On the Number of Full Levels in Tries*

February 7, 2004

Charles Knessl
Dept. Mathematics, Statistics & Computer Science
University of Illinois at Chicago
Chicago, Illinois 60607-7045
U.S.A
knessl@uic.edu

Wojciech Szpankowski†
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

We study the asymptotic distribution of the fill-up level in a binary trie built over n independent strings generated by a biased memoryless source. The fill-up level is the number of full levels in a tree. A level is full if it contains the maximum allowable number of nodes (e.g., in a binary tree level k can have up to 2^k nodes). The fill-up level finds many interesting applications, e.g., in the internet IP lookup problem and in the analysis of *level compressed tries* (LC tries). In this paper, we present a complete asymptotic characterization of the fill-up distribution. In particular, we prove that this distribution concentrates on one or two points around the most probably value $k = \lfloor \log_{1/q} n - \log\log\log n + 1 + \log\log(p/q) \rfloor$ where $p > q = 1 - p$ is the probability of generating the more likely symbol (while $q = 1 - p$ is the probability of the less likely symbol). We derive our results by analytic methods such as generating functions, Mellin transform, the saddle point method and analytic depoissonization. We also present some numerical verification of our results.

Key Words: Trie, fill-up distribution, Mellin transform, saddle point method, depoissonization.

*The work was supported by NSF Grants CCR-0208709 and DMS-02-02815, by the NIH Grant R01 GM068959-01, and NSA Grant MDA 904-03-1-0036

1 Introduction

A wide variety of problems in computer science, information theory and applied probability can be reduced to the study of recursion and/or differential equations. They are thus susceptible to analysis by all the analytical methods available for such problems. Methods commonly used include generating functions and techniques for the asymptotic evaluation of integrals and sums, including singularity analysis, the Laplace and saddle point methods, and the Euler-MacLaurin and Poisson summation formulas (cf. [9, 10, 15]). This class of techniques is most useful if there exists an explicit expression for, say, the generating function of a sequence. Then recovering the sequence itself involves the evaluation of the Cauchy integral, be it exact, asymptotic, or numerical. For many important applied problems, it is possible to obtain a recurrence relation for the quantity of interest and/or differential (or functional-differential) equation for its transform (generating function, Laplace transform, Poisson transform). However, obtaining an exact solution to the problem is difficult if not impossible. It then becomes useful to analyze the equation(s) asymptotically.

In this paper, we concentrate on a particular non-linear recurrence representing the probability distribution of the so called fill-up level in a digital tree known as a *trie*. In fact, this recurrence appears in a variety of problems in computer science and engineering. To motivate our study, we first discuss one simple but important application.

Let us consider the internet IP addresses lookup problem [11, 14]. It can be briefly described as follows: given a (large) set of IP addresses find a fast algorithm that will allow us to direct an incoming packet with a given IP address to its destination. As a matter of fact, this is the *longest matching prefix* problem, thus standard tries and other digital trees (cf. [10, 15]) can be used. However, the search time is too expensive. If there are n IP addresses in the database, the search time is $O(\log n)$, and this is not acceptable. In order to improve the search time, one may use a novel data structure called a *level compressed trie* or in short LC trie [11], discussed briefly below.

A *trie* is a digital tree built over, say n, strings (the reader is referred to [10, 15] for an in depth discussion of digital trees.) A string is stored in an external node of a trie and the path length to such a node is the shortest prefix of the string that is not a prefix of any other strings. Let us assume a binary alphabet. Then each branching node in a trie is a binary node. It is known that a typical trie is a *full tree* (i.e., there are all nodes at each level) up to a certain level (the root is at level zero) that is called the *fill-up* level. We denote it as F_n (cf. Figure 1).

In the LC trie we replace the root with a node of degree 2^{F_n} (i.e., F_n binary links).

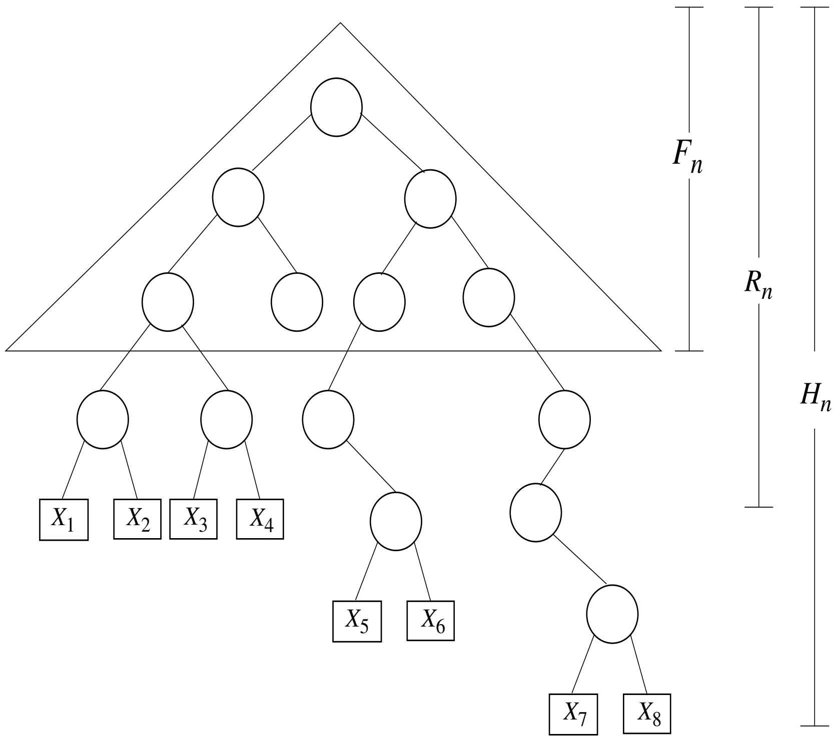

Figure 1: A trie built over eight strings $X_1, \ldots, X_8$ with the fill-up F_n, shortest path R_n, and the height H_n shown.

This is further carried on recursively throughout the whole trie. Knowing the distribution of F_n in a regular trie, one may evaluate the average depth, height and size of such LC tries. Some results, mostly for symmetric tries (i.e., when both symbols are equally likely to occur), can be found in [2, 11, 13].

In passing we should point out that there is another interpretation of the fill-up level that does not refer to tries. Consider n strings. The fill-up F_n is equal to the largest k such that *all* 2^k strings of length k occur as a prefix of at least one of the n strings. Therefore, if $F_n = k$, then at least one of 2^{k+1} strings is not a prefix of initial n strings.

Tries have been studied extensively over the last thirty years (cf. [10, 15]). However, not much is known about LC tries. The first natural problem is to understand the behavior of the fill-up level F_n in regular tries, and this is the topic of our paper. Consider a trie built over n binary strings generated independently by a memoryless source. Let p and $q = 1-p$ denote the probability of generating a "1" and "0", respectively. We are interested in finding the limiting distribution of the fill-up level F_n, that is, the maximal level where the tree is still full (in other words, the tree contains all nodes up to level F_n). Let $f(n,k) = \mathsf{Pr}[F_n \geq k]$.

As shown in Section 2, it satisfies the following recurrence for all $n \geq 0$ and $k \geq 0$

$$f(n, k+1) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} f(i, k) f(n-i, k) \qquad (1)$$

with $f(0,0) = 0$ and $f(n,0) = 1$ for $n \geq 1$. This recurrence is nonlinear. Actually, this recurrence appears in the analysis of other parameters of tries such as the height [3, 4, 10]) and the shortest path in a trie (which is often mistakenly taken to be equal to the fill-up level). We illustrate these parameters in Figure 1. All of these problems have the same recurrence with different initial and boundary conditions (which may lead to quite different solutions). For example, the distribution $r(n,k) = \Pr[R_n \geq k]$ of the shortest path R_n satisfies a slightly modified recurrence, namely

$$r(n, k+1) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} r(i, k) r(n-i, k) - \delta_{n,1}\delta_{k,0}$$

with $r(n,0) = 1$ for $n \geq 0$ where $\delta_{i,j}$ is the Kronecker delta (i.e., $\delta_{i,j} = 1$ for $i = j$ and zero otherwise).

We analyze asymptotically $\Pr[F_n \geq k]$ for various ranges of n and k (cf. Theorems 2 and 3) after finding in Theorem 1 an integral representation for the probability $\Pr[F_n \geq k]$. We present a complete asymptotic characterization of the fill-up distribution. In particular, we prove that this distribution concentrates on one or two points around the most probably value $k = \lfloor \log_{1/q} n - \log\log\log n + 1 + \log\log(p/q) \rfloor$ (cf. Figure 2). We derive our results by analytic methods such as generating functions, Mellin transform, Laplace's method, the saddle point method and analytic depoissonization. We also present some numerical verification of our results. Finally, we would like to point out that extension of our analysis to non-binary strings seems feasible, while extension to Markov sources and suffix trees is much more challenging.

The fill-up level for symmetric tries was analyzed by Devroye [1], and for asymmetric tries by Pittel [12], who studied the distribution of F_n in the central region where all mass is concentrated. In particular, Pittel [12] using probabilistic tools, found the asymptotic accumulation of probability mass at one or two points.

This paper is organized as follows. In the next section we formulate our main results. Most proofs are delayed until Section 3. In Section 4 we provide some numerical studies.

2 Problem Statement and Summary of Results

We let F_n be the number of full levels (fill-up level) in a trie constructed from n binary strings. In each string a "1" occurs with probability $p \in (0,1)$ and a "0" occurs with probability $q = 1 - p$. We let

$$f(n,k) = \mathsf{Pr}[F_n \geq k]$$

be the (complementary) probability distribution of F_n. It satisfies the recurrence relation

$$f(n,k+1) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} f(i,k) f(n-i,k); \qquad k,n \geq 0 \tag{2}$$

with the initial condition(s)

$$f(0,0) = 0 \ ; \qquad f(n,0) = 1, \ n \geq 1. \tag{3}$$

We also have $f(n,k) = 0$ for $n < 2^k$, since we need at least 2^k strings to completely fill k levels. The above recurrence follows from the fact that with probability $\binom{n}{i} p^i q^{n-i}$, i strings end up in the left subtree while $n-i$ move to the right subtree. In order to have the fill-up level of the whole tree equal at least to $k+1$ we must assure that the fill-up levels in both subtrees are at least is equal to k.

We can compute $f(n,k)$ explicitly for small values of k and $n - 2^k$. From (2) we directly find that

$$\begin{aligned}
f(n,1) &= 1 - p^n - q^n, \quad n \geq 1, \\
f(n,2) &= 1 - (1-p^2)^n - (1-q^2)^n - 2(1-pq)^n \\
&+ 2(p^n + q^n) + (p^2+q^2)^n + 2^n (pq)^n - (p^n + q^n)^2, \quad n \geq 1.
\end{aligned}$$

When $p = q = 1/2$, the above simplifies to

$$f(n,2) = 1 - 4\left(\frac{3}{4}\right)^n + 6\left(\frac{1}{2}\right)^n - 4\left(\frac{1}{4}\right)^n, \quad n \geq 1.$$

Finding explicit values of $f(n,k)$ on the other extreme, where n is near 2^k, is a little harder. Observing that for $n = 2^k$ the sum (2) collapses to a single term, and defining $a_k = f(2^k,k)/(2^k!)$ we arrive at the following non-linear recurrence for $k \geq 1$

$$a_k = (\sqrt{pq})^{2^k} a_{k-1}^2$$

with $a_0 = 1$. This can be solved by taking logarithms, leading to

$$f(2^k,k) = (2^k)!(\sqrt{pq})^{k2^k}.$$

In a similar fashion we find that

$$\begin{aligned}
f(2^k+1,k) &= (2^k+1)!\frac{1}{2}(\sqrt{pq})^{k2^k}, \\
f(2^k+2,k) &= (2^k+2)!(\sqrt{pq})^{k2^k}\left[\frac{1}{8} + \frac{1}{24}(p^2+q^2)^k\right].
\end{aligned}$$

From the above explicit formulas it is clear that getting a nice closed-form solution for $f(n,k)$ is rather hopeless. Nevertheless, the recurrence (2) and (3) can be solved using generating functions together with Poisson and Mellin transforms. Then inverting the Poisson transform leads to an integral representation for $f(n,k)$ that we summarize in the theorem below.

First, however, let us recall some notation. We write $[z^n]F(z)$ to denote the coefficient of z^n in $F(z)$. Furthermore, the Mellin transform $F^*(s) = \mathcal{M}[f(x);s]$ of a real function $f(x)$ is

$$F^*(s) = \mathcal{M}[f(x);s] = \int_0^\infty f(x)x^{s-1}dx,$$

and its inverse Mellin transform is

$$f(x) = \frac{1}{2\pi i}\int_{Br} F^*(s)x^{-s}ds,$$

where Br is a vertical contour in the s-half plane where $F^*(s)$ is defined. Now we can formulate our first result.

Theorem 1 *The distribution $f(n,k)$ has the representation*

$$f(n,k) = n![z^n]\exp\left[H_k(z) + z\right],$$

so that by Cauchy's formula

$$f(n,k) = \frac{n!}{2\pi i}\oint \frac{e^z}{z^{n+1}}\exp[H_k(z)]dz. \tag{4}$$

Here the Mellin transform of $H_k(z)$ is

$$\mathcal{M}[H_k(z);s] = -\Gamma(s)\zeta(s+1)(p^{-s}+q^{-s})^k, \qquad \Re(s) > 0,$$

and by inversion we have

$$H_k(z) = -\frac{1}{2\pi i}\int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} z^{-s}\Gamma(s)\zeta(s+1)(p^{-s}+q^{-s})^k ds. \tag{5}$$

Here $\Gamma(\cdot)$ is the Gamma function and $\zeta(\cdot)$ is the Riemann zeta function.

An alternate form for $H_k(z)$ is

$$\exp[H_k(z)] = \prod_{m=0}^{k}\left[1-\exp(-zp^mq^{k-m})\right]^{\binom{k}{m}}, \tag{6}$$

and we also have

$$H_k(z) = 2^k\log z + 2^{k-1}k\log(pq) - \frac{z}{2} + \sum_{\ell=1}^{\infty}\frac{B_{2\ell}}{2\ell}\frac{z^{2\ell}}{(2\ell)!}(p^{2\ell}+q^{2\ell})^k, \tag{7}$$

where $B_{2\ell}$ are the even Bernoulli numbers. The series converges for $|z| < 2\pi p^{-k}$, if $p > q$.

In the symmetric case $p = q = 1/2$, we have

$$f(n,k) = \frac{n!}{2\pi i} \oint \frac{e^z}{z^{n+1}} \left[1 - \exp\left(-\frac{z}{2^k}\right)\right]^{2^k} dz. \tag{8}$$

Proof. We only sketch the proof, referring the reader to [15] for an in-depth discussion of the Poisson transform and Mellin transform. Let $F_k(z)$ be the exponential generating function of $f(n,k)$, that is,

$$F_k(z) = \sum_{n=0}^{\infty} f(n,k) \frac{z^n}{n!}.$$

We denote the Poisson transform by $\widetilde{F}_k(z) := F_k(z)e^{-z}$. Observe that from (2) we obtain

$$F_{k+1}(z) = F_k(pz)F_k(qz)$$

with $F_0(z) = e^z - 1$. Now let $\widetilde{F}_k(z) = \exp[H_k(z)]$. Then we find the following functional equation

$$H_{k+1}(z) = H_k(pz) + H_k(qz)$$

with $H_0(z) = \log(1 - e^{-z})$.

Let now $M_k(s) = \mathcal{M}[H_k(z); s]$. Clearly (cf. [15])

$$M_{k+1}(s) = (p^{-s} + q^{-s})M_k(s)$$

and

$$M_0(s) = \int_0^{\infty} z^{s-1} \log(1 - e^{-z}) dz = -\Gamma(s)\zeta(s+1).$$

Therefore,

$$M_k(s) = -\Gamma(s)\zeta(s+1)(p^{-s} + q^{-s})^k$$

which proves (5). The rest follows from standard analysis and is omitted. ∎

To gain more insight into the structure of $f(n,k)$, we consider the limit $n \to \infty$ for various ranges of k. We first summarize our results for the **symmetric case** when $p = q$.

Theorem 2 *For $p = q = 1/2$ and $k, n \to \infty$ the distribution $f(n,k)$ has the following asymptotic expansions:*

(a) $k, n \to \infty$ *with* $n = 2^k + \ell$, $\ell = O(1)$, $\ell \geq 0$

$$f(n,k) \sim \sqrt{2\pi} e^{-2^k} 2^{k(\ell+1/2)} \frac{1}{2^\ell \ell!} \sim \sqrt{2\pi n} e^{-n} n^\ell \frac{e^\ell}{2^\ell \ell!}. \tag{9}$$

(b) *$k, n \to \infty$ with $n2^{-k} \in (c_1, c_2)$, $c_1 > 1$ and $c_2 < \infty$*

$$\begin{aligned} f(n,k) &\sim \sqrt{\frac{n}{2^k}} \left(\frac{n}{2^k}\right)^n e^{-n} \zeta_0^{-n} (e^{\zeta_0} - 1)^{2^k} \frac{e^{\zeta_0/2} - e^{-\zeta_0/2}}{\sqrt{\zeta_0}\sqrt{e^{\zeta_0} - 1 - \zeta_0}} \\ &= \sqrt{\frac{e^{\zeta_0} - 1}{e^{\zeta_0} - 1 - \zeta_0}} \left(\frac{e^{\zeta_0}}{e^{\zeta_0} - 1}\right)^n e^{-n} (e^{\zeta_0} - 1)^{2^k}. \end{aligned} \tag{10}$$

Here ζ_0 is defined implicitly as the solution to

$$\frac{\zeta_0 e^{\zeta_0}}{e^{\zeta_0} - 1} = n2^{-k}\ ,\ \ 0 < \zeta_0 < \infty$$

which satisfies

$$\begin{aligned} \zeta_0 &\sim 2(n2^{-k} - 1), & n2^{-k} &\to 1 \\ \zeta_0 &\sim n2^{-k}[1 - \exp(-n2^{-k})], & n2^{-k} &\to \infty. \end{aligned}$$

(c) *$n \to \infty$ with $k = O(1)$*

$$1 - f(n,k) \sim 2^k (1 - 2^{-k})^n. \tag{11}$$

We note that the condition $n - 2^k = O(1)$ for $n \to \infty$ can be satisfied only if n is close to a power of 2. We can easily verify that the three expansions in Theorem 2 asymptotically match in appropriate intermediate limits. For example, if we expand the result in part (a) for $\ell \to \infty$, it agrees with the expansion of (b) as $n2^{-k} \to 1^+$. Similarly (b) as $n2^{-k} \to \infty$ agrees with (c) as $k \to \infty$.

For a fixed large n we wish to compute those values of k where f undergoes the transition from $f \approx 1$ to $f \approx 0$. This is where most of the probability mass accumulates, and this occurs in the matching region between cases (b) and (c) above. We approximate f in this region and obtain the following corollary.

Corollary 1 *For $p = q = 1/2$ we define L to be an integer such that*

$$k = \lfloor \log_2 n - \log_2(\log n) \rfloor + L = \log_2 n - \log_2(\log n) - \beta + L$$

where

$$\beta = \beta(n) = \langle \log_2 n - \log_2(\log n) \rangle$$

with $\langle x \rangle = x - \lfloor x \rfloor$ denoting the fractional part of x. Then in the matching region

$$\mathsf{Pr}[F_n \geq k] \approx \exp\left(-\frac{n}{\log n} 2^{L-\beta} n^{-2^{\beta - L}}\right). \tag{12}$$

as $n \to \infty$.

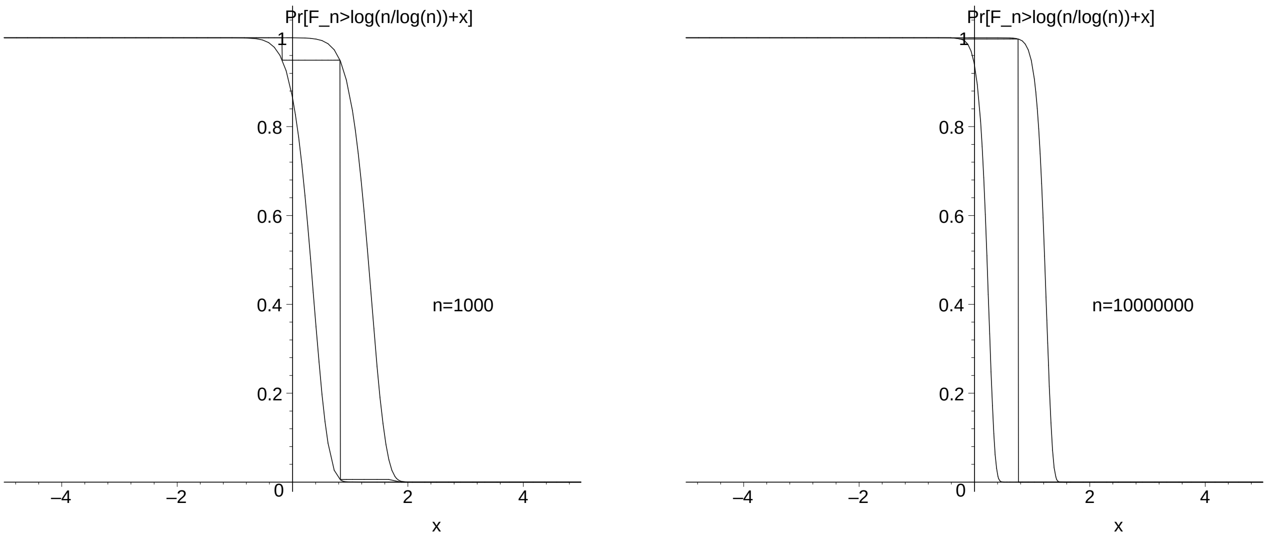

Figure 2: The probability $\mathsf{Pr}[F_n \geq \log_2(n/\log(n)) + x]$ and their corresponding lower and upper envelopes for tries with $n = 1000$ (left) and $n = 10000000$ (right).

We note that for a fixed $\beta \in (0,1)$ and $L = 1$, the expression in (12) is exponentially small for $n \to \infty$. When $\beta \in (0,1)$ and $L = 0$ we have $f(n,k) \sim 1$. This means that if we let $n \to \infty$ along integer subsequences that have $\beta = \beta(n)$ bounded away from 0 and 1, $\mathsf{Pr}[F_n = \lfloor \log_2(n/\log n) \rfloor] \sim 1$, i.e., all the probability mass accumulates at a single point. This was previously shown by Pittel [12]. Expression (12) also shows that along certain subsequences that have either $\beta \to 0^+$ or $\beta \to 1^-$, we can have mass accumulating at two points. This is illustrated in Figure 2. Similar results were shown to hold for the height of other digital trees, namely PATRICIA tries [8] and digital search trees [7]. From a numerical perspective, it is best to use the formula in (12) (or that in Theorem 2(b)) as the approximation to $f(n,k)$ in the range where there is significant mass.

For the mean, (12) suggests the following approximation

$$\begin{aligned} \mathbf{E}[F_n] &= \log_2 n - \log_2(\log n) + \Delta(n) + O(n^{-1}), & (13) \\ \Delta(n) &= -\beta + \exp\left[-\frac{n}{\log n} 2^{-\beta} n^{-2^{\beta}}\right] + \exp\left[-\frac{n}{\log n} 2^{1-\beta} n^{-2^{\beta-1}}\right]. & (14) \end{aligned}$$

The function $\Delta(n)$ fluctuates, as shown in Figure 3. One can find its Fourier series representation, as discussed in [15] (cf. also [7, 8]).

We next consider the **non-symmetric case**, taking $p > q$. We also assume that $p - \frac{1}{2}$ and $1 - p$ remain fixed positive quantities as $n \to \infty$. This excludes, for example, "nearly symmetric" situations where we would have $n \to \infty$ and $p \downarrow \frac{1}{2}$ simultaneously. Now the asymptotics become considerably more complex and there are several ranges to consider.

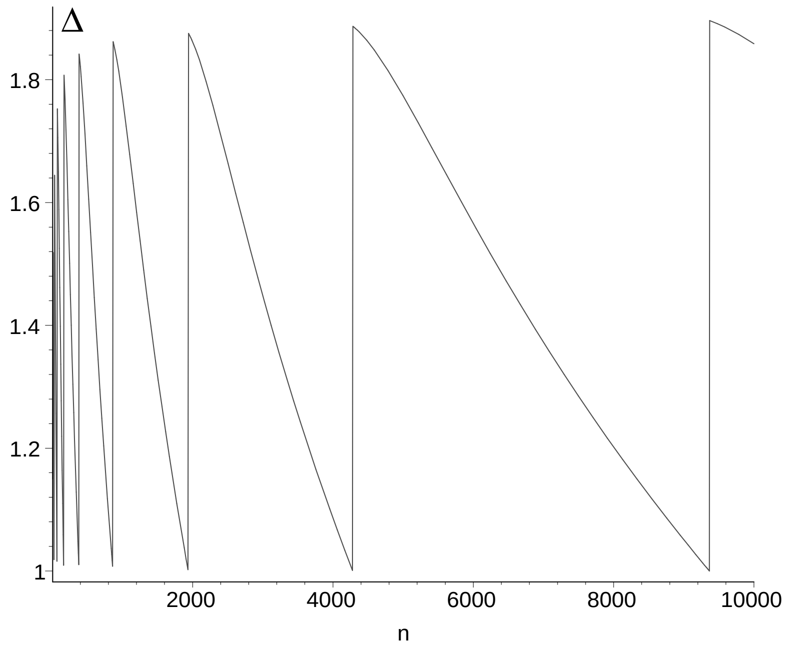

Figure 3: The function $\Delta(n)$ defined in (14) as a function of n.

Theorem 3 *For $p \in \left(\frac{1}{2}, 1\right)$ with $r = p/q$, and $k, n \to \infty$, the distribution $f(n,k)$ has the following asymptotic expansions:*

(a) $k, n \to \infty$ *with* $n = 2^k + \ell$, $\ell = O(1)$, $\ell \geq 0$

$$f(n,k) \sim \sqrt{2\pi} e^{-2^k} 2^{k(\ell+1/2)} (2\sqrt{pq})^{k2^k} \frac{1}{2^\ell \ell!} \sim \frac{n!}{2^\ell \ell!} (\sqrt{pq})^{k2^k}. \tag{15}$$

(b) $k, n \to \infty$ *with* $n - 2^k = (p^{-p} q^{-q})^k e^{b\sqrt{k}}, \quad b = O(1)$

$$f(n,k) = \frac{n!}{(n-2^k)^{n-2^k}} \frac{e^{n-2^k}}{\sqrt{2\pi(n-2^k)}} (\sqrt{pq})^{k2^k} [I(b)]^{n-2^k}, \tag{16}$$

$$I(b) = 1 - \frac{1}{2\sqrt{2\pi}} \int_{b'}^{\infty} e^{-u^2/2} du + O\left(\frac{1}{\sqrt{k}}\right), \qquad b' = \frac{b}{\sqrt{pq}\log(p/q)}. \tag{17}$$

(c) $k, n \to \infty$ *with* $n - 2^k = B^k$, $B \in (p^{-p} q^{-q}, 1/\sqrt{pq})$

$$f(n,k) = \frac{n!}{(n-2^k)!} (\sqrt{pq})^{k2^k} \exp\left(-\Lambda_1(n,k)[1 + O(k^{-1})]\right), \tag{18}$$

where

$$\Lambda_1 = (p^{-S_*} + q^{-S_*})^k (n-2^k)^{-S_*} \frac{\Gamma(S_*)\zeta(S_*+1)(r^{S_*/2} + r^{-S_*/2})}{(\log r)\sqrt{2\pi k}},$$

$$S_* = S_*(B) = -\frac{1}{\log r}\log\left[\frac{-\log(qB)}{\log(pB)}\right].$$

(d) $k, n \to \infty$ *with* $n = (pq)^{-k/2}e^{a\sqrt{k}}$, $a = O(1)$

$$f(n,k) = \exp\left(2^{k-1}\sqrt{k}(\log r)\frac{1}{\sqrt{2\pi}}\left[a'\int_{a'}^{\infty} e^{-u^2/2}du - e^{-(a')^2/2}\right][1+O(k^{-1})]\right), \tag{19}$$

where $a' = 2a/\log r$.

(e) $k, n \to \infty$ *with* $n = A^k$, $A \in (1/\sqrt{pq}, 1/q)$

$$f(n,k) = \exp\left(-\Lambda(n,k)[1+O(k^{-1})]\right), \tag{20}$$

$$\Lambda = (p^{-S_*} + q^{-S_*})^k n^{-S_*}\frac{\Gamma(S_*)\zeta(S_*+1)(r^{S_*/2} + r^{-S_*/2})}{(\log r)\sqrt{2\pi k}},$$

$$S_* = S_*(A) = -\frac{1}{\log r}\log\left[\frac{-\log(qA)}{\log(pA)}\right].$$

(f) $k, n \to \infty$ *with* $nq^k \le c$, $c > 0$

$$f(n,k) = \exp\left(-\frac{1}{\sqrt{1+k(\log^2 r)\bar{S}r^{-\bar{S}}}}\exp\left[-\bar{S} + \left(\bar{S} + \frac{1}{\log r}\right)\log\left(\frac{\bar{S}}{nq^k}\right)\right][1+O(\bar{S}^{-1})]\right) \tag{21}$$

where $\bar{S} = \bar{S}(n,k)$ *is defined implicitly by*

$$\bar{S} = nq^k \exp[k(\log r)r^{-\bar{S}}]. \tag{22}$$

(g) $n \to \infty$ *with* $k = O(1)$

$$1 - f(n,k) \sim (1-q^k)^n. \tag{23}$$

Remark: Here $\approx$ means that $\log \mathsf{Pr}[F_n \ge k]$ is asymptotically equal to the expression on the right hand side, that is, $\mathsf{Pr}[F_n \ge k] \approx \exp(a_n)$ should be read as $\log \mathsf{Pr}[F_n \ge k] \sim a_n$.

We note that the condition $n - 2^k = O(1)$ can again be satisfied only if n is large and equal to or slightly exceeds a power of 2. For $p \ne 1/2$ we have $2^h = p^{-p}q^{-q} < 2 < 1/\sqrt{pq}$ (where $h = -p\log_2 p - q\log_2 q$ is the entropy) and thus the condition under which (c) applies has $n \sim 2^k$ for $B \in (p^{-p}q^{-q}, 2)$ and $n \sim B^k$ for $B \in (2, 1/\sqrt{pq})$. Part (b) and part (c) for $B < 2$ also apply only if n is close to a power of 2. Given a large n, the conditions under which (d)-(g) (and also (c) for $B > 2$) apply can be satisfied for certain ranges of k. If we view the problem as starting with a large fixed k and then varying n ($\ge 2^k$), then we pass through each of the ranges in Theorem 3.

Theorem 3 suggests that an important scale is to have $k,\ n\to\infty$ with $k/\log n$ fixed and $< 1/(\log 2)$. We note that (d) applies in the limit $k = \log n/[-\log(\sqrt{pq})] + O(\sqrt{\log n})$ while (e) has $k = \nu \log n$ with $\nu \in (-1/\log q, -2/\log(pq))$. In parts (a)-(c) we may further approximate $n!$ and $(n-2^k)!$ by Stirling's formula. It is interesting to note that the scale $k = \log_2 n + O(1)$, which was the most important scale for the symmetric case, plays no role in the non-symmetric case. This range is contained in part (c) of Theorem 3. The formulas in Theorem 3 are all explicit functions of n and k, with the exception of part (f), which is in terms of the solution $\bar{S}$ to the transcendental equation (22).

Finally, we note that in parts (b)-(f) we really give only the leading term for $\log[f(n,k)]$. For example, in part (d) the exponent has the order of magnitude $O(2^k\sqrt{k})$. Higher order terms in the asymptotic series would have orders $O(2^k\sqrt{k}k^{-N/2})$, $N\geq 2$ and obtaining the leading term for $f(n,k)$ itself would require that we obtain (at least) the full asymptotic series in the exponent.

We again ask where $f(n,k) = \mathsf{Pr}[F_n \geq k]$ undergoes the transition from $f(n,k)\approx 1$ to $f(n,k)\approx 0$. This occurs in the range where part (f) applies and we obtain the following result.

Corollary 2 *For $p \in \left(\frac{1}{2}, 1\right)$ we define L as an integer such that*

$$\begin{aligned} k &= \left\lfloor \tfrac{1}{\log(1/q)}[\log n - \log\log\log n + \log(e\log r)]\right\rfloor + L \\ &= \tfrac{1}{\log(1/q)}[\log n - \log\log\log n + \log(e\log r)] - \beta_* + L \end{aligned}$$

where β_* denotes the fractional part

$$\beta_* = \beta_*(n) = \left\langle \frac{1}{\log(1/q)}[\log n - \log\log\log n + \log(e\log r)]\right\rangle.$$

Then we have

$$\begin{aligned} \mathsf{Pr}[F_n \geq k] \quad &\approx \quad \exp\left(-\frac{1}{\sqrt{\log\log n}}\frac{1}{\sqrt{1+(\beta_*-L)\log q}}\left(\frac{e\log r}{\log(1/q)[(\beta_*-L)\log q+1]}\right)^{\frac{1+(\beta_*-L)\log q}{\log r}}\right. \\ &\times \quad \left.(\log n)^{\frac{(\beta_*-L)\log q}{\log r}}\left(1+O\left(\frac{1}{\log\log n}\right)\right)\right), \qquad L > L_c \qquad (24) \end{aligned}$$

and

$$1-\mathsf{Pr}[F_n\geq k]\approx\left(\frac{1}{\log n}\right)^{\frac{q^{L-\beta_*}e^{-1}}{\log r}}, \qquad L\leq L_c \qquad (25)$$

where

$$L_c = \left\lfloor \beta_* + \frac{1}{\log q}\right\rfloor.$$

We note that the integer L_c can only take on the three values –2, –1, and 0. By examining the various possibilities for L_c, we can show that the probability mass concentrates at $L = 0$, i.e., $\Pr[F_n = \lfloor \log_{1/q}(en \log r / \log \log n) \rfloor] \sim 1$, except for some special subsequences of n that lead to mass at two consecutive points. This again agrees with results of Pittel [12]. While the expressions in (24) and (25) are more explicit than that in (21), the latter may be preferable for numerical purposes; this is further discussed in Section 4.

Finally, Corollary 2 suggests the following formula for the mean $\mathbf{E}[F_n]$

$$\mathbf{E}[F_n] = \log_{1/q} n - \log \log \log n + O(1)$$

where the $O(1)$ term can be made more explicit (as in (13)). Observe that the second term in the mean in (13) for the symmetric case is $-\log \log n$ while for the asymmetric case it is $-\log \log \log n$. It should be possible also to consider a limit where $n \to \infty$ and $p \downarrow \frac{1}{2}$ that interpolates between these two behaviors of F_n (e.g., see [6]).

3 Derivations

We establish Theorems 2 and 3 by asymptotically evaluating the integrals in Theorem 1. Since this mainly involves using standard methods (e.g., the saddle-point and Laplace methods), we only sketch the main ideas.

3.1 Methodology

We begin by reviewing some facts about Laplace-type integrals. Setting

$$I(\lambda) = \int_{-\infty}^{\infty} e^{\lambda f(x)} h(x) dx \tag{26}$$

we have the standard estimate for $\lambda \to +\infty$ (cf. [15, 16])

$$I(\lambda) = e^{\lambda f(x_*)} h(x_*) \sqrt{\frac{2\pi}{\lambda |f''(x_*)|}} \left(1 + O(\lambda^{-1})\right), \tag{27}$$

where f and h are assumed smooth functions with f having a unique maximum at $x = x_*$, with $f''(x_*) < 0$ and $h(x_*) \neq 0$.

Now consider a modified version of (26) of the form

$$J(\lambda) = \int_{-\infty}^{\infty} \exp[\lambda f(x) + \lambda^a g(x)] h(x) dx, \tag{28}$$

where $a \in (0, 1)$. We can estimate (28) for $\lambda \to +\infty$ by two approaches. First, we can define x_0 as the solution to

$$f'(x_0) + \lambda^{a-1} g'(x_0) = 0 \tag{29}$$

and then apply (27) to obtain

$$J(\lambda) = e^{\lambda f(x_0)} e^{\lambda^a g(x_0)} h(x_0) \sqrt{\frac{2\pi}{\lambda |f''(x_0)|}} \left(1 + O(\lambda^{a-1})\right). \tag{30}$$

This has the disadvantage that if f is simple and g is complicated, x_0 and (30) may also be complicated.

Alternately we can define x_* as the solution to $f'(x_*) = 0$ (as in (27)) and then expand the integrand in (28) about $x = x_*$. This ultimately yields

$$J(\lambda) = h(x_*) \sqrt{\frac{2\pi}{\lambda |f''(x_*)|}} \exp\left(\lambda f(x_*) + \lambda^a g(x_*) - \frac{1}{2}\lambda^{2a-1} \frac{(g'(x_*))^2}{f''(x_*)} + O(\lambda^{3a-2})\right). \tag{31}$$

If $a \in \left(0, \frac{1}{2}\right)$ the above can be obtained by replacing f by $f + \lambda^{a-1} g$ in (26). To relate (30) and (31) we can expand x_0 about x_* in (29), to find that

$$x_0 = x_* - \frac{g'(x_*)}{f''(x_*)} \lambda^{a-1} + O(\lambda^{2a-2}), \ \lambda \to +\infty. \tag{32}$$

Using (32) in (30) regains (31).

Next we consider a Cauchy integral in the form

$$f(n) = \frac{n!}{2\pi i} \oint \frac{e^z}{z^{n+1}} \widetilde{F}(z) dz.$$

where $\widetilde{F}(z)$ is the Poisson transform of $f(n)$, that is,

$$\widetilde{F}(z) = e^{-z} \sum_{n \geq 0} f(n) \frac{z^n}{n!}.$$

In many applications, including the one in this paper, we can estimate $\widetilde{F}(z)$ for $z \to \infty$ in a cone near the real axis. Can we infer from this asymptotic expansion of $\widetilde{F}(z)$ an asymptotic expansion of $f(n)$? The answer is in the affirmative under certain conditions, as proved by Jacquet and Szpankowski [5] (cf. also [15]). The process of recovering $f(n)$ from $\widetilde{F}(z)$ is called (analytic) *depoissonization.*

Theorem 4 (Jacquet and Szpankowski, 1998) *Let $f(n)$ be a sequence whose Poisson transform is $\widetilde{F}(z) = e^{-z} \sum_{n \geq 0} f(n) \frac{z^n}{n!}$ where $\widetilde{F}(z)$ is an entire function. Furthermore, by $\mathcal{S}_\theta = \{z : \ \arg(z) \leq \theta, \ |\theta| < \pi/2\}$ we denote a linear cone around the real axis.*

(i) *Assume that for $z \to \infty$ the following holds:*

(I) For $z \in \mathcal{S}_\theta$

$$|\widetilde{F}(z)| \leq A \exp(B|z|^\nu) \tag{33}$$

where $0 \leq \nu < 1/2$, *and* $A, B > 0$ *are constants.*
(O) For $z \notin \mathcal{S}_\theta$

$$|\widetilde{F}(z)e^z| \leq A_1 \exp(\omega|z|) \tag{34}$$

for $\omega < 1$ *and* $A_1 > 0$. *Then for* $n \to \infty$

$$f(n) = \widetilde{F}(n) + O\left(n^{-(1-2\nu)} \exp(Bn^\nu)\right). \tag{35}$$

(ii) *Let conditions (I) and (O) of (33) and (34) hold again except that now* ν *in (I) satisfies* $\frac{1}{2} < \nu < 1$. *Then*

$$\log f(n) = \log \widetilde{F}(n) + O(n^{2\nu-1}) \tag{36}$$

for $n \to \infty$.

In the present application we shall only know $\widetilde{F}(z)$ asymptotically, and for $z \to \infty$ it will have an expansion of the form $\exp\left[z^\nu \sum_{L=0}^{\infty} d_L (\log z)^{-L-\delta}\right]$, where it will prove feasible to calculate only the first few d_L. Thus the error in (36) will be smaller than $n^\nu (\log n)^{-L}$ for any L.

3.2 Proof of Theorem 2

We first establish Theorem 2. For the limit, $k, n \to \infty$ with $n - 2^k = \ell = O(1)$ we use the expansion

$$\begin{aligned} z^{-n-1}\left[1 - \exp\left(-\frac{z}{2^k}\right)\right]^{2^k} &= z^{-n-1} z^{2^k} 2^{-k2^k} \left[1 - \frac{z}{2^{k+1}} + O\left(\frac{z^2}{4^k}\right)\right]^{2^k} \\ &= z^{-\ell-1} 2^{-k2^k} e^{-z/2}[1 + O(2^{-k})]. \end{aligned} \tag{37}$$

Part (a) of Theorem 2 follows upon using (37) in (8) and expanding $n! = (2^k + \ell)!$ by Stirling's formula. To obtain part (c), we simply use

$$1 - \left[1 - \exp\left(-\frac{z}{2^k}\right)\right]^{2^k} \sim 2^k \exp\left(-\frac{z}{2^k}\right), \qquad z \to \infty$$

and the fact that

$$1 - f(n,k) = \frac{n!}{2\pi i} \int_C \frac{e^z}{z^{n+1}} \left(1 - \left[1 - \exp\left(-\frac{z}{2^k}\right)\right]^{2^k}\right) dz.$$

Now consider the limit $n, k \to \infty$ with $k = \log_2 n + O(1)$, where Theorem 2(b) applies. We scale $z = 2^k \zeta$ and obtain from (8)

$$f(n,k) = n! 2^{-kn} \frac{1}{2\pi i} \oint \frac{1}{\zeta} \exp\{2^k[\zeta + \log(1 - e^{-\zeta}) - n2^{-k} \log \zeta]\} d\zeta.$$

The integrand above has a saddle-point where

$$\frac{d}{d\zeta}[\zeta + \log(1 - e^{-\zeta}) - n2^{-k} \log \zeta] = 0,$$

that is,

$$\frac{e^{\zeta}}{e^{\zeta} - 1} = \frac{n2^{-k}}{\zeta}.$$

The above defines $\zeta = \zeta_0(n2^{-k})$, which satisfies $\zeta_0 \to 0^+$ as $n2^{-k} \to 1$ and $\zeta_0 \to +\infty$ as $n2^{-k} \to \infty$. The directions of steepest descent are $\arg(\zeta - \zeta_0) = \pm\pi/2$ and thus the standard estimate of the integral is

$$n!2^{-kn}\frac{1}{\zeta_0}\left[\frac{n2^{-k}}{\zeta_0^2} - \frac{e^{\zeta_0}}{(e^{\zeta_0} - 1)^2}\right]^{-1/2} \frac{2^{-k/2}}{\sqrt{2\pi}} e^{2^k\zeta_0}(1 - e^{-\zeta_0})^{2^k} \zeta_0^{-n}.$$

Using Stirling's formula to expand $n!$ the above simplifies to (10).

We can easily verify that the expansions in parts (a) and (b) asymptotically match in an intermediate limit where $\ell = n - 2^k \to \infty$, but $n2^{-k} \to 1^+$. Similarly parts (b) and (c) match if $k \to \infty$ and $n2^{-k} \to \infty$, and in the matching region we have

$$f(n,k) \sim \exp[-2^k \exp(-n2^{-k})].$$

Then Corollary 1 follows by choosing k so to make the right side of the above equal to $O(1)$, or as close to this as possible for integer k.

3.3 Proof of Theorem 3

Next we consider the non-symmetric case and sketch the derivation of Theorem 3. To obtain part (a) where $n - 2^k = \ell = O(1)$, we expand the integrand for z small using (7). Hence

$$z^{-n-1}\exp[H_k(z)] \sim z^{-\ell-1}(\sqrt{pq})^{k2^k} e^{-z/2}$$

and Theorem 3(a) follows by integrating the above.

3.3.1 Asymptotic Evaluation of $H_k(z)$

We next establish the asymptotics of $H_k(z)$ in (5) for $z \to \infty$ and various ranges of k. We recall that $p > q$. We consider $k, z \to \infty$ with $k/\log z$ fixed. We note that the integrand has a double pole at $s = 0$ and simple poles at $s = -1, -2, -3, \ldots$. Writing $z^{-s}(p^{-s} + q^{-s})^k = \exp[-s \log z + k \log(p^{-s} + q^{-s})]$ we see that in the indicated limit there are saddle-points where

$$\frac{d}{ds}[-s \log z + k \log(p^{-s} + q^{-s})] = 0,$$

i.e.,

$$\frac{\log z}{k} = \frac{p^{-s}\log(1/p) + q^{-s}\log(1/q)}{p^{-s}+q^{-s}}.$$

Solving the above explicitly leads to

$$s = S_* \equiv \frac{-1}{\log(p/q)} \log\left(\frac{-\log z - k\log q}{\log z + k\log p}\right).$$

The saddle lies on the real axis for $p^{-k} < z < q^{-k}$ (i.e., $k/\log z \in (-1/\log q, -1/\log p)$). It coalesces with the pole at $s = 0$ when $z = z_0 \equiv (pq)^{-k/2}$ and with the pole at $s = -1$ when $z = z_1 \equiv (p^{-p}q^{-q})^k$. We also note that $S_* \to +\infty$ as $z \uparrow q^{-k}$ and $S_* \to -\infty$ as $z \downarrow p^{-k}$.

We now evaluate $H_k(z)$ in seven ranges of z:

Case A: $(pq)^{-k/2} < z < q^{-k}$

For $(pq)^{-k/2} < z < q^{-k}$ we have $S_* \in (0,\infty)$ and then the asymptotics of $H_k(z)$ follow by shifting the contour in (5) into another vertical contour that goes through the saddle. Setting

$$\begin{aligned}\Delta_0 &= \frac{d^2}{ds^2}[-s\log z + k\log(p^{-s}+q^{-s})]\bigg|_{s=S_*} \\ &= \frac{k(\log r)^2}{(r^{S_*}+1)(r^{-S_*}+1)}, \qquad r = \frac{p}{q} > 1,\end{aligned}$$

we obtain

$$H_k(z) \sim H^{\text{saddle}}(k,z) \equiv -\frac{1}{\sqrt{2\pi\Delta_0}}\Gamma(S_*)\zeta(S_*+1)z^{-S_*}(p^{-S_*}+q^{-S_*})^k, \tag{38}$$

with an error factor $1 + O(k^{-1})$.

Case B: $(p^{-p}q^{-q})^k < z < (\sqrt{pq})^{-k}$

When $z \to \infty$ with $(p^{-p}q^{-q})^k < z < (\sqrt{pq})^{-k}$, the saddle satisfies $-1 < S_* < 0$ and hence the contribution of the pole at $s = 0$ must be considered. Since $\Gamma(s) = \frac{1}{s} - \gamma + O(s)$ and $\zeta(s+1) = \frac{1}{s} + \gamma + O(s)$, we can explicitly compute the residue at $s = 0$ (see also (7)) and obtain

$$H_k(z) \sim 2^k \log[z(\sqrt{pq})^k] + H^{\text{saddle}}(k,z). \tag{39}$$

Case C: $z < (p^{-p}q^{-q})^k$

For $z < (p^{-p}q^{-q})^k$ the pole at $s = -1$ also contributes and we obtain

$$H_k(z) \sim 2^k \log[z(\sqrt{pq})^k] - \frac{z}{2} + H^{\text{saddle}}(k,z). \tag{40}$$

Case D: Saddle Close to the Pole $s = -1$

We consider transition ranges where the saddle lies close to one of the poles. For $(\log z)/k = -p\log p - q\log q + O(1/\sqrt{k})$ we introduce w via the scaling

$$z = (p^{-p}q^{-q})^k e^{w\sqrt{k}},$$

expand the integrand about $s = -1$, and set $s+1 = T/\sqrt{k}$ with $T = O(1)$. We have

$$\begin{aligned} -s\log z + k\log(p^{-s}+q^{-s}) &= \log z - w\sqrt{k}(s+1) + \frac{1}{2}qpk(\log r)^2(s+1)^2 + O(k(s+1)^3) \\ &= \log z - wT + \frac{1}{2}qp(\log r)^2T^2 + O(k^{-1/2}) \end{aligned}$$

and $\Gamma(s)\zeta(s+1) \sim -\frac{1}{s+1}\zeta(0) = \frac{1}{2(s+1)}$. Then (5) becomes, upon shifting the line of integration Br to the range $\Re(s) \in (-1, 0)$,

$$\begin{aligned} H_k(z) &\sim 2^k \log[(\sqrt{pq})^k] - \frac{z}{2}\frac{1}{2\pi i}\int_{Br_+} e^{-wT}e^{qp(\log r)^2T^2/2}\frac{dT}{T} \qquad (41) \\ &= 2^k\log[z(\sqrt{pq})^k] - \frac{z}{2}\frac{1}{\sqrt{2\pi}}\int_{w'}^{\infty} e^{-u^2/2}du[1+O(k^{-1})], \qquad w' = \frac{w}{\sqrt{pq}\log r}. \end{aligned}$$

Here we assume that on BR_+ $\Re(T) > 0$.

Case E: Saddle Close to the Double Pole $s = 0$

Next, we consider the situation where the saddle is close to the double pole at $s = 0$. For $s \to 0$ we have $\Gamma(s)\zeta(s+1) = s^{-2} + O(1)$ and

$$-s\log z + k\log(p^{-s}+q^{-s}) = k\log 2 - [\log z + k\log(\sqrt{pq})]s + \frac{k}{8}(\log r)^2 s^2 + O(ks^3).$$

We assume that now $(\log z)/k = -\frac{1}{2}\log(pq) + O(1/\sqrt{k})$ and introduce the scaling

$$z = (pq)^{-k/2}e^{\alpha\sqrt{k}} \qquad \alpha = O(1).$$

Then setting $s = u/\sqrt{k} = O(k^{-1/2})$ we obtain

$$\begin{aligned} H_k(z) &= -\frac{2^k}{\sqrt{k}}\frac{1}{2\pi i}\int_{Br_+}\left[\frac{k}{u^2}+O(1)\right]\exp\left[-\alpha u + \frac{1}{8}(\log r)^2u^2\right]du \qquad (42) \\ &= 2^k\sqrt{k}\frac{\log r}{2}\frac{1}{\sqrt{2\pi}}\int_{\alpha'}^{\infty}(\alpha'-v)e^{-v^2/2}dv[1+O(k^{-1})], \qquad \alpha' = \frac{2\alpha}{\log r}. \end{aligned}$$

Case F: Saddle $S_* \to \infty$

With (38)-(42) we have one or two terms expansions for $H_k(z)$ as $z \to \infty$ for any $z \in (p^{-k}, q^{-k})$. Since S_* is an explicit function of k and z these results are completely explicit. However when $(\log z)/k \approx -\log q$ the arguments leading to (38) are no longer valid, since $S_* \to +\infty$. Now the factor $\Gamma(s)$ in (5) begins to affect the location of the saddle-point. We thus approximate the integrand for s large. Since $p > q$ and $\zeta(s) = 1 + O(2^{-s})$ as $s \to \infty$, we approximate $\Gamma(s)$ by Stirling's formula and obtain

$$\Gamma(s)\zeta(s+1)(p^{-s} + q^{-s})^k \sim \sqrt{\frac{2\pi}{s}} s^s e^{-s} q^{-sk} \exp(kr^{-s}).$$

Using the above in (5), the saddle-point equation now becomes

$$\frac{d}{ds}[s \log s - s - s \log z - sk \log q + kr^{-s}] = 0$$

or

$$s = zq^k \exp[k(\log r)r^{-s}].$$

We denote the solution to the above as $s = S_0(z,k)$ and then a standard saddle-point calculation yields

$$\begin{aligned} H_k(z) &\sim -\frac{1}{\sqrt{1 + k(\log r)^2 S_0 r^{-S_0}}} \exp[-S_0 + k(1 + S_0 \log r) r^{-S_0}] \qquad (43) \\ &= -\frac{1}{\sqrt{1 + (\log r) S_0 \log[S_0/(zq^k)]}} \exp\left(-S_0 + \left[S_0 + \frac{1}{\log r}\right] \log\left(\frac{S_0}{zq^k}\right)\right). \end{aligned}$$

Case G: $k = O(1)$

Finally, for $z \to \infty$ with $k = O(1)$, we obtain

$$H_k(z) \sim \exp(-zq^k). \qquad (44)$$

This can be obtained either by re-examining (5) in this limit, or as a limiting case of (43).

Up to now we have assumed that $z \to +\infty$ along real values. However we can easily extend the results to complex z by writing $z^{-s} = |z|^{-s} e^{-is\theta}$ where $\theta = \arg(z)$. Then the saddle-point calculations are as before if we replace z by $|z|$ and multiply the final results by $e^{-iS_*\theta}$. This also shows that $H_k(z)$ is maximal as $z \to \infty$ along the positive real axis.

3.3.2 Evaluation of the Cauchy Integral (4)

Having approximated $H_k(z)$ for various ranges of z we proceed to evaluate the integral in (4). The factors $e^z z^{-n}$ suggests there is a saddle-point at $z \approx n$, but we must estimate the size of $\exp[H_k(z)]$ to see to what extent this factor affects the location of the saddle.

From (38) we see that for $k, z \to \infty$ with $(\log z)/k$ fixed, $H_k(z)$ grows exponentially with k and algebraically with z. We thus compare the algebraic growth rate of $H_k(z)$ with $z = z^1$. Solving for z in terms of S_* using (38) yields

$$z = \exp\left(-k\left[\frac{\log p + r^{S_*}\log q}{r^{S_*}+1}\right]\right).$$

Comparing the above to the exponential growth rate of the saddle-point approximation in (38), we see that the former dominates the latter if

$$\frac{-k}{r^{S_*}+1}(\log p + r^{S_*}\log q) > k\left[\log(1+r^{S_*}) - \frac{S_*\log r}{r^{S_*}+1}r^{S_*}\right]. \tag{45}$$

Upon setting $\Omega = r^{S_*}$ and noting that $S_* > -1$ for $\Omega > q/p = 1/r$, we see that (45) holds if

$$\log\left(\frac{1}{p}\right) + \Omega\log\left(\frac{1}{q}\right) + \Omega\log\Omega - (1+\Omega)\log(1+\Omega) > 0. \tag{46}$$

But the derivative of (46) is $\log\{\Omega/[q(1+\Omega)]\}$, which is positive for all $\Omega > q/p$. Also, (46) becomes an equality for $\Omega = q/p$ and hence the inequality holds for $\Omega > q/p$. It follows that (38) grows less rapidly than z, as long as $S_* > -1$.

For $n = A^k$ with $1/\sqrt{pq} < A < 1/q$ we simply apply (the exponential of) (36), since $H_k(z)$ grows like z^ν with $\nu < 1$ in view of (45). We thus obtain part (e) of Theorem 3. We can dilate the contour to the circle $|z| = n$, as the saddle lies in this range. To obtain part (c) of Theorem 3, we use (39) and dilate the contour to $|z| = n - 2^k$. Then we can use (36) with n replaced by $n - 2^k$. The transition range where part (d) applies follows from (42) and (36). We note that in each case our approximation to $H_k(z)$ involves error factors of $1 + O(k^{-1})$ or $1 + O(k^{-1/2})$, and since $k = O(\log n)$, these are larger than the error term in (36), arising from the depoissonization Theorem 4(ii). Indeed we would need to compute the full asymptotic series for $H_k(z)$ before considering the error in (36).

To obtain part (a) of Theorem 3, we use (6) and the analog of (37) for the non-symmetric case. This result also follows by using (7) in (4). Part (g) follows by using (44) and $1 - \exp[H_k(z)] \sim H_k(z)$ in (4). Theorem 3(b) follows by using (41), and 3(f) follows from (43) and (36). Evaluating the Cauchy integral thus corresponds to replacing $S_0(z, k)$ by $\bar{S}(n, k)$. This complete the (sketched) derivation of Theorem 3.

3.4 Proof of Corollary 2

Finally we obtain Corollary 2 by finding where $f(n, k)$ changes from ≈ 1 to ≈ 0. We examine the equation in (22) defining $\bar{S} = \bar{S}(n, k)$. To make $\log f$ (and thus f) in (21)

$O(1)$, we need $\bar{S} \sim nq^k e$. Using this in (22) then yields $1 \sim k(\log r)r^{-\bar{S}}$ or

$$\bar{S} = \frac{1}{\log r}[\log k + \log\log r + o(1)].$$

Equating the above to $nq^k e$ and taking the logarithm, yields

$$1 + k\log q + \log n = \log[\log(k \log r)] - \log\log r + o(1), \;\; n \to \infty$$

and this implies

$$k = \frac{1}{\log(1/q)}[\log n - \log\log\log n + \log(e\log r) + o(1)]. \tag{47}$$

But since k is an integer, (47) cannot be satisfied. Thus we consider those integer values of k that are closest to (47), by setting

$$k = \frac{1}{\log(1/q)}[\log n - \log\log\log n + \log(e\log r)] - \beta_*(n) + L \tag{48}$$

where β_* denotes the fractional part of the first term in the right side of (48), as in Corollary 2. Thus L is an integer and we examine (22).

Using (48) we have

$$nq^k = q^{L-\beta_*}\frac{\log\log n}{e\log r}$$

and then (22) becomes, upon taking logarithms,

$$\log \bar{S} = (L-\beta_*)\log q + \log\log\log n - \log(e\log r) + k(\log r)r^{-\bar{S}}. \tag{49}$$

We solve this equation asymptotically for $n \to \infty$ with $L, \beta_* = O(1)$. If $(L-\beta_*)\log q - 1 > 0$ (i.e., $q^{L-\beta_*}e^{-1} > 1$), the last term in the right side of (49) is asymptotically negligible since $k(\log r)r^{-\bar{S}} = O((\log n)^{1-\nu_0})$ where $\nu_0 = q^{L-\beta_*}e^{-1}/\log r > 1$. Then we have

$$\bar{S} \sim q^{L-\beta_*}\frac{\log\log n}{e\log r}, \;\; n \to \infty. \tag{50}$$

If on the other hand $(L-\beta_*)\log q - 1 < 0$, we have

$$\bar{S} = \frac{1}{\log r}\left(\log\log n + \log\log r - \log\log(1/q) - \log[(\beta_* - L)\log q + 1] + o(1)\right) \tag{51}$$

for $n \to \infty$. With (51) we obtain

$$k(\log r)r^{-\bar{S}} \sim (\beta_* - L)\log q + 1.$$

Since L is an integer, we define L_c by

$$L_c = \left\lfloor \beta_*(n) + \frac{1}{\log q} \right\rfloor. \tag{52}$$

Then (50) holds for all $L \le L_c$ while (51) is valid for $L > L_c$. Since $0 \le \beta_* < 1$ and $q \in \left(0, \frac{1}{2}\right)$, we see from (52) that $\beta_* + (\log q)^{-1} \in \left(-\frac{1}{\log 2}, 1\right)$ and thus L_c can take on only the three values

$$L_c = 0, -1, -2.$$

We use (50)-(52) to simplify (21) for L fixed. For $L > L_c$ we have

$$1 + k(\log^2 r)\bar{S} r^{-\bar{S}} \sim (\log r)\bar{S}[(\beta_* - L)\log q + 1] \sim (\log\log n)[(\beta_* - L)\log q + 1] \tag{53}$$

and

$$\begin{aligned} \bar{S}\log\left(\frac{\bar{S}}{nq^k e}\right) &+ \frac{1}{\log r}\log\left(\frac{\bar{S}}{nq^k}\right) = \bar{S}\log\left[\frac{(\log r)\bar{S}}{q^{L-\beta_*}(\log\log n)}\right] + \frac{1}{\log r} \\ &= \frac{1}{\log r}(\beta_* - L)(\log q)\log\log n + \frac{1 + (\beta_* - L)\log q}{\log r}(\log(e\log r) \\ &- \log\log(1/q) - \log[(\beta_* - L)\log q + 1]) + o(1). \end{aligned} \tag{54}$$

Upon using (53) and (54) in (21) we obtain the expression in Corollary 2 valid for $L > L_c$.

For $L \le L_c$, (50) holds. Then from (49) we can refine the estimate (50) to

$$\bar{S} = \frac{q^{L-\beta_*}e^{-1}}{\log r}\log\log n[1 + \varepsilon] \tag{55}$$

where

$$\varepsilon \sim \frac{\log r}{\log(1/q)}(\log n)^{1 - e^{-1}q^{L-\beta_*}}.$$

We note that $\varepsilon = o(1)$ for $L \le L_c$, and that $\log[\bar{S}/(nq^k)] \sim \varepsilon$. With (55) we have $1 + k(\log^2 r)\bar{S} r^{-\bar{S}} \sim 1$ and

$$1 - f(n,k) \sim e^{-\bar{S}} \sim \left(\frac{1}{\log n}\right)^{\frac{q^{L-\beta_*}e^{-1}}{\log r}}, \; L \le L_c. \tag{56}$$

We have thus obtained (25) in Corollary 2.

From Corollary 2 we also observe that if $L_c = -2$ we have $f(n,k) \sim 0$, if $L = L_c + 3 = 1$ and $f(n,k) \sim 1$, if $L = L_c + 2 = 0$. Thus all the mass accumulates at $L = 0$. The same conclusion holds if $L_c = -1$ and $L_c = 0$. If $L_c = 0$, we can use (56) to conclude that $f(n,k) \sim 1$, if $L = 0$ and (24) to conclude that $f(n,k) \sim 0$ if $L = 1$.

4 Numerical Results

We test the numerical accuracy of the results in Theorems 2 and 3. We first consider the symmetric case $p = q = 1/2$ and we denote by f the exact (numerical) values, and by $f^{(a)}$,

Table 1:

$n = 1024$, $p = q = 0.5$

k	f	$f^{(a)}$	$f^{(b)}$	$f^{(c)}$
6	.9999936521	–	.9999942135	.9999936524
7	.95918	–	.95919	.95838
8	.7584(–2)	–	.7585(–2)	< 0
9	.1810(–39)	.7457(–154)	.1810(–39)	–
10	.1537(–442)	.1537(–442)	–	–

$f^{(b)}$, and $f^{(c)}$ the asymptotic approximations in parts (a), (b), and (c) of Theorem 2. For part (a), we use the first formula in (9).

Table 1 has $n = 1024$ and our results predict that the probability will be concentrated at $k = \lfloor \log_2(n/\log n) \rfloor = 7$. Table 1 shows that indeed about 95% of the mass is at this point (cf. also Figure 2). We compare asymptotic and numerical results for $k = 6 - 10$, and the table clearly shows the necessity of having three expansions for the three ranges. The saddle-point approximation $f^{(b)}$ is the most robust, but it is not defined at $n2^{-k} = 1$, and is surpassed in accuracy by $f^{(c)}$ for $k \le 6$.

The three asymptotic ranges are easier to see by fixing k and varying n. In Table 2 we take $k = 8$ and give f, $f^{(a)}$, and $f^{(b)}$ for various n in the range [256,1000]. We see that already at $n = 260$ the saddle-point approximation becomes superior to $f^{(a)}$, which applies for $n - 2^k = \ell = O(1)$. Even at $n = 1000$, f is still fairly small ($\approx .0046$). Corollary 1 shows that the mass concentrates at $L = 0$, which corresponds to $n \approx k2^k(\log 2)2^{\beta}$, where $\beta \in [0, 1)$. When $k = 8$, this suggests that there should be appreciable mass for n in the range [1420,2839], and this window is outside of the range of Table 2. We therefore decrease k to 4 and Table 3 considers n in the range [16,200]. Now the mass should be in the range $[k2^k \log 2, k2^{k+1} \log 2] \approx [44, 89]$, which is well within the range of the table. Table 3 clearly shows the transition between the three expansions in Theorem 2.

Next we consider the non-symmetric case $p > q$. Now $f^{(a)}$ — $f^{(g)}$ will denote the asymptotic formulas in the seven parts of Theorem 2. Our analysis showed that we have the leading term for f for case (a), and for $1 - f$ for case (g). However, for cases (b)-(f) we only have the leading term for $\log(f)$, in view of the error terms in the various exponents in Theorem 3. Thus we would expect our approximations for f to be much less accurate

Table 2:

$k = 8$, $p = q = 0.5$

n	f	$f^{(a)}$	$f^{(b)}$
256	.2654(–109)	.2653(–109)	–
257	.3410(–107)	.3396(–107)	.3697(–107)
258	.2202(–105)	.2173(–105)	.2295(–105)
259	.9533(–104)	.9274(–104)	.9798(–104)
260	.3110(–102)	.2967(–102)	.3174(–102)
300	.6514(–69)	–	.6524(–69)
400	.1434(–35)	–	.1434(–35)
500	.9218(–21)	–	.9220(–21)
600	.9775(–13)	–	.9777(–13)
700	.5172(–8)	–	.5172(–8)
800	.4208(–5)	–	.4209(–5)
900	.2947(–3)	–	.2947(–3)
1000	.4589(–2)	–	.4590(–2)

Table 3:

$k = 4,\ p = q = 0.5$

n	f	$f^{(a)}$	$f^{(b)}$	$f^{(c)}$
16	.1134(–5)	.1128(–5)	–	
17	.9640(–5)	.9026(–5)	.1040(–4)	
18	.4428(–4)	.3610(–4)	.4596(–4)	
19	.1459(–4)	.9628(–5)	.1495(–4)	
20	.3868(–3)	.1925(–3)	.3936(–3)	
30	.4629(–1)	–	.4650(–1)	
40	.2428	–	.2435	< 0
50	.4999	–	.5009	.3651
60	.7046	–	.7057	.6670
70	.8355	–	.8364	.8253
80	.9111	–	.9118	.9084
90	.95268	–	.95318	.95196
100	.97499	–	.97534	.97480
150	.99900067	–	.99904084	.99900043
200	.99996033	–	.99996362	.99996033

Table 4:
$p = 0.75$, $q = 0.25$

		f	$f^{(a)}$
$k = 4$	$n = 16$	.1139(–9)	.1139(–9)
	17	.9684(–9)	.9684(–9)
	18	.4579(–8)	.4258(–8)
	19	.1590(–7)	.1380(–7)
	20	.4524(–7)	.3450(–7)
$k = 8$	$k = 256$	.3066(–237)	.3066(–237)
	257	.3940(–235)	.3940(–235)
	258	.2561(–233)	.2541(–233)
	259	.1122(–231)	.1097(–231)
	260	.3732(–230)	.3566(–230)

than was the case when $p = q$. We take

$$p = .75 \quad , \quad q = .25$$

so that

$$p^{-p}q^{-q} = 1.7547\ldots, \quad 1/\sqrt{pq} = 2.3094\ldots, \quad \text{and} \quad 1/q = 4.0. \tag{57}$$

We computed numerically $f(n, k)$ over the range $n \leq 1024$ and $k \leq 10$.

In Table 4 we test the accuracy of Theorem 3(a) using the second formula in (15). Note that in view of the formulas derived at the beginning of Section 2, this is not only asymptotic but exact, when $n = 2^k$ and $n = 2^k + 1$. We consider both $k = 4$ and $k = 8$ with $\ell = n - 2^k$ in the range [0,5]. Table 4 shows that $f^{(a)}$ is an excellent approximation to f, though the probabilities are extremely small in this range, especially when $k = 8$.

In Table 5 we compare $1 - f$ to $1 - f^{(g)}$ for $k = 4$ and n in the range [400,1000]. The approximation is quite good, even when f is not particularly close to 1, which is certainly true for $n = 400$. For this case the error is about 13%.

Next we test the accuracy of the approximation in part (f) in Theorem 3. We recall that this applies in the range where f undergoes the critical transition from $f \approx 0$ to $f \approx 1$. In Table 6 we take $n = 100$ and $n = 1024$, and compare f to $f^{(f)}$ for the few values of k where there is appreciable mass. We see that when $n = 100$, about 70% of the mass concentrates at $k = 3$. Our analysis shows that the mass should be at $L = 0$ which corresponds to

Table 5:

$k = 4$, $p = 0.75$, $q = 0.25$		
n	$1 - f$	$1 - f^{(g)}$
400	.2370	.2089
500	.1507	.1412
600	.09860	.09552
700	.06556	.06458
800	.04397	.04366
900	.02962	.02952
1000	.01999	.01996

$k = \lfloor (\log_{1/q}(ne \log r / \log \log n) \rfloor$, and this is equal to $k = 3$ (resp., $k = 5$) for $n = 100$ (resp., $n = 1024$). We also have $\beta_*(n) = .805\ldots$ for $n = 100$ and $\beta_*(n) = .312\ldots$ for $n = 1024$. Table 6 shows that when $n = 1024$, there is actually about comparable mass at $k = 4$ and $k = 5$, and thus there is less mass concentration then is the case when $n = 100$. However, this can be explained by the inherent oscillations due to the fractional part $\beta_*(n)$. When $n = 1024$, β_* is much smaller then when $n = 100$, and $\beta_*(n) \to 0$ as $n \to \infty$ leads to mass at two successive points. Table 6 also shows that $f^{(f)}$ becomes a poor approximation if f is close to 0 or 1, and this is predicted by Theorem 3, which suggests that $f^{(a)-(e)}$ or $f^{(g)}$ should then provide better approximations.

To see the various ranges in Theorem 3, it is preferable to fix k and increase n. Then we are guaranteed that the conditions under which the various parts of the theorem apply will indeed be satisfied. If we fix a large n and vary k (from 1 to $\lfloor \log_2 n \rfloor$), then we may never hit some of the cases in Theorem 3. In Table 7, we take $k = 4$ and compare the exact probabilities to the asymptotic result in part (f), and also to that in part (g). We see that $f^{(f)}$ is quite accurate, especially if we consider the fact (in view of the error factor $1 + (\bar{S}^{-1})$ in the exponent in (21)) that Theorem 3, part (f) gives only the leading term for $\log(f)$. We also see that $f^{(g)}$ becomes a better approximation that $f^{(f)}$ for n exceeding about 450.

We proceed to consider parts (b)-(e) of Theorem 3. Here we compare $-\log(f)$ to the asymptotic results. In view of (57) for $k = 8$, the asymptotic result $f^{(c)}$ should apply for $n \in (2^k + (p^{-p}q^{-q})^k,\ 2^k + (pq)^{-k/2}) \approx [346, 1065]$, while $f^{(e)}$ should hold for $n \in$

Table 6:

$p = 0.75,\ q = 0.25$

		f	$f^{(f)}$
$n = 100$	$k = 2$	.998425	.998082
	3	.7731	.7823
	4	.05940	.1202
$n = 1024$	$k = 3$	.9999999005	.9999999975
	4	.9818	.9827
	5	.4894	.5445
	6	.0008498	.004492

Table 7:

$k = 4,\ p = 0.75,\ q = 0.25$

n	f	$f^{(f)}$	$f^{(g)}$
260	.5261	.5799	.6385
270	.5496	.6000	.6524
280	.5719	.6192	.6657
290	.5931	.6374	.6785
300	.6132	.6546	.6909
400	.7629	.7845	.7910
500	.8492	.8618	.8587
600	.9013	.9093	.9044
700	.9344	.9395	.9354
800	.9560	.9591	.9563
900	.9703	.9722	.9704
1000	.9800	.9810	.9800

Table 8:

$k = 8$, $p = 0.75$, $q = 0.25$

n	b	$-\log(f)$	$-\log f^{(b)}$
260	–1.10	530.5	530.5
300	–.252	449.8	443.5
340	–.0239	403.1	385.0
380	.113	368.8	338.1
420	.212	341.7	298.4
460	.289	319.3	264.0
500	.353	300.4	233.4
540	.406	284.1	206.0
580	.453	269.8	181.1

$((pq)^{-k/2}, q^{-k}) \approx [809, 65536]$. In Table 8, we compare the transition result in Theorem 3, part (b), to exact values of $-\log(f)$. Since this result holds for "$b = O(1)$", we also give the values of $b = k^{-1/2}\log[(n-2^k)p^{pk}q^{qk}]$. We see that for $b < 0$, the approximation is quite accurate, but quickly becomes worse as b increases through positive values. We note that $b = 0$ corresponds to $n \approx 346$. The asymptotics become worse for increasing b, partly because we are moving out of region (b) and into region (c), and partly because as $b \to -\infty$, we are moving into the asymptotic matching region between regions (b) and (a), and in the latter region we have not only the leading term for $\log(f)$, but also that for f itself.

In Table 9 we compare $-\log(f)$ to $\log(f^{(c)})$, also giving the values of $S_*(B)$. We recall that $S_*(B)$ gave the location of the saddle-point and that Theorem 3(c) applies only for $S_*(B) \in (-1, 0)$. From the table we see that the approximation is poor for all values of n. The best agreement occurs when $n = 400$ and even here $-\log(f^{(c)})$ over-estimates the true value by 25%. While clearly $-\log(f)$ is a decreasing function of n for a fixed k, Table 9 shows that the asymptotic approximation reaches a minimum value at about $n = 450$, and quickly becomes large as n becomes such that $S_*(B)$ gets close to –1 or to 0. This behavior is certainly consistent with the asymptotic analysis, since as $n \downarrow 346$ ($S_* \downarrow -1$), the saddle coalesces with a simple pole and as $n \uparrow 1065$ ($S_* \uparrow 0$), the saddle coalesces with a double pole. In the former limit $f^{(c)}$ transitions to $f^{(b)}$ (see also Table 8) and in the latter limit $f^{(c)}$ transitions to $f^{(d)}$.

Next we consider the accuracy of $-\log(f^{(d)})$, which applies for $a = O(1)$, where $a =$

Table 9:

$k = 8,\ p = 0.75,\ q = 0.25$

n	$S_*(B)$	$-\log(f)$	$-\log f^{(c)}$
350	–.975	393.6	944.2
400	–.755	354.5	438.5
450	–.613	324.6	428.5
500	–.509	300.4	458.9
550	–.427	280.4	526.9
600	–.358	263.2	643.4
650	–.300	248.8	830.3
700	–.250	235.4	1128
750	–.205	223.8	1614
800	–.164	213.5	2449

$k^{-1/2}\log[n(pq)^{k/2}]$. In Table 10 we consider again $k = 8$, and n in the range [700,1000]. The center of the transition region has $a = 0$ and thus $n \approx 809$. The table shows that $-\log(f^{(d)})$ tends to under-estimate the true value by about 25%, which is consistent with an $O(k^{-1})$ error term in the exponent in (19). By comparing $n = 700$, 750 and 800 in Tables 9 and 10, we clearly see the importance of constructing the transition region expansion for $a = O(1)$. We also note that the numerical values of $|a|$ throughout Table 10 are quite small, which shows that the transition region expansion applies over a large range.

Since we computed f only for $n \leq 1024$, we cannot fully see the range ([809,65536]) where part (e) of Theorem 3 applies when $k = 8$. We thus decrease k to 6 and in Table 11, we compare $-\log(f^{(e)})$ and also $-\log(f^{(d)})$ to the exact values. Now the interval $[(pq)^{-k/2}, q^{-k}]$ corresponds to [152,4096]. We see that the transition range approximation in Theorem 3(d) under-estimates the true value, while the approximation in part (e) over-estimates it. Even when $n = 1000$, we have $a = .769$ which is "O(1)" and thus once again the transition region is quite "thick". When $n = 152$, the approximation arising from part (e) over-estimates the true value of $-\log(f)$ by a factor of about 10^6. This is because the saddle $S_*(A) = .00107$ is very close to the double pole of the integrand in (5) at $s = 0$. As n increases, we see from Table 11 that the approximation in part (e) slowly becomes superior to that in part (d).

To summarize, the asymptotics are very accurate in the symmetric case. In the non-

Table 10:

$k = 8$, $p = 0.75$, $q = 0.25$

n	a	$-\log(f)$	$-\log(f^{(d)})$
700	–.0512	235.4	177.9
750	–.0268	223.8	168.5
800	–.00399	213.5	160.1
850	.0174	204.1	152.4
900	.0376	195.6	145.4
950	.0567	187.8	138.9
1000	.0749	180.7	133.0

Table 11:

$k = 6$, $p = 0.75$, $q = 0.25$

n	a	$S_*(A)$	$-\log(f)$	$-\log(f^{(d)})$	$-\log(f^{(e)})$
152	.000796	.00107	49.84	34.29	.1633(8)
200	.112	.153	39.35	26.23	808.9
300	.278	.382	27.26	16.85	127.6
400	.395	.551	20.59	11.88	59.67
500	.486	.690	16.35	8.866	37.12
600	.561	.808	13.41	6.880	26.29
700	.624	.914	11.25	5.495	20.05
800	.678	1.01	9.618	4.487	16.04
900	.726	1.10	8.328	3.730	13.28
1000	.769	1.18	7.289	3.146	11.26

symmetric case, parts (a) and (g) of Theorem 3 lead to accurate approximations to f and $1-f$, respectively. The part (f) result, which applies where most of the mass concentrates, is quite robust, especially considering the fact that we only have the leading term for $\log(f)$. However, Tables 8-11 show that the approximations for cases (b)-(e) are not very accurate when $k = 6$ and $k = 8$, even as approximations to $\log(f)$. Nevertheless, the asymptotic analysis leading to Theorem 3 is able to explain some of the trends in the numerical data.

References

[1] L. Devroye, A Note on the Probabilistic Analysis of Patricia Tries, *Random Structures & Algorithms*, 3, 203–214, 1992.

[2] L. Devroye, Analysis of Random LC Tries, *Random Structures & Algorithms*, 19, 359–375, 2001.

[3] P. Flajolet, On the Performance Evaluation of Extendible Hashing and Trie Searching, *Acta Informatica*, 20, 345-369, 1983.

[4] P. Jacquet and M. Régnier, Trie Partitioning Process: Limiting Distributions, Lecture Notes in Computer Science, **214**, 196-210, Springer Verlag, New York 1986.

[5] P. Jacquet, and W. Szpankowski, Analytical Depoissonization and Its Applications, *Theoretical Computer Science*, 201, 1–62, 1998.

[6] C. Knessl, A Note on the Asymptotic Behavior of the Depth of Tries, *Algorithmica*, 22, 547–560, 1998.

[7] C. Knessl and W. Szpankowski, Asymptotic Behavior of the Height in a Digital Search Tree and the Longest Phrase of the Lempel-Ziv Scheme, *SIAM J. Computing*, 30, 923-964, 2000.

[8] C. Knessl and W. Szpankowski, Limit Laws for Heights in PATRICIA Tries, *J. Algorithms*, 44, 63-97, 2002.

[9] D. E. Knuth, *Selected Papers on Analysis of Algorithms*, CSLI, Stanford, 2000.

[10] H. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.

[11] S. Nilsson, *Radix Sorting & Searching*, PhD Thesis, Lund University, 1996.

[12] B. Pittel, Paths in a Random Digital Tree: Limiting Distributions, *Adv. in Applied Probability*, 18, 139–155, 1986.

[13] Y. Reznik, Some Results on Tries with Adaptive Branching, *Theoretical Computer Science*, 289, 1009-1026, 2002.

[14] V. Srinivasan and G. Varghese, Fast Address Lookups using Controlled Prefix Expansions, *ACM SIGMETRICS'98*, 1998.

[15] W. Szpankowski *Average Case Analysis of Algorithms on Sequences*, John Wiley, New York, 2001.

[16] R. Wong, *Asymptotic Approximations of Integrals*, Academic Press, Boston, 1989.