ANALYSIS OF A SPLITTING PROCESS ARISING IN PROBABILISTIC COUNTING AND OTHER RELATED ALGORITHMS *

July 23, 1996

Peter Kirschenhofer
Dept. Algebra&Discrete Math.
Technical University of Vienna
A-1040 Vienna
Austria

Helmut Prodinger
Dept. Algebra&Discrete Math.
Technical University of Vienna
A-1040 Vienna
Austria

Wojciech Szpankowski†
Dept. Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.

Abstract

We present an analytical method of analyzing a class of "splitting algorithms" that include probabilistic counting, selecting the leader, estimating the number of questions necessary to identify distinct objects, searching algorithms based on digital tries, approximate counting, and so forth. In our discussion we concentrate on the analysis of a generalized probabilistic counting algorithm. Our technique belongs to the toolkit of the analytical analysis of algorithms, and it involves solutions of functional equations, analytical poissonization and depoissonization as well as Mellin transform. In particular, we deal with an instance of the functional equation $g(z) = \beta a(z)g(z/2) + b(z)$ where $a(z)$ and $b(z)$ are given functions, and $\beta < 1$ is a constant. With respect to our generalized probabilistic counting algorithm, we obtain asymptotic expansions of the first two moments of an estimate of the cardinality of a set that is computed by the algorithm. We also derive the *asymptotic* distribution of this estimate, and observe that it actually fluctuates, leading to a conclusion that its *limiting* distribution does not exist.

*A preliminary version of parts of this paper was presented in *ICALP 92*, Vienna [26]. This work was supported by NSF Grants INT-8912631 and CCR-9201078.

†This author was additionally was supported by NSF Grants NCR-9206315 and NCR-9415491, and in part by NATO Collaborative Grant CRG.950060.

1. INTRODUCTION

In several database algorithms a major determinant of efficiency is the *cardinality* of the underlying set. Therefore, one needs an efficient way of estimating the cardinality of large (multi)sets of data. Clearly, the trivial method of building a list of elements without replications is unacceptable due to the cost of disk access and auxiliary memory. Knowing this fact, Flajolet and Martin [8] proposed an algorithm that is probabilistic in its nature. It works as follows: In order to estimate the cardinality N of a set $\mathcal{M}$ (with replications) every element $x \in \mathcal{M}$ is hashed into a binary string of size m (the choice of m is easy, and $m = 5 + \log N$ suffices [4]). The bitwise OR-composition of the modified hashed strings in which only the least significant (rightmost) "1" survives is used to build the so called `bitmap` and to obtain the estimate R_N of the cardinality N of $\mathcal{M}$. More precisely, the position R_N of the leftmost zero in the `bitmap` string (where we start enumerating positions with 0) approximates $\log_2 N$ as shown by Flajolet and Martin in [8] (cf. also [4]). Observe that, equivalently, R_N is the length of the longest run of ones at the beginning of the bitmap string.

In fact, probabilistic counting is a special case of a *splitting process* that also includes such algorithms as selecting the loser [7, 31], estimating the number of questions necessary to identify distinct objects [30], searching algorithms based on digital tries [17, 18, 19, 26, 28, 32, 36], approximate counting [24], conflict resolution algorithms for multiaccess communication [18, 34, 35] and so forth. Using a digital tree representation one can describe the probabilistic counting in terms of this splitting process as follows: Imagine N persons flipping a fair coin, and those who get 1 (a hit) discontinue throwing (move to the right in the digital tree representation) while the rest continue throwing (i.e., move to the left in the digital tree) until they get a hit. The process continues until all remaining persons flip a 0. It should be clear that the number of rounds in the flipping procedure equals one plus the quantity R_N from above.

Let us now consider the following *more general version* of the splitting process, where the coin flipping ends as soon as at most d persons have flipped a 1 in the same round, where d is a given parameter. This time we denote the number of rounds by $1 + R_{N,d}$. Observe that $d = 0$ corresponds to the original situation. Again we will expect $R_{N,d}$ to be an estimate of $\log_2 N$. We should point out that a similar extension of the primary leader election algorithm was recently considered by Grabner [14].

Returning to the probabilistic counting algorithm: We can interpret the generalized situation as follows: We start from an empty `bitmap` string, that is, with all positions filled by zeros. We further assume that N different objects (e.g., data, persons, etc.) can randomly insert (hit) a 1 at any position of the `bitmap`, however, the probability of hitting the $j \geq 0$ position is equal to 2^{-j-1}. (In terms of the modified hashed strings discussed above this means that the probability of the occurrence of the pattern like $0^j 1 \cdots$ is equal to 2^{-j-1} since 0 and 1 are equally likely.) Every object can hit only one time. We count the number of hits in any position of the `bitmap`, but we count the number of hits only up to the value $d+1$, where d is a given parameter. In other words, the `bitmap` is a $d+2$-ary string now. Clearly, the `bitmap` will contain many $d+1$'s at the beginning of the string. This is due to the fact that the probability of a hit decreases exponentially fast with the increase of the position in the `bitmap`. It is easy to see that the length of the longest run of $d+1$'s in the front of the `bitmap` equals $R_{N,d}$ defined above. More precisely, we have:

$$R_{N,d} = \min\{k : \ \texttt{bitmap}(k) < d+1 \text{ and for all } 0 \leq i < k \ \ \texttt{bitmap}(i) = d+1\} \ . \tag{1}$$

Observe that this time replications in the object set are not allowable since they might influence the result of the bitmap procedure if $d \geq 1$.

For example, for $d = 3$ we may have $\texttt{bitmap} = 44444444444434100000$. As the estimate $R_{N,d}$ we take the position of the leftmost value smaller than $d+1$ (e.g., in the example above we have $R_{N,3} = 11$).

In the present paper we analyze the probabilistic behavior of the parameter $R_{N,d}$. In particular, we prove that $ER_{N,d} = \log_2 N - 1 - C_d + P(\log_2 N) + O(N^{-\frac{1}{2}+\varepsilon})$ where C_d is a constant such that $C_d = \log_2 d - \frac{1}{2} + O(1/d)$ as $d \to \infty$, and $P(x)$ is a periodic function of period 1, mean zero and small amplitude (cf. Theorem 1(i)). More importantly, we show that the variance of $R_{N,d}$ behaves for large N like a periodic function in $\log_2 N$, the mean of which is asymptotically equal to $\frac{1}{\sqrt{\pi d}}$ as $d \to \infty$ (cf. Theorem 1(ii)). Numerical evaluation of the variance for small values of d is presented in Table 1 (e.g., the case $d = 0$ coincides with the value derived in [8]). Finally, we derive the asymptotic distribution of $R_{N,d}$ which confirms our believes that the estimate $R_{N,d}$ is well concentrated around its mean (cf. Theorem 2). But, we observe that the limiting distribution of $R_{N,d}$ does not exist due to a tiny oscillation of the asymptotic distribution. A preliminary version of some of these results can be found in [26].

It must be stressed that the novelty of our paper also lies in the area of the mathematical analysis of algorithms, and we think of the splitting process (in particular, the probabilistic counting) as a motivation for a more general study. In fact, the proposed technique of analysis can be applied to several other algorithms. As indicated above, the study of the splitting process can be reduced to a problem on *digital trees*. Evaluating a parameter of such a tree, we obtain the following functional equation

$$g(z) = \beta a(z/2)g(z/2) + b(z) \ . \qquad (2)$$

In general, this equation does not have an explicit finite solution unless $a(z)$ is a very particular function. For example, $a(z) = e^z$ was discussed in Knuth [28] (cf. also [36]), Flajolet and Saheb [10], and Szpankowski [35], [36] (cf. also [9, 28]) investigated the case $a(z) = e^z + 1$. For these two cases explicit finite solutions of (2) were obtained, but it is rather hopeless to expect the existence of a useful closed form solution for general $a(z)$ and $b(z)$. An asymptotic solution to (2) was suggested in Jacquet and Szpankowski [18] who analyzed the case $a(z) = 1 + (1+z)e^{-z}$. More recently, Flajolet and Richmond [11] investigated asymptotically the above equation with $a(z) = 2z^b/(1+z)^b$ for some integer $b > 0$. In the analysis of the algorithms introduced in this paper, we must investigate among others the equation (2) with $a(z) = 1 - e_d(z)e^{-z}$ where $e_d(z) = \sum_{i=0}^{d} z^i/i!$. Even more complicated functions for $a(z)$ and $b(z)$ are involved in the analysis of the variance and the asymptotic distribution (cf. Sections 2 and 3).

As mentioned above, the case $d = 0$ was analyzed by Flajolet and Martin [8]. In addition, Greenberg *et al.* [15] investigated the instance $d = 1$ in the analysis of a conflict resolution algorithm for multiple channels. It must be stressed that the analysis of Flajolet and Martin [8] is based on the inclusion-exclusion rule, and seems to be very unlikely extendable to $d \geq 1$. This is partially evident from the analysis in [15].

The plan for the present paper is as follows: In the next section we present our main results concerning the generalized counting algorithm and the associated splitting process, and their algorithmic consequences. The proofs are delayed till Section 3 where we use *analytical poissonization* and *analytical depoissonization* to solve first a poissonized version of the problem, and then depoissonize it to find a solution of the original problem. This approach turned to be very useful in many other situations encountered in the analysis of algorithms (cf. [7, 12, 18, 33, 32]).

2. MAIN RESULTS

Let $F_N(u) = Eu^{R_{N,d}}$ be the probability generating function of the estimator $R_{N,d}$. That is, $[u^j]F_N(u)$ (the jth coefficient at $F_N(u)$) is the probability that either the splitting process terminates after $j+1$ rounds or that the result of the `bitmap` is $(d+1)^j c \ldots$ where $c \leq d$, and $(d+1)^j$ denotes j consecutive values "$d+1$". To derive the recurrence for $F_N(u)$ we observe: If at the first step k persons flip 1, then either the contribution to the generating function $F_N(u)$ is $u^0 \sum_{k \leq d} \binom{N}{k} 2^{-N}$ for $k \leq d$, or the contribution becomes

$u \sum_{d+1 \le k \le N} \binom{N}{k} 2^{-N} F_{N-k}(u)$ for $k > d$. In short, we obtain the following recurrence

$$F_N(u) = \sum_{k=0}^{d} \binom{N}{k} 2^{-N} + u \sum_{k=0}^{N-d-1} \binom{N}{k} 2^{-N} F_k(u) \ , \tag{3}$$

which is valid for all $N \ge 0$.

Using (3) we can derive recurrences for the first and second factorial moments of $R_{N,d}$, namely: $ER_{N,d} = F'_N(1)$ and $ER_{N,d}(R_{N,d}-1) = F''_N(1)$. Observe that Var $R_{N,d} = F''_N(1) + F'_N(1) - (F'_N(1))^2$. In particular, let $L(z)$ and $W(z)$ be the exponential generating function for $F'_N(1)$ and $F''_N(1)$, respectively. That is,

$$\begin{aligned} L(z) &= \sum_{N \ge 0} F'_N(1) z^N / N! \ , \\ W(z) &= \sum_{N \ge 0} F''_N(1) z^N / N! \ . \end{aligned}$$

Define also $\widetilde{L}(z) = e^{-z} L(z)$ and $\widetilde{W}(z) = e^{-z} W(z)$ as the *Poisson generating functions.* These Poisson generating functions represent the first two moments in a variation of the probabilistic counting problem in which the number of people (objects) is not fixed but is a random variable distributed according to a Poisson law with mean z. Then, after some algebra (3) implies

$$\widetilde{L}(z) = f_d(z/2) \widetilde{L}(z/2) + f_d(z/2) \ , \tag{4}$$

$$\widetilde{W}(z) = f_d(z/2) \widetilde{W}(z/2) + 2 \widetilde{L}(z/2) f_d(z/2) \tag{5}$$

where

$$f_d(z) = 1 - e_d(z) e^{-z} \qquad \text{and} \qquad e_d(z) = 1 + \frac{z^1}{1!} + \cdots + \frac{z^d}{d!} \ . \tag{6}$$

The above functional equations are solved in the next section (cf. Section 3.1). Their solutions are expressed in terms of the following function that is needed to articulate our main results below:

$$\varphi(x) = \prod_{j=0}^{\infty} f_d(x 2^j) = \prod_{j=0}^{\infty} \left(1 - e_d(x 2^j) e^{-x 2^j}\right) \ . \tag{7}$$

We now can present our first main result which is proved in Section 3.2.

Theorem 1. *Consider the generalized splitting algorithms described above. Then the estimator $R_{N,d}$ behaves as follows:*

(i) *The average value of $R_{N,d}$ becomes asymptotically as $N \to \infty$*

$$ER_{N,d} = \log_2 N - 1 - C_d + P_1(\log_2 N) + O\left(N^{-1/2+\varepsilon}\right) \tag{8}$$

for any $\varepsilon > 0$ *with*

$$\begin{aligned} C_d &= \frac{1}{L^2}\int_0^\infty e^{-x} e_d(x)\varphi(2x)\frac{\log x}{x}dx \\ &= \log_2 d - \frac{1}{2} + O\left(\frac{1}{d}\right) \qquad \text{as} \quad d \to \infty \end{aligned} \tag{9}$$

where $L = \log 2$ *and the function* $\varphi(x)$ *is defined in (7). Furthermore, with* $\chi_k = 2k\pi i/L$,

$$P_1(x) = -\frac{1}{L^2}\sum_{k\neq 0} \Phi^{*\prime}(-\chi_k)e^{2\pi i k x} \tag{10}$$

where $\Phi^*(s)$ *is the Mellin transform of* $\varphi(2x) - \varphi(x)$ *and* $\Phi^{*\prime}(s)$ *is the first derivative of* $\Phi(s)$ *(cf. Section 3 for details). In particular,* $P_1(x)$ *is a periodic function of period 1, mean zero and small amplitude for small values of* d.

(ii) *The variance of* $R_{N,d}$ *is*

$$\text{Var } R_{N,d} = D_d - C_d^2 - [P_1^2]_0 + P_2(\log_2 N) + O\left(N^{-1/2+\varepsilon}\right) \tag{11}$$

for any $\varepsilon > 0$*, where*

$$D_d - C_d^2 - \left[P_1^2\right]_0 = \frac{1}{\sqrt{\pi d}} + O\left(\frac{1}{d}\right) \quad \text{for} \qquad d \to \infty \ , \tag{12}$$

and $P_2(x)$ *is a periodic function with period 1 and mean zero. More precisely*

$$\begin{aligned} D_d &= \frac{1}{L^3}\int_0^\infty e_d(x)e^{-x}\varphi(2x)\frac{\log^2 x}{x}dx \\ &= -C_d - 1/6 + \log_2^2 d + O\left(\frac{1}{d}\right) \qquad \text{as} \qquad d \to \infty \ , \end{aligned} \tag{13}$$

so that

$$D_d - C_d^2 = \frac{1}{12} + O\left(\frac{1}{d}\right) \qquad \text{as} \quad d \to \infty \ . \tag{14}$$

The quantity $[P_1^2]_0$ *is the (integral) mean of the square of* $P_1(x)$*, and*

$$[P_1^2]_0 = \frac{1}{12} - \frac{1}{\sqrt{\pi d}} + O\left(\frac{1}{d}\right) \tag{15}$$

for $d \to \infty$.

Using numerical integration, we computed for some values of d the constant terms (for $N \to \infty$) in $ER_{N,d} - \log_2 N - P_1(\log_2 N)$ and Var $R_{N,d} - P_2(\log_2 N)$. They are displayed in Table 1. We note that the variance initially rapidly decreases (for $d \leq 2$) and then the decrease slows down. Observe that the case $d = 0$ coincides with the value derived in [8].

As an immediate consequence of Theorem 1, we observe that $R_{N,d}$ converges *in probability* (pr.) to $ER_{N,d}$. Indeed, by Chebyshev's inequality we have

$$\Pr\{|R_{N,d}/ER_{N,d}-1|>\varepsilon\} \le \frac{\mathrm{Var}\ R_{N,d}}{\varepsilon^2 E^2 R_{N,d}} = O\left(\frac{1}{\log^2 N}\right) \to 0$$

as $N\to\infty$. This proves the announced convergence, but does not warrant the almost sure (a.s.) convergence since the *Borel-Cantelli Lemma* cannot be applied. However, observe that $R_{N,d}$ is a nondecreasing function of N, in the sense that on every sample path we have $R_{N+1,d}\ge R_{N,d}$. To prove the (a.s.) convergence one should apply the Borel-Cantelli Lemma along a doubly exponentially increasing sequence such as $N=2^{2^k}$ (cf. [21]), thus $R_{N,d}/\log_2 N \to 1$ (a.s.), too. Indeed, to see it, let us choose integers s and k such that $s2^{2^k}\le N\le (s+1)2^{2^k}$. Then,

$$\limsup_{n\to\infty}\frac{R_{N,d}}{\log_2 N} \le \limsup_{k,s\to\infty}\frac{R_{(s+1)2^{2^k},d}}{\log_2(s+1)2^{2^k}}\,\frac{\log(s+1)2^{2^k}}{\log s2^{2^k}} = 1 \quad \text{(a.s.)}\ .$$

In a similar manner, we can bound $\liminf$ due to monotonicity of $R_{N,d}$, and one obtains almost sure convergence of $R_{N,d}/\log_2 N$ along $N=2^{2^k}$.

Our technique also allows to deal with the asymptotic distribution of $R_{N,d}$. For this, we introduce the Poisson transform of $F_N(z)$ as

$$\widetilde{G}(z,u)=e^{-z}\sum_{N=0}^{\infty}F_N(u)\frac{z^N}{N!}\ . \tag{16}$$

Observe that $\widetilde{G}(z,u)$ can be interpreted as the generating function of $R_{N,d}$ when the cardinality of a set is Poisson distributed with mean z.

The Poisson generating function $\widetilde{G}(z,u)$ satisfies the following functional equation

$$\widetilde{G}(z,u)=uf_d(z/2)\widetilde{G}(z/2,u)+(u-1)(f_d(z/2)-1)\ . \tag{17}$$

Using the technique from Section 3.1, we can solve it for $z\to\infty$, and then by depoissonization we prove in Section 3.3 our second main finding.

Theorem 2. *Let t be a fixed real number. Then, for large N the asymptotic distribution of $R_{N,d}$ becomes (d fixed)*

$$\Pr\{R_{N,d}\le \log_2 N+t-1\} = 1-\varphi\left(2^{-t-\{\log_2 N+t\}}\right)+O(N^{-1/2+\varepsilon})\ , \tag{18}$$

for any $\varepsilon>0$, where $\varphi(x)$ is defined in (7), and $\{\log_2 N+t\}=\log_2 N+t-\lfloor\log_2 N+t\rfloor$. Since $\{\log_2 N+t\}$ is dense in $[0,1]$ but not uniformly distributed in $[0,1]$ the limiting distribution

Table 1: Performance of the algorithm for small values of d (neglecting fluctuations of mean zero).

d	$ER_{N,d} - \log_2 N$	Var $R_{N,d}$
0	-0.37	1.26
1	-1.40	0.78
2	-2.00	0.59
3	-2.42	0.49
4	-2.74	0.43
5	-3.02	0.38
6	-3.24	0.35
7	-3.43	0.32

of $R_{N,d}$ does not exist. In fact,

$$\liminf_{N\to\infty} \Pr\{R_{N,d} \le \log_2 N + t - 1\} = 1 - \varphi\left(2^{-t-1}\right) \tag{19}$$

$$\limsup_{N\to\infty} \Pr\{R_{N,d} \le \log_2 N + t - 1\} = 1 - \varphi\left(2^{-t}\right) \tag{20}$$

for real t and all $d \ge 0$.

Remark. With a more involved proof, using the *complex Mellin transform* it can be shown that the remainder term in (18) is $O(N^{-1/2})$, and the error terms in (8) and (11) are $O(\log N/N)$ and $O(\log^2 N/N)$, respectively. □

The nonexistence of the limiting distribution could be expected. Observe that the estimate $R_{N,d}$ is related to an extreme statistics of a set of N discrete random variables (in fact, geometrically distributed). Anderson [3] in 1970 predicted such a behavior for $M_n = \max\{X_1, \ldots, X_N\}$ where X_i are i.i.d. discrete random variables. In fact, when X_i are geometrically distributed with parameter $1/2$, then the asymptotic distribution of $M_n - \log_2 N$ oscillates between two *double exponential* distributions, namely $e^{-2^{-t-1}}$ and $e^{-2^{-t}}$. The bounds (19) and (20) resemble the double exponential distribution. Indeed, consider $d = 0$, and take only the first two terms of the infinite product $\varphi(2^{-t})$. Then, $\varphi(2^{-t}) \approx \sum_{j\ge 0} e^{-2^{-t+j}}$. This is illustrated in Figure 1 where we show the upper "envelope" for the asymptotic distribution of $R_{N,d} - \log_2 N$ (i.e., $1 - \varphi(2^{-t})$) for $d = 0$ and $d = 3$). In passing, we should point out that recently Fill *et. al.* [7] used the depoissonization

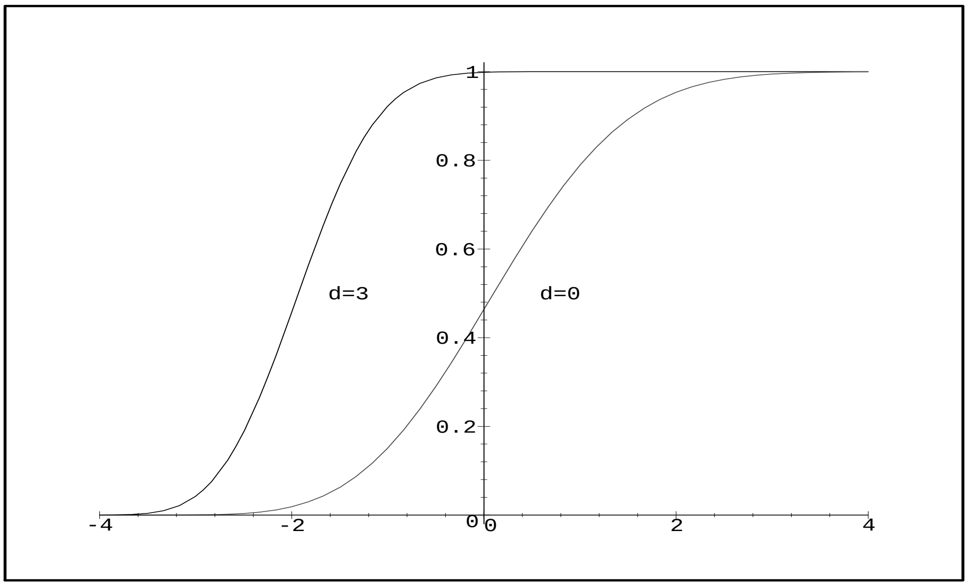

Figure 1: The upper "envelope" for the "shifted" asymptotic distribution $1 - \varphi(2^{-t})$ for $d = 0$ and $d = 3$.

technique to obtain a similar oscillation in the height of the incomplete trie. A thorough account on analytical depoissonization can be found in Jacquet and Szpankowski [20].

Our technique might be applied to analyze several other algorithms that are based on the "splitting process" described at the beginning of this section. More precisely, a splitting algorithm divides randomly N objects according to some rule. For example, consider a multiaccess system with a large number of users. The following "conflict resolution" algorithm found to be successful for some multiaccess systems (cf. [15], [18]). Let N be the size of a conflict, that is, the number of users that simultaneously sent packets. To split it, every user flips a fair coin, and those who get 1 "move right" in a digital tree represented this process, while the others "move left". This splitting process continues until all users sent successfully their packets (i.e., subtrees containing conflicting users are of size one). This algorithm can be analyzed in a unified manner by our approach through the functional equation of type (21).

Here is another example that generalizes PATRICIA tries (cf. [28], [36]). Consider N infinite strings that are built over a V-ary alphabet. Using a trie construction as in Knuth [28] (see also [36]) we can split all strings in such a manner that no two of them will share the common prefix. A trie built in such a manner may contain some internal nodes that are unary. To avoid such a waste of storage, one may compress this trie to obtain a PATRICIA trie. In the PATRICIA all internal nodes have degree greater or equal to two. But a generalization of a PATRICIA trie might be also explored by imposing that all internal nodes have degree of size greater or equal to $d \le V$. What is the search time (i.e., the length of a path to an external node) in such a trie? What about the average height, etc.? The analysis of such a data structure seems to be a nontrivial task, and our approach can be applied to solve some of these problems.

3. ANALYSIS

In this section we present the proof of Theorem 1 (cf. Section 3.2) and Theorem 2 (cf. Section 3.3). To streamline our analysis, in Section 3.1 we discuss a general solution of a functional equation that arises on many occasions in this paper (cf. (2)), and in general in the analysis of algorithms (cf. [4, 7, 10, 11, 15, 22, 24, 25, 28, 31, 32, 33, 34, 35, 36, 37]). In addition, we offer a unified approach to depoissonization: a technique that allows to analyze the Poisson model instead of the more difficult Bernoulli model (i.e., when the number of objects is fixed). Poissonization is a standard probabilistic technique (cf. Aldous [2]), however, depoissonization causes usually some problems. We cope with it in Section 3.1 where we present an analytical approach to the depoissonization (cf. also [7, 18, 19, 20, 32, 33]).

3.1 General Solution and Depoissonization

The functional equations (4) and (5) for the Poisson generating functions of $\widetilde{L}(z)$ and $\widetilde{W}(z)$, as well as (17) on $\widetilde{G}(z,u)$ satisfy the following general functional equation

$$\widetilde{g}(z,u) = \beta a(z/2,u)\widetilde{g}(z/2,u) + b(z,u) \ , \tag{21}$$

where $|\beta| \le 1$, and u is either fixed (cf. (4) and (5)) and in this case we simplify the notation to $\widetilde{g}(z)$, or u belongs to a compact neighborhood $\mathcal{U}(u_0)$ of u_0 (cf. in (17) we have $u_0 = 0$). The factor β may depend on u. Iterating (21), we obtain a general solution in the form of (cf. [18, 35])

$$\widetilde{g}(z,u) = \sum_{n=0}^{\infty} \beta^n b(z2^{-n},u) \prod_{k=1}^{n} a(z2^{-k},u) \tag{22}$$

provided the series converges. Define

$$\varphi(z,u) = \prod_{j=0}^{\infty} a(z2^j,u) \ , \tag{23}$$

provided the infinite product in (23) converges. Then, for $\varphi(z,u) \neq 0$, the general solution (22) can be rewritten as

$$\widetilde{g}(z,u)\varphi(z,u) = \sum_{n=0}^{\infty} \beta^n b(z2^{-n},u)\varphi(z2^{-n},u) \ . \tag{24}$$

The last form is the clue to our asymptotic solution of the functional equation (22), and in particular to solve (4), (5) and (17). This is due to the fact that the sum in (24) falls under the so called *harmonic sum* that can be easily handled by the Mellin transform technique (cf. [12, 34]): Using the Mellin transform we obtain the asymptotic behavior of the above function for $z \to \infty$, and then we "depoissonize" it to recover the original sequence.

In view of this, the important part of our analysis relies on depoissonization that we discuss next. To assure enough generality, let $g_N(k)$ be a double sequence of $N \geq 0$ and $k \geq 0$ (e.g., $g_N(k) = \Pr\{R_{N,d} = k\}$). Define $g_N(u) = \sum_{k\geq 0} u^k g_N(k)$, the exponential generating function $g(z,u) = \sum_{N\geq 0} g_N(u) z^N/N!$, and the Poisson generating function $\widetilde{g}(z,u) = e^{-z} g(z,u)$. When the sequence depends only on N (e.g., $g_N = ER_{N,d}$), we write $\widetilde{g}(z) = e^{-z}\sum_{N\geq 0} g_N z^N/N!$. Often the Poisson generating function $\widetilde{g}(z,u)$ (or $\widetilde{g}(z)$) is easier to handle, and one can obtain the asymptotic behavior of $\widetilde{g}(z,u)$ for $z \to \infty$. The question is how to recover $g_N(u)$ (or g_N). This reverse process is called *depoissonization*, and the lemma below was proved in Rais *et. al.* [32] (cf. also [17, 19, 33]).

Depoissonization Lemma. *Let $\mathcal{S}_\theta$ be a cone $\mathcal{S}_\theta = \{z : \ |\arg z| < \theta \ , \ 0 < \theta < \pi/2\}$. As $z \to \infty$ the following two conditions are assumed to hold for all u in a compact set $\mathcal{U}$:*
(I) For $z \in \mathcal{S}_\theta$

$$\widetilde{g}(z,u) = O(|z|^{\varepsilon}) \tag{25}$$

for some $\varepsilon > 0$;
(O) For $z \notin \mathcal{S}_\theta$

$$\widetilde{g}(z,u)e^{z} = O(e^{\alpha|z|}) \tag{26}$$

for some $\alpha < 1$.

Then for large N uniformly in $u \in \mathcal{U}$ the generating function $g_N(u) = [z^N/N!]\widetilde{g}(z,u)e^z$ satisfies

$$g_N(u) = \widetilde{g}(N,u) + O(N^{-1/2+\varepsilon}) \ . \tag{27}$$

Remark (i) The error term in (27) depends only on the bounds on $\widetilde{g}(z,u)$ (cf. (25) and (26)) and does not depend directly on the function $\widetilde{g}(z,u)$ itself.

(ii) A more careful analysis reveals that the error term in (27) is $O(N^{-1+\varepsilon})$ uniformly in $u \in \mathcal{U}$ (cf. [20]). □

Using the Depoissonization Lemma, we can "depoissonize" functional equation (21), and prove the following general result.

Theorem 3. *Consider the functional equation (21) for u in a compact set $\mathcal{U}$ such that $\widetilde{g}(z,u)$ is bounded in $\mathcal{U}$. Choose large $B > 0$ such that for $|z| > B$ and any constants $\beta_1, \beta_2, \delta, \delta_1 > 0$ the following conditions hold:*
(I) for $z \in \mathcal{S}_\theta$

$$\frac{|\beta a(z/2,u)|}{2^\varepsilon} < 1-\delta \quad , \qquad |b(z,u)| < \beta_1 \delta |z|^\varepsilon \ ; \tag{28}$$

(O) for $z \notin \mathcal{S}_\theta$

$$|\beta a(z/2,u)e^{z/2}| < (1-\delta_1)e^{\alpha|z/2|} \quad , \qquad |b(z,u)e^z| < \beta_2\delta_1 e^{\alpha|z|} \ , \tag{29}$$

for $\alpha < 1$. Then,

$$g_N(u) = \widetilde{g}(N,u) + O(N^{-1/2+\varepsilon}) \tag{30}$$

uniformly in $u \in \mathcal{U}$.

Proof. It suffices to prove that under (28) and (29) the two conditions (25) and (26) of the Depoissonization Lemma hold. The proof is by induction along the so called *increasing domains* as used in [17] and defined precisely in [19]: Suppose $1 < \lambda < 2$ is a given real number, and then define increasing domains $\mathcal{D}_m$ for $m = 0, 1, \ldots$ as

$$\mathcal{D}_m = \{z : \ B \le |z| < B\lambda^{m+1}\}$$

for the constant B defined in the theorem. Note that if $z \in \mathcal{D}_{m+1} - \mathcal{D}_m$, then $z/2 \in \mathcal{D}_m$, for $m = 0, 1, \ldots$. Thus, we can apply induction with respect to m.

We first show that (25) holds provided inequalities (28) are fulfilled for $z \in \mathcal{S}_\theta$. Define $\widehat{\mathcal{D}} = \mathcal{D}_m \cap \mathcal{S}_\theta$. Clearly (25) holds in $\widehat{\mathcal{D}}_0$ as long as a solution of the functional equation is bounded in a compact set. Let us now assume that condition (25) is satisfied in $\widehat{\mathcal{D}}_m$, and we prove that it also holds in $\widehat{\mathcal{D}}_{m+1}$. If $z \in \widehat{\mathcal{D}}_m$, then we are done. If $z \in \widehat{\mathcal{D}}_{m+1} - \widehat{\mathcal{D}}_m$, then $z/2 \in \widehat{\mathcal{D}}_m$, and we can apply induction hypothesis, that is, $|g(z/2,u)| \le \beta_1 |z|^\varepsilon / 2^\varepsilon$ for large $z \in \mathcal{S}_\theta$ and some constant $\beta_1 > 0$. Then, from the induction hypothesis, inequalities (28), and equation (21) we have

$$|\widetilde{g}(z,u)| \le \beta_1(1-\delta)|z|^\varepsilon + \beta_1\delta|z|^\varepsilon = \beta_1|z|^\varepsilon \ .$$

Thus, (25) holds in the larger domain $\mathcal{D}_{m+1}$, and by induction, in the whole cone $\mathcal{S}_\theta$.

In a similar manner we prove that the second set of inequalities (29) imply (26) by noting that $|e^z| = e^{\Re(z)} \le e^{\alpha|z|}$ where $\alpha = \cos\theta < 1$. This time we use domains $\overline{\mathcal{D}}_m = \mathcal{D}_m \cap \overline{\mathcal{S}}_\theta$ where $\overline{\mathcal{S}}_\theta$ is the complimentary set to $\mathcal{S}_\theta$. The rest follows the same line of arguments as above. This completes the proof of Theorem 3. ■

3.2 Analysis of Moments

We now prove Theorem 1. We consider two cases: bounded d, and large d ($d \to \infty$).

CASE A: BOUNDED d

We start with the average value $ER_{N,d}$. Its Poisson generating function $\widetilde{L}(z)$ satisfied the functional equation (4) which we repeat below

$$\widetilde{L}(z) = f_d(z/2)\widetilde{L}(z/2) + f_d(z/2) \tag{31}$$

where $f_d(z) = 1 - e_d(z)e^{-z}$ and $e_d(z) = 1 + z^1/1! + \cdots + z^d/d!$. This equation falls under the general functional equation (21), thus by (24) we have

$$\widetilde{L}(z)\varphi(z) = \sum_{n=0}^{\infty} \varphi(z2^{-n-1}) \ , \tag{32}$$

where

$$\varphi(z) = \prod_{j=0}^{\infty} f_d(z2^j) = \prod_{j=0}^{\infty} \left(1 - e_d(z2^j)e^{-z2^j}\right) \ . \tag{33}$$

Observe that, by d'Alembert's criterion, the final expression converges absolutely for all complex numbers z.

The idea of our analysis is to obtain the asymptotic behavior of $\widetilde{L}(z)$ as $z \to \infty$, and then to apply depoissonization. To obtain the asymptotics of $\widetilde{L}(z)$ we make use of the Mellin transform technique. The Mellin approach is by now a standard technique in the analysis of algorithms. The interested reader can find more on Mellin transform in a recent survey [12] (cf. also [34]). For the reader's convenience, however, we provide below some properties of the transform that we shall use in our analysis.

The following properties are used in this paper:

(P1) DEFINITION AND INVERSE OF THE REAL MELLIN TRANSFORM

$$f^*(s) = \int_0^{\infty} f(x)x^{s-1}dx \quad , \quad f(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} f^*(s)x^{-s}ds \tag{34}$$

where c belongs to the *fundamental strip* defined below.

(P2) FUNDAMENTAL STRIP

The Mellin transform of $f(x)$ exists in the *fundamental strip* $\Re(s) \in <-\alpha, -\beta>$ where

$$f(x) = \begin{cases} O(x^{\alpha}) & x \to 0 \\ O(x^{\beta}) & x \to \infty \ . \end{cases}$$

(P3) Harmonic sum

$$f(x) = \sum_{k \geq 0} \lambda_k g(\mu_k x) \qquad \Leftrightarrow \qquad f^*(s) = g^*(s) \sum_{k \geq 0} \lambda_k \mu_k^{-s} \tag{35}$$

provided the above sum converges. In the above, we used the following property

$$f(x) = g(ax) \qquad \Leftrightarrow \qquad f^*(s) = a^{-s} g^*(s) \tag{36}$$

in the fundamental strip of $f^*(s)$.

(P4) Asymptotic Expansion
If $f^*(s)$ satisfies certain smallness-conditions towards $i\infty$ (cf. [12]) for $-\beta < \Re(s) \leq M$, $(M > 0)$ and

$$f^*(s) = \sum_{k=0}^{K} \frac{d_k}{(s-b)^{k+1}} , \tag{37}$$

then as $x \to \infty$

$$f(x) = \sum_{k=0}^{K} \frac{d_k}{k!} x^{-b} (-\log x)^k + O(x^{-M}) . \tag{38}$$

(P5) Mellin Transform in Complex Plane (cf. [5, 6, 13])
If $F(z)$ is analytic in a cone $\theta_1 \leq \arg(z) \leq \theta_2$ with $\theta_1 < 0 < \theta_2$, then the Mellin transform $F^*(s)$ can be defined by replacing the path of integration $[0, \infty[$ by any curve starting at $z = 0$ and going to ∞ inside the cone, and it is identical with the real transform $f^*(s)$ of $f(z) = F(z)\Big|_{z \in R}$ from (34). In particular, if $f^*(s)$ fulfills an asymptotic expansion (37), then (38) holds for $F(z)$ as $z \to \infty$ in the cone.

Now we are ready to solve asymptotically (32) since we recognize it as a harmonic sum. Let $Q_1(z) = \widetilde{L}(z)\varphi(z)$ and $Q_1^*(s)$ be its Mellin transform which due to (P2) exists in the strip $\Re(s) \in (-\infty, 0)$. By the harmonic sum formula (P3) we have $Q_1(s) = 2^s/(1-2^s)\varphi^*(s)$ where the Mellin $\varphi^*(s)$ of $\varphi(x)$ exists in $\Re(s) \in (-\infty, 0)$. There is a problem, however, since the Mellin transform of $\varphi(s)$ does not exists at $s = 0$, and we need a more precise estimate of $\varphi^*(s)$ at $s = 0$.

Thus, we proceed as follows. Define

$$\Phi(x) = \varphi(2x) - \varphi(x) = e_d(x) e^{-x} \varphi(2x) . \tag{39}$$

Observe now that the Mellin transform of $\Phi(x)$ exists in $(-\infty, \infty)$, and by (36) we finally obtain

$$Q_1^*(s) = \frac{2^s}{1-2^s} \varphi^*(s) = \left(\frac{2^s}{1-2^s} \right)^2 \Phi^*(s) , \tag{40}$$

where $\Phi^*(s)$ is an entire function, as we noticed above, and defined as

$$\Phi^*(s) = \int_0^\infty e_d(x) e^{-x} \varphi(2x) x^{s-1} dx \ . \tag{41}$$

To assess asymptotically $\widetilde{L}(z)$ we use property (P4) of the Mellin transform, that is, we must estimate residues of (40). Note that $\chi_k = 2\pi i k/L$ ($k = 0, \pm 1, \pm 2, \ldots$) are the solutions of $1 - 2^{-s} = 0$, and the main contribution to the asymptotics comes from $\chi_0 = 0$. Since $\varphi(z) \sim 1 + O(z^{-M})$ for any $M > 0$, we obtain, for $z \to \infty$

$$\widetilde{L}(z) = \log_2 z \cdot \frac{\Phi^*(0)}{L} - \frac{\Phi^*(0)}{L} - \frac{\Phi^{*\prime}(0)}{L^2} + P_1(\log_2 z) + O(z^{-M}) \tag{42}$$

where

$$\begin{aligned} \Phi^*(0) &= \int_0^\infty e^{-x} e_d(x) \varphi(2x) \frac{dx}{x} \ , &(43)\\ \Phi^{*\prime}(0) &= \int_0^\infty e^{-x} e_d(x) \varphi(2x) \frac{\log x}{x} dx \ , &(44) \end{aligned}$$

and

$$P_1(x) = \sum_{k \neq 0} \left(\frac{\Phi^*(-\chi_k)(x-1)}{L} - \frac{\Phi^{*\prime}(-\chi_k)}{L^2} \right) e^{2k\pi i x} \ . \tag{45}$$

To complete the proof, we need to evaluate $\Phi^*(0)$,$\Phi^*(\chi_k)$, $\Phi^{*\prime}(0)$, and $\Phi^{*\prime}(\chi_k)$. The first one is rather easy, and by properties of the Mellin transform we find out that $\Phi^*(0) = L = \log 2$. The derivatives are harder to estimate. In order to treat the other constants, we first present another evaluation of $\Phi^*(0)$.

Define the function $\mathbf{1}(x)$ as follows

$$\mathbf{1}(x) = \begin{cases} 1 & \text{if } x \geq 1 \\ 0 & \text{if } x < 1 \ . \end{cases} \tag{46}$$

Then, we can write

$$\begin{aligned} \Phi^*(0) &= \int_0^\infty (\varphi(2x) - \varphi(x)) \frac{dx}{x} \\ &= \int_0^\infty (\varphi(2x) - \mathbf{1}(2x)) \frac{dx}{x} + \int_0^\infty (\mathbf{1}(2x) - \mathbf{1}(x)) \frac{dx}{x} + \int_0^\infty (\mathbf{1}(x) - \varphi(x)) \frac{dx}{x} &(47) \\ &= \int_0^\infty (\mathbf{1}(2x) - \mathbf{1}(x)) \frac{dx}{x} = \int_{1/2}^1 \frac{dx}{x} = L \ , \end{aligned}$$

since after the substitution $u = 2x$ in the second line of the above display, the first and the second integrals cancel. A similar derivation shows that $\Phi^*(\chi_k) = 0$, thus one proves formula (10) for $P_1(x)$ in Theorem 1(i).

To complete the proof of Theorem 1(i), we must establish depoissonization of $\widetilde{L}(z)$ which follows directly from Theorem 3. For the reader convenience, we repeat the arguments below: Observe that due to (42) condition (27) of the Depoissonization Lemma (growth estimate inside a cone S_θ) is automatically satisfied. Assume now $z \notin \mathcal{S}_\theta$ for some $0 < \theta < \pi/2$, and define $1 > \alpha > \cos\theta := c$. We apply induction and the increasing domains as in the proof of Theorem 3. Consider (31), and by induction hypothesis assume that $|\widetilde{L}(z/2)e^{z/2}| \le \beta e^{\alpha|z/2|}$ for some $\beta > 1$. Then:

$$
\begin{aligned}
|\widetilde{L}(z)e^{z}| &\le |f_d(z/2)e^{z/2}||\widetilde{L}(z/2)e^{z/2}| + |f_d(z/2)e^{z}| \\
&\le |f_d(z/2)|e^{c|z/2|}\beta e^{\alpha|z/2|} + |f_d(z/2)|e^{c|z|} \\
&\le \beta e^{\alpha|z/2|}
\end{aligned}
$$

where the last inequality holds as long as we choose ξ such that for $|z| > \xi$ we have $|f_d(z/2)| \le e^{(\alpha-c)|z/2|}$. This completes the depoissonization argument, and after an application of the depoissonization Lemma Theorem 1(i) is finally proved.

The proof of Theorem 1(ii) concerning the *variance* of $R_{N,d}$ is more intricate but it proceeds along the same lines as the above derivation. In particular, we deal now with the functional equation (5) which we repeat below

$$
\widetilde{W}(z) = f_d(z/2)\widetilde{W}(z/2) + 2\widetilde{L}(z/2)f_d(z/2) \ . \tag{48}
$$

Using the idea from Section 3.1 (cf. (21) – (24)) we find immediately an explicit solution to (48), namely

$$
\widetilde{W}(z)\varphi(z) = 2\sum_{n=0}^{\infty} \widetilde{L}(z2^{-n-1})\varphi(z2^{-n-1}) \ ,
$$

which becomes, after inserting solution (32) for $\widetilde{L}(z)$

$$
\widetilde{W}(z)\varphi(z) = 2\sum_{j=1}^{\infty}\sum_{n=1}^{\infty} \varphi(z2^{-n-j}) \ . \tag{49}
$$

To establish the asymptotics of $\widetilde{W}(z)$, thus also for $F_N''(1)$, we proceed as before through the Mellin transform and depoissonization. Let $Q_2(z) = \widetilde{W}(z)\varphi(z)$ and $Q_2^*(s)$ be its Mellin transform. Then, from the harmonic sum formula

$$
Q_2^*(s) = 2\left(\frac{2^s}{1-2^s}\right)^2 \varphi^*(s) = 2\left(\frac{2^s}{1-2^s}\right)^3 \Phi^*(s)
$$

where $\Phi^*(s)$ is defined as before. Noting that

$$
\frac{1}{(1-2^{-s})^3} = \frac{1}{s^3L^3} + \frac{3}{2s^2L^2} + \frac{1}{sL} + O(1) \ ,
$$

and computing residues, we easily obtain the following for $z \to \infty$ and any $M > 0$

$$Q_2(z) = \log_2^2 z - (3 + 2C_d)\log_2 z + 2 + 3C_d + D_d + P_2(\log_2 z) + O(z^{-M})$$

with C_d from (8), $P_2(x)$ is a periodic function with period 1, and

$$D_d = \frac{1}{L^3}\Phi^{*\prime\prime}(0) = \frac{1}{L^3}\int_0^\infty e_d(x)e^{-x}\varphi(2x)\frac{\log^2 x}{x}dx \; . \tag{50}$$

The rest is easy. We first depoissonize $Q_2(z)$ using the same arguments as above for $\widetilde{L}(z)$. Since $\varphi(z) = 1 + O(z^{-M})$ for any $M > 0$, we easily find the asymptotics of the second factorial moment of $R_{N,d}$ using the Depoissonization Lemma. Subtracting the square of the expectation (which was analyzed asymptotically above), the proof of Theorem 1, apart from the asymptotics of the constant term for $d \to \infty$, is completed.

Case B: Large d

Now, we analyze the case when $d \to \infty$. Considering more carefully $\Phi^{*\prime}(0)$, we obtain the following (with $\mathbf{1}(x)$ defined in (49))

$$\begin{aligned}\Phi^{*\prime}(0) &= -L^2/2 + L\int_0^\infty (\mathbf{1}(x) - \varphi(x))\frac{dx}{x} \\ &= -L^2/2 + L\int_0^1 \varphi(x)\frac{dx}{x} + L\int_1^\infty (1-\varphi(x))\frac{dx}{x} \; .\end{aligned} \tag{51}$$

We now estimate the above two integrals of (51) for large d. The first one is easy. Note that $\varphi(x) \le 1 - e_d(x)e^{-x} = e^{-x}\sum_{k=1}^\infty x^{d+k}/(d+k)!$. Thus,

$$0 \le \int_0^1 \varphi(x)\frac{dx}{x} \le \sum_{k=1}^\infty \frac{1}{(d+k)!(d+k)} = O\left(\frac{1}{d!}\right) \; .$$

In order to estimate the second integral of (51) – which we call $I(d)$ – we recognize its similarity to the *incomplete gamma* function (cf. [1]). We shall derive separately a lower and an upper bound on $I(d)$. For the lower bound, we observe that

$$\begin{aligned}I(d) &= \int_1^\infty (1-\varphi(x))\frac{dx}{x} \ge \int_1^\infty e_d(x)e^{-x}\frac{dx}{x} = \int_1^\infty \frac{e^{-t}}{t}dt + \sum_{i=1}^d \frac{1}{i!}(\Gamma(i) - \gamma(i,1)) \\ &= E_1(1) + \sum_{i=1}^d \frac{1}{i!}(\Gamma(i) - \gamma(i,1))\end{aligned} \tag{52}$$

where $E_1(1) = -\gamma - \sum_{n\ge 1}(-1)^n/(nn!)$ is a particular value of the exponential integral $E_1(z) = \int_z^\infty t^{-1}e^{-t}\,dt$, and $\gamma(a,z) = \int_0^z e^{-t}t^{a-1}dt$ the incomplete gamma function (cf. [1]).

Now we must estimate the second term of (52). Note that

$$\frac{\gamma(i+1,1)}{i!} = \frac{1}{e}\sum_{k=i+1}^\infty \frac{1}{k!} \; ,$$

and (cf. Knuth [27] Ex. 1.2.7–1.2.13),

$$\frac{1}{e}\sum_{k=1}^{\infty}\frac{H_k}{k!} = e\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{nn!},$$

where H_k is the kth harmonic number. Hence

$$\begin{aligned}
I(d) &\geq H_d - \gamma + \frac{1}{e}\sum_{k\geq 1}\frac{H_k}{k!} - \frac{1}{e}\sum_{i=1}^{d}\frac{1}{i}\sum_{k\geq i}\frac{1}{k!} \\
&= H_d - \gamma + \frac{1}{e}\sum_{k>d}(H_k - H_d) \\
&\geq H_d - \gamma = \log d + O(d^{-1}), \qquad d \to \infty.
\end{aligned}$$

An upper bound on $I(d)$ is more intricate. First, we rewrite $I(d)$ as

$$I(d) = \int_1^{d+1}(1-\varphi(x))\frac{dx}{x} + \int_{d+1}^{\infty}(1-\varphi(x))\frac{dx}{x}\,. \tag{53}$$

The first integral above denoted as $I_1(d)$ fulfills

$$I_1(d) \leq \int_1^{d+1}\frac{dx}{x} = \log(d+1) = \log d + O(d^{-1}), \qquad d\to\infty.$$

In the following we estimate the second integral of (53) which we denote as $I_2(d)$. We start from the observation that for $0 \leq a_k \leq 1$

$$1 - \sum_{k\geq 0} a_k \leq \prod_{k\geq 0}(1-a_k),$$

which can easily be proved taking logarithms. Therefore

$$I_2(d) \leq \sum_{k\geq 0} I_{2,k}(d),$$

with

$$I_{2,k}(d) = \int_{d+1}^{\infty} e_d(x2^k)e^{-x2^k}\frac{dx}{x} = \int_{(d+1)2^k}^{\infty} e_d(u)e^{-u}\frac{du}{u}\,.$$

Now, for $u \geq d$,

$$e_d(u) \leq (d+1)\frac{u^d}{d!},$$

so that

$$I_{2,k}(d) \leq \frac{d+1}{d!}\int_{(d+1)2^k}^{\infty} u^{d-1}e^{-u}\frac{du}{u}\,.$$

Since the function $f(u) = u^d e^{-u}$ takes its maximum for $u = d$ and is monotonically decreasing for $u \geq d$, we have for $u \geq (d+1)2^k$

$$u^{d-1}e^{-u} \leq u^{-2}\left((d+1)2^k\right)^{d+1} e^{-(d+1)2^k}\,.$$

Hence

$$\begin{aligned} I_{2,k}(d) &\leq \frac{d+1}{d!}\left((d+1)2^k\right)^{d+1} e^{-(d+1)2^k} \int_{(d+1)2^k}^{\infty} \frac{du}{u^2} \\ &= \frac{d+1}{d!}\left(\frac{d+1}{e}\right)^d 2^{k(d+1)} e^{-(d+1)(2^k-1)} \\ &\leq \frac{d+1}{d!}\left(\frac{d+1}{e}\right)^d \left(\frac{2}{e}\right)^{k(d+1)} , \end{aligned}$$

so that $\sum_{k\geq 0} I_{2,k}(d)$ is exponentially small in $d \to \infty$. Altogether we have proved that

$$I(d) = \log d + O(d^{-1}), \qquad d \to \infty \ . \tag{54}$$

Putting everything together, we finally obtained the asymptotics for C_d described in Theorem 1(i), that is,

$$C_d = \frac{\Phi^{*\prime}(0)}{L^2} = \log_2 d - 1/2 + O(d^{-1}) \ . \tag{55}$$

Now, can turn our attention to establishing Theorem 1(ii) for large d. The asymptotics of D_d (cf.(50)) for large d can be derived in a similar manner as (55). Nevertheless there remains a technical problem that we have to solve: We recall that to compute the variance of $R_{N,d}$ we need a precise estimate of $(\widetilde{L}(N))^2$. But the asymptotic formula on $(\widetilde{L}(N))^2$ involves $P_1^2(\log_2 N)$, where the zeroth term in the Fourier expansion (i.e., the integral mean) of $P_1^2(x)$ is *not* equal to zero (in contrast to $P_1(x)$). To recover this zeroth Fourier coefficient $[P_1^2]_0$, we need a new representation of $P_1(x)$ which we recall below

$$P_1(x) = -\frac{1}{L^2}\sum_{k\neq 0} \Phi^{*\prime}(-\chi_k) e^{2k\pi i x} \ .$$

Proceeding as with $\Phi^{*\prime}(0)$ above we find

$$\frac{1}{L^2}\Phi^{*\prime}(\chi_k) = -\frac{1}{L\chi_k} - \frac{1}{L}\int_0^{d+1} \varphi(t)\frac{dt}{t^{1+\chi_k}} + \frac{1}{L}\int_0^{d+1} (1-\varphi(t))\frac{dt}{t^{1+\chi_k}} \ .$$

Using integration by parts the whole expression turns out to be

$$\frac{1}{L^2}\Phi^{*\prime}(\chi_k) = -\frac{1}{L\chi_k}\int_0^{\infty} \varphi'(t)\frac{dt}{t^{\chi_k}}. \tag{56}$$

For later use we note that in the last integral $\varphi'(t)$ can be replaced by $f_d'(t) = \frac{t^d}{d!}e^{-t}$ with an exponentially small error in d: We have

$$\int_0^{\infty} \varphi'(t) t^{-\chi_k} dt = \int_0^{\infty} f_d'(t) t^{-\chi_k} dt + R(k,d),$$

with

$$R(k,d) = \int_0^{\infty} f_d'(t)(\varphi(2t)-1) t^{-\chi_k} dt + 2\int_0^{\infty} \varphi_0(t)\varphi'(2t) t^{-\chi_k} dt \ ,$$

so that, integrating by parts,

$$|R(k,d)| \le 2\int_0^\infty f_d'(t)(\varphi(2t)-1)dt\ .$$

Splitting up the range of integration according to the regions where either one of the factors of the integrand is small, $|R(k,d)|$ can be estimated similarly to $\Phi^{*\prime}(0)$, and $R(k,d)$ turns out to be exponentially small in d (uniformly in k). Now,

$$\int_0^\infty f_d'(t)t^{-\chi_k}dt = \frac{1}{d!}\int_0^\infty t^{d-\chi_k}e^{-t}dt = \frac{\Gamma(d+1-\chi_k)}{d!}\ ,$$

so that

$$-\frac{1}{L^2}\Phi^{*\prime}(-\chi_k) = \frac{1}{L\chi_k}\left(\frac{\Gamma(d+1+\chi_k)}{d!} + R(k,d)\right)$$

and we finally obtain

$$P_1(x) = \frac{1}{L}\sum_{k\neq 0}\frac{1}{\chi_k}\left(\frac{\Gamma(d+1+\chi_k)}{d!} + R(k,d)\right)e^{2k\pi ix}\ . \tag{57}$$

Finally, we can return to the evaluation of the zeroth Fourier coefficient $[P_1^2]_0$ of $P_1^2(x)$. According to the uniform exponential smallness of $R(k,d)$ for d getting large we have

$$[P_1^2]_0 = \frac{1}{L^2 d!^2}\sum_{k\neq 0}\frac{1}{|\chi_k|^2}|\Gamma(d+1+\chi_k)|^2\ \ + \text{exp. small terms in } d\ . \tag{58}$$

In order to evaluate the series occurring in (58) we observe that

$$\frac{1}{\chi_k}\frac{\Gamma(d+1+\chi_k)}{d!} = \binom{d+\chi_k}{d}\Gamma(\chi_k) = \sum_{\lambda=0}^{d}\binom{\lambda+\chi_k-1}{\lambda}\Gamma(\chi_k)\ . \tag{59}$$

Therefore the main term of $P_1(x)$ coincides perfectly with the periodic function $P_1(x)$ analyzed in [25] (cf. Theorem 3 of [25]) if d is replaced by $d+1$. The mean $[P_1^2]_0$ of the square of the latter fluctuating function can be analyzed by different methods. In [25] this quantity is treated using Hankel contour integrals with the "kernel"

$$\frac{\Gamma(j+z)\Gamma(j-z)}{e^{Lz}-1}\ .$$

An earlier approach, using transformation results due to Ramanujan, may be found in [23]. From [25] we have the asymptotic result

$$[P_1^2]_0 = \frac{1}{12} - \frac{1}{\sqrt{\pi d}} + O(\frac{1}{d}), \qquad d\to\infty\ .$$

This completes the proof of part (ii) of Theorem 1.

3.3 Asymptotic Limiting Distributions

In this subsection, we prove Theorem 2 concerning the asymptotic distribution of $R_{N,d}$. The Poisson generating function $\widetilde{G}(z,u)$ of $R_{N,d}$ satisfies the functional equation (17). But, the Mellin transform of $\widetilde{G}(z,u)$ with respect to z does not exist since $\widetilde{G}(z,u) = O(1)$ for both $z \to \infty$ and $z \to 0$. Therefore, we introduce the new function $\widetilde{H}(z,u) = \widetilde{G}(z,u) - 1$ whose Mellin transform exists in $\Re(s) \in (-1,0)$. Observe that $\widetilde{H}(z,u)$ fulfills the following functional equation

$$\widetilde{H}(z,u) = uf_d(z/2)\widetilde{H}(z/2,u) + (u-1)f_d(z/2) \ . \tag{60}$$

This equation falls under our general equation (21) from Section 3.1. In particular, using the same arguments as before, we can write the solution for $\widetilde{H}(z,x)$ as

$$\varphi(z)\widetilde{H}(z,u) = (u-1)\sum_{n=0}^{\infty} u^n \varphi(z2^{-n-1}) \ , \tag{61}$$

where $\varphi(z)$ is defined in (33). As an immediate consequence of the above we obtain

$$\widetilde{h}_k(z) = [u^k]\frac{\widetilde{H}(z,u)}{u-1} = e^{-z}\sum_{N=0}^{\infty}\Pr\{R_{N,d} > k\}\frac{z^N}{N!} = \frac{\varphi(z2^{-k-1})}{\varphi(z)} = \prod_{j=1}^{k+1} f_d(z/2^j) \tag{62}$$

with $f_d(z) = 1 - e_d(z)e^{-z}$. In the following we depoissonize (62).

We start with an estimate on $\widetilde{h}_k(z)$ inside the cone, thus we assume $z \in S_\theta$. Then, $\Re z = |z|\cos(\arg z) \geq |z|\cos\theta$, and

$$\begin{aligned} |f_d(z)| &= |1 - e_d(z)e^{-z}| \leq 1 + e_d(|z|)|e^{-z}| = 1 + e_d(|z|)e^{-\Re z} \\ &\leq 1 + e_d(|z|)e^{-|z|\cos\theta} \ . \end{aligned}$$

If we assume $\theta \leq \frac{\pi}{3}$, the last estimate reads

$$|f_d(z)| \leq 1 + e_d(|z|)e^{-|z|/2} \qquad \text{for all } z \in S_\theta \ .$$

Let $\varepsilon > 0$ be arbitrarily small. Then there exists $K = K(\varepsilon) \geq 1$ such that $e_d(x)e^{-x/2} < \frac{\varepsilon}{2}$ for all $x > K$, resp.

$$|f_d(z)| \leq 1 + \frac{\varepsilon}{2} \qquad \text{for } z \in S_\theta,\ |z| > K(\varepsilon) \ . \tag{63}$$

Let us now consider the region $|z| \leq K$: Since $0 \leq f_d(x) = 1 - e_d(x)e^{-x} \leq f_d(K) < 1$ for $x \in [0,K]$ and the function $f_d(z)$ is continuous, there exists an open neighborhood $\mathcal{U}(x)$ of each point $x \in [0,K]$ with $|f_d(z)| \le 1$ for $z \in \mathcal{U}(x)$. Due to the compactness of $[0,K]$, it follows that we can choose $\theta = \theta(\varepsilon)$ small enough such that

$$|f_d(z)| \le 1 \qquad \text{for } z \in S_{\theta(\varepsilon)},\ |z| \le K(\varepsilon)\ . \tag{64}$$

By (64) it follows that

$$|f_d(z/2^j)| \le 1 \qquad \text{for } j > \log_2(|z|/K(\varepsilon)),\ z \in S_{\theta(\varepsilon)},\ |z| \le K(\varepsilon)\ . \tag{65}$$

To estimate $\widetilde{h}_k(z)$ we split the product from (62) into the regions $1 \le j \le \min(k+1, \log_2(|z|/K(\varepsilon))$ and $j > \min(k+1, \log_2(|z|/K(\varepsilon))$, and we obtain

$$|\widetilde{h}_k(z)| \le \left(1+\frac{\varepsilon}{2}\right)^{\log_2(|z|/K(\varepsilon))} = \left(\frac{|z|}{K(\varepsilon)}\right)^{\log_2(1+\frac{\varepsilon}{2})}$$

i.e.,

$$|\widetilde{h}_k(z)| \le |z|^{\varepsilon} \qquad \text{for } z \in S_{\theta(\varepsilon)},\ |z| \ge 1\ .$$

To complete the depoissonization, we need to estimate $\widetilde{h}_k(z)$ outside the cone. Thus, assume $z \notin S_\theta$, so that $\Re z = |z|\cos(\arg z) \le |z|\cos\theta$. Therefore, with $C_1 = C_1(d)$

$$|e^z f_d(z)| = |e^z - e_d(z)| \le e^{\Re z} + e_d(|z|) \le (1+C_1)e^{|z|\cos\theta}\ . \tag{66}$$

Again we need a sharper estimate for $|e^{z/2^j} f_d(z/2^j)|$ if j is large: Since $e^z f_d(z)$ is continuous with $e^0 f_d(0) = 0$, there exists a constant $C_3 = C_3(d) > 0$ such that

$$|e^z f_d(z)| < 1 \qquad \text{for } |z| < C_3. \tag{67}$$

From (67) it follows that

$$|e^{z/2^j} f_d(z/2^j)| < 1 \qquad \text{for } j > \log_2(|z|/C_3)\ . \tag{68}$$

Now we proceed as follows: Since $\sum_{j=1}^{k+1} \frac{1}{2^j} = 1 - \frac{1}{2^{k+1}}$, we have

$$\left|e^z \prod_{j=1}^{k+1} f_d(z/2^j)\right| = |e^{z/2^{k+1}} \prod_{j=1}^{k+1} (e^{z/2^j} - e_d(z/2^j))|\ , \tag{69}$$

which is estimated by

$$e^{\Re z/2^{k+1}} \prod_{1 \le j \le K} (1+C_1)\, e^{(|z|\cos\theta)/2^j}\ ,$$

where $K = \min(k+1, \log_2(|z|/C_3))$. Therefore the estimate is less than or equal to

$$\begin{aligned}
&e^{(|z|\cos\theta)/2^{k+1}} (1+C_1)^K e^{|z|\cos\theta\,(1-\frac{1}{2^k})} \\
&\le (1+C_1)^{\log_2(|z|/C_3)} e^{|z|\cos\theta\,(1+\frac{1}{2^{k+1}}-\frac{1}{2^k})} \\
&\le \left(\frac{|z|}{C_3}\right)^{\log_2(1+C_1)} e^{|z|\cos\theta} < e^{\alpha|z|}
\end{aligned}$$

for $\cos\theta < \alpha < 1$ and $z \to \infty$, $z \notin S_\theta$.

Thus, we have established the conditions of the Depoissonization Lemma, and consequently get that

$$\Pr\{R_{N,d} > k\} = \frac{\varphi(N2^{-k-1})}{\varphi(N)} + O\left(N^{-1/2+\varepsilon}\right) ,$$

for any $\varepsilon > 0$ where the O–constant is independent of k.

To complete the proof of Theorem 2, it suffices to set $k = \lfloor \log_2 N + t \rfloor - 1$, and notice that $\Pr\{R_{N,d} > \log_2 N + t - 1\} = \Pr\{R_{N,d} > \lfloor \log_2 N + t \rfloor - 1\}$.

ACKNOWLEDGEMENT

We thank Philippe Jacquet (INRIA, Rocquencourt) for several valuable discussions concerning the depoissonization issues. We are also grateful to referees for a careful reading of the manuscript that led to a better presentation.

References

[1] M. Abramowitz, and I. Stegun (Eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, John Wiley & Sons, New York 1972.

[2] D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*, Springer-Verlag, New York 1989.

[3] C. Anderson, Extreme Value Theory for a Class of Discrete Distributions with Applications to Some Stochastic Processes. *J. Appl. Probability*, 7, 99–113, 1970.

[4] G. Brassard and P. Bratley, *Algorithmics. Theory and Practice*, Prentice Hall, Englewood Cliffs 1988.

[5] B. Davies, *Integral Transforms and Their Applications*, Springer-Verlag, New York 1978.

[6] G. Doetsch, *Handbuch der Laplace Transformation,* Vol. 1–3, Birkhäuser-Verlag, Basel 1955.

[7] J. Fill, H. Mahmoud and W. Szpankowski, On the Distribution for the Duration of a Randomized Leader Election Algorithm, *Annals of Applied Probability*, to appear.

[8] P. Flajolet and G.N. Martin, Probabilistic Counting Algorithms for Data Base Applications, *J. Comp. Syst. Sci.*, 31, 182-209, 1985.

[9] P. Flajolet, M. Régnier and D. Sotteau, Algebraic Methods for Trie Statistics, *Annals of Discrete Mathematics*, 25, 145-188, 1985.

[10] P. Flajolet and N. Saheb, The Complexity of Generating an Exponentially Distributed Variate, *J. Algorithms*, 7, 463-488, 1986.

[11] P. Flajolet and B. Richmond, Generalized Digital Trees and Their Difference-Differential Equations, *Random Structures and Algorithms*, 3, 305-320, 1992.

[12] P. Flajolet, X. Gourdon and P. Dumas, Mellin Transforms and Asymptotics: Harmonic Sums, *Theoretical Computer Science*, 144, 3-58, 1995.

[13] P. Flajolet, P. Jacquet, and W. Szpankowski, Mellin Transform of a Complex Variable: A Simple Extension, Unpublished Manuscript, 1995.

[14] P. Grabner, Searching for losers, *Random Structures and Algorithms*, 4, 99–110, 1993.

[15] A. Greenberg, P. Flajolet and R. Ladner, Estimating the Multiplicities of Conflicts to Speed Their Resolution in Multiple Access Channels, *JACM*, 34, 289-325, 1987.

[16] P. Henrici, *Applied and Computational Complex Analysis*, vol. 2, John Wiley *&* Sons, New York 1977.

[17] P. Jacquet, M. Régnier, Trie Partitioning Process: Limiting Distributions. *Lecture Notes in Computer Science*, vol. 214, 196–210, Springer, New York 1986.

[18] P. Jacquet and W. Szpankowski, Ultimate Characterizations of the Burst Response of an Interval Searching Algorithm: A Study of a Functional Equation, *SIAM J. Computing*, 18, 777-791, 1989.

[19] P. Jacquet and W. Szpankowski, Asymptotic behavior of the Lempel-Ziv Parsing Scheme and Digital Search Trees, *Theoretical Computer Science*, 144, 161-197, 1995.

[20] P. Jacquet and W. Szpankowski, Analytical Depoissonization and Its Applications, preprint.

[21] J.F.C. Kingman, *Subadditive processes*, in Ecole d'Eté de Probabilités de Saint-Flour V-1975, *Lecture Notes in Mathematics*, 539, Springer-Verlag, Berlin 1976.

[22] P. Kirschenhofer and H. Prodinger, On the Analysis of Probabilistic Counting, *Lecture Notes in Mathematics*, 1452 (Eds. E. Hlawka and R. Tichy), 117-120, Springer-Verlag, Berlin 1990.

[23] P. Kirschenhofer and H. Prodinger, On Some applications of Formulæ of Ramanujan in the Analysis of Algorithms, *Mathematika*, 38, 14-33, 1991.

[24] P. Kirschenhofer and H. Prodinger, Approximate Counting: An Alternative Approach, *Informatique Théorique et Applications/Theoretical Informatics and Applications*, 25, 43-48, 1991.

[25] P. Kirschenhofer and H. Prodinger, A Result in Order Statistics Related to Probabilistic Counting, *Computing*, 51, 15-27, 1993.

[26] P. Kirschenhofer, H. Prodinger, and W. Szpankowski, How to Count Quickly and Accurately: A Unified Analysis of Probabilistic Counting and Other Related Problems, *International Conference on Automata, Languages, and Programming (ICALP'92)*, Vienna, *Lecture Notes in Computer Science*, 623 (Ed. W. Kuich), 211-222, Springer-Verlag, Berlin 1992.

[27] D. Knuth, *The Art of Computer Programming: Fundamental Algorithms,* vol. 1., Addison-Wesley, Reading, MA 1973.

[28] D.E. Knuth, *The Art of Computer Programming. Sorting and Searching*, vol. 3., Addison-Wesley, Reading, MA 1973.

[29] R. Morris, Counting Large Numbers of Events in Small Registers, *Comm. ACM*, 21, 840-842, 1978.

[30] B. Pittel and H. Rubin, How Many Random Questions Are Necessary to Identify n Distinct Objects?, *J. Combinatorial Theory, Ser. A*, 55, 292-312, 1990.

[31] H. Prodinger, How to Select a Loser, *Discrete Math.*, 120, 149–159, 1993.

[32] B. Rais, P. Jacquet, and W. Szpankowski, Limiting Distribution for the Depth in Patricia Tries, *SIAM J. Discrete Mathematics*, 6, 197-213, 1993.

[33] M. Régnier and P. Jacquet, Normal Limiting Distribution of the Size of Tries, *Proc. Performance'87*, Amsterdam: North-Holland, 209-223, 1987.

[34] U. Schmid, The Average CRI-length of a Tree Collision Resolution Algorithm in Presence of Multiplicity-Dependent Capture Effects, *International Conference on Automata, Languages, and Programming (ICALP'92)*, Vienna, *Lecture Notes in Computer Science*, 623 (Ed. W. Kuich), 223-234, Springer-Verlag, Berlin 1992.

[35] W. Szpankowski, Solution of a Linear Recurrence Equation Arising in the Analysis of Some Algorithms, *SIAM J. Alg. Disc. Methods*, 8, 233-250, 1987.

[36] W. Szpankowski, Patricia Tries Again Revisited, *JACM*, 37, 691-711, 1990.

[37] W. Szpankowski and V. Rego, Yet Another Application of a Binomial Recurrence. Order Statistics, *Computing*, 43, 401-410, 1990.