ is a 1-periodic function (see Figures 3 and 4).

We devote the rest of this subsection to the proof of Theorem 2.

3.4.1 Two-step saddle-point method

We outline here the main steps of the proof of Theorem 2. The approach may be called a two-step saddle-point method since the saddle-point method is applied twice. First, we start from the the Mellin integral (6) and apply the saddle-point method to obtain precise asymptotics of $\tilde{M}_k(re^{i\theta})$ for small θ (i.e., around the *real axis*) and large r. The proof here is complicated by the fact that

$$\left|p^{-\rho-it} + q^{-\rho-it}\right| = p^{-\rho} + q^{-\rho}, \tag{38}$$

when $t = t_j$, $j \in \mathbb{Z}$, which implies that the number of saddle-points with the same real part is *infinite*, yielding the 1-periodic function $G_1(\rho;x)$.

This first application of the saddle-point method yields a good approximation to $\tilde{M}_k(z)$ for z large and near the real axis; then we *de-Poissonize* $\tilde{M}_k(z)$ by another application of the saddle-point method and establish that $\mu_{n,k} \sim \tilde{M}_k(n)$. Ultimately, we will use the de-Poissonization result of Proposition 1, however, in the first approximation we do de-Poissonization by "bare hands" by applying the argument already used in the proof of Proposition 1, namely (17) and (18). Thus we focus on the evaluation of the Cauchy integral (13) but with $|\theta| \le n^{-2/5}$ (the first integral of (25)).

3.4.2 Location of saddle-points

The integrand $z^{-s}\Gamma(s+1)g(s)\,(p^{-s}+q^{-s})^k$ of the integral in (6) has simple poles at $s=-j$, $j=2,3,\ldots$, the rightmost (dominant) one being at $s=-2$; it also has saddle-points, which are the zeros of the equation

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(\Gamma(s+1)n^{-s}(p^{-s}+q^{-s})^k\right)=0; \tag{39}$$

note that $g(s)$ is uniformly bounded for all s. In view of (38), there are infinitely many saddle-points of the form $\rho+it_j/\log(p/q)$ ($j=0,\pm1,\ldots$), where the real part ρ satisfies (39). Also it is easy to see that

$$\begin{cases} \rho\to+\infty, & \text{if } \frac{k}{L_n}\downarrow\frac{1}{\log(1/q)},\\ \rho\to-\infty, & \text{if } \frac{k}{L_n}\uparrow\frac{1}{\log(1/p)}. \end{cases}$$

We distinguish between two cases $\rho\geq 1$ and $-2<\rho<1$. In the former case, the saddle-points are determined by the whole equation (39) (using Stirling's formula for Gamma function), while in the latter case $\Gamma(\rho+1)$ is uniformly bounded.

Consider first the case when $\rho\geq 1$ (the choice of 1 being arbitrary). In this case, by (36) and by applying Stirling's formula, we obtain

$$\frac{k}{L_n-\log\rho}=\frac{p^{-\rho}+q^{-\rho}}{p^{-\rho}\log(1/p)+q^{-\rho}\log(1/q)},$$

which can be written in the form

$$\rho=\frac{1}{\log(p/q)}\log\left(\frac{L_n-\log\rho-k\log(1/p)}{k\log(1/q)-L_n+\log\rho}\right),$$

whenever $L_n-\log\rho<k\log(1/q)$, which will be seen to be the case when k satisfies (34).

On the other hand, when $\rho\leq 1$, the term $\Gamma(s+1)$ does not contribute significantly to the saddle-point location, and therefore we consider the second equation in (36) or

$$\frac{k}{L_n}=\frac{p^{-\rho}+q^{-\rho}}{p^{-\rho}\log(1/p)+q^{-\rho}\log(1/q)},$$

which is solved to be

$$\rho=\frac{1}{\log(p/q)}\log\left(\frac{L_n-k\log(1/p)}{k\log(1/q)-L_n}\right). \tag{40}$$

It follows that if k satisfies (34), then

$$\rho\leq\frac{1}{\log(p/q)}\left(LL_n-\log K_n+\log\frac{\log(p/q)}{\log(1/q)}+o(1)\right), \tag{41}$$

implying, in particular, that $\rho=O(LL_n)$. Also if $k=\alpha_1(L_n-LLL_n+\log\log(p/q)+K_n)$, then

$$\rho=\frac{1}{\log(p/q)}\left(LL_n-\log K_n+\log\frac{\log(p/q)}{\log(1/q)}+O(K_n^{-1})\right).$$

However, if k is close to the right boundary of (34), more precisely, $k=\alpha_2(1-\varepsilon_n)L_n$, where $\varepsilon_n=o(1)$, then

$$\rho=-2+\frac{\varepsilon_n}{\alpha_2\beta_2(-2)}+O(\varepsilon_n^2).$$

Thus $\rho=O(1)$.

From (41), we see that if $\rho \geq 1$ and k satisfies (34), then $k\beta_2(\rho) = \Theta(k(p/q)^\rho)$ and

$$k(p/q)^\rho \geq \frac{K_n}{\log(p/q)} + o(1);$$

on the other hand, if $\rho \geq -2 + K_n L_n^{-1/2}$, then $k(\rho+2)^2 \geq K_n^2$. Thus the O-term in (35) is small if we choose K_n sufficiently large. Before we present a formal proof of Theorem 2, we first discuss the behavior of $\mu_{n,k}$ when k is close to the boundaries of k.

3.4.3 Boundary behaviors of $\mu_{n,k}$

We can derive more transparent asymptotic approximations to $\mu_{n,k}$ when k is sufficiently away or close to the boundaries.

The central range: $\alpha \in [\alpha_1 + \varepsilon, \alpha_2 - \varepsilon]$. In this case, G_1 is bounded and $G_1(\rho; x) \sim G_1(\rho'; x)$, where

$$\rho' := \frac{1}{\log(p/q)} \log\left(\frac{1 - \alpha\log(1/p)}{\alpha\log(1/q) - 1}\right); \tag{42}$$

also $\beta_2(\rho) \sim \beta_2(\rho')$. Note that $g(\rho + it_j) = 1 - p^{it_j}/(p^{-\rho} + q^{-\rho})$ and

$$G_1\left(\rho; \log_{p/q} p^k n\right) = G_1\left(\rho; \log_{p/q} q^k n\right).$$

More precisely, if $k = \alpha(L_n + x\sqrt{\alpha\beta_2(\rho')L_n})$, where $\alpha \in [\alpha_1 + \varepsilon, \alpha_2 - \varepsilon]$ and $x = o(L_n^{1/6})$, then

$$\mu_{n,k} = G_1\left(\rho'; \log_{p/q} p^k n\right) \frac{n^{-\rho'}\left(p^{-\rho'} + q^{-\rho'}\right)^k}{\sqrt{2\pi\alpha\beta_2(\rho')L_n}}\, e^{-x^2/2}\left(1 + O\left(\frac{1+|x|^3}{\sqrt{L_n}}\right)\right),$$

uniformly in x. In particular, when $\alpha = 1/\mathrm{h}$, where $\mathrm{h} := p\log(1/p) + q\log(1/q)$ is the entropy of the Bernoulli variate, then $\rho' = -1$, and it follows that

$$\mu_{n,k} = \frac{\sqrt{\mathrm{h}}\, G_1\left(-1; \log_{p/q} p^k n\right)}{\log(p/q)\sqrt{2\pi pq}} \cdot \frac{n}{\sqrt{L_n}}\, e^{-x^2/2}\left(1 + O\left(\frac{1+|x|^3}{\sqrt{L_n}}\right)\right), \tag{43}$$

uniformly for $x = o(L_n^{1/6})$. Other approximations can be derived for $L_n^{1/6} \ll x = o(\sqrt{L_n})$. Thus $\mu_{n,k}$ reaches the maximum for k near $L_n/\mathrm{h} + O(1)$; also $\mu_{n,k}$ increases with k when $\alpha < 1/\mathrm{h}$ and decreases with k when $\alpha > 1/\mathrm{h}$; see Figure 2. See also Figure 3 for a plot of $G_1(-1; x)$ for a few p's.

The left boundary: $\rho \to -2^+$ and $\rho + 2 \gg L_n^{-1/2}$. In this case, the dominant periodicity vanishes because

$$G_1(\rho; x) \sim \frac{|g(-2)|}{\rho + 2} = \frac{2pq}{(p^2 + q^2)(\rho + 2)};$$

thus

$$\mu_{n,k} \sim \frac{2}{\sqrt{2\pi}\,\log(p/q)(\rho+2)}\, k^{-1/2} n^{-\rho}\left(p^{-\rho} + q^{-\rho}\right)^k. \tag{44}$$

The right boundary: $k/L_n \to 1/\log(1/q)^+$**.** In this case, $\rho \to \infty$ and $\rho = O(LL_n)$. The periodicity in the leading term of (35) does not vanish because we have

$$G_1(\rho; x) \sim \sum_{j\in\mathbb{Z}} \Gamma(\rho + 1 + i t_j) e^{-2j\pi i x},$$

and G_1 is not bounded. Indeed, the periodicity becomes more pronounced for increasing ρ since

$$\left|\frac{\Gamma(\rho + 1 + i t)}{\Gamma(\rho + 1)}\right| = O\left(e^{-t^2/(2\rho)+O(t^4/\rho^3)}\right),$$

for large ρ and $t = o(\rho)$; see Figure 4. This estimate also implies that

$$G_1(\rho; x) = O\left(\sum_{j\in\mathbb{Z}} |\Gamma(\rho + 1 + i t_j)|\right) = O\left(e^{-\rho}\rho^{\rho+1}\right) = O\left(\rho^{1/2}\Gamma(\rho + 1)\right).$$

The order is tight. This means that even if we normalize $G_1(\rho; x)$ by $\Gamma(\rho + 1)$, it still goes to infinity with ρ.

3.4.4 Proof of Theorem 2

In view of (25) (more generally, de-Poissonization Proposition 1), we only need to evaluate $\tilde{M}_k(n)$ and obtain precise local expansions for $\tilde{M}_k(ne^{i\theta})$ when $|\theta| \le \theta_0$ in order to estimate the first integral of (25). We first focus on estimating $\tilde{M}_k(n)$, and then extend the same approach to derive the asymptotics of $\tilde{M}_k(ne^{i\theta})$. This suffices to prove that $\mu_{n,k} \sim \tilde{M}_k(n)$. Later in Subsection 3.8 we refine this analysis to obtain a better error term.

In order to evaluate $\tilde{M}_k(n)$ by the inverse Mellin transform, we move first the line of integration of (6) to $\Re(s) = \rho$, so that

$$\tilde{M}_k(n) = \frac{1}{2\pi}\int_{-\infty}^{\infty} J_k(n; \rho + i t)\mathrm{d}t, \tag{45}$$

where $\rho > -2$ is the saddle-point chosen according to (36) and $J_k(n; s) := n^{-s}\Gamma(s + 1)g(s)(p^{-s} + q^{-s})^k$. We now show that the above integral is small for $|t| \ge \sqrt{L_n}$, and then assess the main contribution of saddle-points falling into the range $|t| \le \sqrt{L_n}$.

Smallness of the integral when $|t| \ge \sqrt{L_n}$**.** Assume from now on that ρ is chosen as described above in (36). We show that the integral in (45) with $|t| \ge \sqrt{L_n}$ is asymptotically negligible.

Since our $\rho > -2$ satisfies (40), we have, by (9),

$$\begin{aligned}\frac{1}{2\pi}\int_{|t|\ge\sqrt{L_n}} J_k(n; \rho + i t)\mathrm{d}t &= O\left(n^{-\rho}(p^{-\rho} + q^{-\rho})^k \int_{\sqrt{L_n}}^{\infty} |\Gamma(\rho + 1 + i t)|\mathrm{d}t\right)\\ &= O\left(n^{-\rho}(p^{-\rho} + q^{-\rho})^k \int_{\sqrt{L_n}}^{\infty} t^{\rho+1/2}e^{-\pi t/2}\mathrm{d}t\right)\\ &= O\left(L_n^{\rho/2+1/4}e^{-\pi\sqrt{L_n}/2}n^{-\rho}(p^{-\rho} + q^{-\rho})^k\right).\end{aligned}$$

On the other hand, since $\rho = O(LL_n)$ and $\rho \ge -2 + K_n L_n^{-1/2}$, we then obtain

$$L_n^{\rho/2+1/4}e^{-\pi\sqrt{L_n}/2} = O\left(e^{-\pi\sqrt{L_n}/2+O(LL_n^2)}\right) = O\left(\Gamma(\rho + 2)e^{-\sqrt{L_n}}\right),$$

for large enough n; the last O-term holds uniformly for $\rho \ge -2 + K_n L_n^{-1/2}$ and ρ satisfying (41).

Contribution from each saddle-point. Let j_0 be the largest integer j for which $2j\pi/\log(p/q) \le \sqrt{L_n}$. Then we can split the integral over $\int_{|t|\le\sqrt{L_n}}$ as follows.

$$\int_{|t|\le\sqrt{L_n}} J_k(n;\rho+it)\mathrm{d}t = \sum_{|j|<j_0} \int_{|t-t_j|\le\pi/\log(p/q)} J_k(n;\rho+it)\mathrm{d}t + \int_{t_{j_0}\le|t|\le\sqrt{L_n}} J_k(n;\rho+it)\mathrm{d}t.$$

The last integral is bounded above by

$$O\left(\Gamma(\rho+2)n^{-\rho}(p^{-\rho}+q^{-\rho})^k e^{-\sqrt{L_n}}\right),$$

by the same argument used above. It remains to evaluate the integrals

$$T_j := \frac{1}{2\pi}\int_{|t-t_j|\le\pi/\log(p/q)} J_k(n;\rho+it)\mathrm{d}t,$$

for $|j| < j_0$.

We derive first a uniform bound for $|p^{-\rho-it}+q^{-\rho-it}|$. By the elementary inequalities (29) and

$$\sqrt{1-x} \le 1-\frac{x}{2} \qquad (x\in[0,1]),$$

we have

$$\begin{aligned}
\left|p^{-\rho-it}+q^{-\rho-it}\right| &= \left(p^{-\rho}+q^{-\rho}\right)\sqrt{1-\frac{2p^{-\rho}q^{-\rho}}{(p^{-\rho}+q^{-\rho})^2}\left(1-\cos\left(t\log(p/q)\right)\right)} \\
&\le \left(p^{-\rho}+q^{-\rho}\right)\left(1-\frac{p^{-\rho}q^{-\rho}}{(p^{-\rho}+q^{-\rho})^2}\left(1-\cos\left((t-t_j)\log(p/q)\right)\right)\right) \\
&\le \left(p^{-\rho}+q^{-\rho}\right)\left(1-\frac{2p^{-\rho}q^{-\rho}}{\pi^2(p^{-\rho}+q^{-\rho})^2}(t-t_j)^2\log(p/q)^2\right) \\
&\le \left(p^{-\rho}+q^{-\rho}\right)e^{-c_0(t-t_j)^2}, \qquad (46)
\end{aligned}$$

uniformly for $|t-t_j| \le \pi/\log(p/q)$, where

$$c_0 = c_0(\rho) := \frac{2p^{-\rho}q^{-\rho}\log(p/q)^2}{\pi^2(p^{-\rho}+q^{-\rho})^2} = \frac{2}{\pi^2}\beta_2(\rho).$$

We now take

$$v_0 := \begin{cases} k^{-2/5}, & \text{if } -2<\rho\le 1; \\ (c_0k)^{-2/5}, & \text{if } \rho\ge 1, \end{cases}$$

and split the integration range into two parts: $|t-t_j| \le v_0$ and $v_0 < |t-t_j| \le \pi/\log(p/q)$. (We assume that k is so large that $v_0 < \pi/\log(p/q)$.)

Consider first the case when $-2 < \rho \le 1$. From the inequality (46), it follows that

$$\begin{aligned}
T_j'' &:= \frac{1}{2\pi}\int_{v_0\le|t-t_j|\le\pi/\log(p/q)} J_k(n;\rho+it)\mathrm{d}t \qquad (47) \\
&= O\left(|\Gamma(\rho+2+it_j)|n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\int_{k^{-2/5}}^{\infty} e^{-c_0kv^2}\mathrm{d}v\right) \\
&= O\left(n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k k^{-3/5}e^{-c_0k^{1/5}}\times\begin{cases}|\Gamma(\rho+1+it_j)|, & \text{if } j\ne 0\\ 1, & \text{if } j=0\end{cases}\right),
\end{aligned}$$

for each $|j| \le j_0$.

When $\rho \geq 1$ and satisfies (34), we have

$$\begin{aligned} T_j'' &= O\left(|\Gamma(\rho+1+it_j)|n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k \int_{(c_0k)^{-2/5}}^{\infty} e^{-c_0kv^2}\mathrm{d}v\right) \\ &= O\left(|\Gamma(\rho+1+it_j)|n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k (c_0k)^{-3/5}e^{-(c_0k)^{1/5}}\right), \end{aligned}$$

for $|j| \leq j_0$.

The dominant terms. It remains to evaluate the integrals T_j for t in the range $|t-t_j| \leq v_0$. Note that by our choice of t_j,

$$p^{-\rho-it_j}+q^{-\rho-it_j} = p^{-it_j}\left(p^{-\rho}+q^{-\rho}\right) = q^{-it_j}\left(p^{-\rho}+q^{-\rho}\right),$$

so that

$$\begin{aligned} \frac{p^{-\rho-it}+q^{-\rho-it}}{p^{-\rho-it_j}+q^{-\rho-it_j}} &= 1+\sum_{\ell\geq 1}\frac{i^\ell(t-t_j)^\ell}{\ell!}\cdot\frac{p^{-\rho-it_j}\log(1/p)^\ell+q^{-\rho-it_j}\log(1/q)^\ell}{p^{-\rho-it_j}+q^{-\rho-it_j}} \\ &= 1+\sum_{\ell\geq 1}\frac{i^\ell(t-t_j)^\ell}{\ell!}\cdot\frac{p^{-\rho}\log(1/p)^\ell+q^{-\rho}\log(1/q)^\ell}{p^{-\rho}+q^{-\rho}}. \end{aligned}$$

It follows that

$$\log\left(p^{-\rho-it}+q^{-\rho-it}\right) = \log\left(p^{-\rho-it_j}+q^{-\rho-it_j}\right)+\sum_{\ell\geq 1}\frac{\beta_\ell(\rho)}{\ell!}\,i^\ell(t-t_j)^\ell,$$

where, in particular,

$$\beta_1(\rho) = \frac{p^{-\rho}\log(1/p)+q^{-\rho}\log(1/q)}{p^{-\rho}+q^{-\rho}}.$$

The remaining manipulation by using the saddle-point method is then straightforward. We use the local expansions

$$\left(\frac{p^{-\rho-it}+q^{-\rho-it}}{p^{-\rho-it_j}+q^{-\rho-it_j}}\right)^k = \exp\left(k\sum_{1\leq\ell\leq 3}\frac{\beta_\ell(\rho)}{\ell!}i^\ell(t-t_j)^\ell+O(k|\beta_4(\rho)||t-t_j|^4)\right),$$

and

$$\Gamma(\rho+1+it)g(\rho+it) = \begin{cases} C_0+C_1i(t-t_j)+O\left(\dfrac{(t-t_j)^2}{(\rho+2)^2}\right), & \text{if } -2<\rho\leq 1; \\ \Gamma(\rho+1+it_j)e^{(\log\rho)i(t-t_j)}\left(1+C_2i(t-t_j)+O(|C_2|^3|t-t_j|^2)\right) & \\ \times\left(g(\rho+it_j)+g'(\rho+it_j)i(t-t_j)+O\left(|t-t_j|^2\right)\right), & \text{if } \rho\geq 1, \end{cases}$$

where

$$\begin{cases} C_0 := \Gamma(\rho+1+it_j)g(\rho+it_j); \\ C_1 := g(\rho+it_j)\Gamma(\rho+1+it_j)\psi(\rho+1+it_j)+g'(\rho+it_j)\Gamma(\rho+1+it_j), \end{cases}$$

$\psi(s) = \Gamma'(s)/\Gamma(s)$ being the logarithmic derivative of the Gamma function, and

$$C_2 := \psi(\rho+1+it_j)-\log\rho \qquad (\rho\geq 1).$$

Here C_0 and C_1 are defined to be their limits when $\rho=-1$ and $j=0$, namely,

$$\begin{cases} C_0 := p\log(1/p)+q\log(1/q); \\ C_1 := -\frac{2p-1}{2}\left(p\log(p)^2-q\log(q)^2\right)-C_0\gamma-2pq\log(p)\log(q). \end{cases}$$

Note that $\psi(\rho+1+it_j)-\log\rho = O(\log(1+|t_j|))$. It follows that for $|j|<j_0$

$$T_j = \frac{g(\rho+it_j)}{\sqrt{2\pi\beta_2(\rho)k}}\,\Gamma(\rho+1+it_j)n^{-\rho-it_j}\left(p^{-\rho}+q^{-\rho}\right)^k p^{-ikt_j} \times\left(1+O\left(\frac{1}{k\beta_2(\rho)}+\frac{1}{k(\rho+2)^2}\right)\right).$$

Summing over all $|j|<j_0$ and collecting all estimates, we obtain

$$\tilde{M}_k(n) = \frac{n^{-\rho}\,(p^{-\rho}+q^{-\rho})^k}{\sqrt{2\pi\beta_2(\rho)k}}\sum_{|j|<j_0} g(\rho+it_j)\Gamma(\rho+1+it_j)(p^k n)^{-it_j} \times\left(1+O\left(\frac{1}{k(p/q)^\rho}+\frac{1}{k(\rho+2)^2}\right)\right).$$

An asymptotic approximation to $\tilde{M}_k(z)$. To complete the de-Poissonization, we need to a more precise expansion for $\tilde{M}_k(ne^{i\theta})$ for small θ. The above proof by the saddle-point method can be easily extended *mutatis mutandis* to $\tilde{M}_k(z)$ for complex values of z lying in the right half-plane since we can write (7) as

$$\tilde{M}_k(ne^{i\theta}) = \frac{1}{2\pi i}\int_{(\rho)} n^{-s}e^{-i\theta s}\Gamma(s+1)g(s)\left(p^{-s}+q^{-s}\right)^k \mathrm{d}s,$$

where $\rho>-2$ and $|\theta|\le\pi/2-\varepsilon$. The result is

$$\tilde{M}_k(ne^{i\theta}) = \frac{(p^{-\sigma}+q^{-\sigma})^k}{\sqrt{2\pi\beta_2(\rho)k}}\sum_{|j|<j_0} g(\sigma+it_j)\Gamma(\sigma+1+it_j)(ne^{i\theta})^{-\rho-it_j}\,p^{-ikt_j} \times\left(1+O\left(\frac{1}{k(p/q)^\rho}+\frac{1}{k(\rho+2)^2}\right)\right), \tag{48}$$

uniformly for $|\theta|\le\pi/2-\varepsilon$ and k lying in the range (34). Note that the index of the sum can be extended to infinity, but it is easier to manipulate a finite sum than an infinite series since we substitute the right-hand side into the Cauchy integral (13) and then integrate term by term. This completes the proof of (35).

3.5 Range (IV): A singularity analysis

We consider the range (IV) first, leaving the analysis in the transitional range when $k=\alpha_2 L_n+o(L_n^{2/3})$ to the next subsection.

We show that for $k\ge\alpha_2 L_n+K_n\sqrt{L_n}$, the asymptotics of the expected profile $\tilde{M}_k(n)$ is dictated by the simple pole at $s=-2$, or structurally, by the number of pairs of input-strings sharing the same prefixes of length at least k.

Theorem 3. *If*

$$k\ge\alpha_2\left(L_n+K_n\sqrt{\alpha_2\beta_2(-2)L_n}\right), \tag{49}$$

where β_2 is defined in (37), then

$$\mu_{n,k} = 2pqn^2(p^2+q^2)^{k-1}\left(1+O\left(K_n^{-1}e^{-K_n^2/2+O(K_n^3/\sqrt{L_n})}\right)\right), \tag{50}$$

uniformly for $1\ll K_n=o(\sqrt{L_n})$.

Proof. To prove (50), we move the line of integration (by absolute convergence of the integral) of the integral in (6) to $\Re(s) = \rho$, where

$$\rho := -2 - \frac{K_n}{\sqrt{\alpha_2\beta_2(-2)L_n}}.$$

Thus $\tilde{M}_k(ne^{i\theta})$ equals the residue of the integrand at $s = -2$ (the dominant term in (50)) plus the integral along $\Re(s) = \rho$.

$$\tilde{M}_k(ne^{i\theta}) = |g(-2)|n^2e^{2i\theta}(p^2+q^2)^k + \frac{1}{2\pi}\int_{-\infty}^{\infty} J_k(ne^{i\theta};\rho+it)\mathrm{d}t,$$

where $|g(-1)| = 2pq/(p^2+q^2)$. We need only to estimate the last integral. By the same analysis used for T_j'' (see (47)) and the inequality (46), we have

$$\begin{aligned}
&\frac{1}{2\pi}\left(\int_{|t|\le\pi/\log(p/q)} + \sum_{|j|\ge1}\int_{|t-t_j|\le\pi/\log(p/q)}\right) J_k(ne^{i\theta};\rho+it)\mathrm{d}t\\
&\quad= O\left(|\Gamma(\rho+1)|n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\int_{|t|\le\pi/\log(p/q)} e^{-c_0kt^2}\mathrm{d}t\right)\\
&\qquad+ O\left(n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\sum_{|j|\ge1}\left|\Gamma\left(\rho+1+\frac{2|j|-1}{\log(p/q)}\pi i\right)\right| e^{(2|j|+1)\pi|\theta|/\log(p/q)}\right.\\
&\qquad\qquad\left.\times\int_{|t-t_j|\le\pi/\log(p/q)} e^{-c_0k(t-t_j)^2}\mathrm{d}t\right)\\
&\quad= O\left(\frac{k^{-1/2}}{|\rho+2|}n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\right)\\
&\quad= O\left(K_n^{-1}n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\right),
\end{aligned}$$

where we used (9) to bound the sum

$$\begin{aligned}
&\sum_{|j|\ge1}\left|\Gamma\left(\rho+1+\frac{2|j|-1}{\log(p/q)}\pi i\right)\right| e^{(2|j|+1)\pi|\theta|/\log(p/q)}\\
&\qquad= O\left(\sum_{|j|\ge1}(2|j|-1)^{\rho+1/2}\exp\left(-\frac{\pi^2(2|j|-1)}{2\log(p/q)}+\frac{(2|j|+1)\pi|\theta|}{\log(p/q)}\right)\right)\\
&\qquad= O(1),
\end{aligned}$$

uniformly for $|\theta| \le \pi/2 - \varepsilon$.

By our choice of ρ and by straightforward expansion, we have

$$\begin{aligned}
\frac{K_n^{-1}n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k}{n^2\left(p^2+q^2\right)^k} &= O\left(K_n^{-1}e^{-L_n(\rho+2)+\frac{k}{\alpha_2}(\rho+2)+\frac{k}{2}\beta_2(-2)(\rho+2)^2+O\left(k|\rho+2|^3\right)}\right)\\
&= O\left(K_n^{-1}e^{-K_n^2/2+O(K_n^3/\sqrt{L_n})}\right).
\end{aligned}$$

Thus

$$\tilde{M}_k(ne^{i\theta}) = |g(-2)|(ne^{i\theta})^2(p^2+q^2)^k\left(1+O\left(K_n^{-1}e^{-K_n^2/2+O(K_n^3/\sqrt{L_n})}\right)\right), \tag{51}$$

uniformly for $|\theta| \le \pi/2 - \varepsilon$. Substituting this into (25), we deduce the desired result (50). □

Remarks. (i) When $K_n \ge \varepsilon\sqrt{L_n}$, we can either take $K_n = \varepsilon\sqrt{L_n}$ or refine the analysis to give a better error term.
(ii) The asymptotic approximation (50) can also be derived from the exact expression (10) by using only elementary arguments.
(iii) Also note that the range (49) implies that the saddle-point ρ satisfies $\rho \le -2 - K_n/\sqrt{L_n}$, but the contribution from this saddle-point is asymptotically negligible (compared to the polar singularity).

3.6 Range (III): A uniform analysis

We consider in this subsection the transitional range $k = \alpha_2 L_n + o(L_n^{2/3})$ and show that the transitional behavior of $\mu_{n,k}$ in this range is bridged by a Gaussian distribution function.

Theorem 4. *If*

$$k = \alpha_2\left(L_n + \xi\sqrt{\alpha_2\beta_2(-2)L_n}\right), \tag{52}$$

where $\xi = \xi_{n,k} = o(L_n^{1/6})$, *then*

$$\mu_{n,k} = |g(-2)|\Phi(\xi)n^2\left(p^2+q^2\right)^k\left(1 + O\left(\frac{1+|\xi|^3}{\sqrt{L_n}}\right)\right), \tag{53}$$

uniformly in ξ, *where* $\Phi(\xi) = (2\pi)^{-1/2}\int_{-\infty}^{\xi} e^{-t^2/2}\mathrm{d}t$.

Proof. We assume first that k satisfies (52) and $k < \alpha_2 L$ (or $\xi < 0$). We move the line of integration of the integral in (11) to $\Re(s) = \rho$, where ρ is taken to be of the same form as in (40); asymptotically

$$\rho = -2 - \frac{\xi}{\sqrt{\alpha_2\beta_2(-2)L_n}} + O\left(\xi^2 L_n^{-1}\right). \tag{54}$$

By a similar analysis as the proof of Theorem 3, we obtain

$$\begin{aligned}\tilde{M}_k(ne^{i\theta}) = \frac{1}{2\pi}\int_{|t|\le L_n^{-2/5}} J_k(ne^{i\theta};\rho+it)\mathrm{d}t + O\left(|\Gamma(\rho+1+iL_n^{-2/5})|n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k e^{-c_0L_n^{1/5}}\right)\\ + O\left(k^{-1/2}n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\right),\end{aligned}$$

where $|\theta| < \pi/2$. By (54), we have

$$|\Gamma(\rho+1+iL_n^{-2/5})| = O\left(\frac{1}{|\xi|L_n^{-1/2}+L_n^{-2/5}}\right) = O(L_n^{2/5}).$$

It follows that

$$\tilde{M}_k(ne^{i\theta}) = \frac{1}{2\pi}\int_{|t|\le L_n^{-2/5}} J_k(ne^{i\theta};\rho+it)\mathrm{d}t + O\left(k^{-1/2}n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\right).$$

Note that since $s \mapsto \Gamma(s+1) + 1/(s+2)$ is analytic for $|s+2| \le 1-\varepsilon$, we have

$$\tilde{M}_k(ne^{i\theta}) = \frac{|g(-2)|}{2\pi}\int_{|t|\le L_n^{-2/5}} \frac{n^{-\rho-it}e^{-i\theta(\rho+it)}}{\rho+2+it}\left(p^{-\rho-it}+q^{-\rho-it}\right)^k \mathrm{d}t + O\left(k^{-1/2}n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\right).$$

The integral on the right-hand side is evaluated as follows:

$$\begin{aligned}
&\frac{|g(-2)|}{2\pi}\int_{|t|\le L_n^{-2/5}} \frac{n^{-\rho-it}e^{-i\theta(\rho+it)}}{\rho+2+it}\left(p^{-\rho-it}+q^{-\rho-it}\right)^k \mathrm{d}t\\
&\quad=\frac{|g(-2)|}{2\pi}n^{-\rho}e^{-i\theta\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\int_{|t|\le L_n^{-2/5}}\frac{e^{\theta t-\beta_2(\rho)kt^2/2+O(k|t|^3)}}{\rho+2+it}\mathrm{d}t\\
&\quad=\frac{|g(-2)|}{2\pi}n^{-\rho}e^{-i\theta\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\int_{-\infty}^{\infty}\frac{e^{-t^2/2}}{\xi_0+it}\left(1+O\left(\frac{|t|+|t|^3}{\sqrt{L_n}}\right)\right)\mathrm{d}t, \qquad (55)
\end{aligned}$$

where

$$\xi_0 := (\rho+2)\sqrt{\beta_2(\rho)k} > 0.$$

Note that $\xi_0 = -\xi + O\left(\xi^2 L_n^{-1/2}\right)$ by (52) and (54). Since $\xi_0 > 0$, we have

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-t^2/2}}{\xi_0+it}\mathrm{d}t &= \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-t^2/2}\int_0^{\infty}e^{-v(\xi_0+it)}\mathrm{d}v\mathrm{d}t\\
&= \frac{1}{2\pi}\int_0^{\infty}e^{-v\xi_0}\int_{-\infty}^{\infty}e^{-t^2/2-itv}\mathrm{d}t\mathrm{d}v\\
&= \frac{1}{\sqrt{2\pi}}\int_0^{\infty}e^{-v^2/2-v\xi_0}\mathrm{d}v\\
&= e^{\xi_0^2/2}\Phi(-\xi_0).
\end{aligned}$$

The error term in (55) is estimated similarly and satisfies

$$L_n^{-1/2}\int_{-\infty}^{\infty}\frac{(|t|+|t|^3)e^{-t^2/2}}{|(\rho+2)\sqrt{\beta_2(\rho)k}+it|}\mathrm{d}t = O\left(L_n^{-1/2}\int_0^{\infty}(v+v^3)e^{-v^2/2-v\xi_0}\mathrm{d}v\right).$$

Observe that

$$e^{x^2/2}\Phi(-x) = \begin{cases} O\left(x^{-1}\right), & \text{if } x\to\infty;\\ O\left(e^{x^2/2}\right), & \text{if } x\to-\infty. \end{cases} \qquad (56)$$

Also

$$\int_0^{\infty}(v+v^3)e^{-v^2/2-vx}\mathrm{d}v = \begin{cases} O\left(x^{-2}\right), & \text{if } x\to\infty;\\ O\left(|x|^3e^{x^2/2}\right), & \text{if } x\to-\infty, \end{cases}$$

so that

$$\int_0^{\infty}v^3e^{-v^2/2-vx}\mathrm{d}v = O\left(e^{x^2/2}\Phi(-x)(1+|x|^3)\right).$$

Thus

$$\begin{aligned}
\tilde{M}_k(ne^{i\theta}) = |g(-2)|(ne^{i\theta})^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k e^{\xi_0^2/2}\Phi(-\xi_0)\left(1+O\left(\frac{1+|\xi|^3}{\sqrt{L_n}}\right)\right)\\
+O\left(k^{-1/2}n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k\right), \qquad (57)
\end{aligned}$$

uniformly for $|\theta|\le\pi/2-\varepsilon$. Substituting this in (25) and using the expansions

$$\begin{aligned}
n^{-\rho}\left(p^{-\rho}+q^{-\rho}\right)^k &= n^2\left(p^2+q^2\right)^k e^{-\xi^2/2+O(|\xi|^3L_n^{-1/2})},\\
e^{\xi_0^2/2}\Phi(\xi_0) &= e^{\xi^2/2}\Phi(\xi)\left(1+O\left(|\xi|^3L_n^{-1/2}\right)\right),
\end{aligned}$$

we deduce (53) when $\xi < 0$.

The restriction that $\xi < 0$ can now be removed by continuity (when $\xi_0 = 0$ the integral path has to be properly indented) or by a similar analysis. This proves (53).

One can easily check, by (56), that the asymptotic estimate (53) coincides with the two estimates (44) and (50) when $\xi \to -\infty$ and $\xi \to \infty$, respectively. □

Remark. The appearance of the normal distribution function is typical when a saddle-point coalesces with a simple pole; see [86]. Also, the polynomial order (4) of $\mu_{n,k}$ now follows from (35), (50) and (53).

3.7 The range where the expected profile grows unbounded

An important consequence of the preceding results is the following characterization of the range where $\mu_{n,k} \to \infty$, which will be seen to be the range where $B_{n,k}$ is asymptotically normally distributed.

Theorem 5. *Define*

$$m_0 := \left\lceil \frac{1}{p/q - 1} \right\rceil \quad \textit{and} \quad \alpha_3 := \frac{2}{\log \frac{1}{p^2+q^2}}.$$

Then $\mu_{n,k} \to \infty$ iff

$$\alpha_1 \left(L_n - LLL_n - \log m_0 + m_0 \log(p/q) - \frac{LLL_n - K_n}{m_0 LL_n} \right) \le k \le \alpha_3 \left(L_n - K_n \right).$$

Proof. Consider first the upper bound. If $k \le \alpha_3 L_n - x$, then

$$n^2 (p^2 + q^2)^k \ge (p^2 + q^2)^{-x},$$

which tends to infinity if $x \to \infty$; on the other hand, if $k \ge \alpha_3 L_n - x$, then the reverse inequality holds and the right-hand side remains bounded if x is less than a positive constant.

For the lower bound, we use the estimate (24). First, if $k \le k_0$ (see (20)), then

$$\mu_{n,k} = \Theta(q^k n e^{-q^k n}) = o(1).$$

Next, if $k_{m-1} \le k \le k_m$, $m \ge 1$, then by (24)

$$\mu_{n,k} = \Theta(S_{n,k,m}) = \Theta(L_n^{m - e^\eta/(p/q-1)} LL_n),$$

where k is written in the form (31). Since $\eta \in [0, \log(p/q)]$, we have

$$m - \frac{p}{p-q} \le m - \frac{e^\eta}{p/q - 1} \le m - \frac{q}{p-q}.$$

Also, by the definition of m_0, we have the inequalities

$$m_0 - 1 < \frac{q}{p-q} = \frac{1}{p/q - 1} \le m_0.$$

Thus if $m \le m_0 - 1$, then

$$m - \frac{e^\eta}{p/q - 1} \le m - \frac{q}{p-q} < m - m_0 + 1 \le 0,$$

implying that $\mu_{n,k} \to 0$ for $k \le k_{m_0 - 1}$. Similarly, since

$$m_0 \le \frac{p}{p-q} < m_0 + 1,$$

we have

$$m - \frac{e^\eta}{p/q-1} \geq m - \frac{p}{p-q} > m - m_0 - 1 \geq 0,$$

when $m \geq m_0 + 1$. Therefore, $\mu_{n,k} \to \infty$ if $k \geq k_{m_0}$ (and remains in the range $k \leq \alpha_1(L_n - LLL_n + K_n)$).

It remains the range $k_{m_0-1} \leq k \leq k_{m_0}$ in which $\mu_{n,k} = \Theta(L_n^{m_0 - e^\eta/(p/q-1)} LL_n)$, where

$$k = \alpha_1(L_n - LLL_n + \log(p/q-1) + m_0 \log(p/q) - \eta).$$

We distinguish three cases: (*i*) if

$$\eta \geq \log m_0 + \log(p/q-1) + \frac{LLL_n + K_n}{m_0 LL_n},$$

then $\mu_{n,k} = \Theta(L_n^{m - e^\eta/(p/q-1)} LL_n)$ and

$$L_n^{m - e^\eta/(p/q-1)} LL_n \leq e^{-K_n} \to 0;$$

(*ii*) if

$$\eta = \log m_0 + \log(p/q-1) + \frac{LLL_n + x}{m_0 LL_n},$$

then

$$\mu_{n,k} \sim S_{n,k,m_0} \sim \frac{\alpha_1^{m_0}}{(m_0-1)!}\, e^{-x},$$

uniformly for $x = O(1)$; and (*iii*) if

$$\eta \leq \log m_0 + \log(p/q-1) + \frac{LLL_n - K_n}{m_0 LL_n},$$

then $\mu_{n,k} = \Theta(L_n^{m - e^\eta/(p/q-1)} LL_n)$ and

$$L_n^{m - e^\eta/(p/q-1)} LL_n \geq e^{K_n} \to \infty.$$

Thus $\mu_{n,k}$ is bounded away from zero and infinity in the second case.

This proves the theorem when k lies in Ranges (I) and (IV). The remaining cases follow easily from (35) and (53). □

Let $\{x\}$ denote the fractional part of x. The lower bound can be further refined as follows.

Corollary 3. *Let*

$$\hat{k} := \alpha_1 \left(L_n - LLL_n - \log m_0 + m_0 \log(p/q) - \frac{LLL_n}{m_0 LL_n} \right), \tag{58}$$

where $m_0 = \lceil 1/(p/q-1) \rceil$. *Then (i)* $\mu_{n,k} \to \infty$ *for* $\lceil \hat{k} \rceil \leq k \leq \alpha_3(L_n - K_n)$; *(ii)* $\mu_{n,k} \to 0$ *for* $k \leq \lceil \hat{k} \rceil - 2$, *and (iii)*

$$\mu_{n,\lceil \hat{k} \rceil - 1} \begin{cases} = \Theta(1), & \text{if } \{\hat{k}\} = O(LL_n^{-1}); \\ \to 0, & \text{otherwise.} \end{cases}$$

Proof. The proof is similar to that of Theorem 5. We consider only the last case. Write first

$$\lceil \hat{k} \rceil - 1 = \hat{k} - \{\hat{k}\} = \alpha_1 \left(L_n - LLL_n + \log(p/q-1) + m_0 \log(p/q) - \eta' \right),$$

where

$$\eta' = \log m_0 + \log(p/q - 1) + \{\hat{k}\}/\alpha_1 + \frac{LLL_n}{m_0 LL_n}.$$

(We assume that $\hat{k}$ is not an integer.) Then we follow the same proof as above by distinguishing into three cases. In particular, the case when $\hat{k}$ is an integer is also covered by the bounded case. □

The result is to be compared with Pittel's result in [67], which says that the probability that the shortest path equals either $\langle \kappa_n \rangle$ or $\langle \kappa_n \rangle + 1$ tends to 1, where $\langle x \rangle$ denotes the nearest integer to x and

$$\kappa_n = \alpha_1 \left(L_n - LLL_n - \log \max_{j \geq 1} j(q/p)^j \right).$$

Note that

$$-\log \max_{j \geq 1} j(q/p)^j = -\log m_0 + m_0 \log(p/q).$$

Our result implies a slightly more precise location; see Section 8.

3.8 Refinement of $\mu_{n,k}$ by de-Poissonization

All expansions for $\mu_{n,k}$ we derived so far are in terms of slowly decreasing powers of L_n^{-1} or LL_n^{-1}, which will turn out to be insufficient for the asymptotics of the variance because of cancelation of dominant terms. Thus in this section we derive a more effective expansion for $\mu_{n,k}$ in terms of $\tilde{M}_k(n)$ and its higher derivatives; namely, we derive an expression of the form (19). The major difference here is that we do not substitute the asymptotic expansions for $\tilde{M}_k(n)$ into the Cauchy integral representation for $\mu_{n,k}$, resulting in less explicit asymptotic approximation to $\mu_{n,k}$ but with a much better error term.

We start with a lemma.

Lemma 4. *Define*

$$\rho_0 := \begin{cases} q^k n, & \text{if } 1 \leq k \leq k_0; \\ \rho, & \text{if } \rho \geq 1 \text{ and } k \geq k_0 \\ 1, & \text{if } \rho \leq 1, \end{cases} \tag{59}$$

where ρ is given by (36). Then

$$\tilde{M}_k^{(\ell)}(ne^{i\theta}) = O\left(\rho_0^\ell n^{-\ell} \tilde{M}_k(n)\right), \tag{60}$$

uniformly for $\theta = o(LL_n^{-1/2})$.

Proof. If $\ell \geq 1$, then, by (8),

$$\begin{aligned} \tilde{M}_k^{(\ell)}(z) &= \sum_{0 \leq j < k} \binom{k-1}{j} (p^j q^{k-1-j})^\ell \tilde{M}_1^{(\ell)}(z) \\ &= \sum_{0 \leq j < k} \binom{k-1}{j} (p^j q^{k-j})^{\ell+1} z e^{-p^j q^{k-j} z} \left(1 + O\left(|z|^{-1}\right)\right), \end{aligned}$$

as $|z| \to \infty$ and $\Re(z) > 0$. If $1 \leq k \leq k_0$ (see (20)), then a proof similar to (and simpler than) that of (30) shows that

$$\tilde{M}_k^{(\ell)}(ne^{i\theta}) = O\left(q^{k(\ell+1)} n e^{-q^k n \cos\theta}\right) = O\left(\rho_0^\ell n^{-\ell} \tilde{M}_k(n)\right),$$

uniformly for $\theta = o(LL_n^{-1/2})$. If $k_{m-1} \le k \le k_m$, where $m \ge 1$, then the proof of (30) is also easily amended and we obtain

$$\tilde{M}_k^{(\ell)}(ne^{i\theta}) = O\left(k^m (p^m q^{k-m})^{\ell+1} n e^{-p^m q^{k-m} n \cos\theta}\right) = O\left(LL_n^{\ell} n^{-\ell} \tilde{M}_k(n)\right),$$

uniformly for $\theta = o(LL_n^{-1/2})$. Note that $\rho = O(LL_n)$ when $k_{m-1} \le k \le k_m$, $m \ge 1$. For the proof of (60) in the remaining ranges of k, we use the integral representation

$$\tilde{M}_k^{(\ell)}(z) = \frac{(-1)^\ell}{2\pi i} \int_{(\rho)} s(s+1)\cdots(s+\ell-1) z^{-s-\ell} \Gamma(s+1) g(s) \left(p^{-s} + q^{-s}\right)^k \mathrm{d}s,$$

and a simpler analysis than that given above for $\tilde{M}_k(z)$. In particular, when k lies in the saddle-point range (II) and $\rho \ge 1$, we have, by the same analysis used for (46),

$$\begin{aligned}
\tilde{M}_k^{(\ell)}(ne^{i\theta}) &= O\left(n^{-\rho-\ell} \sum_{j\in\mathbb{Z}} \left|\rho + i t_j\right|^\ell |\Gamma(\rho+1+it_j)| \int_{|t-t_j|\le \pi/\log(p/q)} \left|(p^{-\rho-it} + q^{-\rho-it})^k\right| \mathrm{d}t\right) \\
&= O\left(k^{-1/2}(q/p)^{\rho/2}(p^{-\rho}+q^{-\rho})^k n^{-\rho-\ell}\rho^\ell \sum_{j\in\mathbb{Z}} |1+it_j|^\ell |\Gamma(\rho+1+it_j)|\right) \\
&= O\left(k^{-1/2}(q/p)^{\rho/2}(p^{-\rho}+q^{-\rho})^k n^{-\rho-\ell}\rho^\ell\right) \\
&= O\left(\rho^\ell n^{-\ell} \tilde{M}_k(n)\right),
\end{aligned}$$

uniformly for $|\theta| \le \pi/2 - \varepsilon$. The other cases are treated similarly. Alternatively, we can use the estimates (48), (51) and (57) for $\tilde{M}_k(ne^{i\theta})$ and the integral formula

$$\tilde{M}_k^{(\ell)}(z) = \frac{\ell!}{2\pi i} \int_{|w-z|\le \varepsilon|z|/\rho_0} \frac{\tilde{M}_k(w)}{(w-z)^{\ell+1}} \mathrm{d}w,$$

following a standard analysis (referred to as Ritt's theorem in [63, pp. 9–10]). □

An application of Proposition 1 (analytic de-Poissonization) and the above lemma leads to our refinement.

Theorem 6. *If $q^{2k} n \to 0$, then*

$$\mu_{n,k} = \tilde{M}_k(n) - \frac{n}{2} \tilde{M}_k''(n) + O\left(\rho_0^4 n^{-2} \tilde{M}_k(n)\right), \tag{61}$$

where ρ_0 is given in (59).

Proof. By (26) and (60), we can take $\delta(n) = \rho_0/n$, which is $o(n^{-1/2})$ if $q^{2k} n \to 0$. □

Remark. The condition that $q^{2k} n \to 0$ is also necessary for $\mu_{n,k} \sim \tilde{M}_k(n)$ because otherwise $\mu_{n,k} \sim q^k n (1-q^k)^{n-1}$, which is not asymptotically equivalent to $\tilde{M}_k(n)$. Note also that (61) and (60) imply that $\mu_{n,k} = \tilde{M}_k(n)\left(1 + O(\rho_0^2 n^{-1})\right)$.

4 Variance of the external profile

Asymptotic approximations to $\sigma_{n,k}^2 := \mathbb{V}(B_{n,k})$ are derived in this section. It turns out that the variance is of the same order as the mean in all ranges, implying that the standard deviation is small; therefore we expect asymptotic normality when the variance tends to infinity with n. The calculations here are more involved due to the cancelation of dominant orders of $\mu_{n,k}^2$. The key idea is a suitable manipulation of the corresponding de-Poissonized approximations for the mean and the second moments.

4.1 Recurrence and generating functions of the second moment

Let $\nu_{n,k} := \mathbb{E}(B_{n,k}^2)$ denote the second moment of $B_{n,k}$. By (1), we have the recurrence

$$\nu_{n,k} = \sum_{0\le j\le n} \binom{n}{j} p^j q^{n-j} \left(\nu_{j,k-1} + \nu_{n-j,k-1}\right) + \omega_{n,k},$$

for $n, k \ge 1$ with $\nu_{n,0} = \delta_{n,1}$, where

$$\omega_{n,k} := 2 \sum_{0\le j\le n} \binom{n}{j} p^j q^{n-j} \mu_{j,k-1}\mu_{n-j,k-1}.$$

Generating functions. Let $N_k(z) := \sum_n \nu_{n,k} z^n/n!$. Then $N_k(z)$ satisfies

$$N_k(z) = e^{qz} N_{k-1}(pz) + e^{pz} N_{k-1}(qz) + \omega_k(z) \qquad (k \ge 2),$$

with $N_1(z) = 2pqz^2 + M_1(z)$, where $\omega_k(z) := 2M_{k-1}(pz)M_{k-1}(qz)$. It follows that the Poisson generating function $\tilde{N}_k(z) := e^{-z}N_k(z)$ satisfies

$$\tilde{N}_k(z) = \tilde{N}_{k-1}(pz) + \tilde{N}_{k-1}(qz) + \tilde{\omega}_k(z),$$

where $\tilde{\omega}_k(z) = 2\tilde{M}_{k-1}(pz)\tilde{M}_{k-1}(qz)$. By iterating this functional equation, we obtain

$$\tilde{N}_k(z) = \sum_{0\le \ell<k} \binom{k-1}{\ell} \tilde{N}_1(p^\ell q^{k-1-\ell} z) + \sum_{0\le m\le k-2} \sum_{0\le \ell\le m} \binom{m}{\ell} \tilde{\omega}_{k-m}(p^\ell q^{m-\ell} z),$$

for $k \ge 2$.

Regularity of $N_k(z)$. The following estimate is useful in justifying the application of the saddle-point method and the de-Poissonization procedure.

Lemma 5. *Let $z = re^{i\theta}$, where $r \ge 0$ and $|\theta| \le \pi$. Then the estimate*

$$|N_k(z)| \le N_k(r) e^{-cr\theta^2} \tag{62}$$

holds for $r \ge 0$ and $|\theta| \le \pi$ for some constant c independent of r, k and θ.

Proof. We start from

$$N_k(z) = e^z \sum_{0\le \ell<k} \binom{k-1}{\ell} \tilde{N}_1(p^\ell q^{k-1-\ell} z) + \omega_k(z) + e^z \sum_{1\le m\le k-2} \sum_{0\le \ell\le m} \binom{m}{\ell} \tilde{\omega}_{k-m}(p^\ell q^{m-\ell} z),$$

and apply Lemma 2 to the first sum. For the second term, we observe first that, by (26),

$$|\omega_k(z)| \le 2M_{k-1}(pr)M_{k-1}(qr)e^{-cr\theta^2} = \omega_k(r)e^{-cr\theta^2},$$

uniformly for $r \ge 0$ and $|\theta| \le \pi$. It remains to estimate the last sum

$$\left| e^z \sum_{1\le m\le k-2} \sum_{0\le \ell\le m} \binom{m}{\ell} \tilde{\omega}_{k-m}(p^\ell q^{m-\ell} z) \right|,$$

for which we apply the same argument as that used in the proof of Lemma 2, yielding an estimate of the type (28). Collecting the three parts gives (62). □

An auxiliary function for asymptotic variance. Define $\tilde{V}_k(z) := \tilde{N}_k(z) - \tilde{M}_k^2(z)$ as the Poisson variance. Then $\tilde{V}_k(z)$ satisfies

$$\tilde{V}_k(z) = \tilde{V}_{k-1}(pz) + \tilde{V}_{k-1}(qz) \qquad (k \geq 2),$$

which, after iterations, yields

$$\tilde{V}_k(z) = \sum_{0 \leq j < k} \binom{k-1}{j} \tilde{V}_1(p^j q^{k-1-j} z), \tag{63}$$

where

$$\tilde{V}_1(z) = \tilde{M}_1(z) + 2pqz^2 e^{-z} - \tilde{M}_1^2(z).$$

It follows that

$$\tilde{V}_k(z) = \frac{1}{2\pi i} \int_{(\rho)} z^{-s} \Gamma(s+1) h(s) \left(p^{-s} + q^{-s}\right)^k \mathrm{d}s, \tag{64}$$

where $\rho > -2$ and

$$h(s) := 1 - \frac{1}{p^{-s} + q^{-s}} - \frac{s+1}{p^{-s} + q^{-s}} \left(\frac{p^{-s} + q^{-s} + 1}{2^{s+2}} - \frac{2p}{(1+p)^{s+2}} - \frac{2q}{(1+q)^{s+2}} \right).$$

Observe that the Poisson variance $\tilde{V}_k(z)$ differs from the Poisson mean $\tilde{M}_k(z)$ only by the appearance of $h(s)$ instead of $g(s)$.

4.2 Asymptotics of $\sigma_{n,k}^2$

In this section we show that the variance $\sigma_{n,k}^2 := \mathbb{V}(B_{n,k})$ is asymptotically equivalent to $\mu_{n,k}$ when k lies in range (I), to $2\mu_{n,k}$ when k in ranges (III) and (IV), and of the same order as $\mu_{n,k}$ in the central range (II).

Theorem 7. (i) *If* $1 \leq k \leq \alpha_1(1 + o(1))L_n$, *then*

$$\sigma_{n,k}^2 \sim \mu_{n,k}. \tag{65}$$

(ii) *If* $\alpha_1(L_n - LLL_n + K_n) \leq k \leq \alpha_2(L_n - K_n\sqrt{L_n})$, *then*

$$\sigma_{n,k}^2 = G_2\left(\rho; \log_{p/q} p^k n\right) \frac{n^{-\rho}\,(p^{-\rho} + q^{-\rho})^k}{\sqrt{2\pi\beta_2(\rho)k}} \left(1 + O\left(\frac{1}{k(p/q)^\rho} + \frac{1}{k(\rho+2)^2}\right)\right), \tag{66}$$

where $\rho = \rho(n,k) > -2$ *is given in (36) and*

$$G_2(\rho; x) = \sum_{j \in \mathbb{Z}} h(\rho + it_j)\Gamma(\rho + 1 + it_j) e^{-2j\pi i x} \qquad (t_j := 2j\pi/\log(p/q)).$$

(iii) *If* $k \geq \alpha_2(1 - o(1))L_n$, *then*

$$\sigma_{n,k}^2 \sim 2\mu_{n,k}. \tag{67}$$

Proof. Since most details are similar to those for $\mu_{n,k}$, only the key differences will be highlighted. We separate the analysis into two overlapping cases: $1 \leq k \leq k_0 = \alpha_1(L_n - LLL_n + \log(p/q - 1))$ and $q^{2k}n \to 0$.

Consider the first case when $1 \leq k \leq k_0$. In this range, $\tilde{M}_k(ne^{i\theta}) \to 0$ for $\theta = o(LL_n^{-1/2})$ by (30) and (33). By (63) and the same proof of (30), we have

$$\tilde{V}_k(ne^{i\theta}) = \tilde{M}_k(ne^{i\theta}) \left(1 + O\left(q^k n e^{-q^k n \cos\theta}\right)\right),$$

uniformly for $\theta = o(LL_n^{-1/2})$. Then since

$$\nu_{n,k} = n![z^n]N_k(z) = n![z^n]e^z\left(\tilde{V}_k(z) + \tilde{M}_k^2(z)\right),$$

and $\mu_{n,k} \to 0$, it is straightforward to show, by (30), (62) and the proof of (22), that $\sigma_{n,k}^2 \sim \mu_{n,k}$ in this case, which established (65).

We now consider the range $q^{2k}n \to 0$ that will cover the other two cases. We will show that $\mathbb{V}(B_{n,k}) \sim \tilde{V}_k(n)$ that will imply (66) and (67) (indeed, $|h(-2)| = 2|g(-2)| = 4pq/(p^2+q^2)$).

In this case, by the integral representation (64) and the same method of proof for $\tilde{M}_k(z)$, we have

$$\tilde{V}_k^{(\ell)}(ne^{i\theta}) = O\left(\rho_0^\ell n^{-\ell}\tilde{V}_k(n)\right), \tag{68}$$

uniformly for $\theta = o(LL_n^{-1/2})$ whenever $q^{2k}n \to 0$. On the other hand, since $\tilde{M}_k(z)$ satisfies the estimate (60), we have

$$\frac{\mathrm{d}^\ell}{\mathrm{d}z^\ell}\tilde{M}_k^2(z)\Big|_{z=ne^{i\theta}} = O\left(\rho_0^\ell n^{-\ell}\tilde{M}_k^2(n)\right) \qquad (\ell = 0, 1, \dots),$$

uniformly for $\theta = o(LL_n^{-1/2})$. Thus $\tilde{N}_k(z) = \tilde{V}_k(z) + \tilde{M}_k^2(z)$ also satisfies condition (15) of Proposition 1. Therefore, by (19) of Proposition 1 we have

$$\nu_{n,k} = \tilde{N}_k(n) - \frac{n}{2}\tilde{N}_k''(n) + O\left(\rho_0^4 n^{-2}\tilde{N}_k(n)\right),$$

for $k \geq k_0$. Note that $\tilde{N}_k(n) = \Theta(\mu_{n,k}^2)$ when $\mu_{n,k} \to \infty$ but $\tilde{N}_k(n) = \Theta(\mu_{n,k})$ when $\mu_{n,k} \to 0$.

On the other hand, by (61),

$$\mu_{n,k}^2 = \tilde{M}_k^2(n) - n\tilde{M}_k(n)\tilde{M}_k''(n) + O\left(\rho_0^4 n^{-2}\tilde{M}_k^2(n)\right).$$

Therefore

$$\sigma_{n,k}^2 = \tilde{V}_k(n)\left(1 + O(\rho_0^2 n^{-1}\mu_{n,k})\right),$$

whenever $q^{2k}n \to 0$. Note that the O-term is at most of order $LL_n^2 L_n^{-1/2}$. In fact, a further refinement (see (16 or [38]) shows that

$$\sigma_{n,k}^2 = \tilde{V}_k(n) - n\tilde{M}_k'(n)^2 - \frac{n}{2}\tilde{V}_k''(n) + O\left(\rho_0^4 n^{-2}\tilde{N}_k(n)\right).$$

It remains to derive asymptotic approximations to $\tilde{V}_k(n)$, which follow the same methods of proof used for $\tilde{M}_k(n)$, the only difference being changing all occurrences of $g(s)$ to $h(s)$. In particular, $G_2(\rho; x) \sim G_1(\rho; x)$ when $\rho \to \infty$, which corresponds to $k \leq \alpha_1(1 + o(1))L_n$; also $|h(-2)| = 2|g(-2)| = 4pq/(p^2+q^2)$. This proves (66) and (67). □

We conclude this section with two corollaries.

Corollary 4. *The variance $\sigma_{n,k}^2 \to \infty$ iff the mean $\mu_{n,k} \to \infty$.*

Corollary 5. *If $\mu_{n,k} \to \infty$, then $B_{n,k}/\mu_{n,k} \to 1$ in probability.*

Proof. This follows from Theorem 7 and Chebyshev's inequality. □

5 Limiting distribution

We prove in this section that the limiting distribution of $B_{n,k}$ is normal if $\sigma_{n,k} \to \infty$ and is Poisson if the variance remains bounded. Since the mean and the variance are asymptotically of the same order, the conditions can also be stated by replacing $\sigma_{n,k}$ by $\mu_{n,k}$. These results cover the range when $k \geq \alpha_1(L_n - LL_n + O(1))$ and $k \leq \alpha_2(L_n + O(1))$. Outside this range, $\mu_{n,k} \to 0$, so $B_{n,k} \to 0$ in probability.

Theorem 8. (i) *If $\sigma_{n,k} \to \infty$, then*

$$\frac{B_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \xrightarrow{d} \mathcal{N}(0,1), \tag{69}$$

where $\mathcal{N}(0,1)$ denotes a standard normal random variable and $\xrightarrow{d}$ stands for convergence in distribution.
(ii) *If $\sigma_{n,k} = \Theta(1)$, then*

$$\begin{cases} \mathbb{P}\left(B_{n,k} = 2m\right) = \dfrac{\lambda_0^m}{m!}\, e^{-\lambda_0} + o(1), \\ \mathbb{P}\left(B_{n,k} = 2m+1\right) = o(1), \end{cases} \tag{70}$$

uniformly for (finite) $m \geq 0$, where $\lambda_0 := pqn^2(p^2+q^2)^{k-1}$.

Note that (70) implies that $B_{n,k}$ takes asymptotically only even numbers when the mean is bounded. This indeed holds in the wider range when

$$k \geq \frac{3}{\log(1/(p^3+q^3))}(L_n + K_n).$$

Intuitively, this is the range where $\binom{n}{j}(p^j+q^j)^k \to 0$ for all $j \geq 3$, where $\binom{n}{j}(p^j+q^j)^k$ is the expected number of groups of j input-strings having common prefixes of length at least k; since $\sum_{j\geq 3}\binom{n}{j}(p^j + q^j)^k \to 0$, all nodes appearing at levels $\geq k$ are most likely only in pairs; see [33] for more precise local limit theorems for $B_{n,k}$.

Let $\tilde{\sigma}_{n,k} := \sqrt{\tilde{V}_k(n) - n\tilde{M}_k'(n)^2}$ (see Theorem 7). We will prove, by extending the above de-Poissonization procedure, that

$$\mathbb{E}\exp\left(\frac{B_{n,k} - \tilde{M}_k(n)}{\tilde{\sigma}_{n,k}}\, i\varphi\right) = e^{-\varphi^2/2}\left(1 + O\left(\frac{1+|\varphi|^3}{\sigma_{n,k}}\right)\right), \tag{71}$$

uniformly for $\varphi = o(\sigma_{n,k}^{1/5})$, which implies (69) by Lévy's continuity theorem since $\mu_{n,k} \sim \tilde{M}_k(n)$ and $\sigma_{n,k} \sim \tilde{\sigma}_{n,k}$ when $\mu_{n,k} \to \infty$. Note that as far as the central limit theorem is concerned, the validity of (71) in the range $\varphi = O(1)$ suffices; observe also that centering $B_{n,k}$ by the exact mean or normalizing $B_{n,k}$ by the exact variance will result in a poorer error term.

Our method of proof of (71) is roughly as follows. We start from deriving a closed-form expression for the bivariate generating function $\mathcal{P}_k(z,y) = \sum_{n\geq 0} P_{n,k}(y)z^n/n!$ by using the recurrence (1). We then will apply the Cauchy integral representation to prove (71), for which we need, as in the analytic de-Poissonization used above, a crude estimate for $|\mathcal{P}_k(ne^{i\theta}, e^{i\varphi})|$ for $|\theta|$ away from zero, as well as a more precise local expansion when $|\theta|$ is very close to zero. The proof for the Poisson limit law (70) is similar.

5.1 An exact expression for $\mathcal{P}_k(z, y)$

By (1), we have the functional equation

$$\mathcal{P}_k(z, y) = \mathcal{P}_{k-1}(pz, y)\mathcal{P}_{k-1}(qz, y) \qquad (k \geq 2),$$

with

$$\mathcal{P}_1(z, y) = e^z + (y-1)z\left(p(e^{qz}-1) + q(e^{pz}-1)\right) + pq(y-1)^2z^2.$$

By iterating this functional equation, we obtain

$$\mathcal{P}_k(z, y) = \prod_{0\leq j<k} \mathcal{P}_1(p^j q^{k-1-j} z, y)^{\binom{k-1}{j}} \qquad (k \geq 1). \tag{72}$$

This expression is, although explicit, less transparent from an asymptotic viewpoint.

5.2 A uniform estimate for $|\mathcal{P}_k(re^{i\theta}, y)|$

We first prove a uniform bound on $|\mathcal{P}_k(re^{i\theta}, y)|$ that is necessary for the proof of Theorem 8.

Proposition 3. *Uniformly for $k \geq 1$, $r \geq 0$, $|\theta| \leq \pi$ and $|y| = 1$*

$$|\mathcal{P}_k(re^{i\theta}, y)| \leq e^{r - cr\theta^2}, \tag{73}$$

for some constant $c > 0$ independent of k, r and θ.

In order to prove the above proposition we need a lemma.

Lemma 6. *If $z = re^{i\theta}$, where $r \geq 0$ and $|\theta| \leq \pi$, then*

$$|e^z - 1 - z| \leq (e^r - 1 - r)e^{-c_1 r\theta^2}, \tag{74}$$

where $c_1 := 2/(3\pi^2)$. On the other hand, if $r \geq r_0$, where $r_0 \approx 1.37$ solves the equation $e^r - r = e^{r/3} + 1$, then

$$|e^z - z| \leq (e^r - r)e^{-c_1 r\theta^2/2} \qquad (|\theta| \leq \pi). \tag{75}$$

Proof. The first inequality is a special case of Pittel's inequality (see [67])

$$\left|e^z - \sum_{0\leq j<m} \frac{z^j}{j!}\right| \leq \left(e^r - \sum_{0\leq j<m} \frac{r^j}{j!}\right) e^{-2r\theta^2/(\pi^2(m+1))} \qquad (r \geq 0; |\theta| \leq \pi).$$

A simple proof of (74) (following [67]) is as follows.

$$\begin{aligned}
|e^z - 1 - z| &= |e^{z/3}|\left|e^{2z/3} - (1+z)e^{-z/3}\right| \\
&= e^{r\cos(\theta)/3}\left|\sum_{j\geq 2} \frac{z^j}{j!3^j}\left(2^j + (-1)^j(3j-1)\right)\right| \\
&\leq e^{r\cos(\theta)/3}(e^{2r/3} - (1+r)e^{-r/3}),
\end{aligned}$$

since $2^j + (-1)^j(3j-1) \geq 0$ for $j \geq 2$. Thus (74) follows from (29).

For the proof of the inequality (75), we have

$$\begin{aligned}|e^z - z| &\le |e^z - 1 - z| + 1\\ &\le (e^r - 1 - r)e^{-c_1 r\theta^2} + 1\\ &\le (e^r - r)e^{-c_1 r\theta^2/2},\end{aligned}$$

since the last inequality is equivalent to

$$1 - e^{-c_1 r\theta^2} \le (e^r - r)e^{-c_1 r\theta^2/2}\left(1 - e^{-c_1 r\theta^2/2}\right),$$

or $e^{c_1 r\theta^2/2} + 1 \le e^r - r$, which follows from our choice of r in view of the inequalities $e^{c_1 r\theta^2/2} + 1 \le e^{r/3} + 1 \le e^r - r$. □

Proof of Proposition 3. We separate the proof into two cases: $r \le r_0$ and $r \ge r_0$. In the first case, we use the expansion

$$\mathcal{P}_1(z, y) = 1 + z + \frac{z^2}{2}\left(1 - 2pq(1 - y^2)\right) + \sum_{j\ge 2}\frac{z^j}{j!}\left(1 - j(pq^{j-1} + qp^{j-1})(1 - y)\right),$$

which yields

$$\begin{aligned}|\mathcal{P}_1(re^{i\theta}, e^{i\varphi})| &\le |1 + re^{i\theta}| + \sum_{j\ge 2}\frac{r^j}{j!}\\ &\le e^r - \frac{2r\theta^2}{\pi^2(1 + r)}\\ &\le e^{r - c_2 r\theta^2},\end{aligned} \tag{76}$$

uniformly for $0 \le r \le r_0$ and $|\theta| \le \pi$, where we used again (29) and $c_2 := 2/(\pi^2(1 + r_0)^2 e^{r_0})$.

Assume now $r \ge r_0$. We can write $\mathcal{P}_1(z, y)$ as follows.

$$\mathcal{P}_1(z, y) = a_1(pz)a_1(qz) + z + (qza_2(pz) + pza_2(qz))y + pqz^2y^2,$$

where $a_1(z) := e^z - z$ and $a_2(z) := e^z - 1 - z$. Note that $\mathcal{P}_1(z, 1) = e^z$. By applying the two inequalities (74) and (75), we have

$$\begin{aligned}\left|\mathcal{P}_1(re^{i\theta}, e^{i\varphi})\right| &\le a_1(pr)a_1(qr)e^{-c_1 r\theta^2/2} + r + qra_2(pr)e^{-c_1 pr\theta^2} + pra_2(qr)e^{-c_1 qr\theta^2} + pqr^2\\ &\le \left(e^r - r - pqr^2\right)e^{-c_1 qr\theta^2} + r + pqr^2\\ &\le e^{r - c_1 qr\theta^2/2},\end{aligned} \tag{77}$$

the last inequality following from an argument similar to the proof of (75). Indeed, it is equivalent to

$$(r + pqr^2)(e^{c_1 qr\theta^2/2} + 1) \le e^r,$$

but the left-hand side is less than $(r + r^2/4)(e^{r/6} + 1)$, which is in turn less than e^r for $r \ge 0.99$.

Collecting the two inequalities (76) and (77), we obtain

$$|\mathcal{P}_1(re^{i\theta}, e^{i\varphi})| \le e^{r - cr\theta^2} \qquad (c = \min\{c_1, c_2\}),$$

uniformly for $r \ge 0$ and $|\theta| \le \pi$. This implies (73) by (72). □

5.3 Local expansion of $\mathcal{P}_k(re^{i\theta}, e^{i\varphi})$

Recall that $\theta_0 := n^{-2/5}$ and ρ_0 is defined in (59).

Proposition 4. *Assume that $\mu_{n,k} \to \infty$. Then uniformly for $|\theta| \le \theta_0$ and $\varphi = o(\sigma_{n,k}^{-2/3})$*

$$\mathcal{P}_k(ne^{i\theta}, e^{i\varphi}) = \exp\left(n - \frac{n}{2}\,\theta^2 + \tilde{M}_k(n) i\varphi - n\tilde{M}_k'(n)\varphi\theta - \frac{\tilde{V}_k(n)}{2}\,\varphi^2 + O(E_4)\right), \tag{78}$$

where

$$E_4 := n|\theta|^3 + \rho_0^2\sigma_{n,k}^2|\varphi|\theta^2 + \rho_0\sigma_{n,k}^2\varphi^2|\theta| + \sigma_{n,k}^2|\varphi|^3.$$

Proof. Define

$$Q(z, y) := \log e^{-z}\mathcal{P}_1(z, y) = \log\left(1 + a_3(z)(y-1) + a_4(z)(y-1)^2\right),$$

where $a_3(z) := pze^{-pz} + qze^{-qz} - ze^{-z}$ and $a_4(z) := pqz^2e^{-z}$. Let

$$Q_k(z, y) := \sum_{0\le j<k} \binom{k-1}{j} Q(p^j q^{k-1-j} z, y) = \log e^{-z}\mathcal{P}_k(z, y).$$

First, we prove in Lemma 7 of Appendix B that $\mathcal{P}_1(re^{i\theta}, e^{i\varphi})$ is away from zero for $r \ge 0$ and $|\theta| \le \varepsilon$, implying that $Q_k(z, y)$ is well-defined when $|\arg(z)| \le \varepsilon$.

Then since $\mu_{n,k} \to \infty$, we need only to consider $k \ge k_0$. To that purpose, we start from the expansion

$$Q(z, y) = \begin{cases} pq(y^2-1)z^2 + O(|y-1||z|^3), & \text{as } z \to 0; \\ q(y-1)ze^{-qz}\left(1 + O(e^{-(p-q)\Re(z)})\right), & \text{as } z \to \infty, |\arg(z)| \le \varepsilon. \end{cases} \tag{79}$$

By (79), we have

$$Q_k(z, y) = \frac{1}{2\pi i}\int_{(\rho)} z^{-s} Q^*(s, y)(p^{-s} + q^{-s})^{k-1}\mathrm{d}s, \tag{80}$$

where $\rho > -2$ and $Q^*(s, y) := \int_0^\infty z^{s-1}Q(z, y)\mathrm{d}z$ is well-defined for $\Re(s) > -2$. Note that

$$Q(z, y) = a_3(z)(y-1) + \frac{2a_4(z) - a_3(z)^2}{2}\,(y-1)^2 + \bar{Q}(z, y)(y-1)^3,$$

where by Taylor's remainder formula

$$\bar{Q}(z, y) := \int_0^1 (1-t)^2(a_3(z) + 2a_4(z)(y-1)t) \times \frac{(a_3(z)^2 - 3a_4(z) + a_3(z)a_4(z)(y-1)t + a_4(z)^2(y-1)^2t^2)}{(1 + a_3(z)(y-1)t + a_4(z)(y-1)^2t^2)^3}\,\mathrm{d}t.$$

The exact form is of less importance here; we need instead the estimates $\bar{Q}(z, y) = O(|z|^4) = O(|z|^2)$ as $z \to 0$ and

$$\bar{Q}(z, y) = O\left(|z|^3 e^{-3q\Re(z)}\right) = O\left(|z|e^{-q\Re(z)}\right),$$

as $z \to \infty$ in the sector $\{z : |\arg(z)| \le \varepsilon\}$. This expansion gives

$$Q_k(z, y) = \tilde{M}_k(z)(y-1) + \frac{\tilde{V}_k(z) - \tilde{M}_k(z)}{2}\,(y-1)^2 + \bar{Q}_k(z, y)(y-1)^3,$$

where

$$\bar{Q}_k(z,y) := \sum_{0\le j<k} \binom{k-1}{j} \bar{Q}(p^j q^{k-1-j} z, y).$$

An application of Lemma 8 presented in Appendix C yields, with $z = ne^{i\theta}$,

$$Q_k(z,y) = \tilde{M}_k(z)(y-1) + \frac{\tilde{V}_k(z) - \tilde{M}_k(z)}{2}(y-1)^2 + O\left(|y-1|^3 |\tilde{M}_k(ne^{i\theta})|\right),$$

where the O-term holds uniformly for $|\theta| \le \varepsilon$ and $|y-1| = o(1)$. Since $\sigma_{n,k}^2 = \Theta(\mu_{n,k}) \to \infty$, this leads to (78) by expansions of $\tilde{M}_k(ne^{i\theta})$ and $\tilde{V}_k(ne^{i\theta})$ at $\theta = 0$, using the estimates (60) and (68). This completes the proof of Proposition 4. □

5.4 Proof of Theorem 8.

We are now ready to prove Theorem 8.

Proof of the central limit theorem (69). By Cauchy's integral formula and the two estimates (73) and (78), we have, similar to (25),

$$\begin{aligned}\mathbb{E}\left(e^{B_{n,k} i\varphi}\right) &= \frac{n!}{2\pi i}\int_{|z|=n} z^{-n-1}\mathcal{P}_k(z,y)\mathrm{d}z \\ &= \frac{n! n^{-n}}{2\pi} e^{n+\tilde{M}_k(n)i\varphi - \tilde{V}_k(n)\varphi^2/2}\int_{-\theta_0}^{\theta_0} e^{-n\theta^2/2 - n\tilde{M}_k'(n)\varphi\theta}\left(1+O(E_4)\right)\mathrm{d}\theta + O\left(n^{-1/10}e^{-cn^{1/5}}\right),\end{aligned}$$

since $E_4 \to 0$ in the range of integration and when $\varphi = o(\sigma_{n,k}^{-4/5})$. Applying Stirling's formula, extending the integration limits to $\pm\infty$ and making the change of variables $\theta \mapsto \theta n^{-1/2}$, we obtain

$$\begin{aligned}\mathbb{E}\left(e^{B_{n,k} i\varphi}\right) = \frac{e^{\tilde{M}_k(n)i\varphi - \tilde{\sigma}_{n,k}^2\varphi^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(\theta+\sqrt{n}\,\tilde{M}_k'(n)\varphi)^2/2} \\ \times\left(1 + O\left(\frac{1+|\theta|^3}{\sqrt{n}} + \frac{\theta^2}{n}\rho_0^2\sigma_{n,k}^2|\varphi| + \frac{|\theta|}{\sqrt{n}}\rho_0\sigma_{n,k}^2\varphi^2 + \sigma_{n,k}^2|\varphi|^3\right)\right)\mathrm{d}\theta,\end{aligned}$$

uniformly in φ. A straightforward evaluation of the integral gives (71). This completes the proof of (69).

Proof of the Poisson limit theorem (70). The proof of (70) is similar as above but proceeds slightly differently. We first show that

$$Q_k(ne^{i\theta}, y) := \log e^{-z}\mathcal{P}_k(z,y) = \lambda_0(y^2-1)e^{2i\theta} + O\left(|y-1|n^{-e(p)}\right), \tag{81}$$

uniformly for $|y| = 1$ and $|\theta| \le \varepsilon$, where $e(p) := 2\log(p^3+q^3)/\log(p^2+q^2) - 3 \in (0,1)$ for $p \in (1/2, 1)$.

This follows from the Mellin inversion integral (80) since the Mellin transform $Q^*(s,y)$ has a simple pole at $s = -2$ with residue $pq(y^2-1)$ and can be meromorphically continued into the whole s-plane. Indeed, by moving the line of integration of the integral in (80) to $\Re(s) = -3-\varepsilon$, we obtain

$$Q_k(ne^{i\theta}, y) = \lambda_0(y^2-1)e^{2i\theta} + O\left(|y-1|n^3(p^3+q^3)^k\right),$$

whenever $|\theta| \le \pi/2 - \varepsilon$ and

$$k \ge \frac{L_n + K_n}{\log((p^2+q^2)/(p^3+q^3))}.$$

Since $\mu_{n,k} = \Theta(1)$, we know that $k = \alpha_3 L_n + O(1)$ and for such values of k, we have $n^3(p^3+q^3)^k = \Theta(n^{-e(p)})$. Thus (81) follows.

By (81) and the same choice of θ_0 and (73), we then deduce that

$$\begin{aligned}\mathbb{E}\left(e^{B_{n,k}i\varphi}\right) &= \frac{e^{\lambda_0(e^{2i\varphi}-1)}}{\sqrt{2\pi}}\int_{-n^{1/10}}^{n^{1/10}} e^{-\theta^2/2}\left(1+O\left(n^{-e(p)}|\varphi|\right)\right)\\ &\qquad\times\left(1+\frac{12\lambda_0(e^{2i\varphi}-1)i\theta - i\theta^3}{6\sqrt{n}}+O\left(\frac{1+\lambda_0|\varphi|\theta^2+\theta^6}{n}\right)\right)\mathrm{d}\theta\\ &= e^{\lambda_0(e^{2i\varphi}-1)}\left(1+O(E_5)\right),\end{aligned}$$

uniformly for $|\varphi|\le\pi$, where $E_5 := |\varphi|n^{-e(p)} + (1+\lambda_0|\varphi|)n^{-1}$. Thus

$$\mathbb{P}\left(\frac{B_{n,k}}{2}=m\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-im\varphi+\lambda_0(e^{i\varphi}-1)}\left(1+O(E_5)\right)\mathrm{d}\varphi,$$

from which the even case in (70) follows since, by (50), $\mu_{n,k}\sim 2\lambda_0$. The odd case is similar. □

Remark. Note that λ_0 is periodic in nature since $k = \alpha_3 L_n + O(1) \in \mathbb{Z}$; indeed, we can write

$$k = \lfloor \alpha_3 L_n\rfloor + \ell = \alpha_3 L_n + \ell - \{\alpha_3 L_n\} \qquad (\ell\in\mathbb{Z}),$$

so that

$$n^2(p^2+q^2)^k = \exp\left(-\frac{2}{\alpha_3}\left(\ell - \{\alpha_3 L_n\}\right)\right).$$

This is why we did not state the Poisson convergence (70) in the usual form: if $\lambda_0\to\lambda<\infty$, then $B_{n,k}$ converges in distribution to $2\mathrm{Po}(\lambda)$, where $\mathrm{Po}(\lambda)$ denotes a Poisson variate with parameter λ.

6 The internal profile

We consider the internal profile in this section. All asymptotic results follow the same footsteps as in the proof we used for $B_{n,k}$; details will thus be omitted. The main differences are that $\mathbb{E}(I_{n,k})$ and $\mathbb{V}(I_{n,k})$ are not of the same order for all ranges of k, and $I_{n,k}$ assumes both odd and even values when $k=\alpha_3 L_n + O(1)$. These are intuitively clear since most levels close to the root are full and internal nodes do not necessarily appear in pairs near the fringes of the tree.

Let $P^{[I]}_{n,k}(y) = \mathbb{E}(y^{I_{n,k}})$ be the probability generating function of $I_{n,k}$. Then

$$P^{[I]}_{n,k}(y) = \sum_{0\le j\le n}\binom{n}{j}p^j q^{n-j}P^{[I]}_{j,k-1}(y)P^{[I]}_{n-j,k-1}(y) \qquad (n\ge 2; k\ge 1), \tag{82}$$

with the initial conditions

$$P^{[I]}_{n,k}(y) = \begin{cases} y, & \text{if } n\ge 2 \text{ and } k=0;\\ 1, & \text{if } n\le 1 \text{ and } k\ge 0.\end{cases}$$

From this, we obtain, defining $\mathcal{P}^{[I]}_k(z,y) := \sum_n P^{[I]}_{n,k}(y)z^n/n!$,

$$\mathcal{P}^{[I]}_k(z,y) = \prod_{0\le j\le k}\mathcal{P}^{[I]}_0(p^jq^{k-j}z,y)^{\binom{k}{j}} \qquad (k\ge 0), \tag{83}$$

with $\mathcal{P}_0^{[I]}(z,y) = ye^z + (1-y)(1+z)$. This suggests that we consider

$$\bar{I}_{n,k} := 2^k - I_{n,k},$$

so that the bivariate generating function $\mathcal{P}_k^{[\bar{I}]}(z,y) := \sum_n P_{n,k}^{[\bar{I}]}(y)z^n/n!$ is given by

$$\mathcal{P}_k^{[\bar{I}]}(z,y) = \prod_{0\le j\le k} \mathcal{P}_0^{[\bar{I}]}(p^j q^{k-j} z, y)^{\binom{k}{j}} \qquad (k \ge 0),$$

with $\mathcal{P}_0^{[\bar{I}]}(z,y) = e^z + (y^{-1}-1)(1+z)$.

6.1 Expected internal profile

We state without proof the asymptotics of $\mathbb{E}(I_{n,k})$ in this subsection. By (82) or (83), we deduce that the Poisson generating function

$$\tilde{M}_k^{[I]}(z) := e^{-z} \sum_{n\ge 0} \frac{\mathbb{E}(I_{n,k})}{n!} z^n,$$

satisfies

$$\tilde{M}_k^{[I]}(z) = 2^k - \sum_{0\le j\le k} \binom{k}{j} \tilde{M}_0^{[I]}(p^j q^{k-j} z) \tag{84}$$

$$= 2^k - \frac{1}{2\pi i} \int_{(\rho)} z^{-s}(s+1)\Gamma(s)(p^{-s}+q^{-s})^k \mathrm{d}s, \tag{85}$$

where $\rho > 0$ and $\tilde{M}_0^{[I]}(z) = (1+z)e^{-z}$. Thus, in particular,

$$\mathbb{E}(I_{n,k}) = 2^k - \sum_{0\le j\le k} \binom{k}{j} \left(1 + p^j q^{k-j}(n-1)\right) \left(1 - p^j q^{k-j}\right)^{n-1}.$$

Due to the presence of 2^k or the simple pole at $s = 0$, we have an additional phase transition for $\mathbb{E}(I_{n,k})$ at $\rho = 0$ or equivalently at $k \sim \alpha_0 L_n$, where

$$\alpha_0 := \frac{2}{\log(1/p) + \log(1/q)}.$$

We now list asymptotic approximations of $\mathbb{E}(I_{n,k})$ for various ranges of k (without proofs since they follow the same lines as the derivations presented above for the external profile).

Asymptotics of $\mathbb{E}(I_{n,k})$ when $1 \le k \le \alpha_1(1+o(1))L_n$. Since

$$\tilde{M}_1^{[I]}(z) = 2^k - (1+pz)e^{-pz} + (1+qz)e^{-qz} = 2^k - qze^{-qz}\left(1 + O(|z|^{-1})\right),$$

as $|z| \to \infty$ in the sector $|\arg(z)| \le \pi/2 - \varepsilon$, we see immediately that in this range

$$\mathbb{E}(I_{n,k}) = 2^k - \mathbb{E}(B_{n,k})(1+o(1)), \tag{86}$$

uniformly in k.

Asymptotics of $\mathbb{E}(I_{n,k})$ when $\alpha_1(L_n - LLL_n + K_n) \le k \le \alpha_0(L_n - K_n\sqrt{L_n})$. By applying the saddle-point method to the Mellin inversion integral in (85) and then de-Poissonizing, we deduce that in this range

$$\mathbb{E}(I_{n,k}) = 2^k - G_3\left(\rho; \log_{p/q} p^k n\right) \frac{n^{-\rho}\,(p^{-\rho}+q^{-\rho})^k}{\sqrt{2\pi\beta_2(\rho)k}} \left(1 + O\left(\frac{1}{k(p/q)^\rho} + \frac{1}{k\rho^2}\right)\right),$$

where $\rho = \rho(n,k) > 0$ satisfies the saddle-point equation (36), $\beta_2(\rho)$ is the same as in (37) and

$$G_3(\rho; x) = \sum_{j\in\mathbb{Z}} (\rho + 1 + it_j)\Gamma(\rho + it_j) e^{-2j\pi i x}$$

where $t_j := 2j\pi/\log(p/q)$.

Asymptotics of $\mathbb{E}(I_{n,k})$ when $k = \alpha_0(L_n + o(L_n^{2/3}))$. In this range, we write

$$k = \alpha_0(L_n + \xi\sqrt{\alpha_0\beta_2(0)L_n}),$$

where $\alpha_0\beta_2(0) = 2(\log(1/p) + \log(1/q))/\log(p/q)^2$ and $\xi = o(L_n^{1/6})$. The same uniform asymptotic analysis we used for proving (53) gives

$$\mathbb{E}(I_{n,k}) = 2^k \Phi(-\xi)\left(1 + O\left(\frac{1+|\xi|^3}{\sqrt{L_n}}\right)\right),$$

uniformly in ξ, where $\Phi(x)$ denotes the standard normal distribution function.

Asymptotics of $\mathbb{E}(I_{n,k})$ when $\alpha_0(L_n + K_n\sqrt{L_n}) \le k \le \alpha_2(L_n - K_n\sqrt{L_n})$. The same saddle-point method and de-Poissonization procedure yield

$$\mathbb{E}(I_{n,k}) = G_3\left(\rho; \log_{p/q} p^k n\right) \frac{n^{-\rho}\,(p^{-\rho}+q^{-\rho})^k}{\sqrt{2\pi\beta_2(\rho)k}} \left(1 + O\left(\frac{1}{k(p/q)^\rho} + \frac{1}{k(\rho+2)^2}\right)\right), \tag{87}$$

with ρ, $\beta_2(\rho)$ and G_3 as defined above.

Asymptotics of $\mathbb{E}(I_{n,k})$ when $k = \alpha_2(L_n + o(L_n^{2/3}))$. In this case, we write

$$k = \alpha_0(L_n + \xi\sqrt{\alpha_2\beta_2(-2)L_n}),$$

and we have

$$\mathbb{E}(I_{n,k}) = \frac{1}{2}\Phi(\xi) n^2 (p^2+q^2)^k \left(1 + O\left(\frac{1+|\xi|^3}{\sqrt{L_n}}\right)\right),$$

uniformly for $\xi = o(L_n^{1/6})$.

Asymptotics of $\mathbb{E}(I_{n,k})$ when $k \ge \alpha_2(L_n + K_n\sqrt{L_n})$. In this case, the simple pole at $s = -2$ dominates and we have

$$\mathbb{E}(I_{n,k}) = \frac{1}{2} n^2 (p^2+q^2)^k \left(1 + O\left(K_n^{-1} e^{-K_n^2/2 + O(K_n^3 L_n^{-1/2})}\right)\right)$$

as $n \to \infty$.

6.2 Asymptotics of $\mathbb{V}(I_{n,k})$

Since $\mathbb{V}(I_{n,k}) = \mathbb{V}(\bar{I}_{n,k})$, we can apply the same analysis used for proving Theorem 7 to derive asymptotic approximations to $\mathbb{V}(I_{n,k})$. The auxiliary function we need is

$$\tilde{V}_k^{[I]}(z) := e^{-z} \sum_{n\geq 0} \frac{\mathbb{E}(\bar{I}_{n,k}^2)}{n!} z^n - \left(e^{-z} \sum_{n\geq 0} \frac{\mathbb{E}(\bar{I}_{n,k})}{n!} z^n \right)^2,$$

which satisfies

$$\tilde{V}_k^{[I]}(z) = \sum_{0\leq j\leq k} \binom{k}{j} \tilde{V}_0^{[I]}(p^j q^{k-j} z) \qquad (k \geq 0), \tag{88}$$

where $\tilde{V}_0^{[I]}(z) = (1+z)e^{-z}(1-(1+z)e^{-z})$. Thus we have

$$\tilde{V}_k^{[I]}(z) = \frac{1}{2\pi i} \int_{(\rho)} z^{-s}(s+1)\Gamma(s) \left(1 - 2^{-s} - s2^{-s-2}\right) (p^{-s} + q^{-s})^k \mathrm{d}s,$$

where $\rho > -2$.

Asymptotics of $\mathbb{V}(I_{n,k})$ when $1 \leq k \leq \alpha_1(1+o(1))L_n$. In this range, we have

$$\mathbb{V}(I_{n,k}) \sim \mathbb{V}(B_{n,k}) \sim \mathbb{E}(B_{n,k}).$$

Asymptotics of $\mathbb{V}(I_{n,k})$ when $\alpha_1(L_n - LLL_n + K_n) \leq k \leq \alpha_2(L_n - K_n\sqrt{L_n})$. We have

$$\mathbb{V}(I_{n,k}) = G_4\left(\rho; \log_{p/q} p^k n\right) \frac{n^{-\rho}\,(p^{-\rho} + q^{-\rho})^k}{\sqrt{2\pi\beta_2(\rho)k}} \left(1 + O\left(\frac{1}{k(p/q)^\rho} + \frac{1}{k(\rho+2)^2}\right)\right),$$

where $\rho = \rho(n,k) > -2$ satisfies the saddle-point equation (36) and

$$G_4(\rho; x) = \sum_{j\in\mathbb{Z}} (\rho + 1 + i\,t_j)\Gamma(\rho + i\,t_j) \left(1 - 2^{-\rho - i t_j} - (\rho + i\,t_j)2^{-\rho-2-i t_j}\right) e^{-2j\pi i x}.$$

Asymptotics of $\mathbb{V}(I_{n,k})$ when $k \geq \alpha_2(L_n + K_n\sqrt{L_n})$. In this case, the simple pole at $s = -2$ again dominates and we have

$$\mathbb{V}(I_{n,k}) \sim \mathbb{E}(I_{n,k}).$$

Observe that, unlike external profile, the variance of the internal profile is asymptotically equivalent to the mean of the internal profile near the height of a trie.

From these asymptotic estimates and Chebyshev's inequality, we see that $I_{n,k}/\mathbb{E}(I_{n,k}) \to 1$ in probability if $\mathbb{E}(I_{n,k}) \to \infty$; see [15].

6.3 Limiting distributions

The same limiting Gaussian-Poisson behavior for $B_{n,k}$ holds for $I_{n,k}$. We state formally our main result for the internal profile in the following theorem. The proof is indeed simpler than that for Theorem 8 since the base function $\mathcal{P}_0^{[I]}(z,y)$ has a simpler form than $\mathcal{P}_0(z,y)$.

Theorem 9. (i) *If* $\mathbb{V}(I_{n,k}) \to \infty$, *then*

$$\frac{I_{n,k} - \mathbb{E}(I_{n,k})}{\sqrt{\mathbb{V}(I_{n,k})}} \xrightarrow{d} \mathcal{N}(0,1)$$

(ii) *If* $\mathbb{V}(I_{n,k}) = \Theta(1)$, *then, with* $\lambda_1 := n^2(p^2+q^2)/2$,

$$\mathbb{P}(I_{n,k} = m) = \frac{\lambda_1^m}{m!} e^{-\lambda_1} + o(1), \tag{89}$$

for all $m \geq 0$.

The theorem states that asymptotic normality (in the sense of convergence in distribution) holds as long as

$$\lceil \hat{k} \rceil \leq k \leq \alpha_3 L_n - K_n,$$

for any sequence $K_n \to \infty$, where $\hat{k}$ is defined in (58).

On the other hand, $I_{n,k}$ is asymptotically Poisson distributed when $k = \alpha_3 L_n + O(1)$. A result related to (89) was given in [67] by a method of moments, as a key step in deriving the asymptotic distribution of the height.

7 Profiles under the unbiased Bernoulli model

All exact expressions we derived up to now, as well as most asymptotic approximations, also hold when $p = q = 1/2$. The major difference is reflected by the fact that $\alpha_1 = \alpha_2$, so that the saddle-point range between α_1 and α_2 does not exist, and most analysis we give above becomes much simpler. For simplicity of presentation, we omit all error terms in our asymptotic estimates.

Expected external profile. By (8), the Poisson generating function of $\mathbb{E}(B_{n,k})$ is given exactly by

$$\tilde{M}_k(z) = z\left(e^{-z/2^k} - e^{-z/2^{k-1}}\right) \qquad (k \geq 1). \tag{90}$$

From this we deduce, by our de-Poissonization procedures, that

$$\mathbb{E}(B_{n,k}) \sim \begin{cases} n\left(1-2^{-k}\right)^{n-1}, & \text{if } 2^{-k}n \to \infty; \\ \tilde{M}_k(n), & \text{if } 4^{-k}n \to 0, \end{cases} \tag{91}$$

where the condition $4^{-k}n \to 0$ is due to the property

$$\tilde{M}_k^{(\ell)}(z) = O\left(2^{-k\ell}|\tilde{M}_k(z)|\right) \qquad (|\arg(z)| \leq \pi/2 - \varepsilon),$$

and $2^{-k} = o(n^{-1/2})$; see Proposition 1 and compare with (61). In particular,

$$\mathbb{E}(B_{n,k}) \sim \begin{cases} ne^{-t}\left(1-e^{-t}\right), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ 2^{-k}n^2, & \text{if } 2^{-k}n \to 0. \end{cases}$$

Note that these approximations can also be easily derived by the exact formula

$$\mathbb{E}(B_{n,k}) = n\left(1-2^{-k}\right)^{n-1} - n\left(1-2^{1-k}\right)^{n-1}, \tag{92}$$

by (90) or (10). But such an elementary approach becomes more messy for the calculation of the variance. Also

$$\max_k \mathbb{E}(B_{n,k}) \sim \frac{n}{4},$$

which is reached when $k \sim \log_2 n - 1$.

Expected internal profile. In a similar manner, we have, by (84),

$$\tilde{M}_k^{[I]}(z) = 2^k - (2^k + z)e^{-z/2^k} \qquad (k \geq 0).$$

Therefore, we have

$$\mathbb{E}(I_{n,k}) \sim \begin{cases} 2^k - n\left(1 - 2^{-k}\right)^{n-1}, & \text{if } 2^{-k}n \to \infty; \\ \tilde{M}_k^{[I]}(n), & \text{if } 4^{-k}n \to 0. \end{cases}$$

This implies that

$$\mathbb{E}(I_{n,k}) \sim \begin{cases} 2^k\left(1 - (1+t)e^{-t}\right), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ 2^{-k-1}n^2, & \text{if } 2^{-k}n \to 0. \end{cases}$$

Note that

$$\mathbb{E}(I_{n,k}) = 2^k\left(1 - \left(1 - 2^{-k}\right)^n\right) - n\left(1 - 2^{-k}\right)^{n-1},$$

and

$$\max_k \mathbb{E}(I_{n,k}) \sim c_3 n, \tag{93}$$

where $c_3 \approx 0.298$ denotes the maximum value achieved by the function $(1 - (1+x)e^{-x})/x$ for $x \in \mathbb{R}^+$.

Asymptotics of the variances. Similarly, by (63) and (88), we have

$$\tilde{V}_k(z) = z\left(e^{-z/2^k} - e^{-z/2^{k-1}}\right) + 2^{-k}z^2e^{-z/2^{k-1}} - 2^{1-k}z^2\left(e^{-z/2^k} - e^{-z/2^{k-1}}\right)^2,$$
$$\tilde{V}_k^{[I]}(z) = (2^k + z)e^{-z/2^k} - 2^k\left(1 + 2^{-k}\right)^2 e^{-z/2^{k-1}},$$

and, if $n/2^k \to \infty$, then

$$\mathbb{V}(B_{n,k}) \sim \mathbb{V}(I_{n,k}) \sim \mathbb{E}(B_{n,k}) \sim n\left(1 - 2^{-k}\right)^{n-1};$$

and if $n/4^k \to 0$, then

$$\mathbb{V}(B_{n,k}) \sim \tilde{V}_k(n), \quad \text{and} \quad \mathbb{V}(I_{n,k}) \sim \tilde{V}_k^{[I]}(n),$$

uniformly in k. These approximations imply that

$$\mathbb{V}(B_{n,k}) \sim \begin{cases} ne^{-t}\left(1 - (1+t)e^{-t} + 2te^{-2t}(2 - e^{-t})\right), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ 2\mathbb{E}(B_{n,k}) \sim 2^{1-k}n^2, & \text{if } 2^{-k}n \to 0. \end{cases} \tag{94}$$

and

$$\mathbb{V}(I_{n,k}) \sim \begin{cases} 2^k(1+t)e^{-t}\left(1 - (1+t)e^{-t}\right), & \text{if } 2^{-k}n \to t \in (0,\infty); \\ \mathbb{E}(I_{n,k}) \sim 2^{-k-1}n^2, & \text{if } 2^{-k}n \to 0. \end{cases}$$

Limiting distributions. Both Theorems 8 and 9 hold when $p = q = 1/2$ by the same method of proof. Note that both bivariate generating functions become simpler (see (72) and (83))

$$\mathcal{P}_k(z,y) = \left(e^{z/2^{k-1}} + (y-1)\frac{z}{2^{k-1}}\left(e^{z/2^k} - 1\right) + (y-1)^2\frac{z^2}{4^k}\right)^{2^{k-1}},$$
$$\mathcal{P}_k^{[I]}(z,y) = \left(ye^{z/2^k} + (1-y)\left(1 + \frac{z}{2^k}\right)\right)^{2^k}.$$

Observe that $\mathbb{E}(B_{n,k}) \to \infty$ iff $\mathbb{V}(B_{n,k}) \to \infty$ iff $\mathbb{V}(I_{n,k}) \to \infty$ iff

$$\frac{1}{\log 2}\left(L_n - LL_n + \frac{K_n}{L_n}\right) \le k \le \frac{2}{\log 2}(L_n - K_n), \tag{95}$$

for any sequence $K_n \to \infty$; compare Theorem 5 for asymmetric case.

Theorem 10. (i) *If k lies in the range (95), then*

$$\frac{B_{n,k} - \mathbb{E}(B_{n,k})}{\sqrt{\mathbb{V}(B_{n,k})}}, \frac{I_{n,k} - \mathbb{E}(I_{n,k})}{\sqrt{\mathbb{V}(I_{n,k})}} \xrightarrow{d} \mathcal{N}(0,1).$$

(ii) *If $k = 2(L_n + O(1))/\log 2$, then, with $\lambda_2 := 2^{-k-1}n^2$,*

$$\begin{cases} \mathbb{P}(B_{n,k} = 2m) = \dfrac{\lambda_2^m}{m!}e^{-\lambda_2} + o(1), \quad \mathbb{P}(B_{n,k} = 2m+1) = o(1), \\ \mathbb{P}(I_{n,k} = m) = \dfrac{\lambda_2^m}{m!}e^{-\lambda_2} + o(1), \end{cases}$$

uniformly for $m \ge 0$.

Note that when $p = q$, $\lambda_0 = \lambda_1 = \lambda_2$.

8 Applications of results

In this section, we briefly discuss a few properties of some shape characteristics of random tries, as either implied by our results or by our approaches. We consider only depth, height, shortest path, fill-up level, width and right-profile.

Depth. The distribution of the depth D_n is given by $\mathbb{P}(D_n = k) = \mu_{n,k}/n$. Our asymptotic approximations for $\mu_{n,k}$ give very precise results for the distribution of D_n. Consider first the case when $p \ne q$. By definition, we see that the result (4) for the limiting behaviors of $\log \mu_{n,k}/\log n$ also describes those of $-1 + \log \mathbb{P}(D_n = k)/\log n$, or essentially the large deviations of the distribution of D_n.

Furthermore, (43) can be regarded as a local limit theorem for D_n. Indeed, we have, for $k = \mathrm{h}^{-1}(L_n + x\sqrt{\mathrm{h}^{-1}\beta_2(-1)L_n})$, where $\mathrm{h} := p\log(1/p) + q\log(1/q)$ is the entropy rate,

$$\mathbb{P}(D_n = k) = G_1\left(-1; \log_{p/q} p^k n\right) \frac{e^{-x^2/2}}{\sqrt{2\pi\mathbb{V}(D_n)}}\left(1 + O\left(\frac{1+|x|^3}{\sqrt{L_n}}\right)\right), \tag{96}$$

uniformly for $x = o(L_n^{1/6})$, where $\mathbb{V}(D_n) \sim (\mathrm{h}_2 - \mathrm{h}^2)/\mathrm{h}^3 \log n$, with $\mathrm{h}_2 := p\log^2 p + q\log^2 q$, (see [34, 79]) is also rederived below in (97). Because of the appearance of the uncommon periodic function G_1, we see that D_n *satisfies a central limit theorem, but not a local limit theorem (of the usual form).* It can be showed that the right-hand side indeed sums (over k) asymptotically up to 1. The result (96) is new.

If $p = q$, then, by the exact formula (92), we have

$$\mathbb{P}(D_n = k) = \left(1 - 2^{-k}\right)^{n-1} - \left(1 - 2^{1-k}\right)^{n-1},$$

which implies that

$$\mathbb{P}(D_n = \lfloor \log_2 n \rfloor + \ell) = \left(e^{-2^{-\ell+\{\log_2 n\}}} - e^{-2^{1-\ell+\{\log_2 n\}}}\right)\left(1 + O\left(n^{-1}2^{-\ell}\right)\right),$$

uniformly for $\ell \in \mathbb{Z}$, where $\{x\}$ denotes the fractional part of x.

On the other hand, if one is interested in the cumulative distribution functions or tail probabilities, then, by (6) and by partial summation,

$$\mathbb{P}(D_n \le k) = (n-1)![z^n]\frac{e^z}{2\pi i}\int_{(\rho)} z^{-s}\Gamma(s+1)\left(p^{-s}+q^{-s}\right)^k \mathrm{d}s,$$

for $k \ge 1$, where $\rho > -1$. Equivalently, by (11), we have (see [36])

$$\mathbb{P}(D_n \le k) = \frac{1}{2\pi i}\int_{(\rho)} \frac{\Gamma(n)\Gamma(s+1)}{\Gamma(n+1+s)}\left(p^{-s}+q^{-s}\right)^k \mathrm{d}s,$$

where $\rho > -1$. Asymptotics of such integrals can be treated by our approaches, which give not only the central limit theorem of D_n with convergence rate (since there is a simple pole at $s = -1$) but also precise estimates for tail probabilities. Indeed, we have

$$\mathbb{P}\left(D_n \le \mathrm{h}^{-1}(L_n + x\sqrt{\mathrm{h}^{-1}\beta_2(-1)L_n})\right) = \Phi(x)\left(1 + O\left(\frac{1+|x|^3}{\sqrt{L_n}}\right)\right),$$

uniformly for $x = o(L_n^{1/6})$, as already shown in [34, 36]. Furthermore,

$$-\frac{\log \mathbb{P}(D_n \le \alpha L_n)}{\log n} \to \rho' + 1 - \alpha\log(p^{-\rho'}+q^{-\rho'}) \qquad (\alpha_1 \le \alpha \le \mathrm{h}^{-1}),$$

$$-\frac{\log \mathbb{P}(D_n \ge \alpha L_n)}{\log n} \to \begin{cases} \rho' + 1 - \alpha\log(p^{-\rho'}+q^{-\rho'}), & \text{if } \mathrm{h}^{-1} \le \alpha \le \alpha_2; \\ -1 - \alpha\log(p^2+q^2), & \text{if } \alpha_2 \le \alpha \le \alpha_3, \end{cases}$$

both tails being asymptotic to -1 for smaller and larger α, respectively, where ρ' is given in (42). These results imply, in particular, that $\mathbb{E}(D_n) \sim L_n/\mathrm{h}$ and

$$\mathbb{V}(D_n) \sim \beta_2(-1)\mathrm{h}^{-3}L_n = \frac{pq\log^2(p/q)}{(p\log(1/p)+q\log(1/q))^3}L_n = \frac{\mathrm{h}_2 - \mathrm{h}^2}{\mathrm{h}^3}L_n \tag{97}$$

where $\mathrm{h}_2 := p\log^2 p + q\log^2 q$; see [13, 79]. Note that the constant on the right-hand side becomes zero when $p = q$.

Width. The width of tries W_n is defined to be $W_n := \max_k I_{n,k}$, or the size of the most abundant level(s). As a natural lower bound for $\mathbb{E}(W_n)$, we consider $\max_k \mathbb{E}(I_{n,k})$. By (87) and a similar analysis for (43), we have, when $p \ne q$,

$$\mathbb{E}(I_{n,k}) = \frac{\sqrt{\mathrm{h}}\,G_3\left(-1; \log_{p/q} p^k n\right)}{\log(p/q)\sqrt{2\pi pq}} \times \frac{n}{\sqrt{L_n}}\left(1 + O(L_n^{-1/2})\right),$$

uniformly for $k = L_n/\mathrm{h} + O(1)$. This approximation, together with the estimates for $\mathbb{E}(I_{n,k})$ in other ranges given in Section 6.1, yields

$$\mathbb{E}(W_n) \ge \max_k \mathbb{E}(I_{n,k}) = \Theta(nL_n^{-1/2}),$$

when $p \ne q$. Indeed, we have

$$\mathbb{E}(W_n) = \Theta(nL_n^{-1/2}).$$

The upper bound can be proved by applying the arguments used in [16], which start from the inequality

$$\mathbb{E}(W_n) \le \max_k \mathbb{E}(I_{n,k}) + \sum_{|k-L_n/\mathrm{h}|\le \varepsilon L_n^{2/3}} \frac{\mathbb{V}(I_{n,k})}{\mathcal{M}_n - \mathbb{E}(I_{n,k})} + \sum_{|k-L_n/\mathrm{h}|>\varepsilon L_n^{2/3}} \mathbb{E}(I_{n,k}),$$

and then use the asymptotics of $\mathbb{E}(I_{n,k})$ and $\mathbb{V}(I_{n,k})$ given in Section 6.1. Details are omitted here. Finer results for $\mathbb{E}(W_n)$ can be derived, but the proof is more involved due to the presence of the periodic function G_3 (whose parameter involving k).

For symmetric tries, we have easily, by (93) and the trivial bound $\mathbb{E}(W_n) \le n$,

$$\mathbb{E}(W_n) = \Theta(n).$$

Thus *random symmetric tries are "fatter" and most nodes lie near the most abundant levels* $k = \log_2 n + O(1)$.

Height. We next derive an estimate for the height of random tries, as a consequence of our estimates for the external profiles together with the use of the first and second moment methods (see [82]).

Corollary 6. ***(Height of a trie)*** *Let* $H_n := \max\{k : B_{n,k} > 0\}$ *be the height of a random trie. Then* $H_n/\log n \to \alpha_3$ *in probability.*

Proof. Let $k_H := \alpha_3 L_n$. First we derive an upper bound for H_n as follows.

$$\begin{aligned}\mathbb{P}(H_n > (1+\varepsilon)k_H) &\le \mathbb{P}(B_{n,k} \ge 1), \text{ for some } k \ge (1+\varepsilon)k_H \\ &\le \mathbb{E}(B_{n,k}) \to 0,\end{aligned}$$

where the last inequality follows from Theorem 3 when $p \ne q$ and (91) when $p = q$. For the lower bound, we use the second moment method (see [82]) to find

$$\begin{aligned}\mathbb{P}(H_n < (1-\varepsilon)k_H) &\le \mathbb{P}(B_{n,\lceil (1-\varepsilon)k_H\rceil} = 0) \\ &\le \frac{\mathbb{V}(B_{n,\lceil (1-\varepsilon)k_H\rceil})}{(\mathbb{E}(B_{n,\lceil (1-\varepsilon)k_H\rceil}))^2} \\ &= O\left(\frac{1}{\mathbb{E}(B_{n,\lceil (1-\varepsilon)k_H\rceil})}\right) \to 0,\end{aligned}$$

by Theorems 3 and 7 and (94). Combining the two estimates, we obtain the required result. □

Corollary 6 is not new and it was already derived in Devroye [12], Pittel [66, 67] and Szpankowski [80].

Shortest path. The shortest path $S_n := \min\{j : B_{n,j} > 0\}$ of a random trie, discussed next, attracted much less attention than the height (see [82]) in the literature. It is closely related to the behaviors of the external profile in Range (I) near $k = \alpha_1(L_n - LLL_n + O(1))$ as discussed in Theorem 1 and its refinement in Corollary 3.

Define

$$\hat{k} := \begin{cases} \alpha_1\left(L_n - LLL_n - \log m_0 + m_0 \log(p/q) - \dfrac{LLL_n}{m_0 LL_n}\right), & \text{if } p \ne q; \\ \alpha_1(L_n - LL_n), & \text{if } p = q, \end{cases}$$

where $m_0 := \lceil 1/(p/q-1)\rceil$, and

$$k_S := \begin{cases} \lceil \hat{k}\rceil, & \text{if } p \ne q; \\ \lfloor \hat{k}\rfloor, & \text{if } p = q. \end{cases}$$

Corollary 7. ***(Shortest Path Length of Tries)*** *If $p \neq q$, then*

$$S_n = \begin{cases} k_S, & \text{if } \{\hat{k}\}LL_n \to \infty; \\ k_S \text{ or } k_S - 1, & \text{if } \{\hat{k}\}LL_n = O(1); \end{cases}$$

with high probability[2]*; if $p = q = 1/2$, then*

$$S_n = \begin{cases} k_S + 1, & \text{if } \{\hat{k}\}L_n \to \infty; \\ k_S \text{ or } k_S + 1, & \text{if } \{\hat{k}\}L_n = O(1), \end{cases}$$

with high probability.

Proof. Assume $p \neq q$. Consider first the case $\{\hat{k}\}LL_n \to \infty$. In this case, we have, by Corollary 3,

$$\begin{cases} \mu_{n,k_S} \to \infty, \\ \mu_{n,k} \to 0 \text{ for } k \leq k_S - 1. \end{cases}$$

Thus, again by the second moment method,

$$\mathbb{P}(S_n > k_S) \leq \mathbb{P}(B_{n,k_S} = 0) \leq \frac{\mathbb{V}(B_{n,k_S})}{(\mathbb{E}(B_{n,k_S}))^2} = O\left(\frac{1}{\mu_{n,k_S}}\right) \to 0.$$

On the other hand, by using the first moment method, we have

$$\begin{aligned} \mathbb{P}(S_n < k_S) &\leq \mathbb{P}(B_{n,k} \geq 1), \text{ for some } k < k_S \\ &\leq \mu_{n,k} \to 0. \end{aligned}$$

These two estimates imply that $\mathbb{P}(S_n = k_S) \to 1$.

Now if $\{\hat{k}\}LL_n = O(1)$, then, again by Corollary 3,

$$\begin{cases} \mu_{n,k_S} \to \infty, \\ \mu_{n,k_S-1} = \Theta(1), \\ \mu_{n,k} \to 0 \text{ for } k \leq k_S - 2. \end{cases}$$

Thus applying *mutatis mutandis* the same proof gives

$$\mathbb{P}(S_n = k_S) + \mathbb{P}(S_n = k_S - 1) \to 1.$$

The proof for the symmetric case is similar; for, $\mu_{n,k} \to \infty$ when k lies in the range (95) and from this result we deduce that $\mu_{n,k_S+1} \to \infty$, $\mu_{n,k_S-1} \to 0$ and

$$\mu_{n,k_S} \begin{cases} \to 0, & \text{if } \{\hat{k}\}L_n \to \infty; \\ = \Theta(1), & \text{if } \{\hat{k}\}L_n = O(1). \end{cases}$$

This completes the proof. □

[2] We say that an event holds *with high probability*, if it holds with probability tending to 1 as $n \to \infty$.

Fill-up level. We now consider the fill-up level $F_n = \max\{k : I_{n,k} = 2^k\}$ of a random trie, which was also analyzed previously by Devroye [12], Pittel [66, 67] and Knessl and Szpankowski [50].

Corollary 8. ***(Fill-up level of a trie)*** *If $p \neq q$, then*

$$F_n = \begin{cases} k_S - 1, & \text{if } \{\hat{k}\}LL_n \to \infty; \\ k_S - 2 \text{ or } k_S - 1, & \text{if } \{\hat{k}\}LL_n = O(1); \end{cases}$$

with high probability; if $p = q = 1/2$, then

$$F_n = \begin{cases} k_S, & \text{if } \{\hat{k}\}L_n \to \infty; \\ k_S \text{ or } k_S - 1, & \text{if } \{\hat{k}\}L_n = O(1); \end{cases}$$

Proof. Observe that

$$F_n = \max\{k : \bar{I}_{n,k} = 0\} = \min\{k : \bar{I}_{n,k} > 0\} - 1.$$

By (86), we have $\mathbb{E}(\bar{I}_{n,k}) \sim \mu_{n,k}$ when $k \leq \alpha_1(1 + o(1))L_n$. Thus the proof of Corollary 7 applies with little modification. □

Profile enumerating only right branches. We consider the random variable $R_{n,k}$, which denotes the number of external nodes in random tries that are away from the root by k right branches. Since a right branch means a "1" in the input string, $R_{n,k}$ enumerates the number of strings with exactly k 1's; it also has other concrete interpretations in splitting processes and conflict resolution algorithms. All our tools can be extended to $R_{n,k}$, although $R_{n,k}$ exhibits very different behaviors. For example, unlike $B_{n,k}$ or $I_{n,k}$, there is no need to distinguish between symmetric and asymmetric tries, all results being uniform in p; also the Poisson heuristic holds for all $k \geq 0$. This example further reveals the power of our approaches.

The probability generating function $F_{n,k}(y) := \mathbb{E}(y^{R_{n,k}})$ of $R_{n,k}$ satisfies the recurrence

$$F_{n,k}(y) = \sum_{0 \leq j \leq n} \binom{n}{j} p^j q^{n-j} F_{j,k-1}(y) F_{n-j,k}(y) \qquad (n \geq 2; k \geq 0),$$

with the initial conditions $F_{n,k}(y) = 1$ for $n \leq 1$ or $k < 0$ and $F_{2,1}(y) = y$. Thus the bivariate generating function $\mathcal{F}_k(z, y) := \sum_n F_{n,k}(y) z^n/n!$ satisfies

$$\mathcal{F}_k(z, y) = \mathcal{F}_k(qz, y)\mathcal{F}_{k-1}(pz, y) = \prod_{j \geq 0} \mathcal{F}_0(p^k q^j z, y)^{\binom{k+j-1}{j}},$$

where

$$\mathcal{F}_0(z, y) = e^{pz}\mathcal{F}_0(qz, y) + p(1 - p/2)(y - 1)z^2,$$

which is further solved to be

$$\mathcal{F}_0(z, y) = e^z + p(1 - p/2)(y - 1) \sum_{j \geq 0} (q^j z)^2 e^{(1-q^j)z}. \tag{98}$$

From this we deduce that the expected right-profile is given by

$$\mathbb{E}(R_{n,k}) = p(1 - p/2)n![z^n] \frac{e^z}{2\pi i} \int_{(\rho)} z^{-s} \Gamma(s + 2) \frac{p^{-ks}}{(1 - q^{-s})^{k+1}} \mathrm{d}s,$$

where $-2 < \rho < 0$. The integral is not of the same type as (6) but similar, and our methods of proof easily extend. It has simple poles at $s = -2, -3, \ldots$ and poles of order $k + 1$ at $s = 2j\pi i/\log(1/q)$, $j \in \mathbb{Z}$. Thus the asymptotics of $\mathbb{E}(R_{n,k})$ is divided into four overlapping ranges.

- If $0 \le k = o(\log n)$, then the residues of the poles on the imaginary lines are dominant and we have

$$\mathbb{E}(R_{n,k}) \sim p(1-p/2)\frac{(\log p^k n)^k}{k!(\log(1/q))^{k+1}}\left(1+\sum_{j\neq 0}\Gamma(1+\chi_j)(p^k n)^{-\chi_j}\right),$$

  uniformly in k, where $\chi_j := 2j\pi i/\log(1/q)$.

- If $k \to \infty$ and $k \le \alpha^*(L_n - K_n\sqrt{L_n})$, where $K_n \to \infty$ and

$$\alpha^* := \frac{1-q^2}{(1-q^2)\log(1/p) - q^2\log(1/q)},$$

  then by the saddle-point method

$$\mathbb{E}(R_{n,k}) \sim \frac{p(1-p/2)q^{\rho/2}}{\sqrt{2\pi k}\log(1/q)}(p^k n)^{-\rho}(1-q^{-\rho})^{-k}\sum_{j\in\mathbb{Z}}\Gamma(\rho+1+\chi_j)(p^k n)^{-\chi_j},$$

  uniformly in k, where

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

- If $k = \alpha^* L_n + x\sqrt{\alpha^*(1+\alpha^*\log(p/q))(1+\alpha^*\log p)L_n}$, then

$$\mathbb{E}(R_{n,k}) \sim \frac{1}{2}\Phi(x)(p^k n)^2(1-q^2)^{-k},$$

  uniformly for $x = o(L_n^{1/6})$.

- If $k \ge \alpha^* L_n + K_n\sqrt{\alpha^*(1+\alpha^*\log(p/q))(1+\alpha\log p)L_n}$, then

$$\mathbb{E}(R_{n,k}) \sim \frac{1}{2}(p^k n)^2(1-q^2)^{-k}.$$

These results imply that $\mathbb{E}(R_{n,k}) \to \infty$ iff

$$1 \le k \le \frac{2}{\log\frac{2-p}{p}}L_n - K_n,$$

where $K_n \to \infty$. Note that

$$\log e^{-z}\mathcal{F}_0(z,y) = \log(1+(y-1)\tau(z)),$$

where $\tau(z) := p(1-p/2)\sum_{j\ge 0}(q^j z)^2 e^{-q^j z}$ satisfies $\tau(z) = O(|z|^2)$ as $z \to 0$, and, by Mellin transform, $\tau(z) = O(1)$ as $|z| \to \infty$ in a small sector containing the real axis. This implies, by a straightforward modification of our approaches, that $\mathbb{V}(R_{n,k}) = \Theta(\mathbb{E}(R_{n,k}))$ for all $k = k(n) \ge 0$ and that

$$\frac{R_{n,k} - \mathbb{E}(R_{n,k})}{\sqrt{\mathbb{V}(R_{n,k})}} \xrightarrow{d} \mathcal{N}(0,1),$$

whenever $\mathbb{E}(R_{n,k})$ or $\mathbb{V}(R_{n,k}) \to \infty$. Two remaining cases are $k = 0$ and $k = 2L_n/\log\frac{2-p}{p} + O(1)$. In the first case, $R_{n,0}$ by (98) is Bernoulli distributed with mean equal to $\tau(n)$, which is asymptotic to the periodic function

$$\frac{1}{\log(1/q)}\left(1+\sum_{j\neq 0}\Gamma(2-\chi_j)n^{-\chi_j}\right);$$

and in the second case,

$$\mathbb{P}(R_{n,k} = m) = \frac{\lambda_3^m}{m!}e^{-\lambda_3} + o(1),$$

where $\lambda_3 := (p^k n)^2(1-q^2)^{-k}/2$.

Appendix A: Proof of Lemma 3

In this appendix, we prove Lemma 3. For part (i) let $z = ne^{i\theta}$, where $\theta = o(LL_n^{-1/2})$. By (8)

$$\begin{aligned}\tilde{M}_k(z) &= \sum_{0\le j<k} \binom{k-1}{j} p^j q^{k-j} z e^{-p^j q^{k-j} z} \left(1 + O\left(e^{-(p-q)p^j q^{k-1-j} n\cos\theta}\right)\right) \\ &= \sum_{0\le j<k} \binom{k-1}{j} p^j q^{k-j} z e^{-p^j q^{k-j} z} \left(1 + O\left(e^{-(p-q)q^{k-1} n\cos\theta}\right)\right) \qquad (99) \\ &= q^k z e^{-q^k z}\left(1 + O(E_6)\right),\end{aligned}$$

where

$$E_6 := \sum_{1\le j<k} \frac{(p\alpha_1/q)^j}{j!} L_n^j e^{-q^k n((p/q)^j - 1)\cos\theta}.$$

Since $1 \le k \le k_0 - \alpha_1 K_n/LL_n$, we have

$$q^k n \ge \frac{LL_n}{p/q - 1}\, e^{K_n/LL_n}.$$

It follows, by using the inequality

$$\frac{t^j - 1}{t - 1} \ge j \qquad (t > 1; j \ge 1),$$

that

$$\begin{aligned}E_6 &\le \sum_{j\ge1} \frac{(p\alpha_1/q)^j}{j!} L_n^{j - \frac{(p/q)^j - 1}{p/q - 1}\cos(\theta) e^{K_n/LL_n}} \\ &\le \sum_{j\ge1} \frac{(p\alpha_1/q)^j}{j!} L_n^{-j(\cos(\theta)e^{K_n/LL_n} - 1)} \\ &\le \sum_{j\ge1} \frac{(p\alpha_1/q)^j}{j!} e^{-jK + O(j\theta^2 LL_n)} \\ &= O(e^{-K_n}),\end{aligned}$$

since $\theta = o(LL_n^{-1/2})$. This proves (30).

For part (ii), by (99),

$$\tilde{M}_k(z) = S_{k,m}(z)\left(1 + O\left(e^{-(p-q)p^m q^{k-1-m} n\cos\theta} + E_7 + E_8\right)\right),$$

where

$$\begin{aligned}E_7 &:= \sum_{0\le j<m} \left|\frac{S_{k,j}(z)}{S_{k,m}(z)}\right| \left(1 + O\left(e^{-(p-q)p^j q^{k-1-j} n\cos\theta}\right)\right), \\ E_8 &:= \sum_{m<j<k} \left|\frac{S_{k,j}(z)}{S_{k,m}(z)}\right| \left(1 + O\left(e^{-(p-q)p^j q^{k-1-j} n\cos\theta}\right)\right).\end{aligned}$$

By (31), $p^m q^{k-m} n = e^\eta LL_n/(p/q-1)$. It follows, by changing j to $m-j$, that

$$\begin{aligned}
E_7 &= O\left(m! \sum_{1\le j\le m} \frac{(q/p)^j}{(m-j)!}\, k^{-j} \exp\left(p^m q^{k-m} n\cos(\theta)\left(1-(q/p)^j\right)\right)\right)\\
&= O\left(m! \sum_{1\le j\le m} \frac{(q/p\alpha_1)^j}{(m-j)!} L_n^{-j+\frac{1-(q/p)^j}{p/q-1}\, e^\eta\cos\theta}\right)\\
&= O\left(m! \sum_{1\le j\le m} \frac{(q/p\alpha_1)^j}{(m-j)!} L_n^{-j(1-qe^\eta\cos\theta/p)}\right)\\
&= O\left(m_n^{-(1-qe^\eta\cos\theta/p)}\right) = O\left(me^{-K_n}\right),
\end{aligned}$$

since $\eta \le \log(p/q) - K_n/LL_n$ and $\theta = o(LL_n^{-1/2})$.

Similarly,

$$\begin{aligned}
E_8 &= O\left(m! \sum_{j>m} \frac{(p/q)^{j-m}}{j!}\, k^{j-m} \exp\left(-p^m q^{k-m} n\cos(\theta)\left((p/q)^{j-m}-1\right)\right)\right)\\
&= O\left(m! \sum_{j\ge 1} \frac{(p\alpha_1/q)^j}{(j+m)!} L_n^{j-\frac{(p/q)^j-1}{p/q-1}\, e^\eta\cos\theta}\right)\\
&= O\left(m! \sum_{j\ge 1} \frac{(p\alpha_1/q)^j}{(j+m)!} L_n^{-j(e^\eta\cos\theta-1)}\right)\\
&= O\left(m^{-1} L_n^{-(e^\eta\cos\theta-1)}\right) = O\left(m^{-1}e^{-K_n}\right),
\end{aligned}$$

since $\eta \ge K_n/LL_n$. This completes the proof of (32). □

Appendix B: Well-definedness of $Q_k(z, y)$

We prove here the following lemma that is needed for the proof of Proposition 4.

Lemma 7. *The function $Q_k(re^{i\theta}, y)$ is well-defined for $r \ge 0$, $|\theta| \le \varepsilon$ and $|y| = 1$.*

Proof. We first show that

$$|a_3(r)(e^{i\varphi}-1) + a_4(r)(e^{i\varphi}-1)^2| < 1, \tag{100}$$

for $r \ge 0$ and $|y| = 1$. By direct calculation, we have

$$|a_3(r)(e^{i\varphi}-1) + a_4(r)(e^{i\varphi}-1)^2|^2 = a_3(r)^2\upsilon - a_4(r)(a_3(r)-a_4(r))\upsilon^2,$$

where $\upsilon := 2(1-\cos\varphi)$. Since

$$a_3(r) - a_4(r) \ge a_3(r) - 2a_4(r) = e^{-r}\left(pr(e^{qr}-1-qr) + qr(e^{pr}-1-pr)\right) \ge 0,$$

we have

$$|a_3(r)(e^{i\varphi}-1) + a_4(r)(e^{i\varphi}-1)^2| \le \sqrt{2}\, a_3(r).$$

By simple calculus, we have $a_3(r) < 2^{-1/2}$, which implies (100). Indeed, the inequality $a_3(r) < 2^{-1/2}$ is equivalent to

$$pre^{-pr} + qre^{-qr} - re^{-r} < 2^{-1/2} \qquad (r \geq 0),$$

and we have

$$pre^{-pr} + qre^{-qr} - re^{-r} \leq \max_{r\geq 0} re^{-r}(e^{r/2} - 1) \approx 0.52 < 2^{-1/2}.$$

This proves the lemma when $z = r$; the assertion of the lemma follows from analyticity. □

Appendix C: A useful approximation

In the proof of Proposition 4 we need the following lemma.

Lemma 8. *Let $f(z)$ be an entire function satisfying*

$$f(z) = \begin{cases} O(z^2), & \text{as } z \to 0; \\ O(|z|e^{-q\Re(z)}), & \text{as } z \to \infty, |\arg(z)| \leq \varepsilon. \end{cases} \tag{101}$$

Then uniformly for all $k = k(n) \geq 1$ and $z = ne^{i\theta}$, $|\theta| \leq \varepsilon$,

$$f_k(z) := \sum_{0\leq j<k} \binom{k-1}{j} f(p^j q^{k-1-j} z) = \Theta(|\tilde{M}_k(z)|).$$

Proof. If $1 \leq k \leq k_0$, then it is easy to see that $f_k(z) = \Theta(|\tilde{M}_k(z)|)$ for $|\theta| \leq \varepsilon$. When $k \geq k_0$, let $f^*(s) := \int_0^\infty x^{s-1} f(x)\mathrm{d}x$. Then $f^*(s)$ is well-defined in the half-plane $\Re(s) > -2$ by (101). By the estimates in (101) and the same argument used in [24, Proposition 5], we have, assuming $\rho \geq 1$ and $t > 0$,

$$\begin{aligned} f^*(\rho + it) &= \int_0^{e^{i\varepsilon}\infty} x^{\rho+it-1} f(x)\mathrm{d}x \\ &= e^{i\varepsilon(\rho+it)} \int_0^\infty x^{\rho+it-1} f(xe^{i\varepsilon})\mathrm{d}x \\ &= O(e^{-\varepsilon t} \int_0^1 x^{\rho+1}\mathrm{d}x) + O\left(e^{-\varepsilon t} \int_1^\infty x^\rho e^{-qx\cos\varepsilon}\mathrm{d}x\right) \\ &= O\left(e^{-\varepsilon t}\rho^{-1} + e^{-\varepsilon t} q^{-\rho}\rho^{1/2}(\rho/e)^\rho\right), \end{aligned}$$

uniformly in ρ and t. If $t < 0$, then changing $e^{i\varepsilon}$ to $e^{-i\varepsilon}$ gives

$$f^*(\rho + it) = O\left(e^{-\varepsilon|t|}\rho^{-1} + e^{-\varepsilon|t|} q^{-\rho}\rho^{1/2}(\rho/e)^\rho\right).$$

When $-2 < \rho \leq 1$, $f^*(\rho + it) = O(e^{-\varepsilon|t|})$ for large $|t|$ by the same argument. On the other hand, by the first estimate in (101), we also have

$$f^*(s) = O\left(|s + 2|^{-1}\right) \qquad (s \to 2).$$

With these estimates and the Mellin inversion integral

$$f_k(z) = \frac{1}{2\pi i} \int_{(\rho)} z^{-s} f^*(s)(p^{-s} + q^{-s})^{k-1}\mathrm{d}s,$$

we can apply the arguments used for $\tilde{M}_k(z)$ and prove (101). □

References

[1] R. Aguech, N. Lasmar and H. Mahmoud, Distribution of inter-node distances in digital trees, in *2005 International Conference on Analysis of Algorithms*, C. Martínez (ed.), *Discrete Mathematics and Theoretical Computer Science*, Proceedings AD, pp. 1–10, 2005.

[2] D. Aldous and P. Shields, A diffusion limit for a class of randomly-growing binary trees, *Probability Theory and Related Fields*, **79** (1988) 509–542.

[3] B. C. Berndt, *Ramanujan's Notebooks. Part I*, Springer Verlag, New-York, 1965.

[4] J. Bourdon, M. Nebel and B. Valleé, On the stack-size of general tries, *Theoretical Informatics and Applications*, **35** (2001) 163–185.

[5] B. Chauvin, M. Drmota, and J. Jabbour-Hattab, The profile of binary search trees, *Annals of Applied Probability*, **11** (2001) 1042–1062.

[6] J.-D. Choi, K. Lee, A. Loginov, R. O'Callahan, V. Sarkar and M. Sridharan, Efficient and precise datarace detection for multithreaded object-oriented programs, in *Proceedings of the 2002 ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, 2002, pp. 258–269.

[7] C. A. Christophi and H. M. Mahmoud, The oscillatory distribution of distances in random tries, *Annals of Applied Probability*, **15** (2005), 1536–1564.

[8] J. Clément, P. Flajolet and B. Vallée, Dynamical sources in information theory: a general analysis of trie structures, *Algorithmica*, **29** (2001) 307–369.

[9] R. de la Briandais, File searching using variable length keys, in *Proceedings of the AFIPS Spring Joint Computer Conference*. AFIPS Press, Reston, Va., (1959), pp. 295–298.

[10] L. Devroye, A note on the average depth of tries, *Computing*, **28** (1982), 367–371.

[11] L. Devroye, A probabilistic analysis of the height of tries and of the complexity of triesort, *Acta Informatica*, **21** (1984), 229–237.

[12] L. Devroye, A study of trie-like structures under the density model, *Annals of Applied Probability*, **2** (1992) 402–434.

[13] L. Devroye, Universal limit laws for depths in random trees, *SIAM Journal on Computing* **28** (1999) 409–432.

[14] L. Devroye, Laws of large numbers and tail inequalities for random tries and PATRICIA trees, *Journal of Computational and Applied Mathematics* **142** (2002), 27–37.

[15] L. Devroye, Universal asymptotics for random tries and PATRICIA trees, *Algorithmica*, **42** (2005) 11–29.

[16] L. Devroye and H.-K. Hwang, Width and mode of the profile for some random trees of logarithmic height, *Annals of Applied Probability*, **16** (2006), 886–918.

[17] L. Devroye and P. Kruszewski, On the Horton-Strahler number for random tries, *RAIRO Informatique Théorique et Applications*, **30** (1996) 443–456.

[18] M. Drmota, Profile and height of random binary search trees, *Journal of the Iranian Statistical Society*, **3** (2004) 117–138.

[19] M. Drmota and H.-K. Hwang, Bimodality and phase transitions in the profile variance of random binary search trees, *SIAM Journal on Discrete Math.*, **19** (2005) 19–45.

[20] M. Drmota and H.-K. Hwang, Profile of random trees: correlation and width of random recursive trees and binary search trees, *Advances in Applied Probability*, **37** (2005) 321–341.

[21] J. Fayolle and M. D. Ward, Analysis of the average depth in a suffix tree under a Markov model, in *2005 International Conference on Analysis of Algorithms*, C. Martínez (ed.), *Discrete Mathematics and Theoretical Computer Science*, Proceedings AD, pp. 95–104, 2005.

[22] J. A. Fill, H. M. Mahmoud and W. Szpankowski, On the distribution for the duration of a randomized leader election algorithm, *Annals of Applied Probability* **6** (1996) 1260–1283.

[23] P. Flajolet, On the performance evaluation of extendible hashing and trie searching, *Acta Informatica*, **20** (1983) 345–369.

[24] P. Flajolet, X. Gourdon, and P. Dumas, Mellin transforms and asymptotics: harmonic sums, *Theoretical Computer Science* **144** (1995) 3–58.

[25] P. Flajolet, M. Régnier and D. Sotteau, Algebraic methods for trie statistics, in *Analysis and Design of Algorithms for Combinatorial Problems* (Udine, 1982), 145–188, North-Holland Math. Stud., 109, North-Holland, Amsterdam, 1985.

[26] P. Flajolet and R. Sedgewick, Mellin transforms and asymptotics: finite differences and Rice's integrals, *Theoretical Computer Science* **144** (1995) 101–124.

[27] P. Flajolet and J.-M. Steyaert, A branching process arising in dynamic hashing, trie searching and polynomial factorization, in *Lecture Notes in Computer Science*, **140** (1982) 239–251.

[28] E. Fredkin, Trie memory, *Communications of the ACM*, **3** (1960) 490–499.

[29] M. Fuchs, H.-K. Hwang and R. Neininger, Profiles of random trees: Limit theorems for random recursive trees and binary search trees, *Algorithmica*, accepted for publication (2005).

[30] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, Cambridge (1997).

[31] H.-K. Hwang, Asymptotic expansions for the Stirling numbers of the first kind. *Journal of Combinatorial Theory. Series A* **71** (1995) 343–351.

[32] H.-K. Hwang, Profiles of random trees: plane-oriented recursive trees, preprint submitted for publication (2005).

[33] H.-K. Hwang, Local limit theorems for the profiles of random tries, preprint (2006).

[34] P. Jacquet and M. Régnier, Trie partitioning process: limiting distributions, in *Lecture Notes in Computer Science*, **214** (1986) 196–210.

[35] P. Jacquet and M. Reǵnier, Normal limiting distribution of the size of tries, in *Performance'87* (Brussels, 1987), pp. 209–223, North-Holland, Amsterdam, 1988.

[36] P. Jacquet, and W. Szpankowski, Analysis of digital tries with Markovian dependency, *IEEE Transactions on Information Theory*, **37** (1991) 1470–1475.

[37] P. Jacquet, and W. Szpankowski, Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach, *Journal of Combinatorial Theory, Series A*, **66** (1994) 237–269.

[38] P. Jacquet and W. Szpankowski, Analytical depoissonization and its applications, *Theoretical Computer Science*, **201** (1998) 1–62.

[39] S. Janson and W. Szpankowski, Analysis of an asymmetric leader election algorithm, *Electronic Journal of Combinatorics*, **64** R17 (1997) 1–6.

[40] P. Jacquet, W. Szpankowski and J. Tang, Average profile of the Lempel-Ziv parsing scheme for a Markovian source, *Algorithmica*, **31** (2001) 318–360.

[41] P. Kirschenhofer and H. Prodinger, Some further results on digital search trees, in *Lecture Notes in Computer Science*, **226** (1986) 177–185.

[42] P. Kirschenhofer and H. Prodinger, b-tries: a paradigm for the use of number-theoretic methods in the analysis of algorithms, in *Contributions to General Algebra*, Volume 6, 141–154, Hölder-Pichler-Tempsky, Vienna, 1988.

[43] P. Kirschenhofer and H. Prodinger, Further results on digital search trees, *Theoretical Computer Science*, **58** (1988) 143–154.

[44] P. Kirschenhofer, H. Prodinger, and W. Szpankowski, On the variance of the external path in a symmetric digital trie, *Discrete Applied Mathematics*, **25** (1989) 129–143.

[45] P. Kirschenhofer, H. Prodinger, and W. Szpankowski, Analysis of a splitting process arising in probabilistic counting and other related algorithms, *Random Structures and Algorithms*, **9** (1996) 379–401.

[46] P. Kirschenhofer and H. Prodinger, On some applications of formulae of Ramanujan in the analysis of algorithms, *Mathematika* **38** (1991) 14–33.

[47] C. Knessl, A note on the asymptotic behavior of the depth of tries, *Algorithmica*, **22** (1998) 547–560.

[48] C. Knessl and W. Szpankowski, A note on the asymptotic behavior of the heights in b-tries for b large, *Electronic Journal of Combinatorics*, **7** (2000), R39, 16 pp.

[49] C. Knessl and W. Szpankowski, Limit laws for the height in PATRICIA tries, *Journal of Algorithms*, **44** (2002) 63–97.

[50] C. Knessl and W. Szpankowski, On the number of full levels in tries, *Random Structures and Algorithms*, **25** (2004) 247–276.

[51] D. E. Knuth, *The Art of Computer Programming, Volume III: Sorting and Searching*, Second edition, Addison Wesley, Reading, MA, 1998.

[52] D. E. Knuth, *Selected Papers on Analysis of Algorithms*, CSLI, Stanford, 2000.

[53] K. Kukich, Techniques for automatically correcting words in text, *ACM Computing Surveys*, **24** (1992) 377–439.

[54] G. Louchard, Trie size in a dynamic list structure, *Random Structures and Algorithms*, **5** (1994) 665–702.

[55] H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

[56] M. E. Nebel, On the Horton-Strahler number for combinatorial tries, *Informatique Théorique et Applications*, **34** (2000) 279–296.

[57] M. E. Nebel, The stack-size of combinatorial tries revisited, *Discrete Mathematics and Theoretical Computer Science*, **5** (2002) 1–16.

[58] M. E. Nebel, The stack-size of tries: a combinatorial study, *Theoretical Computer Science*, **270** (2002) 441–461.

[59] R. Neininger and L. Rüschendorf, A general limit theorem for recursive algorithms and combinatorial structures. *Annals of Applied Probability*, **14** (2004) 378–418.

[60] M. Nguyen-The, Distribution de valuations sur les arbres, Ph.D. Thesis, LIX, Ecole polytechnique, 2003.

[61] P. Nicodème, Average profiles, from tries to suffix-trees, in *2005 International Conference on Analysis of Algorithms*, C. Martínez (ed.), *Discrete Mathematics and Theoretical Computer Science*, Proceedings AD, pp. 257–266, 2005.

[62] S. Nilsson and M. Tikkanen, An experimental study of compression methods for dynamic tries, *Algorithmica*, **33** (2002) 19–33.

[63] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York-London, 1974.

[64] G. Park and W. Szpankowski, Towards a complete characterization of tries, *SIAM-ACM Symposium on Discrete Algorithms*, 33-42, Vancouver, 2005.

[65] G. Park, Profile of Tries, Ph.D. Thesis, Purdue University, 2006.

[66] B. Pittel, Asymptotic growth of a class of random trees, *Annals of Probability*, **18** (1985) 414–427.

[67] B. Pittel, Paths in a random digital tree: limiting distributions, *Advances in Applied Probability*, **18** (1986) 139-155.

[68] H. Prodinger, How to select a loser, *Discrete Mathematics*, **120** (1993) 149–159.

[69] S. T. Rachev and L. Rüschendorf, Probability metrics and recursive algorithms, *Advances in Applied Probability*, **27** (1995) 770–799.

[70] M. Régnier and P. Jacquet, New results on the size of tries, *IEEE Transactions on Information Theory*, **35** (1989) 203–205.

[71] Y. Reznik, Some results on tries with adaptive branching, *Theoretical Computer Science*, **289** (2002) 1009–1026.

[72] S. Ristov and E. Laporte, Ziv Lempel compression of huge natural language data tries using suffix arrays, *Journal of Discrete Algorithms*, **1** (2000) 241–256.

[73] W. Schachinger, On the variance of a class of inductive valuations of data structures for digital search, *Theoretical Computer Science*, **144** (1995), 251–275.

[74] W. Schachinger, Asymptotic normality of recursive algorithms via martingale difference arrays, *Discrete Mathematics and Theoretical Computer Science*, **4** (2001) 363–397.

[75] W. Schachinger, Concentration of size and path length of tries, *Combinatorics, Probability and Computing*, **13** (2004) 763–793.

[76] R. Sedgewick and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1996.

[77] V. Srinivasan and G. Varghese, Fast address lookups using controlled prefix expansions, *ACM Transactions on Computer Systems*, **17** (1999) 1–40.

[78] W. Szpankowski, Average complexity of additive properties for multiway tries: a unified approach, in *Lecture Notes in Computer Science*, **249** (1987) 13–25.

[79] W. Szpankowski, Some results on V-ary asymmetric tries, *Journal of Algorithms*, **9** (1988) 224–244.

[80] W. Szpankowski, On the height of digital trees and related problems, *Algorithmica*, **6** (1991) 256–277.

[81] W. Szpankowski, A generalized suffix tree and its (un)expected asymptotic behaviors, *SIAM J. Computing*, **22** (1993) 1176–1198.

[82] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.

[83] B. W. Watson, *Taxonomies and Toolkits of Regular Language Algorithms*, Ph. D. Thesis, Department of Mathematics and Computer Science, Eindhoven University of Technology, 1995.

[84] M. D. Ward and W. Szpankowski, Analysis of a randomized selection algorithm motivated by the LZ'77 scheme, in *The First Workshop on Analytic Algorithmics and Combinatorics (ANALCO 04)*, New Orleans, 2004.

[85] M. D. Ward and W. Szpankowski, Analysis of the multiplicity matching parameter in suffix trees, in *2005 International Conference on Analysis of Algorithms*, C. Martínez (ed.), *Discrete Mathematics and Theoretical Computer Science*, Proceedings AD, pp. 307–322, 2005.

[86] R. Wong, Asymptotic Approximations of Integrals, Corrected reprint of the 1989 original, SIAM, Philadelphia, PA, 2001.