Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms

Editors:

Ralph Neininger and Marek Zaionc

Kraków, Poland, July 4–8, 2016

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

AofA'16

27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms

July 4–8, 2016, Kraków, Poland

Confirmed keynote speakers

Jean Bertoin, Zürich Univ., Switzerland
Paweł Błasiak, Polish Academy of Sciences, Kraków, Poland
Hsien-Kuei Hwang, Academia Sinica, Taipeh, Taiwan
Bob Sedgewick, Princeton Univ., USA – FLAJOLET LECTURE
Wojtek Szpankowski, Purdue Univ., USA
Nick Wormald, Monash Univ., Australia

Scientific Committee

Nicolas Broutin, INRIA Rocquencourt, France
Jacek Cichoń, Wrocław UT, Poland
Michael Drmota, TU Vienna, Austria
Hsien-Kuei Hwang, Academia Sinica, Taiwan
Mihyun Kang, TU Graz, Austria
Jakub Kozik, JU Kraków, Poland
Gábor Lugosi, UPF Barcelona, Spain
Markus Nebel, TU Kaiserslautern, Germany
Ralph Neininger, GU Frankfurt, Germany, co-chair
Marc Noy, UPC Barcelona, Spain
Daniel Panario, Carleton Univ., Canada
Alois Panholzer, TU Vienna, Austria
Bruno Salvy, INRIA Lyon, France
Michèle Soria, Paris 6, France
Henning Sulzbach, McGill, Montréal, Canada
Brigitte Vallée, Caen Univ., France
Stephan Wagner, Stellenbosch Univ., South Africa
Marek Zaionc, JU Kraków, Poland, co-chair

Organizing Committee

Marek Zaionc, JU Kraków, Poland
Ralph Neininger, GU Frankfurt, Germany
Katarzyna Grygiel, JU Kraków, Poland
Maciej Bendkowski, JU Kraków, Poland
Agnieszka Łupińska, JU Kraków, Poland
Łukasz Lachowski, JU Kraków, Poland

Steering Committee

Bob Sedgewick (Chair)
Nicolas Broutin
Michael Drmota
Conrado Martinez
Michèle Soria
Brigitte Vallée

Conference office

Grupa A-05 Meetings & Events
31-101 Kraków, pl. Na Groblach 14/2
+48 12 429 62 23, info@a05.pl

Important dates

Submission deadline: **February 12, 2016**
Notification: **April 22, 2016**
Final version: **May 13, 2016**
Early registration: **May 13, 2016**

www.aofa2016.meetings.pl

Preface

The present volume collects the proceedings of AofA 2016, the 27^{th} International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms held at the Jagiellonian University in Kraków, during July 4–8, 2016. The conference builds on the communities of the former series of conferences 'Mathematics and Computer Science' and 'Analysis of Algorithms', and aims at studying rigorously the combinatorial objects which appear in the analysis of data structures and algorithms, as well as the essential ubiquitous combinatorial structures. The program committee selected submissions covering this wide range of topics. These regular papers, that were presented in 30-minute talks, appear in the present volume. The conference included poster and software demo sessions.

The conference also presented six invited plenary lectures. The 'Flajolet lecture' was given by Bob Sedgewick (Princeton University, USA) and the five invited talks were the following:

- Jean Bertoin, Zürich University, Switzerland,
- Paweł Błasiak, Polish Academy of Sciences, Kraków, Poland,
- Hsien-Kuei Hwang, Academia Sinica, Taipeh, Taiwan,
- Wojtek Szpankowski, Purdue University, USA,
- Nick Wormald, Monash University, Australia.

We thank the members of the steering and program committees for their involvement. We also thank the invited speakers and the authors of the contributed papers. We express our gratitude to the members of the organizing committee for their invaluable help in making this meeting a great success. Finally, our special thanks go to the sponsors of the conference for their contributions: the Department of Mathematics and Computer Science of Jagiellonian University in Kraków, the City of Kraków and the Polish Academy of Sciences.

Ralph Neininger and Marek Zaionc

Accepted papers

1. **Maxim Arnold, Yuliy Baryshnikov and Yuriy Mileyko**
   Typical representatives of free homotopy classes in a multi-punctured plane
2. **Nicolas Auger, Mathilde Bouvel, Cyril Nicaud and Carine Pivoteau**
   Analysis of Algorithms for Permutations Biased by Their Number of Records
3. **Martin Aumüller, Martin Dietzfelbinger, Clemens Heuberger, Daniel Krenn and Helmut Prodinger**
   Counting Zeros in Random Walks on the Integers and Analysis of Optimal Dual-Pivot Quicksort
4. **Jacek Cichoń, Rafał Kapelko and Dominik Markiewicz**
   On Leader Green Election
5. **Gwendal Collet, Michael Drmota and Lukas Klausner**
   Vertex Degrees in Planar Maps
6. **Jean-François Delmas, Jean-Stéphane Dhersin and Marion Sciauveau**
   Cost functionals for large random trees
7. **Michael Drmota, Abram Magner and Wojciech Szpankowski**
   Asymmetric Renyi Problem and PATRICIA Tries
8. **Michael Drmota and Mehdi Javanian**
   Solutions of First Order Linear Partial Differential Equations and Central Limit Theorems
9. **Sandro Franceschi, Kilian Raschel and Irina Kourkova**
   Analytic approach for reflected Brownian motion in the quadrant
10. **Michael Fuchs and Hsien-Kuei Hwang**
    Dependence between External Path-Length and Size in Random Tries
11. **Jeffrey Gaither and Mark Daniel Ward**
    Variance of the Internal Profile in Suffix Trees
12. **Zhicheng Gao and Edward Bender**
    Locally Restricted Sequential Structures and Runs of a Subcomposition in Integer Compositions
13. **Antoine Genitrini**
    Full asymptotic expansion for Polya structures
14. **Benjamin Hackl, Clemens Heuberger and Daniel Krenn**
    Asymptotic Expansions in SageMath
15. **Benjamin Hackl, Clemens Heuberger and Helmut Prodinger**
    The Register Function and Reductions of Binary Trees and Lattice Paths
16. **Clemens Heuberger and Sara Kropf**
    On the Higher Dimensional Quasi-Power Theorem and a Berry-Esseen Inequality
17. **Pawel Hitczenko and Amanda Lohss**
    Probabilistic consequences of some polynomial recurrences
18. **Cecilia Holmgren, Svante Janson and Matas Sileikis**
    Using Pólya urns to show normal limit laws for fringe subtrees in m-ary search trees
19. **Philippe Jacquet and Wojciech Szpankowski**
    Average Size of a Suffix Tree for Markov Sources
20. **Mihyun Kang and Tamas Makai**
    Bootstrap percolation on G(n,p) revisited
21. **Sara Kropf and Stephan Wagner**
    q-Quasiadditive Functions
22. **Markus Kuba and Alois Panholzer**
    Combinatorial analysis of growth models for series-parallel networks

23. **Amanda Lohss and Pawel Hitczenko**
Corners in Tree–Like Tableaux

24. **Eugenijus Manstavicius and Vytautas Stepas**
Variance of additive functions defined on random assemblies

25. **Robertas Petuchovas**
Recent results on permutations without short cycles

26. **Patricio Poblete and Alfredo Viola**
Robin Hood Hashing really has constant average search cost and variance in full tables

27. **Dimbinaina Ralaivaosaona and Stephan Wagner**
Additive functionals of d-ary increasing trees

28. **Vonjy Rasendrahasina, Andry Rasoanaivo and Vlady Ravelomanana**
The Maximum Block Size of Critical Random Graphs

29. **Takeharu Shiraga**
The Cover Time of Deterministic Random Walks for General Transition Probabilities

30. **Michael Wallner**
A half-normal distribution scheme for generating functions and the unexpected behavior of Motzkin paths

31. **Vytas Zacharovas**
On the exponential decay of the characteristic function of the quicksort distribution

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Analysis of Algorithms for Permutations Biased by Their Number of Records

Nicolas Auger[1], Mathilde Bouvel[2†], Cyril Nicaud[1], Carine Pivoteau[1]

[1] *Université Paris-Est, LIGM (UMR 8049), F77454 Marne-la-Vallée, France*
[2] *CNRS and Institut für Mathematik, Universität Zürich, Zurich, Switzerland*

The topic of the article is the parametric study of the complexity of algorithms on arrays of pairwise distinct integers. We introduce a model that takes into account the non-uniformness of data, which we call the Ewens-like distribution of parameter θ for records on permutations: the weight θ^r of a permutation depends on its number r of records. We show that this model is meaningful for the notion of presortedness, while still being mathematically tractable. Our results describe the expected value of several classical permutation statistics in this model, and give the expected running time of three algorithms: the Insertion Sort, and two variants of the Min-Max search.

Keywords: permutation, Ewens distribution, random generation, analysis of algorithms

1 Introduction

A classical framework for analyzing the average running time of algorithms is to consider uniformly distributed inputs. Studying the complexity of an algorithm under this uniform model usually gives a quite good understanding of the algorithm. However, it is not always easy to argue that the uniform model is relevant, when the algorithm is used on a specific data set. Observe that, in some situations, the uniform distribution arises by construction, from the randomization of a deterministic algorithm. This is the case with Quick Sort for instance, when the pivot is chosen uniformly at random. In other situations, the uniformity assumption may not fit the data very well, but still is a reasonable first step in modeling it, which makes the analysis mathematically tractable.

In practical applications where the data is a sequence of values, it is not unusual that the input is already partially sorted, depending on its origin. Consequently, assuming that the input is uniformly distributed, or shuffling the input as in the case of Quick Sort, may not be a good idea. Indeed, in the last decade, standard libraries of well-established languages have switched to sorting algorithms that take advantage of the "almost-sortedness" of the input. A noticeable example is Tim Sort Algorithm, used in Python (since 2002) and Java (since Java 7): it is particularly efficient to process data consisting of long increasing (or decreasing) subsequences.

In the case of sorting algorithms, the idea of taking advantage of some bias in the data towards sorted sequences dates back to Knuth [9, p. 336]. It has been embodied by the notion of *presortedness*, which

[†]Supported by a Marie Heim-Vögtlin grant of the Swiss National Science Foundation.

quantifies how far from sorted a sequence is. There are many ways of defining measures of presortedness, and it has been axiomatized by Mannila [10] (see Section 2.2 for a brief overview). For a given measure of presortedness m, the classical question is to find a sorting algorithm that is optimal for m, meaning that it minimizes the number of comparisons as a function of both the size of the input and the value of m. For instance, Knuth's Natural Merge Sort [9] is optimal for the measure $r =$"number of runs" , with a worst case running time of $\mathcal{O}(n \log r)$ for an array of length n.

Most measures of presortedness studied in the literature are directly related to basic statistics on permutations. Consequently, it is natural to define biased distributions on permutations that depend on such statistics, and to analyze classical algorithms under these non-uniform models. One such distribution is very popular in the field of discrete probability: the Ewens distribution. It gives to each permutation σ a probability that is proportional to $\theta^{\mathrm{cycle}(\sigma)}$, where $\theta > 0$ is a parameter and $\mathrm{cycle}(\sigma)$ is the number of cycles in σ. Similarly, for any classical permutation statistics χ, a non-uniform distribution on permutations may be defined by giving to any σ a probability proportional to $\theta^{\chi(\sigma)}$. We call such distributions *Ewens-like distributions*. Note that the Ewens-like distribution for the number of inversions is quite popular, under the name of *Mallows distribution* [7, and references therein].

In this article, we focus on the Ewens-like distribution according to $\chi =$ number of records (a.k.a. *left to right maxima*). The motivation for this choice is twofold. First, the number of records is directly linked to the number of cycles by the fundamental bijection (see Section 2.1). So, we are able to exploit the nice properties of the classical Ewens distribution, and have a non-uniform model that remains mathematically tractable. Second, we observe that the number of non-records is a measure of presortedness. Therefore, our distribution provides a model for analyzing algorithms which is meaningful for the notion of presortedness, and consequently which may be more realistic than the uniform distribution. We first study how this distribution impacts the expected value of some classical permutation statistics, depending on the choice of θ. Letting θ depend on n, we can reach different kinds of behavior. Then, we analyze the expected complexity of Insertion Sort under this biased distribution, as well as the effect of branch prediction on two variants of the simultaneous minimum and maximum search in an array.

2 Permutations and Ewens-like distributions

2.1 Permutations as words or sets of cycles

For any integers a and b, let $[a, b] = \{a, \dots, b\}$ and for every integer $n \geq 1$, let $[n] = [1, n]$. By convention $[0] = \emptyset$. If E is a finite set, let $\mathfrak{S}(E)$ denote the set of all permutations on E, *i.e.*, of bijective maps from E to itself. For convenience, $\mathfrak{S}([n])$ is written $\mathfrak{S}_n$ in the sequel. Permutations of $\mathfrak{S}_n$ can be seen in several ways (reviewed for instance in [3]). Here, we use both their representations as words and as sets of cycles.

A permutation σ of $\mathfrak{S}_n$ can be represented as a word $w_1 w_2 \dots w_n$ containing exactly once each symbol in $[n]$: by simply setting $w_i = \sigma(i)$ for all $i \in [n]$. Conversely, any sequence (or word) of n distinct integers can be interpreted as representing a permutation of $\mathfrak{S}_n$. For any sequence $s = s_1 s_2 \dots s_n$ of n distinct integers, the rank $\mathrm{rank}_s(s_i)$ of s_i is defined as the number of integers appearing in s that are smaller than or equal to s_i. Then, for any sequence s of n distinct integers, the *normalization* $\mathrm{norm}(s)$ of s is the unique permutation σ of $\mathfrak{S}_n$ such that $\sigma(i) = \mathrm{rank}_s(s_i)$. For instance, $\mathrm{norm}(8254) = 4132$.

Many permutation statistics are naturally expressed on their representation as words. One will be of particular interest for us: the number of records. If σ is a permutation of $\mathfrak{S}_n$ and $i \in [n]$, there is a *record* at position i in σ (and subsequently, $\sigma(i)$ is a record) if $\sigma(i) > \sigma(j)$ for every $j \in [i-1]$. In the word representation of permutations, records are therefore elements that have no larger elements to their

left. This equivalent definition of records naturally extends to sequences of distinct integers, and for any sequence s of distinct integers, the positions of the records in s and in $\mathrm{norm}(s)$ are the same. A position that is not a record is called a *non-record*.

A cycle of size k in a permutation $\sigma \in \mathfrak{S}_n$ is a subset $\{i_1, \ldots, i_k\}$ of $[n]$ such that $i_1 \overset{\sigma}{\mapsto} i_2 \ldots \overset{\sigma}{\mapsto} i_k \overset{\sigma}{\mapsto} i_1$. It is written $(i_1, i_2, \ldots, i_k)$. Any permutation can be decomposed as the set of its cycles. For instance, the cycle decomposition of τ represented by the word 6321745 is $(32)(641)(75)$.

These two ways of looking at permutations (as words or as set of cycles) are rather orthogonal, but there is still a link between them, provided by the so-called *fundamental bijection* or *transition lemma*. The fundamental bijection, denoted F, is the following transformation:

1. Given σ a permutation of size n, consider the cycle decomposition of σ.
2. Write every cycle starting with its maximal element, and write the cycles in increasing order of their maximal (i.e., first) element.
3. Erasing the parenthesis gives $F(\sigma)$.

Continuing our previous example gives $F\big(\tau\big) = 3264175$. This transformation is a bijection, and transforms a permutation as set of cycles into a permutation as word. Moreover, it maps the number of cycles to the number of records. For references and details about this bijection, see for example [3, p. 109–110].

2.2 The number of non-records as a measure of presortedness

The concept of presortedness, formalized by Mannila [10], naturally arises when studying sorting algorithms which efficiently sort sequences already almost sorted. Let E be a totally ordered set. We denote by $E^\star$ the set of all nonempty sequences of distinct elements of E, and by $\cdot$ the concatenation on $E^\star$. A mapping m from $E^\star$ to $\mathbb{N}$ is a *measure of presortedness* if it satisfies:

1. if $X \in E^\star$ is sorted then $m(X) = 0$;
2. if $X = (x_1, \cdots, x_\ell)$ and $Y = (y_1, \cdots, y_\ell)$ are two elements of $E^\star$ having same length, and such that for every $i, j \in [\ell]$, $x_i < x_j \Leftrightarrow y_i < y_j$ then $m(X) = m(Y)$;
3. if X is a subsequence of Y then $m(X) \leq m(Y)$;
4. if every element of X is smaller than every element of Y then $m(X \cdot Y) \leq m(X) + m(Y)$;
5. for every symbol $a \in E$ that does not occur in X, $m(a \cdot X) \leq |X| + m(X)$.

Classical measures of presortedness [10] are the number of inversions, the number of swaps, ... One can easily see, checking conditions 1 to 5, that $m_{\mathrm{rec}}(s) =$ number of non-records in $s = |s| -$ number of records in s defines a measure of presortedness on sequences of distinct integers. Note that because of condition 2, studying a measure of presortedness on $\mathfrak{S}_n$ is not a restriction with respect to studying it on sequences of distinct integers.

Given a measure of presortedness m, we are interested in optimal sorting algorithms with respect to m. Let $\mathrm{below}_m(n,k) = \{\sigma : \sigma \in \mathfrak{S}_n,\ m(\sigma) \leq k\}$. A sorting algorithm is *m-optimal* (see [10] and [11] for more details) if it performs in the worst case $\mathcal{O}(n + \log|\,\mathrm{below}_m(n,k)|)$ comparisons when applied to $\sigma \in \mathfrak{S}_n$ such that $m(\sigma) = k$, uniformly in k. There is a straightforward algorithm that is m_{rec}-optimal. First scan σ from left to right and put the records in one (sorted) list L_R and the non-records in another list L_N. Sort L_N using a $\mathcal{O}(|L_N| \log |L_N|)$ algorithm, then merge it with L_R. The worst case running time of this algorithm is $\mathcal{O}(n + k \log k)$ for permutations σ of $\mathfrak{S}_n$ such that $m_{\mathrm{rec}}(\sigma) = k$. Moreover, $|\,\mathrm{below}_{m_{\mathrm{rec}}}(n,k)| \geq k!$ for any $k \geq n$, since it contains the $k!$ permutations of the form

$(k+1)(k+2)\dots n\cdot\tau$ for $\tau\in\mathfrak{S}_k$. Consequently, $\mathcal{O}(n+k\log k)=\mathcal{O}(n+\log|\operatorname{below}_{m_{\text{rec}}}(n,k)|))$, proving m_{rec}-optimality.

2.3 Ewens and Ewens-like distribution

The Ewens distribution on permutations (see for instance [1, Ch. 4 & 5]) is a generalization of the uniform distribution on $\mathfrak{S}_n$: the probability of a permutation depends on its number of cycles. Denoting $\operatorname{cycle}(\sigma)$ the number of cycles of any permutation σ, the Ewens distribution of parameter θ (where θ is any fixed positive real number) gives to any σ the probability $\frac{\theta^{\operatorname{cycle}(\sigma)}}{\sum_{\rho\in\mathfrak{S}_n}\theta^{\operatorname{cycle}(\rho)}}$. As seen in [1, Ch. 5], the normalization constant $\sum_{\rho\in\mathfrak{S}_n}\theta^{\operatorname{cycle}(\rho)}$ is $\theta^{(n)}$, where the notation $x^{(n)}$ (for any real x) denotes the *rising factorial* defined by $x^{(n)}=x(x+1)\cdots(x+n-1)$ (with the convention that $x^{(0)}=1$).

Mimicking the Ewens distribution, it is natural (and has appeared on several occasions in the literature, see for instance [4, Example 12]) to define other non-uniform distributions on $\mathfrak{S}_n$, where we introduce a bias according to some statistics χ. The Ewens-like distribution of parameter θ (again θ is any fixed positive real number) for statistics χ is then the one that gives to any $\sigma\in\mathfrak{S}_n$ the probability $\frac{\theta^{\chi(\sigma)}}{\sum_{\rho\in\mathfrak{S}_n}\theta^{\chi(\rho)}}$. The classical Ewens distribution corresponds to $\chi=$ number of cycles. Ewens-like distributions can be considered for many permutations statistics, like the number of inversions, of fixed points, of runs, ... In this article, we focus on the distribution associated with $\chi=$ number of *records*. We refer to it as the Ewens-like distribution for records (with parameter θ). For any σ, we let $\operatorname{record}(\sigma)$ denote the number of records of σ, and define the *weight* of σ as $\mathrm{w}(\sigma)=\theta^{\operatorname{record}(\sigma)}$. The Ewens-like distribution for records on $\mathfrak{S}_n$ gives probability $\frac{\mathrm{w}(\sigma)}{\mathrm{W}_n}$ to any $\sigma\in\mathfrak{S}_n$, where $\mathrm{W}_n=\sum_{\rho\in\mathfrak{S}_n}\mathrm{w}(\rho)$. Note that the normalization constant is $\mathrm{W}_n=\theta^{(n)}$, like in the classical Ewens distribution: indeed, the fundamental bijection reviewed above shows that there are as many permutations with c cycles as permutations with c records. Fig. 1 shows random permutations under the Ewens-like distribution for records, for various values of θ.

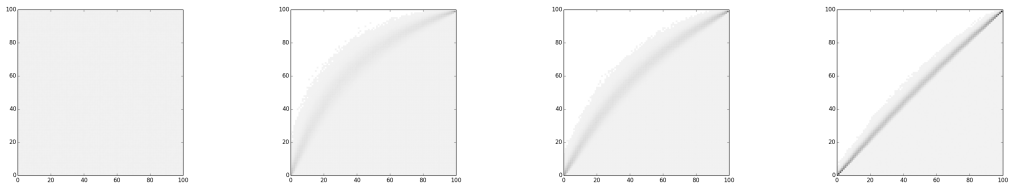


Fig. 1: Random permutations under the Ewens-like distribution on $\mathfrak{S}_{100}$ with, from left to right, $\theta=1$ (corresponding to the uniform distribution), 50, 100, and 500. For each diagram, the darkness of a point (i,j) is proportional to the number of generated permutations σ such that $\sigma(i)=j$, for a sampling of 10000 random permutations.

2.4 Linear random samplers

Efficient random samplers have several uses for the analysis of algorithms in general. They allow to estimate quantities of interest (even when their computation with a theoretical approach is not feasible), and can be used to double-check theoretical results. They are also a precious tool to visualize the objects under study (the diagrams in Fig. 4 were obtained in this way), allowing to define new problems on these objects (for example: can we describe the limit shape of the diagrams shown in Fig. 4?).

As mentioned in [6, §2.1], one can easily obtain a linear time and space algorithm to generate a random permutation according to the Ewens distribution (for cycles), using a variant of the Chinese restaurant process reviewed in what follows. To generate a permutation of size n, we start with an empty array[(i)] σ of length n that is used to store the values of the $\sigma(i)$'s. For i from 1 to n, we choose to either create a new cycle containing only i with probability $\frac{\theta}{\theta+i-1}$ or to insert i in one of the existing cycles with probability $\frac{i-1}{\theta+i-1}$. To create a new cycle, we set $\sigma[i] = i$. To insert i in an existing cycle, we choose uniformly at random an element j in $[i-1]$ to be the element following i in its cycle, and we set $\sigma[i] = j$ and $\sigma[\sigma^{-1}[j]] = i$. To avoid searching for $\sigma^{-1}[j]$ in the array σ, we only need to keep σ^{-1} in a second array while adding the elements in σ.

Starting from this algorithm, we can easily design a linear random sampler for permutations according to the Ewens-like distribution for records, using the fundamental bijection. The first step is to generate a permutation σ in $\mathfrak{S}_n$ with the above algorithm. Then, we write the cycles of σ in reverse order of their maximum, as sequences, starting from the last element and up to exhaustion of the cycle: $n, \sigma[n], \sigma[\sigma[n]], \ldots, \sigma^{-1}[n]$. Each time we write an element i, we set $\sigma[i] = 0$ and each time a cycle is finished, we search the next value of i such that $\sigma[i] \neq 0$ to start the next cycle. This new cycle will be written before the one that has just been written. Note that all these operations can be performed in time complexity $O(1)$ using doubly linked lists for the resulting permutation. In the end, the cycles will be written as sequences starting by their maximum, sorted in increasing order of their maximum, which is the fundamental bijection.

Note that there exists another branching process, known as the *Feller coupling*, to generate permutations according to the Ewens distribution (see for instance [1, p.16]). Although it is less natural than with the Chinese restaurant process, it is also possible to infer linear random samplers from it. Details will be provided in an extended version of this work.

3 Average value of statistics in biased random permutations

Let θ be any fixed positive real number. In this section, we study the behavior of several statistics on permutations, when they follow the Ewens-like distribution for records with parameter θ. Our purpose is mostly to illustrate methods to obtain precise descriptions of the behavior of such statistics. Such results allow a fine analysis of algorithms whose complexity depends on the studied statistics.

Recall that, for any $\sigma \in \mathfrak{S}_n$, $\mathrm{w}(\sigma) = \theta^{\mathrm{record}(\sigma)}$ and the probability of σ is $\frac{\mathrm{w}(\sigma)}{\mathrm{W}_n}$, with $\mathrm{W}_n = \theta^{(n)}$. Recall also that the records of any sequence of distinct integers are well-defined. For any such sequence s we subsequently set $\mathrm{record}(s)$ to be the number of records of s and $\mathrm{w}(s) = \theta^{\mathrm{record}(s)}$. Note that for any such sequence s, $\mathrm{w}(s) = \mathrm{w}(\mathrm{norm}(s))$, because the positions (and hence the number) of records do not change when normalizing.

3.1 Technical lemmas

Seeing a permutation of $\mathfrak{S}_n$ as a word, it can be split (in many ways) into two words as $\sigma = \pi \cdot \tau$ for the usual concatenation on words. Note that here π and τ are not normalized permutations: τ belongs to the set $\mathfrak{S}_{k \text{ in } n}$ of all sequences of k distinct integers in $[n]$ where $k = |\tau|$, and π belongs to $\mathfrak{S}_{n-k \text{ in } n}$. The weight function w behaves well with respect to this decomposition, as shown in the following lemmas.

(i) Note that our array starts at index 1.

Lemma 1 *Let n be an integer, and τ be a sequence of $k \leq n$ distinct integers in $[n]$. Denote by m the number of records in τ whose value is larger than the largest element of $[n]$ which does not appear in τ, and define $\mathrm{w}'_n(\tau)$ as θ^m. For all $\sigma \in \mathfrak{S}_n$, if $\sigma = \pi \cdot \tau$, then $\mathrm{w}(\sigma) = \mathrm{w}(\pi) \cdot \mathrm{w}'_n(\tau)$.*

For instance, the definition of $\mathrm{w}'_n(\tau)$ gives $\mathrm{w}'_9(6489) = \theta^2$ (8 and 9 are records of τ larger than 7) and $\mathrm{w}'_{10}(6489) = 1$ (there are no records in τ larger than 10).

We extend the weight function w to subsets X of $\mathfrak{S}_n$ as $\mathrm{w}(X) = \sum_{\sigma \in X} \mathrm{w}(\sigma)$. For any sequence τ of $k \leq n$ distinct integers in $[n]$, the right-quotient of X with τ is $X/\tau = \{\pi : \pi \cdot \tau \in X\}$. Since $\mathrm{w}(\pi) = \mathrm{w}(\mathrm{norm}(\pi))$ for all sequences π of distinct integers, we have $\mathrm{w}(X/\tau) = \mathrm{w}(\mathrm{norm}(X/\tau))$ for all X and τ as above (As expected, $\mathrm{norm}(Y)$ means $\{\mathrm{norm}(\pi) : \pi \in Y\}$).

For $k \in [n]$, we say that $X \subseteq \mathfrak{S}_n$ is *quotient-stable for k* if $\mathrm{w}(X/\tau)$ is constant when τ runs over $\mathfrak{S}_{k \text{ in } n}$. When X is quotient-stable for k, we denote $\mathrm{w}^{\mathrm{q}}_k(X)$ the common value of $\mathrm{w}(X/\tau)$ for τ as above. For instance, $X = (4321, 3421, 4132, 3142, 4123, 2143, 3124, 1324)$ is quotient-stable for $k = 1$. Indeed,

$$\mathrm{w}^{\mathrm{q}}_1(X) = \mathrm{w}(X/1) = \mathrm{w}(\{432, 342\}) = \mathrm{w}(X/2) = \mathrm{w}(\{413, 314\}) = \\ \mathrm{w}(X/3) = \mathrm{w}(\{412, 214\}) = \mathrm{w}(X/4) = \mathrm{w}(\{312, 132\}) = \theta + \theta^2.$$

Note that $\mathfrak{S}_n$ is quotient-stable for all $k \in [n]$: indeed, for any τ of size k, $\mathrm{norm}(\mathfrak{S}_n/\tau) = \mathfrak{S}_{n-k}$ so that $\mathrm{w}(\mathfrak{S}_n/\tau) = \mathrm{w}(\mathfrak{S}_{n-k})$ for all τ of size k. It follows that $\mathrm{w}^{\mathrm{q}}_k(\mathfrak{S}_n) = \mathrm{w}(\mathfrak{S}_{n-k}) = \theta^{(n-k)}$.

Lemma 2 *Let $X \subseteq \mathfrak{S}_n$ be quotient-stable for $k \in [n]$. Then $\mathrm{w}(X) = \frac{\theta^{(n)}}{\theta^{(n-k)}} \mathrm{w}^{\mathrm{q}}_k(X)$.*

A typical example of use of Lemma 2 is given in the proof of Theorem 3.

Remark: Lemma 2 is a combinatorial version of a simple probabilistic property: Let E_τ be the set of elements of $\mathfrak{S}_n$ that end with τ. If A is an event on $\mathfrak{S}_n$ and if the probability of A given E_τ is the same for every $\tau \in \mathfrak{S}_{k \text{ in } n}$, then it is equal to the probability of A, by the law of total probabilities. ⋄

3.2 Summary of asymptotic results

The rest of this section is devoted to studying the expected behavior of some permutation statistics, under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ. We are especially interested in the asymptotics in n when θ is constant or is a function of n. The studied statistics are: number of records, number of descents, first value, and number of inversions. A summary of our results is presented in Table 3.2. The asymptotics reported in Table 3.2 follow from Corollaries 4, 6, 9, 11 either immediately or using the so-called *digamma* function. The digamma[(ii)] function is defined by $\Psi(x) = \Gamma'(x)/\Gamma(x)$. It satisfies the identity $\sum_{i=0}^{n-1} \frac{1}{x+i} = \Psi(x+n) - \Psi(x)$, and its asymptotic behavior as $x \to \infty$ is $\Psi(x) = \log(x) - \frac{1}{2x} - \frac{1}{12x^2} + o\left(\frac{1}{x^2}\right)$. We also define $\Delta(x, y) = \Psi(x+y) - \Psi(x)$, so that $\Delta(x, n) = \sum_{i=0}^{n-1} \frac{1}{x+i}$ for any positive integer n. In Table 3.2 and in the sequel, we use the notations $\mathbb{P}_n(E)$ (resp. $\mathbb{E}_n[\chi]$) to denote the probability of an event E (resp. the expected value of a statistics χ) under the Ewens-like distribution on $\mathfrak{S}_n$ for records.

Remark: To some extent, our results may also be interpreted on the classical Ewens distribution, via the fundamental bijection. Indeed the number of records (resp. the number of descents, resp. the first value) of σ corresponds to the number of cycles (resp. the number of anti-excedances[(iii)] , resp. the minimum over

[(ii)] For details, see `https://en.wikipedia.org/wiki/Digamma_function` (accessed on April 27, 2016).

[(iii)] An anti-excedance of $\sigma \in \mathfrak{S}_n$ is $i \in [n]$ such that $\sigma(i) < i$. The proof that descents of σ are equinumerous with anti-excedances of $F^{-1}(\sigma)$ is a simple adaptation of the proof of Theorem 1.36 in [3], p. 110–111.

	$\theta = 1$ (uniform)	fixed $\theta > 0$	$\theta := n^\epsilon$, $0 < \epsilon < 1$	$\theta := \lambda n$, $\lambda > 0$	$\theta := n^\delta$, $\delta > 1$	See Cor.
$\mathbb{E}_n[\text{record}]$	$\log n$	$\theta \cdot \log n$	$(1-\epsilon) \cdot n^\epsilon \log n$	$\lambda \log(1 + 1/\lambda) \cdot n$	n	4
$\mathbb{E}_n[\text{desc}]$	$n/2$	$n/2$	$n/2$	$n/2(\lambda + 1)$	$n^{2-\delta}/2$	6
$\mathbb{E}_n[\sigma(1)]$	$n/2$	$n/(\theta + 1)$	$n^{1-\epsilon}$	$(\lambda + 1)/\lambda$	1	9
$\mathbb{E}_n[\text{inv}]$	$n^2/4$	$n^2/4$	$n^2/4$	$n^2/4 \cdot f(\lambda)$	$n^{3-\delta}/6$	11

Tab. 1: Asymptotic behavior of some permutation statistics under the Ewens-like distribution on $\mathfrak{S}_n$ for records. We use the shorthand $f(\lambda) = 1 - 2\lambda + 2\lambda^2 \log(1 + 1/\lambda)$. All the results in this table are asymptotic equivalents.

all cycles of the maximum value in a cycle) of $F^{-1}(\sigma)$. Consequently, Corollary 4 is just a consequence of the well-known expectation of the number of cycles under the Ewens distribution (see for instance [1, §5.2]). Similarly, the expected number of anti-excedances (Corollary 6) can be derived easily from the results of [6]. Those results on the Ewens distribution do not however give access to results as precise as those stated in Theorems 3 and 5, which are needed to prove our results of Section 4. Finally, to the best of our knowledge, the behavior of the third statistics (minimum over all cycles of the maximum value in a cycle) has not been previously studied, and we are not aware of any natural interpretation of the number of inversions of σ in $F^{-1}(\sigma)$. ◇

3.3 Expected values of some permutation statistics

We start our study by computing how the value of parameter θ influences the expected number of records.

Theorem 3 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, for any $i \in [n]$, the probability that there is a record at position i is: $\mathbb{P}_n(\text{record at } i) = \frac{\theta}{\theta+i-1}$.*

Proof: We prove this theorem by splitting permutations seen as words after their i-th element, as shown in Fig. 2. Let $\mathcal{R}_{n,i}$ denote the set of permutations of $\mathfrak{S}_n$ having a record at position i. We claim that the set $\mathcal{R}_{n,i}$ is quotient-stable for $n - i$, and that $\mathrm{w}(\mathcal{R}_{n,i}) = \frac{\theta^{(n)}}{\theta^{(i)}} \cdot \theta^{(i-1)} \cdot \theta$. It will immediately follow that $\mathbb{P}_n(\text{record at } i) = \frac{\mathrm{w}(\mathcal{R}_{n,i})}{\theta^{(n)}} = \frac{\theta^{(i-1)} \cdot \theta}{\theta^{(i)}} = \frac{\theta}{\theta+i-1}$. We now prove the claim. Let τ be any sequence in $\mathfrak{S}_{n-i \text{ in } n}$. Observe that $\mathrm{norm}(\mathcal{R}_{n,i}/\tau) = \mathcal{R}_{i,i}$. Since the number of records is stable by normalization, it follows that $\mathrm{w}(\mathcal{R}_{n,i}/\tau) = \mathrm{w}(\mathcal{R}_{i,i})$. By definition, $\pi \in \mathfrak{S}_i$ is in $\mathcal{R}_{i,i}$ if and only if $\pi(i) = i$. Thus $\mathcal{R}_{i,i} = \mathfrak{S}_{i-1} \cdot i$ in the word representation of permutations. Hence, $\mathrm{w}(\mathcal{R}_{i,i}) = \theta^{(i-1)}\theta$, since the last element is a record by definition. This yields $\mathrm{w}(\mathcal{R}_{n,i}/\tau) = \theta^{(i-1)}\theta$ for any $\tau \in \mathfrak{S}_{n-i \text{ in } n}$, proving that $\mathcal{R}_{n,i}$ is quotient-stable for $n - i$, and that $\mathrm{w}^{\mathrm{q}}_{n-i}(\mathcal{R}_{n,i}) = \theta^{(i-1)}\theta$. By Lemma 2, it follows that $\mathrm{w}(\mathcal{R}_{n,i}) = \frac{\theta^{(n)}}{\theta^{(n-(n-i))}} \cdot \mathrm{w}^{\mathrm{q}}_{n-i}(\mathcal{R}_{n,i}) = \frac{\theta^{(n)}}{\theta^{(i)}} \cdot \theta^{(i-1)} \cdot \theta$. □

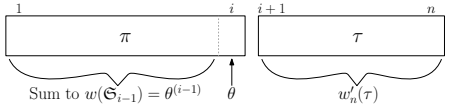


Fig. 2: The decomposition used to compute the probability of having a record at i. This record has weight θ and thus, for any fixed τ, the weights of all possible π sum to $\mathrm{w}(\mathfrak{S}_{i-1}) \cdot \theta = \theta^{(i-1)} \cdot \theta$.

Corollary 4 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected value of the number of records is: $\mathbb{E}_n[\text{record}] = \sum_{i=1}^{n} \frac{\theta}{\theta+i-1} = \theta \cdot \Delta(\theta, n)$.*

Next, we study the expected number of descents. Recall that a permutation σ of $\mathfrak{S}_n$ has a *descent* at position $i \in \{2, \dots, n\}$ if $\sigma(i-1) > \sigma(i)$. We denote by $\mathrm{desc}(\sigma)$ the number of descents in σ. We are interested in descents as they are directly related to the number of increasing runs in a permutation (each such run but the last one is immediately followed by a descent, and conversely). Some sorting algorithms, like Knuth's Natural Merge Sort, use the decomposition into runs.

The following theorem is proved using Lemmas 1 and 2 and the decomposition of Fig. 3.

Theorem 5 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, for any $i \in \{2, \dots, n\}$, the probability that there is a descent at position i is: $\mathbb{P}_n\big(\sigma(i-1) > \sigma(i)\big) = \frac{(i-1)(2\theta+i-2)}{2(\theta+i-1)(\theta+i-2)}$.*

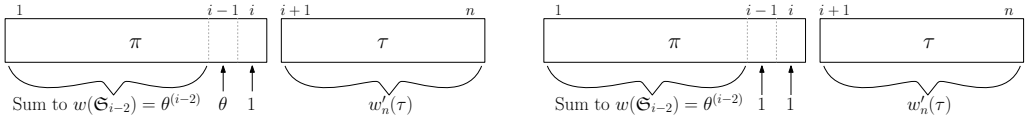


Fig. 3: The two cases for the probability of having a descent at i. We decompose σ as $\pi \cdot \sigma(i-1) \cdot \sigma(i) \cdot \tau$, and we let $\rho = \mathrm{norm}(\pi \cdot \sigma(i-1) \cdot \sigma(i))$. On the left, the case where $\sigma(i-1)$ is a record, that is, $\rho(i-1) = i$: there are $i-1$ possibilities for $\rho(i)$. On the right, the case where $\sigma(i-1)$ is not a record: there are $\binom{i-1}{2}$ possibilities for the values of $\rho(i)$ and $\rho(i-1)$. In both cases, once the images of $j \in \{i-1, \dots n\}$ by σ have been chosen, the weight of all possible beginnings sum to $w(\mathfrak{S}_{i-2}) = \theta^{(i-2)}$.

Corollary 6 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected value of the number of descents is: $\mathbb{E}_n[\mathrm{desc}] = \frac{n(n-1)}{2(\theta+n-1)}$.*

In the second row of Table 3.2, remark that the only way of obtaining a sublinear number of descents is to take very large values for θ.

Finally, we study the expected value of $\sigma(1)$. We are interested in this statistic to show a proof that differs from the ones for the numbers of records and descents: the expected value of the first element of a permutation is not obtained using Lemma 2.

Lemma 7 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, for any $k \in [0, n-1]$, the probability that a permutation starts with a value larger than k is: $\mathbb{P}_n(\sigma(1) > k) = \frac{(n-1)!\theta^{(n-k)}}{(n-k-1)!\,\theta^{(n)}}$.*

Proof: Let $\mathcal{F}_{n,k}$ denote the set of permutations of $\mathfrak{S}_n$ such that $\sigma(1) > k$. Such a permutation can uniquely be obtained by choosing the preimages of the elements in $[k]$ in $\{2, \dots, n\}$, then by mapping bijectively the remaining elements to $[k+1, n]$. Since none of the elements in $[k]$ is a record and since the elements of $[k+1, n]$ can be ordered in all possible ways, we get that $\mathrm{w}(\mathcal{F}_{n,k}) = \binom{n-1}{k} k!\,\theta^{(n-k)}$. Indeeed,there are $\binom{n-1}{k} k!$ ways to position and order the elements of $[k]$, and the total weight of the elements larger than k is $\theta^{(n-k)}$. Hence, $\mathbb{P}_n(\sigma(1) > k) = \frac{\mathrm{w}(\mathcal{F}_{n,k})}{\mathrm{w}(\mathfrak{S}_n)} = \frac{\binom{n-1}{k} k!\theta^{(n-k)}}{\theta^{(n)}} = \frac{(n-1)!\theta^{(n-k)}}{(n-k-1)!\,\theta^{(n)}}$. □

Theorem 8 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, for any $k \in [n]$, the probability that a permutation starts with k is: $\mathbb{P}_n(\sigma(1) = k) = \frac{(n-1)!\,\theta^{(n-k)}\theta}{(n-k)!\theta^{(n)}}$.*

Corollary 9 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected value of the first element of a permutation is:* $\mathbb{E}_n[\sigma(1)] = \frac{\theta+n}{\theta+1}$.

Remark: Our proof of Corollary 9 relies on calculus, but gives a very simple expression for $\mathbb{E}_n[\sigma(1)]$. We could therefore hope for a more combinatorial proof of Corollary 9, but we were not able to find it. $\diamond$

3.4 Number of inversions and expected running time of INSERTIONSORT

Recall that an *inversion* in a permutation $\sigma \in \mathfrak{S}_n$ is a pair $(i, j) \in [n] \times [n]$ such that $i < j$ and $\sigma(i) > \sigma(j)$. In the word representation of permutations, this corresponds to a pair of elements in which the largest is to the left of the smallest. This equivalent definition of inversions naturally generalizes to sequences of distinct integers. For any $\sigma \in \mathfrak{S}_n$, we denote by $\mathrm{inv}(\sigma)$ the number of inversions of σ, and by $\mathrm{inv}_j(\sigma)$ the number inversions of the form (i, j) in σ, for any $j \in [n]$. More formally, $\mathrm{inv}_j(\sigma) = \big|\{i \in [j-1] :\ (i, j) \text{ is an inversion of } \sigma\}\big|$.

Theorem 10 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, for any $j \in [n]$ and $k \in [0, j-1]$, the probability that there are k inversions of the form (i, j) is:* $\mathbb{P}_n\big(\mathrm{inv}_j(\sigma) = k\big) = \frac{1}{\theta+j-1}$ *if $k \neq 0$ and* $\mathbb{P}_n\big(\mathrm{inv}_j(\sigma) = k\big) = \frac{\theta}{\theta+j-1}$ *if $k = 0$.*

Corollary 11 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected value of the number of inversions is:* $\mathbb{E}_n[\mathrm{inv}] = \frac{n(n+1-2\theta)}{4} + \frac{\theta(\theta-1)}{2}\Delta(\theta, n)$.

Recall that the INSERTIONSORT algorithm works as follows: at each step $i \in \{2, \ldots, n\}$, the first $i-1$ elements are already sorted, and the i-th element is then inserted at its correct place, by swapping the needed elements.

It is well known that the number of swaps performed by INSERTIONSORT when applied to σ is equal to the number of inversions $\mathrm{inv}(\sigma)$ of σ. Moreover, the number of comparisons $C(\sigma)$ performed by the algorithm satisfies $\mathrm{inv}(\sigma) \leq C(\sigma) \leq \mathrm{inv}(\sigma) + n - 1$ (see [5] for more information on INSERTIONSORT).

As a direct consequence of Corollary 11 and the asymptotic estimates of the fourth row of Table 3.2, we get the expected running time of INSERTIONSORT:

Corollary 12 *Under the Ewens-like distribution for records with parameter $\theta = \mathcal{O}(n)$, the expected running time of* INSERTIONSORT *is $\Theta(n^2)$, like under the uniform distribution. If $\theta = n^\delta$ with $1 < \delta < 2$, it is $\Theta(n^{3-\delta})$. If $\theta = \Omega(n^2)$, it is $\Theta(n)$.*

4 Expected Number of Mispredictions for the Min/Max Search

4.1 Presentation

In this section, we turn our attention to a simple and classical problem: computing both the minimum and the maximum of an array of size n. The straightforward approach (called *naive* in the sequel) is to compare all the elements of the array to the current minimum and to the current maximum, updating them when it is relevant. This is done[(iv)] in Algorithm 1 and uses exactly $2n - 2$ comparisons. A classical optimization is to look at the elements in pairs, and to compare the smallest to the current minimum and the largest to the current maximum (see Algorithm 2). This uses only $3n/2$ comparisons, which is optimal. However, as reported in [2], with an implementation in C of these two algorithms, the naive algorithm proves to be

[(iv)] Note that, for consistency, our arrays start at index 1, as stated at the beginning of this paper.

the fastest on uniform permutations as input. The explanation for this is a trade-off between the number of comparisons involved and an other inherent but less obvious factor that influences the running time of these algorithms: the behavior of the branch predictor.

Algorithm 1: NAIVEMINMAX(T, n)

```
1 min ← T[1]
2 max ← T[1]
3 for i ← 2 to n do
4     if T[i] < min then
5         min ← T[i]
6     if T[i] > max then
7         max ← T[i]
8 return min, max
```

Algorithm 2: 3/2-MINMAX(T, n)

```
1 min, max ← T[n], T[n]
2 for i ← 2 to n by 2 do
3     if T[i − 1] < T[i] then
4         pMin, pMax ← T[i − 1], T[i]
5     else pMin, pMax ← T[i], T[i − 1] if
      pMin < min then min ← pMin if
      pMax > max then max ← pMax
6 return min, max
```

In a nutshell, when running on a modern processor, the instructions that constitute a program are not executed strictly sequentially but instead, they usually overlap one another since most of the instructions can start before the previous one is finished. This mechanism is commonly described as a *pipeline* (see [8] for a comprehensive introduction on this subject). However, not all instructions are well-suited for a pipelined architecture: this is specifically the case for branching instructions such as an *if* statement. When arriving at a branch, the execution of the next instruction should be delayed until the outcome of the test is known, which stalls the pipeline. To avoid this, the processor tries to predict the result of the test, in order to decide which instruction will enter the pipeline next. If the prediction is right, the execution goes on normally, but in case of a *misprediction*, the pipeline needs to be flushed, which can significantly slow down the execution of a program.

There is a large variety of branch predictors, but nowadays, most processors use *dynamic* branch prediction: they remember partial information on the results of the previous tests at a given *if* statement, and their prediction for the current test is based on those previous results. These predictors can be quite intricate, but in the sequel, we will only consider local *1-bit predictors* which are state buffers associated to each *if* statement: they store the last outcome of the test and guess that the next outcome will be the same.

Let us come back to the problem of simultaneously finding the minimum and the maximum in an array. We can easily see that, for Algorithm 1, the behavior of a 1-bit predictor when updating the maximum (resp. minimum) is directly linked to the succession of records (resp. min-records[(v)]) in the array. As we explain later on, for Algorithm 2, this behavior depends on the "pattern" seen in four consecutive elements of the array, this "pattern" indicating not only which elements are records (resp. min-records), but also where we find descents between those elements. As shown in [2], for uniform permutations, Algorithm 1 outerperforms Algorithm 2, because the latter makes more mispredictions than the former, compensating for the fewer comparisons made by Algorithm 2. This corresponds to our Ewens-like distribution for $\theta = 1$. But when θ varies, the way records are distributed also changes, influencing the performances of both Algorithms 1 and 2. Specifically, when $\theta = \lambda n$, we have a linear number of records (as opposed to a logarithmic number when $\theta = 1$). The next subsections provide a detailed analysis of the number of mispredictions in Algorithms 1 and 2, under the Ewens-like distribution for records, with a particular emphasis on $\theta = \lambda n$ (which exhibits a very different behavior w.r.t. the uniform distribution – see Fig. 4).

[(v)] A *min-record* (a.k.a. left to right minimum) is an element of the array such that no smaller element appears to its left.

4.2 Expected Number of Mispredictions in NaiveMinMax

Theorem 13 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected numbers of mispredictions at lines 4 and 6 of Algorithm 1 satisfy respectively $\mathbb{E}_n[\mu_4] \leq \frac{2}{\theta}\mathbb{E}_n[\mathrm{record}]$ and $\mathbb{E}_n[\mu_6] = 2\theta\Delta(\theta, n-1) - \frac{(2\theta+1)(n-1)}{\theta+n-1}$.*

Consequently, with our previous results on $\mathbb{E}_n[\mathrm{record}]$, the expected number of mispredictions at line 4 is $\mathcal{O}(\log n)$ when $\theta = \Omega(1)$ (*i.e.*, when $\theta = \theta(n)$ is constant or larger). Moreover, using the asymptotic estimates of the digamma function, the asymptotics of the expected number of mispredictions at line 6 is such that (again, for $\lambda > 0$, $0 < \epsilon < 1$ and $\delta > 1$):

	fixed $\theta > 0$	$\theta := n^\epsilon$	$\theta := \lambda n$	$\theta := n^\delta$
$\mathbb{E}_n[\mu_6]$	$\sim 2\theta \cdot \log n$	$\sim 2(1-\epsilon) \cdot n^\epsilon \log n$	$\sim 2\lambda(\log(1+1/\lambda) - 1/(\lambda+1)) \cdot n$	$o(n)$

In particular, asymptotically, the expected total number of mispredictions of Algorithm 1 is given by $\mathbb{E}_n[\mu_6]$ (up to a constant factor when θ is constant).

4.3 Expected Number of Mispredictions in $\frac{3}{2}$MinMax

Mispredictions in Algorithm 2 can arise in any of the three *if* statements. We first compute the expected number of mispredictions at each of them independently. We start with the *if* statement of line 3, which compares $T[i-1]$ and $T[i]$. For our 1-bit model, there is a misprediction whenever there is a descent at $i-2$ and an ascent at i, or an ascent at i and a descent at $i-2$. A tedious study of all possible cases gives:

Theorem 14 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected number of mispredictions at line 3 of Algorithm 2 satisfies*

$$\begin{aligned}\mathbb{E}_n[\nu_3] =& \frac{n-2}{4} + \frac{\theta(\theta-1)^2}{4} + \frac{\theta^2(\theta-1)^2}{12}\left(\frac{1}{\theta+n-1} - \frac{3}{\theta+n-2} - \frac{1}{\theta+1}\right)\\ &+ \frac{\theta^2(\theta-1)^2}{6}\left(\Delta\left(\frac{\theta+1}{2}, \frac{n-2}{2}\right) - \Delta\left(\frac{\theta}{2}, \frac{n-2}{2}\right)\right).\end{aligned}$$

As a consequence, if $\theta = \lambda n$, then $\mathbb{E}_n[\nu_3] \sim \frac{6\lambda^2+8\lambda+3}{12(\lambda+1)^3}\, n$.

Theorem 15 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected number of mispredictions at line 5 of Algorithm 2 satisfies $\mathbb{E}_n[\nu_7] \leq \frac{2}{\theta}\mathbb{E}_n[\mathrm{record}]$. As a consequence, if $\theta = \lambda n$, then $\mathbb{E}_n[\nu_7] = \mathcal{O}(1)$.*

We now consider the third *if* statement of Algorithm 2. If there is a record (resp. no record) at position $i-3$ or $i-2$, then there is a misprediction when there is no record (resp. a record) at position $i-1$ or i. Studying all the possible configurations at these four positions gives the following result.

Theorem 16 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter θ, the expected number of mispredictions at line 5 of Algorithm 2 satisfies*

$$\begin{aligned}\mathbb{E}_n[\nu_8] =& \frac{(n-2)((2\theta^3+\theta^2-9\theta-3)n + 2\theta^4 - 5\theta^2 + 9\theta + 3)}{3(\theta+n-1)(\theta+n-2)}\\ &+ \frac{\theta(2\theta^3+\theta+3)}{3}\,\Delta\left(\frac{\theta+1}{2}, \frac{n-2}{2}\right) - \frac{\theta(2\theta^3+\theta-3)}{3}\,\Delta\left(\frac{\theta}{2}, \frac{n-2}{2}\right).\end{aligned}$$

As a consequence, if $\theta = \lambda n$, then $\mathbb{E}_n[\nu_7] \sim \left(2\lambda \log\left(1+\frac{1}{\lambda}\right) - \frac{\lambda(6\lambda^2+15\lambda+10)}{3(\lambda+1)^3}\right)n$.

It follows from Theorems 14, 15 and 16 that:

Corollary 17 *Under the Ewens-like distribution on $\mathfrak{S}_n$ for records with parameter $\theta = \lambda n$, the total number of mispredictions of Algorithm 2 is*

$$\mathbb{E}_n[\nu] \sim \left(2\lambda \log\left(1+\frac{1}{\lambda}\right) - \frac{24\lambda^3+54\lambda^2+32\lambda-3}{12(\lambda+1)^3}\right) n.$$

Fig. 4 shows that, unlike in the uniform case ($\theta = 1$), Algorithm 2 is more efficient than Algorithm 1 under the Ewens-like distribution for records with $\theta := \lambda n$, as soon as λ is large enough.

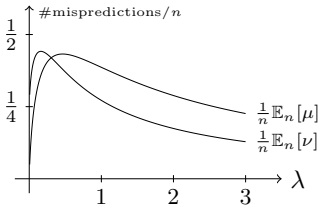


Fig. 4: The expected number of mispredictions produced by the naive algorithm (μ) and for $\frac{3}{2}$-minmax (ν), when $\theta := \lambda n$. We have $\mathbb{E}_n[\mu] \sim \mathbb{E}_n[\nu]$ for $\lambda_0 = \frac{\sqrt{34}-4}{6} \approx 0.305$, and there are fewer mispredictions on average with $\frac{3}{2}$-minmax as soon as $\lambda > \lambda_0$. However, since $\frac{3}{2}$-minmax performs $\frac{n}{2}$ fewer comparisons than the naive algorithm, it becomes more efficient before λ_0. For instance, if a misprediction is worth 4 comparisons, $\frac{3}{2}$-minmax is the most efficient as soon as $\lambda > 0.110$.

Acknowledgments Many thanks to Valentin Féray for providing insight and references on the classical Ewens distribution and to the referees whose comments helped us clarify the presentation of our work.

References

[1] R. Arratia, A. D. Barbour, and S. Tavaré. *Logarithmic combinatorial structures: a probabilistic approach*. EMS Monographs in Mathematics. EMS, Zürich, 2003.

[2] N. Auger, C. Nicaud, and C. Pivoteau. Good predictions are worth a few comparisons. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016, Orléans, France*, pages 12:1–12:14, 2016.

[3] M. Bóna. *Combinatorics of permutations*. Chapman-Hall and CRC Press, 2d edition edition, 2012.

[4] A. Borodin, P. Diaconis, and J. Fulman. On adding a list of numbers (and other one-dependent determinantal processes). *Bulletin of the American Mathematical Society (N.S.)*, 47(4):639–670, 2010.

[5] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and Clifford S. *Introduction to Algorithms*. MIT Press, Cambridge, MA, third edition, 2009.

[6] V. Féray. Asymptotics of some statistics in Ewens random permutations. *Electronic Journal of Probability*, 18(76):1–32, 2013.

[7] A. Gladkich and R. Peled. On the cycle structure of Mallows permutations. Preprint available at `http://arxiv.org/abs/1601.06991`.

[8] J. L. Hennessy and D. A. Patterson. *Computer Architecture, Fifth Edition: A Quantitative Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 5th edition, 2011.

[9] D. E. Knuth. *The Art of Computer Programming, Volume 3: (2nd Ed.) Sorting and Searching*. Addison Wesley Longman Publish. Co., Redwood City, CA, USA, 1998.

[10] Heikki Mannila. Measures of presortedness and optimal sorting algorithms. *IEEE Trans. Computers*, 34(4):318–325, 1985.

[11] O. Petersson and A. Moffat. A framework for adaptive sorting. *Discrete Applied Mathematics*, 59(2):153–179, may 1995.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Counting Zeros in Random Walks on the Integers and Analysis of Optimal Dual-Pivot Quicksort

Martin Aumüller[1], Martin Dietzfelbinger[2], Clemens Heuberger[3‡], Daniel Krenn[4‡], Helmut Prodinger[5§]

[1] *IT University of Copenhagen, Denmark, maau@itu.dk*
[2] *Institut für Theoretische Informatik, Technische Universität Ilmenau, Germany, martin.dietzfelbinger@tu-ilmenau.de*
[3] *Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Austria, clemens.heuberger@aau.at*
[4] *Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Austria, math@danielkrenn.at or daniel.krenn@aau.at*
[5] *Department of Mathematical Sciences, Stellenbosch University, South Africa, hproding@sun.ac.za*

We present an average case analysis of two variants of dual-pivot quicksort, one with a non-algorithmic comparison-optimal partitioning strategy, the other with a closely related algorithmic strategy. For both we calculate the expected number of comparisons exactly as well as asymptotically, in particular, we provide exact expressions for the linear, logarithmic, and constant terms. An essential step is the analysis of zeros of lattice paths in a certain probability model. Along the way a combinatorial identity is proven.

Keywords: Dual-pivot quicksort, lattice paths, asymptotic enumeration, combinatorial identity

1 Introduction

Dual-pivot quicksort [Sed75, WNN15, AD15] is a family of sorting algorithms related to the well-known quicksort algorithm. In order to sort an input sequence $(a_1, \dots, a_n)$ of distinct elements, dual-pivot quicksort algorithms work as follows. (For simplicity we forbid repeated elements in the input.) If $n \le 1$, there is nothing to do. If $n \ge 2$, two input elements are selected as pivots. Let p be the smaller and q be the larger pivot. The next step is to partition the remaining elements into

- the elements smaller than p ("small elements"),
- the elements between p and q ("medium elements"), and
- the elements larger than q ("large elements").

‡C. Heuberger and D. Krenn are supported by the Austrian Science Fund (FWF): P 24644-N26 and by the Karl Popper Kolleg "Modeling–Simulation–Optimization" funded by the Alpen-Adria-Universität Klagenfurt and by the Carinthian Economic Promotion Fund (KWF).
§H. Prodinger is supported by an incentive grant of the National Research Foundation of South Africa.

Then the procedure is applied recursively to these three groups to complete the sorting.

The cost measure used in this work is the number of comparisons between elements. As is common, we will assume the input sequence is in random order, which means that each permutation of the n elements occurs with probability $1/n!$. With this assumption we may, without loss of generality, choose a_1 and a_n as the pivots. Even in this setting there are different dual-pivot quicksort algorithms; their difference lies in the way the partitioning is organized, which influences the partitioning cost. This is in contrast to standard quicksort with one pivot, where the partitioning strategy does not influence the cost—in partitioning always one comparison is needed per non-pivot element. In dual-pivot quicksort, the average cost (over all permutations) of partitioning and of sorting can be analyzed only when the partitioning strategy is fixed.

Only in 2009, Yaroslavskiy, Bentley, and Bloch [Yar09] described a dual-pivot quicksort algorithm that makes $1.9n \log n + O(n)$ comparisons [WNN15].[(i)] This beats the classical quicksort algorithm [Hoa62], which needs $2n \log n + O(n)$ comparisons on average. In [AD15], the first two authors of this article described the full design space for dual-pivot quicksort algorithms with respect to counting element comparisons. Among others, they studied two special partitioning strategies. The first one—we call it "Clairvoyant" in this work—assumes that the number of small and large elements is given (by an "oracle") before partitioning starts. It cannot be implemented; however, it is optimal among all partioning strategies that have access to such an oracle, and hence its cost provides a lower bound for the cost of all algorithmic partitioning strategies. In [AD15] it was shown that dual-pivot quicksort carries out $1.8n \log n + O(n)$ comparisons on average when this partitioning strategy is used. Further a closely related algorithmic partitioning strategy—called "Count" here—was described, which makes only $O(\log n)$ more comparisons on average than "Clairvoyant" and hence leads to a dual-quicksort variant with only $O(n)$ more comparisons.[(ii)]

One purpose of this paper is to make the expected number of comparisons in both variants precise and to determine the exact difference of the cost of these two strategies, both for partitioning and for the resulting dual-pivot quicksort variants.

Already in [AD15] it was noted that the exact value of the expected partitioning cost (i.e., the number of comparisons) of both strategies depends on the expected number of the zeros of certain lattice paths (Parts I and II). A complete understanding of this situation is the basis for our analysis of dual-pivot quicksort, which appears in Part III.

Lattice path enumeration has a long tradition. An early reference is [Moh79]; a recent survey paper is [Kra15]. As space is limited, many proofs and some additional results can be found in an appendix at arXiv:1602.04031v1.

2 Overview and Results

This work is split into three parts. We give a brief overview on the main results of each of these parts here. We use the Iversonian expression

$$[expr] = \begin{cases} 1 & \text{if } expr \text{ is true,} \\ 0 & \text{if } expr \text{ is false,} \end{cases}$$

(i) In this paper "log" denotes the natural logarithm to base e.

(ii) After completing this extended abstract we found a proof that "Count" is optimal among all algorithmic strategies. Details to be given in the full version.

popularized by Graham, Knuth, and Patashnik [GKP94].

The harmonic numbers and their variants will be denoted by

$$H_n = \sum_{m=1}^{n} \frac{1}{m}, \qquad H_n^{\text{odd}} = \sum_{m=1}^{n} \frac{[m \text{ odd}]}{m} \qquad \text{and} \qquad H_n^{\text{alt}} = \sum_{m=1}^{n} \frac{(-1)^m}{m}.$$

Of course, there are relations between these three definitions such as $H_n^{\text{alt}} = H_n - 2H_n^{\text{odd}}$ and $H_n^{\text{odd}} + H_{\lfloor n/2 \rfloor}/2 = H_n$, but it will turn out to be much more convenient to use all three notations.

Part I: Lattice Paths

In the first part we analyze certain lattice paths of a fixed length n. We start on the vertical axis, allow steps/increments $(1, +1)$ and $(1, -1)$ and end on the horizontal axis at $(n, 0)$. To be precise, the starting point on the vertical axis is chosen uniformly at random from the set $\{(0, -n), (0, -n + 2), \ldots, (0, n - 2), (0, n)\}$ of feasible points. Once this starting point is fixed, all paths to $(n, 0)$ are equally likely. We are interested in the number of zeros, denoted by the random variable Z_n, of such paths.

An exact formula for the expected number $\mathbb{E}(Z_n)$ of zeros is derived in two different ways (see identity (2.1) for these formulæ): On the one hand, we use the symbolic method and generating functions (see Appendix A), which gives the result in form of a double sum. This machinery extends well to higher moments and also allows us to obtain the distribution. The exact distribution is given in Appendix E; its limiting behavior as $n \to \infty$ is the discrete distribution

$$\mathbb{P}(Z_n = r) \sim \frac{1}{r(r+1)}.$$

On the other hand, a more probabilistic approach gives the expectation $\mathbb{E}(Z_n)$ as the simple single sum

$$\mathbb{E}(Z_n) = \sum_{m=1}^{n+1} \frac{[m \text{ odd}]}{m} = H_{n+1}^{\text{odd}},$$

see Section 4 for more details. The asymptotic behavior $\mathbb{E}(Z_n) \sim \frac{1}{2} \log n$ can be extracted (Appendix D).

The two approaches above give rise to the identity

$$\sum_{m=1}^{n+1} \frac{[m \text{ odd}]}{m} = \frac{4}{n+1} \sum_{0 \le k < \ell < \lceil n/2 \rceil} \frac{\binom{n}{k}}{\binom{n}{\ell}} + [n \text{ even}] \frac{1}{n+1} \left(\frac{2^n}{\binom{n}{n/2}} - 1 \right) + 1; \tag{2.1}$$

the double sum above equals the single sum of Theorem 4.1 by combinatorial considerations. One might ask about a direct proof of this identity. This can be achieved by methods related to hypergeometric sums and the computational proof is presented in Appendix C. We also provide a completely elementary proof which is "purely human".

Part II: More Lattice Paths and Zeros

The second part acts as connecting link between the lattice paths of fixed length of Part I and the dual-pivot quicksort algorithms of Part III.

The probabilistic model introduced in Section 3 (in Part I) is extended, and lattice paths are allowed to vary in length. For a number n (the number of elements to sort) the length of a path is the number of elements remaining when the two pivots, given by a random set of elements of size two, and the elements between these pivots are cut out.

The number of zeros X_n in this full model is analyzed; we provide again exact as well as asymptotic formulæ for the expectation $\mathbb{E}(X_n)$. Details are given in Section 7. Moreover, more specialized zero-configurations (needed for the analysis of different partitioning strategies in Part III) are considered as well (Section 6).

Part III: Dual-Pivot Quicksort

The main result of this work analyzes comparisons in the dual-pivot quicksort algorithm that uses the optimal (but unrealistic) partitioning strategy "Clairvoyant". Aumüller and Dietzfelbinger showed in [AD15] that this algorithm requires $1.8n \log n + O(n)$ comparisons on average, which improves on the average number of comparisons in quicksort ($2n \log n + O(n)$) and the recent dual-pivot algorithm of Yaroslavskiy et al. ($1.9n \log n + O(n)$, see [WNN15]). However, for real-world input sizes n the (usually negative) factor in the linear term has a great influence on the comparison count. Our asymptotic result is stated as the following theorem.

Theorem. *The average number of comparisons in the dual-pivot quicksort algorithm with a comparison-optimal partitioning strategy is*

$$\frac{9}{5} n \log n + An + B \log n + C + O(1/n)$$

as $n \to \infty$, with $A = \frac{9}{5}\gamma - \frac{1}{5}\log 2 - \frac{89}{25} = -2.659\ldots$.

The constants B and C are explicitly given, too, and more terms of the asymptotics are presented. The precise result is formulated as Corollary 10.2.

In fact, we even get an exact expression for the average comparison count. The precise result is formulated as Theorem 10.1. Moreover the same analysis is carried out for the partitioning strategy "Count", which is an algorithmic variant of the comparison-optimal strategy "Clairvoyant". Aumüller and Dietzfelbinger [AD15] could show that it requires $\frac{9}{5} n \log n + O(n)$ comparisons as well. In this paper we obtain the exact average comparison count (Theorem 10.3). The asymptotic result is again $\frac{9}{5} n \log n + An + O(\log n)$, but now with $A = -2.382\ldots$, so there is only a small gap between the average number of comparisons in the comparison-optimal strategy "Clairvoyant" and its algorithmic variant.

Part I: Lattice Paths

In this first part we analyze lattice paths of a fixed length n. These are introduced in Section 3 by a precise description of our probabilistic model. We will work with this model throughout Part I, and we analyze the number of zeros Z_n.

The outline is as follows: We derive an exact expression for the expected number $\mathbb{E}(Z_n)$ of zeros by the generating functions machinery in Appendix A; a more probabilistic approach can

be found in Section 4. Appendix D deals with asymptotic considerations. Direct proofs of the obtained identity are given in Appendix C and the distribution of Z_n is tackled in Appendix E.

3 Probabilistic Model

We consider paths of a given length n on the lattice $\mathbb{Z}^2$, where only steps $(1,+1)$ and $(1,-1)$ are allowed. These paths are chosen at random according to the rules below.

Let us fix a length $n \in \mathbb{N}_0$. A path P_n ending in $(n,0)$ (no choice for this end-point) is chosen according to the following rules.

1. First, choose a starting point $(0,S)$ where S is a random integer uniformly distributed in $\{-n,-n+2,\ldots,n-2,n\}$, i. e., $S=s$ occurs only for integers s with $|s|\le n$ and $s\equiv n \pmod 2$.

2. Second, a path is chosen uniformly at random among all paths from $(0,S)$ to $(n,0)$.

The conditions on S characterize those starting points from which $(n,0)$ is reachable.

We are interested in the number of intersections with the horizontal axis of a path. To make this precise, we define a *zero* of a path P_n as a point $(x,0)\in P_n$.

Thus, let P_n be a path of length n which is chosen according to the probabilistic model above and define the random variable

$$Z_n = \text{number of zeros of } P_n.$$

In the following sections, we determine the value of $\mathbb{E}(Z_n)$ exactly (Appendix A and Section 4), as well as asymptotically (Appendix D). In Appendix A, we use the machinery of generating functions. This machinery turns out to be overkill if we are just interested in the expectation $\mathbb{E}(Z_n)$. However, it easily allows extension to higher moments and the limiting distribution.

In Section 4, we follow a probabilistic approach, which first gives a result on the probability model that at the first glance looks surprising: the equidistribution at the initial values turns out to carry over to every fixed length of the remaining path. This result yields a simple expression for the expectation $\mathbb{E}(Z_n)$ in terms of harmonic numbers, and thus immediately yields a precise asymptotic expansion for $\mathbb{E}(Z_n)$. The generating function approach, however, gives the expectation in terms of a double sum of quotients of binomial coefficients (the right-hand side of (2.1)), see Appendix A.

Appendix C gives a direct computational proof that these two results coincide. The original expression in [AD15] (a double sum over a quotient of a product of binomial coefficients and a binomial coefficient) is also shown to be equal in Appendix C. Explicit as well as asymptotic expressions for the distribution $\mathbb{P}(Z_n=r)$ can be found in Appendix E.

4 A Probabilistic Approach

Theorem 4.1. *For a randomly (as described in Section 3) chosen path of length n, the expected number of zeros is*

$$\mathbb{E}(Z_n) = H^{\mathrm{odd}}_{n+1}.$$

Before proving the theorem, we consider an equivalent probability model for our random paths formulated as an urn model. A number R from $\{0, \dots, n\}$ is chosen uniformly at random. We place R red balls and $B = n - R$ black balls in an urn. Subsequently, in n rounds the balls are taken from the urn (without replacements), in each round choosing one uniformly at random. The color of the ball chosen in round i is denoted by U_i.

We construct a random walk $(W_i)_{0 \le i \le n}$ on $\{-n, \dots, n\}$ from $U_1, \dots, U_n$ by setting $W_0 = R - B = 2R - n$ and

$$W_i = \begin{cases} W_{i-1} + 1 & \text{if } U_i = \text{black}, \\ W_{i-1} - 1 & \text{if } U_i = \text{red} \end{cases}$$

for $1 \le i \le n$. In each step, W_i equals the difference of the number of remaining red and black balls in the urn. Clearly, then, $W_n = 0$.

One can look at the trajectories of this random walk, represented in the grid $\{0, \dots, n\} \times \{-n, \dots, n\}$ as sequences $((0, W_0), (1, W_1), \dots, (n, W_n))$. Appendix B explains the equivalence between the two models.

In order to prove Theorem 4.1, we need the following property of our paths.

Lemma 4.2. *Let $m \in \mathbb{N}_0$ with $m \le n$. The probability that a random path P_n (as defined in Section 3) runs through $(n - m, k)$ is*

$$\mathbb{P}((n - m, k) \in P_n) = \frac{1}{m + 1} \tag{4.1}$$

for all k with $|k| \le m$ and $k \equiv m \pmod 2$, otherwise 0.

The proof of this lemma can be found in Appendix B.

A closer look reveals that when we reverse the paths, our model is equivalent to a contagion Pólya urn model with two colors, starting with one ball of each color, where we sample with replacement and put another ball of the color just drawn into the urn. In this setting, uniform distribution for feasible points with the same first coordinate and hence the result of the lemma are well-known phenomena. These results and more on the urn model can be found, for example, in Mahmoud [Mah08].

We continue with the actual proof of our theorem.

Proof of Theorem 4.1: By Lemma 4.2, the expected number of zeros of P_n is

$$\mathbb{E}(Z_n) = \sum_{m=0}^{n} \mathbb{P}((n - m, 0) \in P_n) = \sum_{m=0}^{n} \frac{[m \text{ even}]}{m + 1} = \sum_{m=1}^{n+1} \frac{[m \text{ odd}]}{m} = H_{n+1}^{\text{odd}}. \qquad \square$$

5 Additional Results

The expected number of zeros can be evaluated asymptotically. We obtain

Corollary 5.1.

$$\mathbb{E}(Z_n) = \frac{1}{2} \log n + \frac{\gamma + \log 2}{2} + \frac{1 + [n \text{ even}]}{2n} - \frac{2 + 9[n \text{ even}]}{12n^2} + O\Big(\frac{1}{n^3}\Big)$$

asymptotically as n tends to infinity.

The proof of this result uses the well-known asymptotic expansion of the harmonic numbers. The actual asymptotic computations[iii] have been carried out using the asymptotic expansions module [HK15] of SageMath [Dev16], see Appendix D.

By combining the generating function and probabilistic approach we obtain the following identity.

Theorem 5.2. *For $n \geq 0$, we have*

$$\begin{aligned}&\frac{4}{n+1}\sum_{0\leq k<\ell<\lceil n/2\rceil}\frac{\binom{n}{k}}{\binom{n}{\ell}}+[n \text{ even}]\,\frac{1}{n+1}\left(\frac{2^n}{\binom{n}{n/2}}-1\right)+1\\ &=\frac{1}{n+1}\sum_{m=0}^{\lfloor n/2\rfloor}\sum_{\ell=m}^{n-m}\frac{\binom{2m}{m}\binom{n-2m}{\ell-m}}{\binom{n}{\ell}}=H_{n+1}^{\mathrm{odd}}.\end{aligned}$$

The second expression for the expected number of zeros, but without taking the zero at $(n, 0)$ into account, has been given in [AD15, displayed equation after (14)]. In Appendix C we give two direct proofs of the identity above: One of them follows a computer generated proof ("creative telescoping") by extracting the essential recurrence. The second proof is "human" and completely elementary using not more than Vandermonde's convolution.

Furthermore, the generating function machinery allows us to determine the distribution of the number Z_n of zeros. Beside an exact formula (see Appendix E), we get the following asymptotic result.

Theorem 5.3. *Let $0 < \varepsilon \leq \frac{1}{2}$. For positive integers r with $r = O\big(n^{1/2-\varepsilon}\big)$, we have asymptotically*

$$\mathbb{P}(Z_n = r) = \frac{1}{r(r+1)}\left(1 + O\big(1/n^{2\varepsilon}\big)\right)$$

as n tends to infinity.

Part II: More Lattice Paths and Zeros

This second part deals with an analysis of some special zero-configurations, which are needed for the analysis of the partitioning strategies in Part III. Moreover, in Section 7, we extend the model introduced in Section 3 to accommodate lattice paths of variable length. Again expectations are studied exactly and asymptotically.

6 Going to Zero and Coming From Zero

For the analysis of comparison-optimal dual-pivot quicksort algorithms (see Part III) we need the following two variants of zeros on the lattice path.

- An *up-to-zero situation* is a point $(x, 0) \in P_n$ such that $(x - 1, -1) \in P_n$.
- A *down-from-zero situation* is a point $(x, 0) \in P_n$ such that $(x + 1, -1) \in P_n$.

[iii] A worksheet containing the computations can be found at `http://www.danielkrenn.at/downloads/quicksort-paths/quicksort-paths.ipynb`.

We show

$$\mathbb{E}(\text{number of up-to-zero situations on } P_n) = \frac{1}{2}\Big(\mathbb{E}(Z_n) - \frac{[n \text{ even}]}{n+1}\Big) = \frac{1}{2}H_n^{\text{odd}}$$

and

$$\mathbb{E}(\text{number of down-from-zero situations on } P_n) = \frac{1}{2}\left(\mathbb{E}(Z_n) - 1\right) = \frac{1}{2}\big(H_{n+1}^{\text{odd}} - 1\big).$$

Proof idea: The factor $\frac{1}{2}$ stems from symmetry: Up-to-zero situations at $(x, 0)$ occur with the same probability as the symmetric "down-to-zero" situations at $(x, 0)$, similarly for down-from-zero situations. The correction terms $\frac{[n \text{ even}]}{n+1}$ and 1 are caused by the fact that there is a zero, but no up-to-zero situation, at $(0, 0)$, and a zero, but no down-from-zero situation, at $(n, 0)$. The full proofs are in Appendix F. □

7 Lattice Paths of Variable Length

In this section, we use a random variable N' instead of the fixed n above. Let us fix an $n \in \mathbb{N}$ with $n \geq 2$. We choose a path length N' according to the following rules.

1. Choose (P, Q) with $1 \leq P < Q \leq n$ uniformly at random from all $\binom{n}{2}$ possibilities.
2. Let $N' = n - 1 - (Q - P)$.
3. Choose a path of length N' according to Section 3.

Let us denote the number of up-to-zero and down-from-zero situations on the path by $X_n^{\nearrow}$ and $X_n^{\searrow}$, respectively. In Appendix G, we show

$$\mathbb{E}\big(X_n^{\nearrow}\big) = \frac{1}{2\binom{n}{2}} \sum_{n'=0}^{n-2} \sum_{m=1}^{n'} [m \text{ odd}]\, \frac{n'+1}{m} = \frac{1}{2}H_{n-2}^{\text{odd}} - \frac{1}{8} + \frac{(-1)^n}{8(n - [n \text{ even}])}$$

and

$$\mathbb{E}\big(X_n^{\searrow}\big) = \frac{1}{2\binom{n}{2}} \sum_{n'=0}^{n-2} \sum_{m=3}^{n'+1} [m \text{ odd}]\, \frac{n'+1}{m} = \mathbb{E}\big(X_n^{\nearrow}\big) - \frac{1}{2} + \frac{1}{2(n - [n \text{ even}])}.$$

Part III: Dual-Pivot Quicksort

In this third and last part of this work, we finally analyze two different partitioning strategies and the dual-pivot quicksort algorithm itself.

As mentioned in the introduction, the number of comparisons of dual-pivot quicksort depends on the concrete partitioning procedure. For example, if one wants to classify a large element, i. e., an element larger than the larger pivot, comparing it with the larger pivot is unavoidable, but it depends on the partitioning procedure whether a comparison with the smaller pivot occurs, too.

First, in Section 8, we make our set-up precise, fix notions, and start solving the dual-pivot quicksort recurrence (8.1). This recurrence relates the cost of the partitioning step to the total number of comparisons of dual-pivot quicksort.

In Section 9 two partitioning strategies, called "Clairvoyant" and "Count", are introduced and their respective cost is analyzed. It will turn out that the results on lattice paths obtained in Parts I and II are central in determining the partitioning cost exactly.

Everything is put together in Section 10: We obtain the exact comparison count for two versions of dual-pivot quicksort (Theorems 10.1 and 10.3). The asymptotic behavior is extracted out of the exact results (Corollaries 10.2 and 10.4).

8 Solving the Dual-Pivot Quicksort Recurrence

We consider versions of dual-pivot quicksort that act as follows on an input sequence $(a_1, \dots, a_n)$ consisting of distinct numbers: If $n \le 1$, do nothing, otherwise choose a_1 and a_n as pivots, and by one comparison determine $p = \min(a_1, a_n)$ and $q = \max(a_1, a_n)$. Use a partitioning procedure to partition the remaining $n-2$ elements into the three classes *small*, *medium*, and *large*. Then call dual-pivot quicksort recursively on each of these three classes to finish the sorting, using the same partitioning procedure in all recursive calls.

Let P_n, a random variable, denote the *partitioning cost*. This is defined as the number of comparisons made by the partitioning procedure if the input $(a_1, \dots, a_n)$ is assumed to be in random order. Further, let C_n be the random variable that denotes the number of comparisons carried out when sorting n elements with dual-pivot quicksort. The reader should be aware that both P_n and C_n are determined by the partitioning procedure used.

Since the input $(a_1, \dots, a_n)$ is in random order and the partitioning procedure does nothing but compare elements with the two pivots, the inputs for the recursive calls are in random order as well, which implies that the distributions of P_n and C_n only depend on n. In particular we may assume that when the sorting algorithm is called on n elements during recursion, the input is a permutation of $\{1, \dots, n\}$.

The recurrence

$$\mathbb{E}(C_n) = \mathbb{E}(P_n) + \frac{3}{\binom{n}{2}} \sum_{k=1}^{n-2} (n-1-k)\, \mathbb{E}(C_k) \tag{8.1}$$

for $n \ge 0$ describes the connection between the expected sorting cost $\mathbb{E}(C_n)$ and the expected partitioning cost $\mathbb{E}(P_n)$. It will be central for our analysis. Note that it is irrelevant for (8.1) how the partitioning cost $\mathbb{E}(P_n)$ is determined; it need not even be referring to comparisons. The recurrence is simple and well-known; a version of it occurs already in Sedgewick's thesis [Sed75]. For the convenience of the reader we give a brief justification in Appendix H. In Hennequin [Hen91] recurrence (8.1) was solved exactly for $\mathbb{E}(P_n) = an + b$, where a and b are constants. For $\mathbb{E}(P_n) = an + O(n^{1-\varepsilon})$ the solution is $\mathbb{E}(C_n) = \frac{6}{5} an \log n + O(n)$, see [AD15, Theorem 1].

9 Partitioning Algorithms and Their Cost

In Section 8 we saw that in order to calculate the average number of comparisons $\mathbb{E}(C_n)$ of a dual-pivot quicksort algorithm we need the expected partitioning cost $\mathbb{E}(P_n)$ of the partitioning

procedure used. The aim of this section is to determine $\mathbb{E}(P_n)$ for two such partitioning procedures, "Clairvoyant" and "Count", to be described below.

We use the set-up described at the beginning of Section 8. For partitioning we use comparisons to *classify* the $n-2$ elements $a_2, \dots, a_{n-1}$ as *small*, *medium*, or *large*. We will be using the term *classification* for this central aspect of partitioning. Details of a partitioning procedure that concern how the classes are represented or elements are moved around may and will be ignored. (Nonetheless, in Appendix M we provide pseudocode for the considered classification strategies turned into dual-pivot quicksort algorithms.) The cost P_n depends on the concrete classification strategy used, the only relevant difference between classification strategies being whether the next element to be classified is compared with the smaller pivot p or the larger pivot q first. This decision may depend on the whole history of outcomes of previous comparisons. (The resulting abstract classification strategies may conveniently be described as classification trees, see [AD15], but we do not need this model here.)

Two comparisons are necessary for each medium element. Furthermore, one comparison with p is necessary for small and one comparison with q is necessary for large elements. As the input consists of the elements $1, \dots, n$, there are $p-1$ small, $q-p-1$ medium, and $n-q$ large elements. Averaging over all $\binom{n}{2}$ positions of the pivots, we see that on average

$$\frac{4}{3}(n-2)+1 \tag{9.1}$$

necessary comparisons are required no matter how the classification procedure works, see [AD15, (5)]; the summand $+1$ corresponds to the comparison of a_1 and a_n when choosing the two pivots.

We call other comparisons occurring during classification *additional comparisons*. That means, an additional comparison arises when a small element is compared with q first or a large element is compared with p first. In order to obtain $\mathbb{E}(P_n)$ for some classification strategy, we have to calculate the expected number of additional comparisons.

Next we describe two (closely related) classification strategies from [AD15]. Let s_i and ℓ_i denote the number of elements that have been classified as small and large, respectively, in the first i classification rounds. Set $s_0 = \ell_0 = 0$.

Strategy "Clairvoyant". *Assume the input contains $s = p-1$ small and $\ell = n-q$ large elements. When classifying the ith element, for $1 \le i \le n-2$, proceed as follows: If $s - s_{i-1} \ge \ell - \ell_{i-1}$, compare with p first, otherwise compare with q first.*

The number of additional comparisons of Clairvoyant is denoted by A_n^{cv}, its partitioning cost P_n^{cv}.

Note that the strategy "Clairvoyant" cannot be implemented algorithmically, since s and ℓ are not known until the classification is completed.

As shown in [AD15, Section 6], this strategy offers the smallest expected classification cost among all strategies that have oracle access to s and ℓ at the outset of a classification round. As such, its expected cost is a lower bound for the cost of all algorithmic classification procedures; hence we call it an *optimal strategy*.

The non-algorithmic strategy "Clairvoyant" can be turned into an algorithmic classification strategy, which is described next. It will turn out that its cost is only marginally larger than that of strategy "Clairvoyant".

Strategy "Count". *When classifying the ith element, for $1 \le i \le n-2$, proceed as follows: If $s_{i-1} \ge \ell_{i-1}$, compare with p first, otherwise compare with q first.*

The number of additional comparisons of this strategy is called A_n^{ct}, its cost P_n^{ct}.

No algorithmic solution for the classification problem can have cost smaller than "Clairvoyant". Strategy "Count" is algorithmic. Thus any cost-minimal algorithmic classification procedure has cost between $\mathbb{E}(P_n^{\mathrm{cv}})$ and $\mathbb{E}(P_n^{\mathrm{ct}})$, and a precise analysis of both will lead to good lower and upper bounds for the cost of such a procedure. It was shown in [AD15] that $\mathbb{E}(P_n^{\mathrm{ct}}) - \mathbb{E}(P_n^{\mathrm{cv}}) = O(\log n)$ and that, as a consequence, both strategies lead to dual-pivot quicksort algorithms that use $\frac{9}{5} n \log n + O(n)$ comparisons on average. In the following, we carry out a precise analysis of $\mathbb{E}(P_n^{\mathrm{cv}})$ and $\mathbb{E}(P_n^{\mathrm{ct}})$, which will make it possible to determine the expected comparison count of an optimal dual-pivot quicksort algorithm up to $0.28n$.

Lemma 9.1. *(a) The expected number of additional comparisons of strategy "Clairvoyant" is*

$$\mathbb{E}(A_n^{\mathrm{cv}}) = \frac{n}{6} - \frac{7}{12} + \frac{1}{4(n - [n \text{ even}])} - \mathbb{E}\big(X_n^{\searrow}\big)\,.$$

(b) The expected number of additional comparisons of strategy "Count" is

$$\mathbb{E}\big(A_n^{\mathrm{ct}}\big) = \frac{n}{6} - \frac{7}{12} + \frac{1}{4(n - [n \text{ even}])} + \mathbb{E}\big(X_n^{\nearrow}\big)\,.$$

Proof ideas: (The full proof can be found in Appendix J. A different proof of a related statement was given in [AD15].)

(a) Noticing that medium elements can be ignored, we consider a reduced input of size $n' = s + \ell$, consisting only of the s small and the ℓ large elements in the input. For $0 \le i \le n'$ let $s_i' = s - s_i$ and $\ell_i' = \ell - \ell_i$ denote the number of small respectively large elements left unclassified after step i. Then $\{(i, s_i' - \ell_i') \mid 0 \le i \le n'\}$ is a lattice path with distribution (including the distribution of n') exactly as in Section 7, so that the results on the expected number of zeros on such paths given there may be applied. We also note that the sign of $s_{i-1}' - \ell_{i-1}'$ decides whether the ith element to be classified is compared with p first or with q first, and that additional comparisons correspond to steps on the path that lead away from the horizontal axis, excepting down-from-zero steps (due to the asymmetry in treating the situation $s - s_i = \ell - \ell_i$ in strategy "Clairvoyant"). For the number of steps away from the horizontal axis one easily finds the expression $\min(s, \ell)$. Averaging over all choices for n' and the two pivots leads to the formula claimed in (a).

(b) Now assume strategy "Count" is applied to $n' = s + \ell$ elements. The set $\{(i, s_i - \ell_i) \mid 0 \le i \le n'\}$ forms a lattice path that starts at $(0, 0)$ and ends at $(n', s - \ell)$. It can be shown that reflection with respect to the vertical line at $n'/2$ maps these paths in a probability-preserving way to the paths from from (a) (and thus from our model), and it turns out that additional comparisons in this strategy correspond to steps away from the horizontal axis and up-to-zero steps. As in (a), averaging leads to the formula claimed in (b). □

Lemma 9.1 allows us to give an exact expression for the average number of comparisons of "Clairvoyant" and "Count" in a single partitioning step. The expressions for $\mathbb{E}(P_n^{\mathrm{cv}})$ and $\mathbb{E}(P_n^{\mathrm{cv}})$ are obtained by adding the expected number of necessary comparisons $\frac{4}{3}(n-2) + 1$ to the cost terms in Lemma 9.1 (see Appendix J).

10 Main Results and their Asymptotic Aspects

In this section we give precise formulations of our main results. We use the partitioning cost from the previous section to calculate the expected number of comparisons of the two dual-pivot quicksort variants obtained by using classification strategies "Clairvoyant" and "Count", respectively. We call these sorting algorithms "Clairvoyant" and "Count" again. Recall that "Clairvoyant" uses an oracle and is comparison-optimal, and that "Count" is its algorithmic version. We validated our main results in experiments which can be found in Appendix L. They show that the error term $O(n^{-4})$ is small already for real-life input sizes n, and that the linear term has a big influence even for larger n.

Theorem 10.1. *For $n \geq 4$, the average number of comparisons in the comparison-optimal dual-pivot quicksort algorithm "Clairvoyant" (with oracle) is*

$$\mathbb{E}(C_n^{\mathrm{cv}}) = \frac{9}{5}nH_n + \frac{1}{5}nH_n^{\mathrm{alt}} - \frac{89}{25}n + \frac{77}{40}H_n + \frac{3}{40}H_n^{\mathrm{alt}} + \frac{67}{800} - \frac{(-1)^n}{10} + r_n$$

where

$$r_n = \frac{[n \text{ even}]}{320}\Big(\frac{1}{n-3} + \frac{3}{n-1}\Big) - \frac{[n \text{ odd}]}{320}\Big(\frac{3}{n-2} + \frac{1}{n}\Big).$$

Corollary 10.2. *The average number of comparisons in the algorithm "Clairvoyant" is*

$$\mathbb{E}(C_n^{\mathrm{cv}}) = \frac{9}{5}n\log n + An + B\log n + C + \frac{D}{n} + \frac{E}{n^2} + \frac{F[n \text{ even}] + G}{n^3} + O\Big(\frac{1}{n^4}\Big)$$

with

$$A = \frac{9}{5}\gamma - \frac{1}{5}\log 2 - \frac{89}{25} = -2.6596412392892\ldots, \qquad B = \frac{77}{40} = 1.925,$$
$$C = \frac{77}{40}\gamma - \frac{3}{40}\log 2 + \frac{787}{800} = 2.042904116393455\ldots, \qquad D = \frac{13}{16} = 0.8125,$$
$$E = -\frac{77}{480} = -0.1604166\ldots, \qquad F = \frac{1}{8} = 0.125, \qquad G = -\frac{19}{400} = -0.0475,$$

asymptotically as n tends to infinity.

Before continuing with the second partitioning strategy, let us make a remark on the (non-)influence of the parity of n. It is noteworthy that in Corollary 10.2 no such influence is visible in the first six terms (down to $1/n^2$); only from $1/n^3$ on the parity of n appears. This is somewhat unexpected, since a term $(-1)^n$ appears in Theorem 10.1.

Theorem 10.3. *The average number of comparisons in the dual-pivot quicksort algorithm "Count" is*

$$\mathbb{E}\big(C_n^{\mathrm{ct}}\big) = \frac{9}{5}nH_n - \frac{1}{5}nH_n^{\mathrm{alt}} - \frac{89}{25}n + \frac{67}{40}H_n - \frac{3}{40}H_n^{\mathrm{alt}} - \frac{83}{800} + \frac{(-1)^n}{10} - r_n$$

where r_n is defined in Theorem 10.1.

Again, the asymptotic behavior follows from the exact result.

Corollary 10.4. *The average number of comparisons in the algorithm "Count" is*

$$\mathbb{E}\big(C_n^{\mathrm{ct}}\big) = \frac{9}{5} n \log n + An + B \log n + C + \frac{D}{n} + \frac{E}{n^2} + \frac{F[n \text{ even}] + G}{n^3} + O\Big(\frac{1}{n^4}\Big)$$

with

$$A = \frac{9}{5}\gamma + \frac{1}{5}\log 2 - \frac{89}{25} = -2.3823823670652\ldots, \qquad B = \frac{67}{40} = 1.675,$$
$$C = \frac{67}{40}\gamma + \frac{3}{40}\log 2 + \frac{637}{800} = 1.81507227725206\ldots, \qquad D = \frac{11}{16} = 0.6875,$$
$$E = -\frac{67}{480} = -0.1395833\ldots, \qquad F = -\frac{1}{8} = -0.125, \qquad G = \frac{31}{400} = 0.0775,$$

asymptotically as n tends to infinity.

The idea of the proofs of Theorems 10.1 and 10.3 is to translate the recurrence (8.1) into a second order differential equation for the generating function $C(z)$ of $\mathbb{E}(C_n)$ in terms of the generating function $P(z)$ of $\mathbb{E}(P_n)$. Integrating twice yields $C(z)$. This generating function then allows extraction of the exact expressions for $\mathbb{E}(C_n)$. The asymptotic results follow. See Appendix K for details.

Appendix

The appendices can be found at arXiv:1602.04031v1.

References

[AD15] Martin Aumüller and Martin Dietzfelbinger. Optimal partitioning for dual-pivot quicksort. *ACM Trans. Algorithms*, 12(2):18:1–18:36, November 2015.

[Dev16] The SageMath Developers. *SageMath Mathematics Software (Version 7.0)*, 2016. `http://www.sagemath.org`.

[GKP94] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics. A foundation for computer science.* Addison-Wesley, second edition, 1994.

[Hen91] P. Hennequin. *Analyse en moyenne d'algorithmes : tri rapide et arbres de recherche.* PhD thesis, Ecole Politechnique, Palaiseau, 1991.

[HK15] Benjamin Hackl and Daniel Krenn. Asymptotic expansions in SageMath. `http://trac.sagemath.org/17601`, 2015. module in SageMath 6.10.beta2.

[Hoa62] Charles A. R. Hoare. Quicksort. *Comput. J.*, 5(1):10–15, 1962.

[Kra15] Christian Krattenthaler. Lattice path enumeration. In *Handbook of enumerative combinatorics*, Discrete Math. Appl. (Boca Raton), pages 589–678. CRC Press, Boca Raton, FL, 2015.

[Mah08] Hosam Mahmoud. *Polya Urn Models.* Chapman & Hall/CRC, 1 edition, 2008.

[Moh79] Sri Gopal Mohanty. *Lattice path counting and applications.* Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London-Toronto, Ont., 1979. Probability and Mathematical Statistics.

[Sed75] Robert Sedgewick. *Quicksort.* PhD thesis, Standford University, 1975.

[WNN15] Sebastian Wild, Markus E. Nebel, and Ralph Neininger. Average case and distributional analysis of dual-pivot quicksort. *ACM Transactions on Algorithms*, 11(3):22, 2015.

[Yar09] Vladimir Yaroslavskiy. Replacement of quicksort in java.util.arrays with new dual-pivot quicksort. `http://permalink.gmane.org/gmane.comp.java.openjdk.core-libs.devel/2628`, 2009. Archived version of the discussion in the OpenJDK mailing list.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

On Leader Green Election

Jacek Cichoń[†], Rafal Kapelko, Dominik Markiewicz

Department of Computer Science
Faculty of Fundamental Problems of Technology
Wrocław University of Technology
Poland

We investigate the number of survivors in the Leader Green Election (LGE) algorithm introduced by P. Jacquet, D. Milioris and P. Mühlethaler in 2013. Our method is based on the Rice method and gives quite precise formulas. We derive upper bounds on the number of survivors in this algorithm and we propose a proper use of LGE.

Finally, we discuss one property of a general urns and balls problem and show a lower bound for a required number of rounds for a large class of distributed leader election protocols.

Keywords: leader election, distributed algorithms, geometric distribution, Rice method, urns and balls model

1 Introduction

In Jacquet et al. (2013) Philippe Jacquet, Dimitris Milioris and Paul Mühlethaler introduced a novel energy efficient broadcast leader election algorithm, which they called, in accordance with the popular fashion in those years, a Leader Green Election (LGE). This algorithm was also presented by P. Jacquet at the conference AofA'13.

We will use the same model as in Jacquet et al. (2013), namely we assume that the communication medium is of the broadcast type and is prone to collisions. We also assume that the time is slotted. Each slot can be empty (the slot does not contain any burst), collision (the slot contains at least two burst) or successful (the slot contains a single burst).

During the investigation of efficiency of LGE algorithm we found a connection of the leader election problem with some properties of the general "'urns and balls"' model. This connection is discussed in Section 3.

1.1 Short Description of LGE

We will give a short description of a slightly simplified version of the LGE algorithm (for example, authors of Jacquet et al. (2013) consider an arbitrary base of numeral systems, but we restrict our considerations only to base 3, since some additional arguments, not presented in this paper, show that base-3 is an optimal choice for our purposes).

[†]This paper was supported by Polish National Science Center (NCN) grant number 2013/09/B/ST6/02258

We assume that the broadcast medium has N connected users (assume $N \approx 10^6$) and that the number of contenders n is always smaller or equal to N. We fix a number $p \in (0,1)$ and we assume that p is not close to one (e.g. $p = 0.01$). We also fix a number $L = \mathrm{O}(\log\log N)$.

Each contender ω selects independently a random number g_ω according to the geometric distribution with parameter p (see next section for details). If $g_\omega \geq 3^{L+1}$ then we put $g_\omega = 0$. The number g_ω is written

$$g_\omega = \sum_{k=0}^{L} b_k \cdot 3^k \ , \tag{1}$$

where $b_k \in \{0,1,2\}$. We fix a function $f : \{0,1,2\} \to \{0,1\}^2$ by $f(0) = 00$, $f(1) = 01$ and $f(2) = 10$, and define the transmission key K_ω for a contender ω as the concatenation

$$K_\omega = f(b_L)||f(b_{L-1})||\ldots||f(b_1)||f(b_0) \ .$$

Notice that lenght(K_ω) = $\mathrm{O}(\log\log N)$. This key K_ω is used in the following algorithm played in discrete rounds:

```
candidate = true
for i=1 to lenght(K_ω) do
    if K_ω(i) = 1 then
        send a beep
    else
        listen
        if you hear a beep then
            candidate = false
            exit loop
        end if
    end if
end for
```

The survivors of this algorithm are those contenders which at the end have the variable "candidate" set to true. In Jacquet et al. (2013) authors propose to repeat this algorithm several times in order to reduce the number of survivors to 1. However we propose in this paper an another approach: we propose to use this algorithm only once (in order to reduce number of survivors to a small number) and then to use other leader election algorithm for final selection a leader.

1.2 Mathematical Background

The core of LGE algorithm is based on properties of extremal statistics of random variables with geometric distributions. Let us recall that a random variable X has a geometric distribution with parameter $p \in [0,1]$ ($X \sim \mathrm{Geo}(p)$) if $P[X = k] = (1-p)^{k-1}p$ for $k \geq 1$. In the first part of LGE, each user chooses independently a random variable with geometric distribution with a fixed parameter p. The winners of this part of LGE are those users who select a maximal number.

Definition 1 *A random variable M has distribution $\mathrm{MGeo}(n,p)$ if there are independent random variables $X_1, \ldots, X_n$ with distribution $\mathrm{Geo}(p)$ such that*

$$M = \max\{X_1, \ldots, X_n\} \ .$$

It is well known (see e.g. Szpankowski and Rego (1990), Cichoń and Klonowski (2013)) that if $M \sim \mathrm{MGeo}(n,p)$ then $\mathbf{E}[M] = \frac{1}{2} + \frac{\mathrm{H_n}}{\ln \frac{1}{1-p}} + P(n) + \mathrm{O}\left(\frac{1}{n}\right)$, where $P(n)$ is a periodic function with small amplitude and $\mathrm{H_n}$ is the n^{th} harmonic number. Let us recall that $\mathrm{H_n} = \ln n + \gamma + \mathrm{O}\left(\frac{1}{n}\right)$, where $\gamma = 0.557\ldots$ is the Euler constant.

The distribution $\mathrm{MGeo}(n,p)$ controls the number of time slots used in LGE algorithm. More precisely, the LGE algorithm requires some upper approximation on the variable with the $\mathrm{MGeo}(n,p)$ distribution. The next Lemma gives some upper bound for it.

Lemma 1 *Let $M \sim \mathrm{MGeo}(n,p)$, $C > 0$ and $Q = \frac{1}{1-p}$. Then*

$$\Pr[M > C\frac{\ln n}{\ln Q}] \leq \frac{1}{n^{C-1}} .$$

Proof: Let $q = 1-p$. Let us recall that if $X \sim \mathrm{Geo}(p)$ and k is an integer then $\Pr[X > k] = q^k$. Therefore $\Pr[M > k] \leq nq^k$, hence $\Pr\left[M > C\frac{\ln n}{\ln Q}\right] \leq nq^{C\frac{\ln n}{\ln Q}} = \frac{1}{n^{C-1}}$. □

We introduce the next distribution which models the number of survivors in LGE algorithm.

Definition 2 *A random variable W has distribution $\mathrm{WMGeo}(n,p)$ if there are independent random variables $X_1, \ldots, X_n$ with distribution $\mathrm{Geo}(p)$ such that*

$$W = \mathrm{card}\left(\{k : X_k = \max\{X_1, \ldots, X_n\}\}\right) .$$

2 Probabilistic Propeties of LGE

The formal analysis of LGE algorithm in Jacquet et al. (2013) is based on the Mellin transform. In this section, we use an approach based on Rice's method (see e.g. Knuth (1998) and Flajolet and Sedgewick (1995)). We shall derive formulas for expected number of survivors and probabilities for the number of survivors. By $W_{n,p}$ we denote a random variable with $\mathrm{WMGeo}(n,p)$ distribution.

Theorem 1 *Let $n \geq 2$, $p \in (0,1)$ and $q = 1-p$. Let $W_{n,p} \sim \mathrm{WMGeo}(n,p)$ and $a \geq 1$. Then*

$$\Pr[W_{n,p} = a] = \binom{n}{a} p^a \sum_{b=0}^{n-a} \binom{n-a}{b} \frac{(-1)^b}{1-q^{a+b}}$$

and

$$\mathbf{E}[W_{n,p}] = \frac{np}{q} \sum_{b=0}^{n-1} \binom{n-1}{b} \frac{(-1)^b}{1-q^{b+1}} .$$

Proof: Let us fix $n \geq 2$, $p \in (0,1)$ and $q = 1-p$. Let $X_1, \ldots, X_n$ be independent random variables with distribution $\mathrm{Geo}(p)$ and let

$$A_{n;k,a} = (\max\{X_1, \ldots, X_n\} = k) \wedge (\mathrm{card}\left(\{i : X_i = k\} = a\right)) .$$

Then $[W_{n,p}=a]=\bigcup_{k\geq 1}A_{n;k,a}$ and $\Pr[A_{n;k,a}]=\binom{n}{a}(q^{k-1}p)^a(1-q^{k-1})^{n-a}$. Therefore,

$$\Pr[W_n=a]=\sum_{k\geq 1}\binom{n}{a}(q^{k-1}p)^a(1-q^{k-1})^{n-a}=$$
$$\binom{n}{a}p^a\sum_{k\geq 0}q^{ka}(1-q^k)^{n-a}=\binom{n}{a}p^a\sum_{k\geq 0}\sum_{b=0}^{n-a}\binom{n-a}{b}(-1)^b q^{kb}q^{ka}=$$
$$\binom{n}{a}p^a\sum_{b=0}^{n-a}\binom{n-a}{b}(-1)^b\sum_{k\geq 0}q^{k(b+a)}=\binom{n}{a}p^a\sum_{b=0}^{n-a}\binom{n-a}{b}\frac{(-1)^b}{1-q^{a+b}}\,,$$

so the first part of the Theorem is proved. Next we have

$$\sum_{a=1}^{n}a\Pr[A_{n;k,a}]=n\sum_{a=1}^{n}\binom{n-1}{a-1}(q^{k-1}p)^a(1-q^{k-1})^{n-a}=$$
$$nq^{k-1}p\sum_{a=1}^{n}\binom{n-1}{a-1}(q^{k-1}p)^{a-1}(1-q^{k-1})^{(n-1)-(a-1)}=$$
$$nq^{k-1}p\sum_{b=0}^{n-1}\binom{n-1}{b}(q^{k-1}p)^b(1-q^{k-1})^{(n-1)-b}=$$
$$nq^{k-1}p(q^{k-1}p+1-q^{k-1})^{n-1}=nq^{k-1}p(1-q^k)^{n-1}\,.$$

Therefore, for fixed n, we have

$$\sum_{k\geq 1}\sum_{a=1}^{n}a\Pr[A_{k,a}]=\sum_{k\geq 1}npq^{k-1}\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b q^{kb}=$$
$$np\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b\sum_{k\geq 1}q^{k-1}q^{kb}=\frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b\sum_{k\geq 1}q^kq^{kb}=$$
$$\frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b\sum_{k\geq 1}q^{k(b+1)}=\frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b q^{b+1}\sum_{k\geq 0}(q^{b+1})^k=$$
$$\frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}\frac{(-1)^b q^{b+1}}{1-q^{b+1}}\,.$$

Since we assumed that $n \geq 2$, we have

$$\frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}\frac{(-1)^b q^{b+1}}{1-q^{b+1}} = \frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b\frac{(q^{b+1}-1)+1}{1-q^{b+1}} =$$

$$-\frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b + \frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}(-1)^b\frac{1}{1-q^{b+1}} =$$

$$-\frac{np}{q}(-1+1)^{n-1} + \frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}\frac{(-1)^b}{1-q^{b+1}} = \frac{np}{q}\sum_{b=0}^{n-1}\binom{n-1}{b}\frac{(-1)^b}{1-q^{b+1}}\,.$$

□

From Theorem 1 we obtain the following equality $\Pr[W_{n,p}=1] = np\sum_{b=0}^{n-1}\binom{n-1}{b}\frac{(-1)^b}{1-q^{1+b}}$. Therefore, we have the following nice equality

$$\mathbf{E}\left[W_{n,p}\right] = \frac{1}{1-p}\Pr[W_{n,p}=1]\,.$$

Remark Quite recently we learned that Theorem 1 and part of results from the next subsection has been proved in Kirschenhofer and Prodinger (1996). Due to the completeness of arguments we decided to leave the proof in this paper. Our new contribution in this section is the Theorem 3.

2.1 Approximations

Let us fix the number $p \in (0,1)$ and let $q = 1-p$. Let $f_a(z) = \frac{1}{1-q^{a+z}}$. We shall consider complex variable functions f_a for such indexes a which are integers such that $a \geq 1$. Notice that the function f_a has singularities at points from the set $\{\zeta_{a,k} : k \in \mathbb{Z}\}$, where $\zeta_{a,k} = -a + \frac{2k\pi\mathbf{i}}{\ln(q)}$. The function f_a is periodic with period $2\pi\mathbf{i}/\ln(q)$, has single poles at points $\zeta_{a,k}$ and

$$\mathrm{Res}(f_a(z) : z = \zeta_{a,k}) = \frac{-1}{\ln q}\,.$$

It is easy to check that $\lim_{x\to\infty}|f_a(x+\mathbf{i}y)| = 1$ and $\lim_{x\to-\infty}|f_a(x+\mathbf{i}y)| = 0$ for each fixed $y \in \mathbb{R}$.

Let $K_n(s) = \frac{n!}{s(s-1)\cdots(s-n)}$. Notice that if $n \geq 1$ then $|K_n(s)| = \mathrm{O}\left(\frac{1}{|s|^2}\right)$ as $|s|$ grows to infinity. Also notice that if $a > 0$ is an integer, then $K_n(-a) = (-1)^{n+1}\frac{1}{a}\binom{n}{a}^{-1}$. Notice also that the sets of singularity points of functions f_a and K_n are disjoint. This fact greatly simplifies the analysis of the singular points of the product of these functions

Lemma 2 *If $m \geq 1$, $a \geq 1$ and $q \in (0,1)$ then*

$$\sum_{b=0}^{m}\binom{m}{b}\frac{(-1)^b}{1-q^{a+b}} = (-1)^m\frac{1}{\ln q}\sum_{k\in\mathbb{Z}}K_m(\zeta_{a,k})\,.$$

Proof: Rice's integrals summation method (see Knuth (1998)) is based on the formula

$$\sum_{b=0}^{m}\binom{m}{b}(-1)^b g(b) = \frac{(-1)^m}{2\pi i}\oint_{\mathcal{C}} g(s)K_m(s)ds\,,$$

where g is analytic in a domain containing $[0,+\infty)$ and $\mathcal{C}$ is a positively oriented closed curve that lies in the domain of analyticity of g and encircles the real interval $[0,m]$.

We use Rice' formula for functions f_a. Notice that

$$\frac{1}{2\pi i}\oint_{\mathcal{C}} f_a(s)K_m(s)ds = \sum_{k=0}^{m} \mathrm{Res}(f_a(z)K_m(z) : z=k) \; .$$

Let C_k be the positively oriented square with corners at points $\pm\eta_{q,k} \pm \eta_{q,k}\mathbf{i}$, where $\eta_{q,k} = (2k+1)\pi/\ln q$. We consider such k that $|\eta_{q,k}| > m$. For such k the interval $[0,m]$ lies inside the square C_k. The mentioned before Lemma 2 properties of the function f_a (periodicity and boundedness on horizontal lines not crossing singular points) and the kernel function K_m imply that

$$\lim_{k\to\infty}\oint_{C_k} f_a(s)K_m(s)ds = 0 \; ,$$

from which we deduce that

$$\sum_{k\in\mathbb{Z}} \mathrm{Res}(f_a(z)K_m(z) : z=\zeta_{a,k}) + \sum_{k=0}^{m} \mathrm{Res}(f_a(z)K_m(z) : z=k) = 0 \; .$$

Therefore,

$$\sum_{b=0}^{m} \binom{m}{b} \frac{(-1)^b}{1-q^{a+b}} = (-1)^{m+1} \sum_{k\in\mathbb{Z}} \mathrm{Res}(f_a(z)K_m(z) : z=\zeta_{a,k}) =$$

$$(-1)^{m+1} \sum_{k\in\mathbb{Z}} \mathrm{Res}(f_a(z) : z=\zeta_{a,k})K_m(\zeta_{a,k}) = (-1)^{m+1} \sum_{k\in\mathbb{Z}} \frac{-1}{\ln q} K_m(\zeta_{a,k}).$$

□

Lemma 3 *Suppose that $a>0$ is an integer and that $b\in\mathbb{C}$. Then*

$$K_m(-a+b) = \frac{(-1)^{m+1}}{a\binom{a+m}{a}} \cdot \frac{1}{\prod_{j=a}^{m+a}(1-\frac{b}{j})} \; .$$

Proof: Directly from the definition of the kernel function K_m we have

$$K_m(-a+b) = m! \prod_{j=0}^{m} \frac{1}{-a+b-j} = (-1)^{m+1} m! \prod_{j=0}^{m} \frac{1}{a+j-b} =$$

$$(-1)^{m+1} m! \prod_{j=a}^{m+a} \frac{1}{j-b} = (-1)^{m+1} m! \prod_{j=a}^{m+a} \frac{1}{j(1-\frac{b}{j})} =$$

$$(-1)^{m+1} m! \frac{(a-1)!}{(m+a)!} \prod_{j=a}^{m+a} \frac{1}{(1-\frac{b}{j})} \; .$$

□

The next Lemma follows directly from Theorem 1, Lemmas 2 and 3:

Lemma 4 *If $n > a$ then*

$$\Pr[W_{n,p} = a] = \frac{p^a}{a \ln \frac{1}{q}} \left(1 + \sum_{k \in \mathbb{Z} \setminus \{0\}} \frac{1}{\prod_{j=a}^{n} (1 - \frac{2k\pi \mathbf{i} / \ln q}{j})} \right) ,$$

where $q = 1 - p$.

Theorem 2 *If $0 < a < n$ then $\Pr[W_{n,p} = a] = \frac{p^a}{a \ln(Q)} + r_n$, where $|r_n| < \frac{(a+1)^2}{12a} p^a \ln(Q)$, where $Q = \frac{1}{1-p}$.*

Proof: Let $\eta_k = \frac{2\pi k \mathbf{i}}{\ln q}$, where $q = 1 - p$. Notice that

$$\left| \prod_{j=a}^{n} (1 - \frac{\eta_k}{j}) \right|^2 = \prod_{j=a}^{n} \left(1 + \frac{|\eta_k|^2}{j^2} \right) \geq \prod_{j=a}^{a+1} \left(1 + \frac{|\eta_k|^2}{j^2} \right) \geq \left(1 + \frac{|\eta_k|^2}{(a+1)^2} \right)^2 .$$

Therefore,

$$\left| \sum_{k \in \mathbb{Z} \setminus \{0\}} \prod_{j=a}^{n} \frac{1}{1 - \frac{\eta_k}{j}} \right| \leq 2 \sum_{k=1}^{\infty} \frac{1}{1 + \frac{|\eta_k|^2}{(a+1)^2}} \leq 2(a+1)^2 \sum_{k=1}^{\infty} \frac{1}{|\eta_k|^2} =$$
$$\frac{(a+1)^2 (\ln q)^2}{2\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{(a+1)^2 (\ln q)^2}{12} ,$$

so the conclusion follows from Lemma 4. □

Let us fix $p \in (0, 1)$, let $Q = \frac{1}{1-p}$. We put

$$\phi_p(a) = \frac{p^a}{a \ln Q} + \frac{(a+1)^2}{12a} p^a \ln Q.$$

Notice that $\Pr[W_{n,p} = a] \leq \phi_p(a)$.

Theorem 3 $\Pr[W_n \geq k] < \frac{\phi(k)}{1-2p}$

Proof: It can be observed that $\frac{\phi_p(a+1)}{\phi_p(a)} < 2p$. Therefore,

$$\Pr[W_n \geq k] = \sum_{a=k}^{n} \Pr[W_n = a] < \sum_{a=k}^{\infty} \phi(a) < \frac{\phi(k)}{1 - 2p} .$$

□

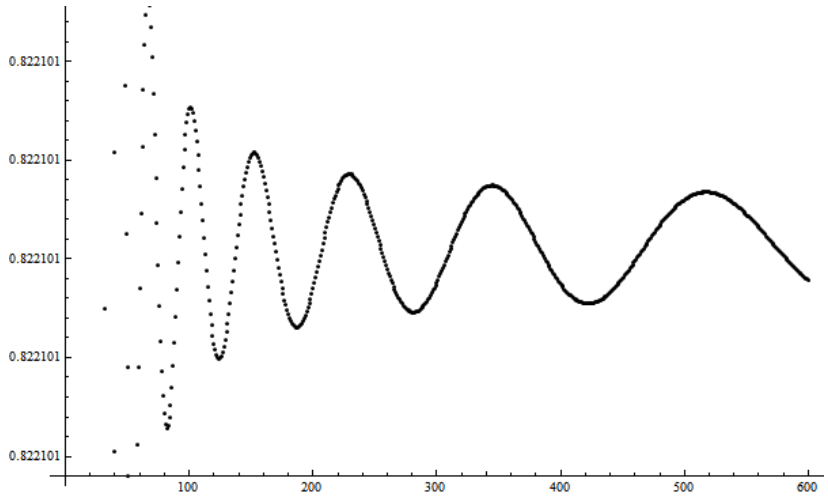


Fig. 1: Plot of $\Pr[W_{n,\frac{1}{3}} = 1]$ for $n = 1, \ldots, 600$.

2.2 Discussion

Let us observe that formulas from Theorem 2 do not depend on the number n. However, small fluctuations (which are very interesting from theoretical point of view) are hidden inside the error term, which can be observed on the Fig. 2.2.

This practical independence of the number n of nodes on the number of survivors is very interesting. However, the number n has an influence on the required number of rounds in LGE. This number may be controlled by Lemma 1: from this lemma we deduce that if $X \sim \mathrm{MGeo}(n,p)$ then $\Pr[X > (\ln 10^{20} + \ln n)/\ln(Q)] < 10^{-20}$ (where $Q = 1/(1-p)$), and hence from a practical point of view it is negligible. This implies that (see Jacquet et al. (2013) for details) the LGE algorithm should run $2 \cdot \left\lceil \log_3 \left(\frac{1}{\ln(Q)} (\ln n + \ln(10^{20}) \right) \right\rceil$ rounds in order to ensure that its probabilistic properties are controlled by the distribution WGeo with probability at least $1 - 10^{-20}$.

From Theorem 2 we deduce that $\Pr[W_{n,p} = 1] = 1 - \frac{p}{2} + \mathrm{O}\left(p^2\right)$ and $\Pr[W_{n,p} = 2] = \frac{p}{2} + \mathrm{O}\left(p^2\right)$. From these formulas we deduce that the probability of failure of one phase of LGE is quite large. However, notice that from Theorem 3 we get $\Pr[W_{n,0.01} > 10] \approx 1.006 \cdot 10^{-19}$. Therefore, the LGE algorithm may be used for quick reduction of potential leaders to a small subgroup. We see that if we use this algorithm with parameter $p = \frac{1}{100}$, then with probability at least $1 - 10^{-19}$, the number of survivors will be less or equal 10. The survivors may then take part in another algorithm (e.g. in an algorithm based on paper Prodinger (1993) or in algorithm based on paper Janson and Szpankowski (1997), Louchard and Prodinger (2009)), which deals better with small sets of nodes, in order to select a leader with high and controllable probability.

3 Lower Bound

In the previous section we recalled that the LGE algorithm should use $\mathrm{O}\left(\ln\ln(n)\right)$ rounds in order to achieve high effectiveness. In this section we prove a general result confirming that this bound is near to an optimal. We use a method applied by D. E. Willard in Willard (1986) for an analysis of resolution protocols in a multiple access channel.

Let us consider a system $(U_i)_{i=1,\ldots,L}$ of L urns and let us fix a number n. We consider a process of throwing an arbitrary number $Q \in \{2,\ldots,n\}$ of balls into these urns. We assume that all balls are thrown independently and that the probability that the ball is thrown into ith urn is equal p_i. This process is fully described by the vector $\vec{p}$ of probabilities from the simplex $\Sigma_L = \{(p_1,\ldots,p_L) \in [0,1]^L : p_1+\ldots+p_L = 1\}$ and the number Q of balls.

The most broadly studied model of urns and balls is the uniform case, i.e. the case when $\vec{p} = (\frac{1}{L},\ldots,\frac{1}{L})$. However, in several papers (see e.g. Flajolet et al. (1992), Boneh and Hofri (1997)) one can find some results for the general case. In this section we are interested in the existence of at least one singleton, i.e. in the existence of an urn U_i with precisely one ball. The problem of estimation of the number of singletons was quite recently analyzed in Penrose (2009).

Let $S_{\vec{p},Q}$ denote the event "there exists at least one urn with a single ball" and let $S_{\vec{p},Q,i}$ denote the event "there is exactly one ball in ith urn". Then, $\Pr[S_{\vec{p},Q,i}] = Qp_i(1-p_i)^{Q-1}$ and $S_{\vec{p},Q} = \bigcup_{i=1}^{L} S_{\vec{p},Q,i}$, therefore, $\Pr[S_{\vec{p},Q}] \leq Q\sum_{i=1}^{L} p_i(1-p_i)^{Q-1}$.

Let us assume that the number Q of balls is unknown but it is bounded by a number n. We are going to show that if the number n is sufficiently large compared to L, then there is no $\vec{p} \in \Sigma_L$ which will guarantee the existence of singleton with a high probability for arbitrary Q from $\{2,\ldots,n\}$. More precisely, let

$$\mathrm{MSP}\,(L,n) = \max_{\vec{p}\in\Sigma_L} \min_{2\leq Q\leq n} \Pr[S_{\vec{p},Q}]\,.$$

(term MSP is an abbreviation of "Maximal Success Probability").

Theorem 4 *For arbitrary $L \geq 1$ and $n \geq 2$, we have*

$$\mathrm{MSP}\,(L,n) < \frac{L-1}{\mathrm{H_n}-1}\,.$$

Proof: Let us observe that if $\vec{p} \in \Sigma_L$ is such that for some i we have $p_i = 1$ and $Q \geq 2$, then $\Pr[S_{\vec{p},Q}] = 0$, so $\min_{2\leq Q\leq n} \Pr[S_{\vec{p},Q}] = 0$. Hence, we may consider only such $\vec{p} \in \Sigma_L$ that $p_i < 1$ for each $i = 1,\ldots,L$.

Let us fix the number L of urns and let us consider the following function (this is the trick which we borrow from Willard (1986)):

$$f(\vec{p}) = \sum_{Q=2}^{n} \frac{\Pr[S_{\vec{p},Q}]}{Q}\,.$$

Then we have

$$f(\vec{p}) \leq \sum_{Q=2}^{n}\sum_{i=1}^{L} \frac{\Pr[S_{\vec{p},Q,i}]}{Q} = \sum_{Q=2}^{n}\sum_{i=1}^{L} p_i(1-p_i)^{Q-1} \leq \sum_{i=1}^{L}\sum_{Q=2}^{\infty} p_i(1-p_i)^{Q-1} =$$

$$\sum_{i=1}^{L} p_i(1-p_i)\frac{1}{1-(1-p_i)} = \sum_{i=1}^{L}(1-p_i) = L - \sum_{i=1}^{L} p_i = L-1\,.$$

On the other side, let $p^* = \min\{\Pr[S_{\vec{p},Q}] : 2 \leq Q \leq n\}$. Then we have

$$f(\vec{p}) \geq \sum_{Q=2}^{n} \frac{p^*}{Q} = p^* \sum_{Q=2}^{n} \frac{1}{Q} = p^*(\mathrm{H_n} - 1) \, .$$

Therefore, we have

$$p^*(\mathrm{H_n} - 1) \leq f(\vec{p}) < L - 1 \, .$$

Hence, if we take Q^* such that $\Pr[S_{\vec{p},Q^*}] = p*$, then

$$\Pr[S_{\vec{p},Q^*}] < \frac{L-1}{\mathrm{H_n} - 1} \, ,$$

so

$$\min_{2 \leq Q \leq n} \Pr[S_{\vec{p},Q}] < \frac{L-1}{\mathrm{H_n} - 1}$$

for arbitrary $\vec{p} \in \Sigma_n$. □

Corollary 1 *If* $1 \leq L \leq \frac{1}{2} \ln n + \frac{1+\gamma}{2}$ *then* $\mathrm{MSP}\,(L, n) < \frac{1}{2}$.

Corollary 2 *If* $n \geq \exp(2L - (1+\gamma))$ *then* $\mathrm{MSP}\,(L, n) < \frac{1}{2}$.

Proof: Both proofs follow directly from Theorem 4 and the inequality $\mathrm{H_n} \geq \ln(n) + \gamma$. □

3.1 Application to Leader Election Problem

Let us consider any oblivious leader election algorithm in which at the beginning each station selects randomly and independently a sequence of bits of length M, and later this station use the sequence in the algorithm in a deterministic way. Let n denote the upper bound on the number of stations taking part in this algorithm and let b_i denote the sequence of bits chosen by the ith station. Observe that if for each i there is $j \neq i$ such that $b_i = b_j$, then the algorithm must fail. Hence, success is possible only if there is a singleton in choices made from the space $\{0,1\}^M$ of all possible sequences of bits. When we use Corollary 1 with $L = 2^M$, then we deduce that if $M \leq \log_2\left(\frac{1}{2}\ln n + \frac{1+\gamma}{2}\right)$ then the probability that the considered algorithm chooses a leader is less than $\frac{1}{2}$. We may say that $\log_2(\frac{1}{2}\ln n)$ random bits are too few for distinguishing an arbitrary collection of $\leq n$ objects with a high probability.

Acknowledgment

The authors wish to thank the anonymous reviewers for their valuable comments and for drawing attention to that a large part of the results from Section 2 of this paper has already been proven in Kirschenhofer and Prodinger (1996) and that the are new results about the asymmetric leader election algorithm (e.g. Louchard and Prodinger (2009)) not mentioned in a previous version of our paper.

References

A. Boneh and M. Hofri. The coupon-collector problem revisited—a survey of engineering problems and computational methods. *Stochastic Models*, 13(1):39–66, 1997.

J. Cichoń and M. Klonowski. On flooding in the presence of random faults. *Fundam. Inform.*, 123(3): 273–287, 2013.

P. Flajolet and R. Sedgewick. Mellin transforms and asymptotics: Finite differences and Rice's integrals. *Theor. Comput. Sci.*, 144(1,2):101–124, 1995. URL http://dblp.uni-trier.de/db/journals/tcs/tcs144.html#FlajoletS95.

P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, Coupon collectors, Caching algorithms and Self-organizing search. *Discrete Appl. Math.*, 39(3):207–229, Nov. 1992. ISSN 0166-218X. doi: 10.1016/0166-218X(92)90177-C. URL http://dx.doi.org/10.1016/0166-218X(92)90177-C.

P. Jacquet, D. Milioris, and P. Mühlethaler. A novel energy efficient broadcast leader election. In *MASCOTS*, pages 495–504. IEEE, 2013.

S. Janson and W. Szpankowski. Analysis of an asymmetric leader election algorithm. *Electr. J. Comb.*, 4(1), 1997. URL http://www.combinatorics.org/Volume_4/Abstracts/v4i1r17.html.

P. Kirschenhofer and H. Prodinger. The number of winners in a discrete geometrically distributed sample. *Ann. Appl. Probab.*, 6(2):687–694, 05 1996. doi: 10.1214/aoap/1034968150. URL http://dx.doi.org/10.1214/aoap/1034968150.

D. E. Knuth. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998. ISBN 0-201-89685-0.

G. Louchard and H. Prodinger. The asymmetric leader election algorithm: Another approach. *Annals of Combinatorics*, 12:449–478, 2009.

M. Penrose. Normal approximation for isolated balls in an urn allocation model. *Electron. J. Probab.*, 14:no. 74, 2155–2181, 2009. ISSN 1083-6489. doi: 10.1214/EJP.v14-699. URL http://ejp.ejpecp.org/article/view/699.

H. Prodinger. How to select a loser. *Discrete Mathematics*, 120(1-3):149–159, 1993. doi: 10.1016/0012-365X(93)90572-B. URL http://dx.doi.org/10.1016/0012-365X(93)90572-B.

W. Szpankowski and V. Rego. Yet another application of a binomial recurrence. Order statistics. *Computing*, 43(4):401–410, Feb. 1990. ISSN 0010-485X. doi: 10.1007/BF02241658. URL http://dx.doi.org/10.1007/BF02241658.

D. E. Willard. Log-logarithmic selection resolution protocols in a multiple access channel. *SIAM J. Comput.*, 15(2):468–477, May 1986. ISSN 0097-5397. doi: 10.1137/0215032. URL http://dx.doi.org/10.1137/0215032.

Vertex Degrees in Planar Maps

Gwendal Collet, Michael Drmota, Lukas Daniel Klausner

Institute of Discrete Mathematics and Geometry
TU Wien
Vienna
Austria

We prove a general multi-dimensional central limit theorem for the expected number of vertices of a given degree in the family of planar maps whose vertex degrees are restricted to an arbitrary (finite or infinite) set of positive integers D. Our results rely on a classical bijection with mobiles (objects exhibiting a tree structure), combined with refined analytic tools to deal with the systems of equations on infinite variables that arise. We also discuss some possible extension to maps of higher genus.

Keywords: Planar maps, Central limit theorem, Analytic combinatorics, Mobiles

1 Introduction and Results

In this paper we study statistical properties of planar maps, which are connected planar graphs, possibly with loops and multiple edges, together with an embedding into the plane. Such objects are frequently used to describe topological features of geometric arrangements in two or three spatial dimensions. Thus, the knowledge of the structure and of properties of "typical" objects may turn out to be very useful in the analysis of particular algorithms that operate on planar maps. We say that map is *rooted* if an edge e is distinguished and oriented. It is called the root edge. The first vertex v of this oriented edge is called the root-vertex. The face to the right of e is called the root-face and is usually taken as the outer (or infinite) face. Similarly, we call a planar map *pointed* if just a vertex v is distinguished. However, we have to be really careful with the model. In rooted maps the root edge *destroys* potential symmetries, which is not the case if we consider pointed maps.

The enumeration of rooted maps is a classical subject, initiated by Tutte in the 1960's, see [11]. Among many other results, Tutte computed the number M_n of rooted maps with n edges, proving the formula

$$M_n = \frac{2(2n)!}{(n+2)!n!}3^n$$

which directly provides the asymptotic formula

$$M_n \sim \frac{2}{\sqrt{\pi}}n^{-5/2}12^n.$$

[0] *Funded by FWF – SFB F50 Algorithmic and Enumerative Combinatorics*

We are mainly interested in planar maps with degree restrictions. Actually, it turns out that this kind of asymptotic expansion is quite universal. Furthermore, there is always a (very general) central limit theorem for the number of vertices of given degree.

Theorem 1. *Suppose that D is an arbitrary set of positive integers but not a subset of $\{1, 2\}$, let $\mathcal{M}_D$ be the class of planar rooted maps with the property that all vertex degrees are in D and let $M_{D,n}$ denote the number of maps in $\mathcal{M}_D$ with n edges. Furthermore, if D contains only even numbers, then set $d = \gcd\{i : 2i \in D\}$; set $d = 1$ otherwise.*

Then there exist positive constants c_D and ρ_D with

$$M_{D,n} \sim c_D n^{-5/2} \rho_D^{-n}, \qquad n \equiv 0 \bmod d. \tag{1}$$

Furthermore, let $X_n^{(d)}$ denote the random variable counting vertices of degree d ($\in D$) in maps in $\mathcal{M}_D$. Then $\mathbb{E}(X_n^{(d)}) \sim \mu_d n$ for some constant $\mu_d > 0$ and for $n \equiv 0 \bmod d$, and the (possibly infinite) random vector $\mathbf{X}_n = (X_n^{(d)})_{d \in D}$ ($n \equiv 0 \bmod d$) satisfies a central limit theorem, that is,

$$\frac{1}{\sqrt{n}} \left(\mathbf{X}_n - \mathbb{E}(\mathbf{X}_n) \right), \qquad n \equiv 0 \bmod d, \tag{2}$$

converges weakly to a centered Gaussian random variable $\mathbf{Z}$ (in ℓ^2).

Note that maps where all vertex degrees are 1 or 2 are very easy to characterize and are not really of interest, and that actually, their asymptotic properties are different from the general case. It is therefore natural to assume that D is not a subset of $\{1, 2\}$.

Since we can equivalently consider dual maps, this kind of problem is the same as considering planar maps with restrictions on the face valencies. This means that the same results hold if we replace *vertex degree* by *face valency*. For example, if we assume that all face valencies equal 4, then we just consider planar quadrangulations (which have also been studied by Tutte [11]). In fact, our proofs will refer just to face valencies.

Theorem 1 goes far beyond known results. There are some general results for the Eulerian case where all vertex degrees are even. First, the asymptotic expansion (1) is known for Eulerian maps by Bender and Canfield [2]. Furthermore, a central limit theorem of the form (2) is known for all Eulerian maps (without degree restrictions) [9]. However, in the non-Eulerian case there are almost no results of this kind; there is only a one-dimensional central limit theorem for $X_n^{(d)}$ for all planar maps [10].

Section 2 introduces planar mobiles which, being in bijection with pointed planar maps, will reduce our analysis to simpler objects with a tree structure. Their asymptotic behaviour is derived in Section 3, first for the simpler case of bipartite maps (i.e., when D contains only even integers), then for families of maps without constraints on D. Section 4 is devoted to the proof of the central limit theorem using analytic tools from [8, 9]. Finally, in Section 5 we discuss the combinatorics of maps on orientable surface of higher genus. The expressions we obtain are much more involved than in the planar case, but it is expected to lead to similar analytic results.

2 Mobiles

Instead of investigating planar maps themselves, we will follow the principle presented in [5], whereby pointed planar maps are bijectively related to a certain class of trees called mobiles. (Their version of

mobiles differ from the definition originally given in [3]; the equivalence of the two definitions is not shown explicitly in [5], but [7] gives a straightforward proof.)

Definition 1. A *mobile* is a planar tree – that is, a map with a single face – such that there are two kinds of vertices (black and white), edges only occur as black–black edges or black–white edges, and black vertices additionally have so-called "legs" attached to them (which are not considered edges), whose number equals the number of white neighbor vertices.
A *bipartite mobile* is a mobile without black–black edges.
The *degree* of a black vertex is the number of half-edges plus the number of legs that are attached to it.
A mobile is called *rooted* if an edge is distinguished and oriented.

The essential observation is that mobiles are in bijection to pointed planar maps.

Theorem 2. *There is a bijection between mobiles that contain at least one black vertex and pointed planar maps, where white vertices in the mobile correspond to non-pointed vertices in the equivalent planar map, black vertices correspond to faces of the map, and the degrees of the black vertices correspond to the face valencies. This bijection induces a bijection on the edge sets so that the number of edges is the same. (Only the pointed vertex of the map has no counterpart.)*

Similarly, rooted mobiles that contain at least one black vertex are in bijection to rooted and vertex-pointed planar maps.

Finally, bipartite mobiles with at least two vertices correspond to bipartite maps with at least two vertices, in the unrooted as well as in the rooted case.

Proof. For the proof of the bijection between mobiles and pointed maps we refer to [7], where the bipartite case is also discussed. It just remains to note that the induced bijection on the edges can be directly used to transfer the root edge together with its direction. ∎

2.1 Bipartite Mobile Counting

We start with bipartite mobiles since they are more easy to *count*, in particular if we consider rooted bipartite mobiles, see [7].

Proposition 1. *Let* $R = R(t, z, x_1, x_2, \ldots)$ *be the solution of the equation*

$$R = tz + z \sum_{i \geq 1} x_{2i} \binom{2i-1}{i} R^i. \tag{3}$$

Then the generating function $M = M(t, z, x_1, x_2, \ldots)$ *of bipartite rooted maps satisfies*

$$\frac{\partial M}{\partial t} = 2\left(R/z - t\right), \tag{4}$$

where the variable t *corresponds to the number of vertices,* z *to the number of edges, and* x_{2i}*,* $i \geq 1$*, to the number of faces of valency* $2i$*.*

Proof. Since rooted mobiles can be considered as ordered rooted trees (which means that the neighboring vertices of the root vertex are linearly ordered and the subtrees rooted at these neighboring vertices are

again ordered trees) we can describe them recursively. This directly leads to a functional equation for R of the form

$$R = \frac{tz}{1 - z\sum_{i\geq 1} x_{2i}\binom{2i-1}{i}R^{i-1}}$$

which is apparently the same as (3). Note that the factor $\binom{2i-1}{i}$ is precisely the number of ways of grouping i legs and $i-1$ edges around a black vertex (of degree $2i$; one edge is already there).

Hence, the generating function of rooted mobiles that are rooted by a white vertex is given by R/z. Since we have to discount the mobile that consists just of one (white) vertex, the generating function of rooted mobiles that are rooted at a white vertex and contain at least two vertices is given by

$$R/z - t = \sum_{i\geq 1} x_{2i}\binom{2i-1}{i}R^i. \tag{5}$$

We now observe that the right hand side of (5) is precisely the generating function of rooted mobiles that are rooted at a black vertex (and contain at least two vertices). Summing up, the generating function of bipartite rooted mobiles (with at least two vertices) is given by

$$2(R/z - t).$$

Finally, if M denotes the generating function of bipartite rooted maps (with at least two vertices) then $\frac{\partial M}{\partial t}$ corresponds to rooted maps, where a non-root vertex is pointed (and discounted). Thus, by Theorem 2 we obtain (4).

∎

Remark 1. It can be easily checked that Formula (4) can be specialized to count M_D, for any subset D of even positive integers: It suffices to set to 0 every x_{2i} such that $2i \in D$.

2.2 General Mobile Counting

We now proceed to develop a mechanism for general mobile counting that is adapted from [5]. For this, we will require Motzkin paths.

Definition 2. A *Motzkin path* is a path starting at 0 and going rightwards for a number of steps; the steps are either diagonally upwards $(+1)$, straight (0) or diagonally downwards (-1). A *Motzkin bridge* is a Motzkin path from 0 to 0. A *Motzkin excursion* is a Motzkin bridge which stays non-negative.

We define generating functions in the variables t and u, which count the number of steps of type 0 and -1, respectively. (Explicitly counting steps of type 1 is then unnecessary, of course.) The ordinary generating functions of Motzkin bridges, Motzkin excursions, and Motzkin paths from 0 to $+1$ shall be denoted by $B(t,u)$, $E(t,u)$ and $B^{(+1)}(t,u)$, respectively.

Continuing to follow the presentation of [5] and decomposing these three types of paths by their last passage through 0, we arrive at the equations:

$$\begin{aligned} E &= 1 + tE + uE^2, \\ B &= 1 + (t + 2uE)B, \\ B^{(+1)} &= EB. \end{aligned}$$

In what follows we will also make use of bridges where the first step is either of type 0 or -1. Clearly, their generating function $\overline{B}$ is given by

$$\overline{B} = tB + uB^{(+1)} = B(t + uE).$$

When Motzkin bridges are not constrained to stay non-negative, they can be seen as a random arrangement of a given number of steps $+1, 0, -1$. It is then possible to obtain explicit expressions for

$$B_{\ell,m} = [t^\ell u^m]B(t,u) = \binom{l+2m}{l,m,m}, \tag{6}$$

$$B^{(+1)}_{\ell,m} = [t^\ell u^m]B^{(+1)}(t,u) = \binom{l+2m+1}{l,m,m+1}, \tag{7}$$

$$\overline{B}_{\ell,m} = [t^\ell u^m]\overline{B}(t,u) = B_{\ell-1,m} + B^{(+1)}_{\ell,m-1} = \frac{l+m}{l+2m}\binom{l+2m}{l,m,m}. \tag{8}$$

Using the above, we can now finally compute relations for generating functions of proper classes of mobiles. We define the following series, where t corresponds to the number of white vertices, z to the number of edges, and x_i, $i \geq 1$, to the number of black vertices of degree i:

- $L(t, z, x_1, x_2, \ldots)$ is the series counting rooted mobiles that are rooted at a black vertex and where an additional edge is attached to the black vertex.
- $Q(t, z, x_1, x_2, \ldots)$ is the series counting rooted mobiles that are rooted at a univalent white vertex, which is not counted in the series.
- $R(t, z, x_1, x_2, \ldots)$ is the series counting rooted mobiles that are rooted at a white vertes and where an additional edge is attached to the root vertex.

Similarly to the above we obtain the following equations for the generating functions of mobiles and rooted maps.

Proposition 2. *Let $L = L(t, z, x_1, x_2, \ldots)$, $Q = Q(t, z, x_1, x_2, \ldots)$, and $R = R(t, z, x_1, x_2, \ldots)$ be the solutions of the equation*

$$\begin{aligned} L &= z\sum_{\ell,m} x_{2m+\ell+1} B_{\ell,m} L^\ell R^m, \\ Q &= z\sum_{\ell,m} x_{\ell+2m+2} B^{(+1)}_{\ell,m} L^\ell R^m, \\ R &= \frac{tz}{1-Q}, \end{aligned} \tag{9}$$

and let $T = T(t, z, x_1, x_2, \ldots)$ be given by

$$T = 1 + \sum_{\ell,m} x_{2m+\ell} \overline{B}_{\ell,m} L^\ell R^m, \tag{10}$$

where the numbers $B_{\ell,m}$, $B^{(+1)}_{\ell,m}$, and $\overline{B}_{\ell,m}$ are given by (6)–(8). Then the generating function $M = M(t, z, x_1, x_2, \ldots)$ of rooted maps satisfies

$$\frac{\partial M}{\partial t} = R/z - t + T, \tag{11}$$

where the variable t corresponds to the number of vertices, z to the number of edges, and x_i, $i \geq 1$, to the number of faces of valency i.

Proof. The system (9) is just a rephrasement of the recursive structure of rooted mobiles. Note that the numbers $B_{\ell,m}$ and $B^{(+1)}_{\ell,m}$ are used to count the number of ways to circumscribe a specific black vertex and considering white vertices, black vertices and "legs" as steps -1, 0 and $+1$. The generating function T given in (10) is then the generating function of rooted mobiles where the root vertex is black.

Finally, the equation (11) follows from Theorem 2 since $R/z - t$ corresponds to rooted mobiles with at least one black vertex where the root vertex is white and T corresponds to rooted mobiles where the root vertex is black. ∎

Remark 2. Note that Proposition 1 is a special case of Proposition 2. We just have to restrict to the terms corresponding to $\ell = 0$ since bipartite mobiles have no black–black edges. In particular, the series for L is not needed any more and the second and third equations from (9) can be used to easily eliminate Q in order to recover the equation (3).

3 Asymptotic Enumeration

In this section we prove the asymptotic expansion (1). It turns out that it is much easier to start with bipartite maps. Actually, the bipartite case has already been treated by Bender and Canfield [2]. However, we apply a slightly different approach, which will then be extended to cover the general case as well the central limit theorem.

3.1 Bipartite maps

Let D be a non-empty subset of even positive integers different from $\{2\}$. Then by Proposition 1 the counting problem reduces to the discussion of the solutions $R_D = R_D(t, z)$ of the functional equation

$$R_D = tz + z \sum_{2i \in D} \binom{2i-1}{i} R_D^i \tag{12}$$

and the generating function $M_D(t, z)$ that satisfies the relation

$$\frac{\partial M_D}{\partial t} = 2\left(R_D/z - t\right). \tag{13}$$

Let $d = \gcd\{i : 2i \in D\}$. Then for combinatorial reasons it follows that there only exist maps with n edges for n that are divisible by d. This is reflected by the fact that the equation (12) can we rewritten in the form

$$\tilde{R} = t + \sum_{2i \in D} \binom{2i-1}{i} z^{i/d} \tilde{R}^i, \tag{14}$$

where we have substituted $R_D(t, z) = z\tilde{R}(t, z^d)$. (Recall that we finally work with R_D/z.)

Lemma 1. *There exists an analytic function $\rho(t)$ with $\rho(1) > 0$ and $\rho'(1) \neq 0$ that is defined in a neighborhood of $t = 1$, and there exist analytic functions $g(t, z)$, $h(t, z)$ with $h(1, \rho(1)) > 0$ that are defined in a neighborhood of $t = 1$ and $z = \rho(1)$ such that the unique solution $R_D = R_D(t, z)$ of the equation (12) that is analytic at $z = 0$ and $t = 0$ can be represented as*

$$R_D = g(t, z) - h(t, z)\sqrt{1 - \frac{z}{\rho(t)}}. \tag{15}$$

Furthermore, the values $z = \rho(t)e(2\pi ij/d)$, $j \in \{0, 1, \ldots, d-1\}$, are the only singularities of the function $z \mapsto R_D(t, z)$ on the disc $|z| \leq \rho(t)$, and there exists an analytic continuation of R_D to the range $|z| < |\rho(t)| + \eta$, $\arg(z - \rho(t)e(2\pi ij/d)) \neq 0$, $j \in \{0, 1, \ldots, d-1\}$.

Proof. From general theory (see [8, Theorem 2.21]), we know that an equation of the form $R = F(t, z, R)$, where F is a power series with non-negative coefficients, has a square-root singularity if there are positive solutions (ρ, R_0) to the following system:

$$R_0 = F(1, \rho, R_0), \qquad 1 = F_R(1, \rho, R_0).$$

It is important to observe that the solutions are inside the region of convergence of F. Besides, one has to check several analytic conditions on the derivatives of F evaluated at this singular point. For a more detailed proof, the reader can refer to the work of Bender and Canfield [2]. ∎

It is now relatively easy to obtain similar properties for $M_D(t, z)$.

Lemma 2. *The function $M = M_D(t, z)$ that is given by (13) has the representation*

$$M_D = g_2(t, z) + h_2(t, z)\left(1 - \frac{z}{\rho(t)}\right)^{3/2} \tag{16}$$

in a neighborhood of $t = 1$ and $z = \rho(1)$, where the functions $g_2(t, z)$, $h_2(t, z)$ are analytic in a neighborhood of $t = 1$ and $z = \rho(1)$ and we have $h_2(1, \rho(1)) > 0$. Furthermore, the values $z = \rho(t)e(2\pi ij/d)$, $j \in \{0, 1, \ldots, d-1\}$, are the only singularities of the function $z \mapsto M_D(t, z)$ on the disc $|z| \leq \rho(t)$, and there exists an analytic continuation of M_D to the range $|z| < |\rho(t)| + \eta$, $\arg(z - \rho(t)e(2\pi ij/d)) \neq 0$, $j \in \{0, 1, \ldots, d-1\}$.

Proof. This is a direct application of [8, Lemma 2.27]. ∎

In particular it follows that $M_D(1, z)$ has the singular representation

$$M_D = g_2(1, z) + h_2(1, z)\left(1 - \frac{z}{\rho(1)}\right)^{3/2}$$

around $z = \rho(1)$. The singular representations are of the same kind around $z = \rho(1)e(2\pi ij/d)$, $j \in \{1, \ldots, d-1\}$ and we have the analytic continuation property. Hence it follows by usual singularity analysis (see for example [8, Corollary 2.15]) that there exists a constant $c_D > 0$ such that

$$[z^n]M_D(1, z) \sim c_D n^{-5/2}\rho(1)^{-n}, \qquad n \equiv 0 \bmod d,$$

which completes the proof of the asymptotic expansion in the bipartite case.

3.2 General Maps

We now suppose that D contains at least one odd number. It is easy to observe that in this case we have $[z^n]M_D(1,z) > 0$ for $n \geq n_0$ (for some n_0), so we do not have to deals with several singularities.

By Proposition 2 we have to consider the system of equations for $L_D = L_D(t,z)$, $Q_D = Q_D(t,z)$, $R_D = R_D(t,z)$:

$$\begin{aligned} L_D &= z\sum_{i\in D}\sum_m B_{i-2m-1,m}L_D^{i-2m-1}R_D^m,\\ Q_D &= z\sum_{i\in D}\sum_m B^{(+1)}_{i-2m-2,m}L_D^{i-2m-2}R_D^m,\\ R_D &= \frac{tz}{1-Q_D}, \end{aligned} \tag{17}$$

and also the function

$$T_D = T_D(t,z) = 1 + \sum_{i\in D}\sum_m \overline{B}_{i-2m,m}L_D^{i-2m}R_D^m.$$

Lemma 3. *There exists an analytic function $\rho(t)$ with $\rho(1) > 0$ and $\rho'(1) \neq 0$ that is defined in a neighborhood of $t = 1$, and there exist analytic functions $g(t,z)$, $h(t,z)$ with $h(1,\rho(1)) > 0$ that are defined in a neighborhood of $t = 1$ and $z = \rho(1)$ such that*

$$R_D/z - t + T_D = g(t,z) - h(t,z)\sqrt{1-\frac{z}{\rho(t)}}. \tag{18}$$

Furthermore, the value $z = \rho(t)$ is the only singularity of the function $z \mapsto R_D/z - t + T_D$ on the disc $|z| \leq \rho(t)$, and there exists an analytic continuation of R_D to the range $|z| < |\rho(t)|+\eta$, $\arg(z-\rho(t)) \neq 0$.

Proof. Instead of a single equation, we have to deal with the strongly connected system (17), which is known to have similar analytic properties (see [8, Theorem 2.33]). As in Lemma 1, the main observation is that the singular point lies within the region of convergence of the equations, which follows directly in the finite case, but gets more technical in the infinite case. ∎

Lemma 3 shows that we are precisely in the same situation as in the bipartite case (actually, it is slightly easier since there is only one singularity on the circle $|z| = \rho(t)$). Hence we immediately get the same property for M_D as stated in Lemma 2 and consequently the proposed asymptotic expansion (1).

4 Central Limit Theorem for Bipartite Maps

Based on this previous result, we now extend our analysis to obtain a central limit theorem. Actually, this is immediate if the set D is finite, whereas the infinite case needs much more care.

Let D be a non-empty subset of even positive integers different from $\{2\}$. Then by Proposition 1 the generating functions $R_D = R_D(t,z,(x_{2i})_{2i\in D})$ and $M_D = M_D(t,z,(x_{2i})_{2i\in D})$ satisfy the equations

$$R_D = tz + z\sum_{2i\in D} x_{2i}\binom{2i-1}{i}R_D^i \tag{19}$$

and

$$\frac{\partial M_D}{\partial t} = 2\left(R_D/z - t\right). \tag{20}$$

If D is finite, then the number of variables is finite, too, and we can apply [8, Theorem 2.33] to obtain a representation of R_D of the form

$$R_D = g(t, z, (x_{2i})_{2i\in D}) - h(t, z, (x_{2i})_{2i\in D})\sqrt{1 - \frac{z}{\rho(t, (x_{2i})_{2i\in D})}}, \tag{21}$$

a proper extension of the transfer lemma [8, Lemma 2.27] (where the variables x_{2i} are considered as additional parameters) leads to

$$M_D = g_2(t, z, (x_{2i})_{2i\in D}) + h_2(t, z, (x_{2i})_{2i\in D})\left(1 - \frac{z}{\rho(t, (x_{2i})_{2i\in D})}\right)^{3/2}, \tag{22}$$

and finally [8, Theorem 2.25] implies a multivariate central limit theorem for the random vector $\mathbf{X}_n = (X_n^{(2i)})_{2i\in D}$ of the proposed form.

Thus, we just have to concentrate on the infinite case. Actually, we proceed there in a similar way; however, we have to take care of infinitely many variables. There is no real problem to derive the same kind of representation (21) and (22) if D is infinite. Everything works in the same way as in the finite case, we just have to assume that the variables x_i are uniformly bounded. And of course we have to use a proper notion of analyticity in infinitely many variables. We only have to apply the functional analytic extension of the above cited theorems that are given in [9]. Moreover, in order to obtain a proper central limit theorem we need a proper adaption of [9, Theorem 3]. In this theorem we have also a single equation $y = F(z, (x_i)_{i\in I}, y)$ for a generating function $y = y(z, (x_i)_{i\in I})$ that encodes the distribution of a random vector $(X_n^{(i)})_{i\in I}$ in the form

$$y = \sum_n y_n \left(\mathbb{E}\prod_{i\in I} x_i^{X_n^{(i)}}\right) z^n,$$

where $X_n^{(i)} = 0$ for $i > cn$ (for some constant $c > 0$) which also implies that all appearing potentially infinite products are in fact finite. (In our case this is satisfied since there is no vertex of degree larger than n if we have n edges.) As we can see from the proof of [9, Theorem 3], the essential part is to provide tightness of the involved normalized random vector, and tightness can be checked with the help of moment conditions. It is clear that asymptotics of moments for $X_n^{(i)}$ can be calculated with the help of derivatives of F, for example $\mathbb{E}X_n^{(i)} = F_{x_i}/(\rho F_z)\cdot n + O(1)$. This follows from the fact all information on the asymptotic behavior of the moments is *encoded* in the derivatives of the singularity $\rho(z, (x_i)_{i\in I})$ and by implicit differentiation these derivatives relate to derivatives of F. More precisely, [9, Theorem 3] says that the following conditions are sufficient to deduce tightness of the normalized random vector:

$$\sum_{i\in I} F_{x_i} < \infty, \qquad \sum_{i\in I} F_{yx_i}^2 < \infty, \qquad \sum_{i\in I} F_{x_ix_i} < \infty,$$

$$F_{zx_i} = o(1), \qquad F_{zx_ix_i} = o(1), \qquad F_{yyx_i} = o(1), \qquad F_{yyx_ix_i} = o(1),$$

$$F_{zzx_i} = O(1), \qquad F_{zyx_i} = O(1), \qquad F_{zyyx_i} = O(1), \qquad F_{yyyx_i} = O(1),$$

$$(i \to \infty),$$

where all derivatives are evaluated at $(\rho, (1)_{i\in I}, y(\rho))$.

The situation is slightly different in our case since we have to work with M_D instead of R_D. However, the only real difference between R_D and M_D is that the critical exponents in the singular representations (21) and (22) are different, but the behavior of the singularity $\rho(t, z, (x_i)_{i\in I})$ is precisely the same. Note that after the integration step we can set $t = 1$. Now tightness for the normalized random vector that is encoded in the function M_D follows in the same way as for R_D. And since the singularity $\rho(1, z, (x_i)_{i\in I})$ is the same, we get precisely the same conditions as in the case of [9, Theorem 3].

This means that we just have to check the above conditions hold for

$$F = F(1, z, (x_{2i})_{2i\in D}, y) = z + z \sum_{2i\in D} x_{2i} \binom{2i-1}{i} y^i,$$

where all derivatives are evaluated at $z = \rho$, $x_{2i} = 1$, and $y = R_D(\rho) < 1/4$. However, they are trivially satisfied since

$$\sum_{i\geq 1} \binom{2i-1}{i} i^K y^i < \infty$$

for all $K > 0$ and for positive real $y < 1/4$.

Remark 3. As stated in Theorem 1, the results and methods extend to the general case as well. The main idea is to reduce the (positive strongly connected) system of two equations (17) to a single functional equation, by applying [8, Theorem 3].

5 Maps of Higher Genus

The bijection used in Section 2 relies solely on the orientability of the surface on which the maps are embedded. Therefore it can easily be extended to maps of higher genus, i.e., embedded on an orientable surface of genus $g \in \mathbb{Z}_{>0}$ (while planar maps correspond to maps of genus 0). The main difference lies in the fact that the corresponding mobiles are no longer trees but rather *one-faced* maps of higher genus, while the other properties still hold.

However, due to the apparition of cycles in the underlying structure of mobiles, another difficulty arises. Indeed, in the original bijection, vertices and edges in mobiles could carry labels (related to the geodesic distance in the original map), subject to local constraints. In our setting, the legs actually encode the *local* variations of these labels, which are thus implicit. Local constraints on labels are naturally translated into local constraints on the number of legs. But the labels have to remain consistent along each cycle of the mobiles, which gives rise to non-local constraints on the repartition of legs.

In order to deal with these additional constraints, and to be able to control the degrees of the vertices at the same time, we will now use a hybrid formulation of mobiles, carrying both labels and legs. As before, we will focus on the simpler case of mobiles coming from bipartite maps.

5.1 g-Mobiles

Definition 3. Given $g \in \mathbb{Z}_{\geq 0}$, a g-mobile is a one-faced map of genus g – embedded on the g-torus – such that there are two kinds of vertices (black and white), edges only occur as black–black edges or black–white edges, and black vertices additionally have so-called "legs" attached to them (which are not considered edges), whose number equals the number of white neighbor vertices.

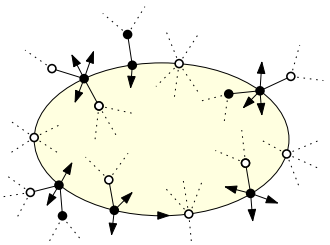


Fig. 1: An oriented cycle in a g-mobile and the constraint on its left (colored area). Notice that a similar constraint holds on its right, but is necessarily satisfied thanks to the properties of a g-mobile.

Furthermore, for each cycle c of the g-mobile, let $n_\circ$, $n_\rightarrow$ and $n_{-\circ}$ respectively be the numbers of white vertices on c, of legs dangling to the left of c and of white neighbours to the left of c. One has the following constraint (see Figure 5.1):

$$n_\rightarrow = n_\circ + n_{-\circ} \tag{23}$$

The *degree* of a black vertex is the number of half-edges plus the number of legs that are attached to it. A *bipartite* g-mobile is a g-mobile without black–black edges. A g-mobile is called *rooted* if an edge is distinguished and oriented.
Notice that a 0-mobile is simply a mobile as described in Definition 1.

Theorem 3. *Given $g \geq 0$, there is a bijection between g-mobiles that contain at least one black vertex and pointed maps of genus g, where white vertices in the mobile correspond to non-pointed vertices in the equivalent map, black vertices correspond to faces of the map, and the degrees of the black vertices correspond to the face valencies. This bijection induces a bijection on the edge sets so that the number of edges is the same. (Only the pointed vertex of the map has no counterpart.)*

Similarly, rooted g-mobiles that contain at least one black vertex are in bijection to rooted and vertex-pointed maps of genus g.

Proof. This generalization of the bijection to higher genus was first given in [6] for quadrangulations and [4] for Eulerian maps, from which we will exploit many ideas in the present section. ∎

5.2 Schemes of g-Mobiles

g-mobiles are not as easily decomposed as planar mobiles, due to the existence of cycles. However, they still exhibit a rather simple structure, based on *scheme* extraction.
The *g-scheme* (or simply the *scheme*) of a g-mobile is what remains when we apply the following operations (see Figure 2): first remove all legs, then remove iteratively all vertices of degree 1 and finally replace any maximal path of degree-2-vertices by a single edge.

Once these operations are performed, the remaining object is still a one-faced map of genus g, with black and white vertices (white–white edges can now occur), where the vertices have minimum degree 3.

To count g-mobiles, one key ingredient is the fact that there is only a finite number of schemes of a given genus. Indeed, let d_i be the number of degree i vertices of a g-scheme:

$$\sum_{k\geq 3}(i-2)d_i = \sum_{k\geq 3} i d_i - 2\sum_{k\geq 3} d_i = 2(\#\text{edges} - \#\text{vertices}) = 4g-2.$$

The number of vertices (respectively edges) is then bounded by $4g-2$ (respectively $6g-3$), where this bound is reached for cubic schemes (see an example in Figure 2).

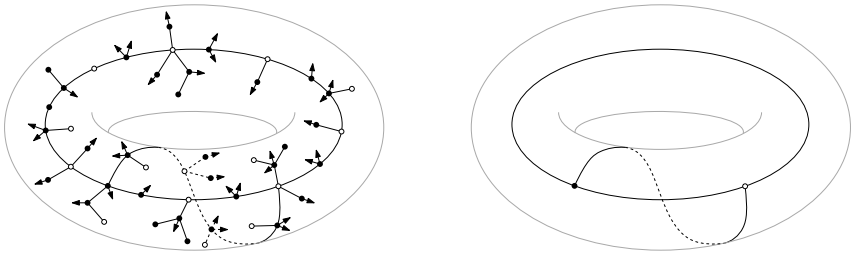


Fig. 2: A 1-mobile on the torus and its scheme.

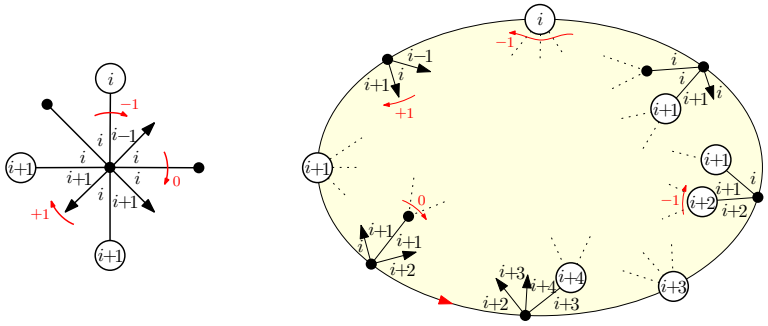


Fig. 3: The variations of labels around a black vertex and along an oriented cycle.

To recover a proper g-mobile from a given g-scheme, one would have to insert a suitable planar mobile into each corner of the scheme and to substitute each edge with some kind of path of planar mobiles. Unfortunately, this cannot be done independently: Around each black vertex, the total number of legs in every corner must equal the number of white neighbors, and around each cycle, (23) must hold.

In order to make these constraints more transparent, we will equip schemes with labels on white vertices and black corners. Now, when trying to reconstruct a g-mobile from a scheme, one has to ensure that the local variations are consistent with the global labelling. To be precise, the label variations are encoded as follows (see Figure 3):

- Around a black vertex of degree d, let $(l_1, \ldots, l_d)$ be the labels of its corners read in clockwise order:

$$\forall i, l_{i+1} - l_i = \begin{cases} +1 & \text{if there is a leg between the two corresponding corners,} \\ 0 & \text{if there is a black neighbor,} \\ -1 & \text{if there is a white neighbor.} \end{cases}$$

- Along the left side of an oriented cycle, the label decreases by 1 after a white vertex or when encountering a white neighbor and increases by 1 when encountering a leg.

The above statements hold for general – as well as bipartite – mobiles. In the following, we will only consider bipartite mobiles, as they are much easier to decompose.

5.3 Reconstruction of Bipartite Maps of Genus g

In the following, it will be convenient to work with rooted schemes. One can then define a canonical labelling and orientation for each edge of a rooted scheme. An edge e now has an origin e_- and an endpoint e_+. The k corners around a vertex of degree k are clockwisely ordered and denoted by $c_1, \dots, c_k$.

Given a scheme S, let $V_\circ, V_\bullet, C_\circ, C_\bullet$ be respectively the sets of white and black vertices and of white and black corners. A *labelled scheme* $(S, (l_c)_{c\in V_\circ \cup C_\bullet})$ is a pair consisting of a scheme S and a labelling on white vertices and black corners, with $l_c \geq 0$ for all c. Labellings are considered up to translation, as they will not affect local variations. For $e \in E_S$, an edge of S, we associate a label to each extremity l_{e_-}, l_{e_+}. If an extremity is a white vertex of label l, its label is l. If the extremity is a black vertex, its label is the same as the next clockwise corner of the black vertex.

Let a *doubly-rooted planar mobile* be a rooted (on a black or white vertex) planar mobile with a secondary root (also black or white). These two roots are the extremities of a path $(v_1, \dots, v_k)$. The *increment* of the doubly-rooted mobile is then defined as $n_\rightarrow - n_\circ - n_{\multimap\circ}$, which is not necessarily 0, as the path is not a cycle.

Similarly as in [4], we present a non-deterministic algorithm to reconstruct a g-mobile:

Algorithm.

(1) *Choose a labelled g-scheme $(S, (l_c)_{c\in V_\circ \cup C_\bullet})$.*

(2) *$\forall v \in V_\bullet$, choose a sequence of non-negative integers $(i_k)_{1\leq k\leq deg(v)}$, then attach i_k planar mobiles and $i_k + l_{c_{k+1}} - l_{c_k} + 1$ legs to c_k (the k^{th} corner of v).*

(3) *$\forall e \in S$, replace e by a doubly-rooted mobile of increment* $incr(e) = l_{e_+} - l_{e_-} + \begin{cases} +1 & \text{if } e_- \text{ is white,} \\ -1 & \text{if } e_- \text{ is black.} \end{cases}$

(4) *On each white corner of S, insert a planar mobile.*

(5) *Distinguish and orient an edge as the root.*

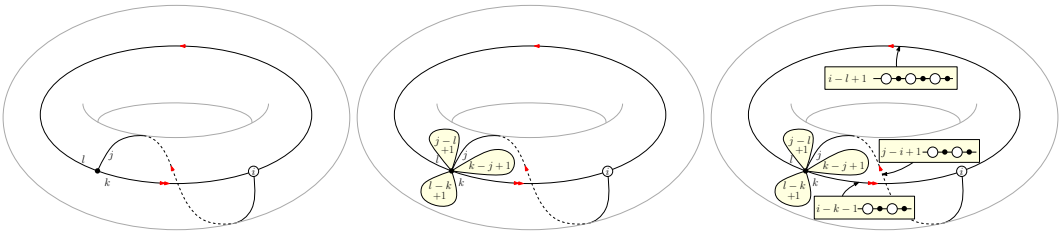


Fig. 4: Steps (1)–(3) of the algorithm.

Proposition 3. *Given $g > 0$, the algorithm generates each rooted bipartite g-mobile whose scheme has k edges in exactly $2k$ ways.*

Proof. One can easily see that the obtained object is indeed bipartite. Attaching planar mobiles and legs added at step (2) in a corner c_k creates new corners, such that:

- The first carries the same label l_{c_k} as c_k, and

- the last carries the label $l_{c_k} + (i_k + l_{c_{k+1}} - l_{c_k} + 1) - i_k = l_{c_{k+1}} + 1$.

The next corner should then be labelled $(l_{c_{k+1}} + 1) - 1 = l_{c_{k+1}}$, due to the next white neighbor, which is precisely what we want.

In the same fashion, at step (3), a simple counting shows that each edge is replaced by a path such that the labels along it evolve according to the scheme labelling.

We thus obtain a well-formed rooted bipartite g-mobile, with a secondary root on its scheme. Since the first root destroys all symmetries, there are exactly $2k$ choices for the secondary root, which would give the same rooted g-mobile. ∎

5.4 g-Mobile Counting

A doubly-rooted bipartite planar mobile can be decomposed along a sequence of elementary cells forming the path between its two roots. Its increment is simply the sum of the increments of its cells.

Definition 4. An *elementary cell* is a half-edge connected to a black vertex itself connected to a white vertex with a dangling half-edge. The white vertex has a sequence of black-rooted mobiles attached on each side. The black vertex has $j \geq 0$ legs and $k \geq 0$ white-rooted mobiles on its left, $l \geq 0$ white-rooted mobiles and $k + l - j + 2$ legs on its right, and its degree is $2(k + l + 2)$. The *increment* of the cell is then $j - k - 1$.

The generating series $P := P(t, z, (x_{2i}), s)$ of a cell, where s marks the increment, is:

$$P(t, z, (x_{2i}), s) = \frac{z^2 R^2}{t} \sum_{j,k,l \geq 0} \binom{j+k}{j} \binom{k + 2l - j + 2}{l} s^{j-k-1} x_{2(k+l+2)} R^{k+l} = \frac{z^2 R^2}{st} \widehat{P}.$$

The generating series $S := S(t, z, (x_{2i}), s)$ of a doubly-rooted mobile depends on the color of its roots (u, v):

$$S_{(u,v)}(t, z, (x_{2i}), s) = \begin{cases} \frac{1}{1-\widehat{P}} & \text{if } (u, v) = (\circ, \bullet) \text{ or } (\bullet, \circ), \\ \frac{z\widehat{P}}{1-\widehat{P}} & \text{if } (u, v) = (\circ, \circ), \\ \frac{zR^2}{st(1-\widehat{P})} & \text{if } (u, v) = (\bullet, \bullet). \end{cases}$$

We can now express the generating series $R_S := R_S(t, z, (x_{2i}))$ of rooted bipartite g-mobiles with scheme S:

$$R_S(t, z, (x_{2i})) = 2 \frac{z\partial}{\partial z} \frac{1}{2|E|} z^{|E|} t^{|V_\circ|} \left(\frac{R}{tz}\right)^{|C_\circ|} \bullet$$

$$\bullet \sum_{(l_c) \text{ labelling}} \left[\prod_{e \in E} [s^{incr(e)}] S_{(e_-, e_+)} \prod_{v \in V_\bullet} \sum_{i_1, \ldots, i_{\deg(v)} \geq 0} \left(\prod_{k=1}^{\deg(v)} \binom{2i_k + l_{c_{k+1}} - l_{c_k} + 1}{i_k} \right) x_{2(\deg(v) + \sum i_k)} \right]. \tag{24}$$

Proposition 4. *The generating series $M_D^{(g)} := M_D^{(g)}(t, z, (x_{2i}))$ for the family of rooted bipartite maps of genus g, where the vertex degrees belong to D, satisfies the relation:*

$$\frac{\partial M_D^{(g)}}{\partial t} = \frac{2}{z} \sum_{\substack{S \text{ scheme} \\ \text{of genus } g}} R_S(t, z, (x_{2i} \mathbb{1}_{\{2i \in D\}})). \tag{25}$$

Proof. This follows directly from Theorem 3 and Equation (24). ∎

6 Conclusion

Theorem 1 confirms the existence of a universal behaviour of planar maps. The asymptotics (with exponent $-5/2$) and this central limit theorem for the expected number of vertices of a given degree are believed to hold for any "reasonable" family of maps. It has also been shown in [6, 4] that a similar phenomenom occurs for maps of higher genus: The generating series of several families (quadrangulations, general and Eulerian maps) of genus g exhibit the same asymptotic exponent $5g/2 - 5/2$.

The expression obtained in Section 5 needs to be properly studied in order to obtain an asymptotic expansion. It refines previous results by controlling the degree of each vertex in the corresponding map.

References

[1] Cyril Banderier and Michael Drmota, *Formulae and asymptotics for coefficients of algebraic functions*, Combinatorics, Probability and Computing **24** (2015), no. 1, 1–53.

[2] E.A. Bender and E.R. Canfield, *Enumeration of degree restricted maps on the sphere.*, Planar graphs. Workshop held at DIMACS from November 18, 1991 through November 21, 1991, Providence, RI: American Mathematical Society, 1993, pp. 13–16.

[3] J. Bouttier, P. Di Francesco, and E. Guitter, *Planar maps as labeled mobiles*, Electron. J. Combin. **11** (2004), no. 1, Research Paper 69, 27.

[4] Guillaume Chapuy, *Asymptotic enumeration of constellations and related families of maps on orientable surfaces*, Combin. Probab. Comput. **18** (2009), no. 4, 477–516.

[5] Guillaume Chapuy, Éric Fusy, Mihyun Kang, and Bilyana Shoilekova, *A complete grammar for decomposing a family of graphs into 3-connected components*, Electron. J. Combin. **15** (2008), no. 1, Research Paper 148, 39.

[6] Guillaume Chapuy, Michel Marcus, and Gilles Schaeffer, *A bijection for rooted maps on orientable surfaces*, SIAM Journal on Discrete Mathematics **23** (2009), no. 3, 1587–1611.

[7] Gwendal Collet and Éric Fusy, *A simple formula for the series of bipartite and quasi-bipartite maps with boundaries*, Discrete Math. Theor. Comput. Sci. (2012), 607–618.

[8] Michael Drmota, *Random trees. An interplay between combinatorics and probability.*, Wien: Springer, 2009.

[9] Michael Drmota, Bernhard Gittenberger, and Johannes F. Morgenbesser, *Infinite systems of functional equations and Gaussian limiting distributions.*, Proceeding of the 23rd international meeting on probabilistic, combinatorial, and asymptotic methods in the analysis of algorithms (AofA'12), Montreal, Canada, June 18–22, 2012, Nancy: The Association. Discrete Mathematics & Theoretical Computer Science (DMTCS), 2012, pp. 453–478.

[10] Michael Drmota and Konstantinos Panagiotou, *A central limit theorem for the number of degree-k vertices in random maps.*, Algorithmica **66** (2013), no. 4, 741–761.

[11] W.T. Tutte, *A census of planar maps.*, Can. J. Math. **15** (1963), 249–271.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Cost functionals for large random trees

Jean-François Delmas[1], Jean-Stéphane Dhersin[2], Marion Sciauveau[1†]

[1] *CERMICS, École des Ponts, UPE, Champs-sur-Marne, France*
[2] *Université Paris 13, LAGA, 93430 Villetaneuse, France*

Additive tree functionals allow to represent the cost of many divide-and-conquer algorithms. We give an invariance principle for such tree functionals for the Catalan model (random tree uniformly distributed among the full binary ordered tree with given number of internal nodes). This relies on the natural embedding of binary trees into the Brownian excursion and then on elementary L^2 computations. We recover results first given by Fill and Kapur (2004) and then by Fill and Janson (2009).

Keywords: random binary tree, cost functional, toll function, Brownian excursion, continuum random tree

1 Introduction

1.1 Additive functionals and toll functions

Additive functionals on binary trees allow to represent the cost of algorithms such as "divide and conquer", see Kapur's PhD thesis [24] and Fill and Kapur [16]. For T a rooted full binary ordered tree, we set $|T|$ its cardinal, $\emptyset$ its root, $L(T)$ and $R(T)$ the left-sub-tree and right-sub-tree of the root of T. A functional F on binary trees is called an additive functional if it satisfies the following recurrence relation:

$$F(T) = F(L(T)) + F(R(T)) + b_{|T|}, \tag{1}$$

for all trees T such that $|T| \geq 1$ and with $F(\emptyset) = 0$. The given sequence $(b_n, n \geq 1)$ is called the toll function. Notice that:

$$F(T) = \sum_{v \in T} b_{|T_v|}, \tag{2}$$

where T_v is the sub-tree above v whose root is v.

We give some examples of commonly used toll functions or index functions related to additive functional. For $v, w \in T$, we say that w is an ancestor of v and write $w \leq v$ if $v \in T_w$. For $u, v \in T$, we denote by $u \wedge v$, the most recent common ancestor of u and v: $u \wedge v$ is the only element of T such that: $w \leq u$ and $w \leq v$ implies $w \leq u \wedge v$. We shall denote by d the graph distance in T.

- The total size of the tree T, $|T|$, corresponds to the additive functional with toll function $b_n = 1$.

†This work is partially supported by DIM RDMath IdF

- The total path length of T is defined by $P(T) = \sum_{v\in T} d(\emptyset, v)$. We have that $P(T) + |T|$ is the additive functional with toll function $b_n = n$ as

$$\sum_{w\in T} |T_w| = \sum_{w\in T}\sum_{v\in T} \mathbf{1}_{\{w\le v\}} = \sum_{v\in T} \big(1 + d(\emptyset, v)\big) = |T| + P(T).$$

- The shape functional of a tree T is the additive functional with toll function $b_n = \log(n)$. (This functional will not be covered by the main results of this paper.)

- The Wiener index of the tree T is defined by $W(T) = \sum_{u,v\in T} d(u, v)$. Notice that $d(u, v) = d(\emptyset, u) + d(\emptyset, v) - 2d(\emptyset, u \wedge v)$. This implies that $W(T) = 2|T| \sum_{w\in T} |T_w| - 2\sum_{w\in T} |T_w|^2$ as

$$\sum_{w\in T} |T_w|^2 = \sum_{w\in T}\sum_{u,v\in T} \mathbf{1}_{\{w\le u\wedge v\}} = \sum_{u,v\in T} \big(1 + d(\emptyset, u \wedge v)\big) = |T|^2 + \sum_{u,v\in T} d(\emptyset, u \wedge v).$$

  According to (2), the functional $\sum_{w\in T} |T_w|^2$ is an additive functional with toll function $b_n = n^2$. And thus the Wiener index of a full binary tree is a combination of two additive functionals.

- The Sackin index (or external path length) of the tree T, used to study the balance of the tree, is similar to the total path length of T when one considers only the leaves: $S(T) = \sum_{v\in\mathcal{L}(T)} d(\emptyset, v)$, where the set of leaves is $\mathcal{L}(T) = \{v \in T; |T_v| = 1\}$. Using that for a full binary tree we have $|T| = 2|\mathcal{L}(T)| - 1$, we deduce that $2S(T) = \sum_{w\in T} |T_w| - 1$. The Colless index of the tree T is defined as $C(T) = \sum_{v\in T} \big||\mathcal{L}(L_v)| - |\mathcal{L}(R_v)|\big|$. Since T is a full binary tree, we get $2|\mathcal{L}(L_v)| - 2|\mathcal{L}(R_v)| = |L_v| - |R_v|$ and $|L_v| + |R_v| = |T_v| - 1$. We obtain that $2C(T) = \sum_{w\in T} |T_w| - |T| - 2\sum_{v\in T} \min(|L_v|, |R_v|)$. The cophenetic index of the tree T, which is used in [27] to study the balance of the tree, is defined by $\mathrm{Co}(T) = \sum_{u,v\in\mathcal{L}(T),\, u\neq v} d(\emptyset, u\wedge v)$. Using again that T is a full binary tree, we get $4\mathrm{Co} = 4\sum_w |\mathcal{L}(T_w)|(|\mathcal{L}(T_w)| - 1) - 4|\mathcal{L}(T)|(|\mathcal{L}(T)| - 1) = \sum_{w\in T} |T_w|^2 - |T|^2 - |T| + 1$.

1.2 Asymptotics for additive functionals in the Catalan model

We consider the Catalan model: let T_n be a random tree uniformly distributed among the set of full binary ordered trees with n internal nodes (and thus $n+1$ leaves), which has cardinal $C_n = (2n)!/[(n!^2)(n+1)]$. In particular, we have:

$$|T_n| = 2n + 1.$$

Recall that T_n is a (full binary) Galton-Watson tree (also known as simply generated tree) conditioned on having n internal nodes. It is well known, see Takàcs [34], Aldous [3, 4] and Janson [21], that $|T_n|^{-3/2} P(T_n)$ converges in distribution, as n goes to infinity, towards $2\int_0^1 B_s\, ds$, where $B = (B_s, 0 \le s \le 1)$ is the normalized Brownian excursion. This result can be seen as a consequence of the convergence in distribution of T_n (in fact the contour process) properly scaled towards the Brownian continuum tree whose contour process is B, see [3] and Duquesne [9], or Duquesne and Le Gall [10] in the setting of Brownian excursion. For a combinatorial approach, which can be extended to other families of trees, see also Fill and Kapur [15, 17] or Fill, Flajolet and Kapur [13].

In [16], the authors considered the toll functions $b_n = n^\beta$ with $\beta > 0$ and they proved that with a suitable scaling the corresponding additive functional $F_\beta(T_n)$ converge in distribution to a limit, say Y_β. The distribution of Y_β is characterized by its moments. (In [16], the authors consider also the toll function $b_n = \log(n)$.) See also Janson and Chassaing [23] for asymptotics of the Wiener index, which is a consequence of the joint convergence in distribution of $(F_1(T_n), F_2(T_n))$ with a suitable scaling and Blum, François and Janson [6] for the convergence of the Sackin and Colless indexes. It is announced in Fill and Janson [14] that for $\beta > 1/2$, Y_β can be represented as a functional of the normalized Brownian excursion. More precisely, for $\beta > 1/2$, Y_β is distributed as $\phi_\beta(B)$, where for any non-negative continuous function h defined on $[0,1]$:

$$\phi_\beta(h) = \beta \int_0^1 [t^{\beta-1} + (1-t)^{\beta-1}] h(t)\, dt - \frac{1}{2}\beta(\beta-1) \int_{[0,1]^2} |t-s|^{\beta-2} [h(t) + h(s) - 2m_h(s,t)]\, dsdt,$$

with

$$m_h(s,t) = \inf_{u \in [s\wedge t, s\vee t]} h(u). \tag{3}$$

Furthermore, for $\beta = 1$, this reduces to $\phi_1(h) = \int_0^1 h$ and for $\beta > 1$ we also have:

$$\phi_\beta(h) = \beta(\beta-1) \int_{[0,1]^2} |t-s|^{\beta-2} m_h(s,t)\, dsdt. \tag{4}$$

We use the natural embedding of T_n into the Brownian excursion, see [4], so that the convergence in distribution of the additive functional is then an a.s. convergence (which holds simultaneously for all $\beta > 1/2$) and also give the fluctuations for this a.s. convergence. From this convergence, we also provide another representation of $\phi_\beta(h)$ which is a natural by-product of the a.s. convergence.

Remark 1.1 *The method presented in this paper based on the embedding of T_n into a Brownian excursion can not be extended directly to other models of trees such as binary search trees, recursive trees or simply generated trees.*

Concerning binary search trees (or random permutation model or Yule trees), see [31] and [32] for the convergence of the external path length (which corresponds in our setting to the Sackin index), [28] for toll function $b_n = n^\beta$, [29] for the Wiener index (and [21] for simply generated trees), [6] (and [18] for other trees) for the Sacking and Colless indexes.

Concerning recursive trees, see [26], [8] for the convergence of the total path length and [29] for the Wiener index. In the setting of recursive tree, then (1) is a stochastic fixed point equation, which can be analyzed using the approach of [33].

Remark 1.2 *One can replace the toll function $b_{|T|}$ in (1) by a function of the tree, say $\mathbf{b}(T)$. For example, if one consider $\mathbf{b}(T) = \mathbf{1}_{\{|T|=1\}}$, then the corresponding additive functional $F(T) = |\mathcal{L}(T)|$ gives the number of leaves. The case of "local" toll function $\mathbf{b}$ (with finite support or fast decreasing rate) has been considered in the study of fringe trees, see [2], [7] for binary search trees, and [22] for simply generated trees and [19] for binary search trees and recursive trees.*

See [20] for the study of the phase transition on asymptotics of additive functionals with toll functions $b_n = n^\beta$ on binary search trees between the "local" regime (corresponding to $\beta \leq 1/2$) and the "global" regime ($\beta > 1/2$). The same phase transition is observed for the Catalan model, see [16]. Our main result, see Theorem 3.1, concerns specifically the "global" regime.

2 Binary trees in the Brownian excursion

We begin by recalling the definition of a real tree, see [12], and some elementary properties of the Brownian continuum random tree, see [25]. A real tree is a metric space $(\mathcal{T}, d)$ which satisfies the following two properties for every $x, y \in \mathcal{T}$:

(i) There exists a unique isometric map $f_{x,y}$ from $[0, d(x, y)]$ into $\mathcal{T}$ such that $f_{x,y}(0) = x$ and $f_{x,y}(d(x, y)) = y$.

(ii) If ϕ is a continuous injective map from $[0, 1]$ into $\mathcal{T}$ such that $\phi(0) = x$ and $\phi(1) = y$, then we have $\phi([0, 1]) = f_{x,y}([0, d(x, y)])$.

Equivalently, a metric space $(\mathcal{T}, d)$ is a real tree if and only if $\mathcal{T}$ is connected and d satisfies the four point condition:

$$d(s, t) + d(x, y) \leq \max(d(s, x) + d(t, y), d(s, y) + d(t, x)) \quad \text{for all} \quad s, t, x, y \in \mathcal{T}.$$

A rooted real tree is a real tree $(\mathcal{T}, d)$ with a distinguished vertex $\emptyset$ called the root. In the following paragraphs, we will only consider compact rooted real trees.

One can use continuous functions to encode compact rooted real trees as follows. Let h be a non-negative continuous function defined on $[0, 1]$ such that $h(0) = h(1) = 0$. For every $x, y \in [0, 1]$, we set $d_h(x, y) = h(x) + h(y) - 2m_h(x, y)$, where m_h is defined in (3). It is easy to check that d_h is symmetric and satisfies the triangle inequality. The relation $\sim_h$ defined on $[0, 1]^2$ by $x \sim_h y \Leftrightarrow d_h(x, y) = 0$ is an equivalence relation. Let $\mathcal{T}_h = [0, 1]/\sim_h$ be the corresponding quotient space. The function d_h on $[0, 1]^2$ induces a function on $\mathcal{T}_h^2$, which we still denoted by d_h, and which is a distance on $\mathcal{T}_h$. It is not difficult to check that $(\mathcal{T}_h, d_h)$ is then a compact real tree. We denote by $\mathbf{p}$ the canonical projection from $[0, 1]$ into $\mathcal{T}_h$. Thus, the metric space $(\mathcal{T}_h, d_h)$ is a compact real tree which can be viewed as a rooted real tree by setting $\emptyset = \mathbf{p}(0)$.

Let $B = (B_t, 0 \leq t \leq 1)$ be a normalized Brownian excursion. Informally, B is just a linear standard Brownian path started from the origin and conditioned to stay positive on $(0, 1)$ and to come back to 0 at time 1. For $\alpha > 0$, let $e = \sqrt{2/\alpha}\, B$ and let $\mathcal{T}_e$ denote the Brownian tree. The continuum random tree introduced by Aldous corresponds to $\alpha = 1/2$ and the Brownian tree associated to the normalized Brownian excursion corresponds to $\alpha = 2$. We shall keep the parameter α so that the two previous cases are easy to read on the results.

Let $(U_n, n \in \mathbb{N}^*)$ be a sequence of independent random variables uniform on $[0, 1]$, independent of e. We denote by $\mathcal{T}_n$ the random tree spanned by the $n + 1$ points $\mathbf{p}(U_1), \ldots, \mathbf{p}(U_{n+1})$ that is the smallest connected subset of $\mathcal{T}_e$ that contains $\mathbf{p}(U_1), \ldots, \mathbf{p}(U_{n+1})$ and the root. The tree $\mathcal{T}_n$ has exactly $2n + 1$ nodes. There is a natural order on $\mathcal{T}_n$ given by the order of its external nodes $\mathbf{p}(U_1), \ldots, \mathbf{p}(U_{n+1})$. Let T_n be the corresponding trees when one forget about the branch lengths. It is well known that T_n is uniform among the full binary ordered trees with n internal nodes. See Figure (1) for an example with $n = 4$.

Let $(h_1, \ldots, h_{2n+1})$ be the branch lengths of the tree $\mathcal{T}_n$ given in the lexicographical order. We recall, see [4], [30] (Theorem 7.9) or [11], that the density of $(h_1, \ldots, h_{2n+1})$ is given by:

$$f_n(h_1, \ldots, h_{2n+1}) = 2\frac{(2n)!}{n!}\frac{\alpha^{n+1}}{L_n}\, \mathrm{e}^{-\alpha L_n^2}\, \mathbf{1}_{\{h_1>0,\ldots,h_{2n+1}>0\}},$$

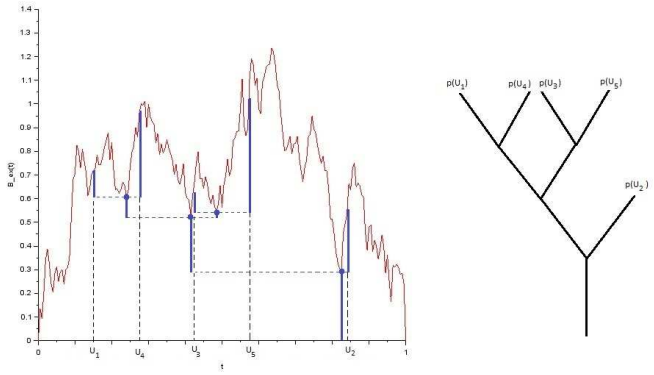


Fig. 1: The Brownian excursion, $\mathcal{T}_n$ (for $n = 4$) and T_n.

where $L_n = \sum_{k=1}^{2n+1} h_k$ denotes the total length of $\mathcal{T}_n$. Notice that the edge-lengths are independent of the shape of the tree T_n. It is then easy to deduce that the density of L_n is given by:

$$f_{L_n}(x) = 2\frac{\alpha^{n+1}}{n!} x^{2n+1}\, \mathrm{e}^{-\alpha x^2}\, \mathbf{1}_{\{x>0\}}.$$

And we have, see [1] that $L_n/\sqrt{n}$ converges a.s. towards $1/\sqrt{\alpha}$. Furthermore, elementary computations give that $(h_1, \ldots, h_{2n+1})$ has the same distribution as $(L_n\Delta_1, \ldots, L_n\Delta_{2n+1})$, where $\Delta_1, \ldots, \Delta_{2n+1}$ represents the lengths of the $2n+1$ intervals obtained by cutting $[0,1]$ at $2n$ points $V_1, \ldots, V_{2n}$, where $V_1, \ldots, V_{2n}$ are $2n$ independent uniform random variables on $[0,1]$ and independent of L_n. We thus deduce the following elementary Lemma.

Lemma 2.1 *The random vector* $(h_1, \ldots, h_{2n+1})$ *has the same distribution as:*

$$\left(L_n \frac{E_1}{S_{2n+1}}, \ldots, L_n \frac{E_{2n+1}}{S_{2n+1}} \right),$$

where $E_1, \ldots, E_{2n+1}$ *are* $2n+1$ *independent exponential random variables, independent of* L_n*, and* $S_{2n+1} = \sum_{k=1}^{2n+1} E_k$.

We end this section with a result on the Brownian excursion. We set m for m_e defined in (3). For $s \in [0,1]$ and $r \in [0, e_s)$, the length of the excursion of e above r straddling s is given by:

$$\sigma_{r,s} = \int_0^1 \mathbf{1}_{\{m(s,t)\geq r\}} dt.$$

For $\beta \geq 0$, we set:

$$Z_\beta = \int_0^1 ds \int_0^{e_s} dr\ \sigma_{r,s}^{\beta-1}. \tag{5}$$

The next result is proved using the representation of Brownian excursion from Biane [5].

Lemma 2.2 *Let $\beta > 0$. We have:*

$$\mathbb{P}\left(Z_\beta < +\infty\right) = \{\, 0 \quad \text{if } \beta \leq 1/2, 1 \quad \text{if } \beta > 1/2.$$

The following result based on elementary computations allows to recover the formulation of our Corollary 3.2 given in [16] and [14], see (4).

Lemma 2.3 *We have $Z_1 = \int_0^1 e_s \, ds$ and for $\beta > 1$:*

$$Z_\beta = \frac{1}{2}\,\beta(\beta-1)\int_{[0,1]^2} |t-s|^{\beta-2}\, m(s,t)\, ds\, dt.$$

3 Results

Inspired by (2), we consider the following random measure A_n associated to the tree T_n defined as follows. For any non-negative function defined on $[0,1]$, we set:

$$A_n(f) = |T_n|^{-\frac{3}{2}} \sum_{v \in T_n} |T_v| f\left(\frac{|T_v|}{|T_n|}\right),$$

where we recall that T_v is the sub-tree above v with root v and $|T_n| = 2n+1$. The case $f(x) = x^{\beta-1}$ corresponds to the additive functional on T_n given by (2) with toll function $b_n = n^\beta$ up to the scaling factor $|T_n|^{-(\frac{1}{2}+\beta)}$.

We define the following random measure associated to the excursion e:

$$\Phi(f) = \sqrt{2\alpha} \int_0^1 ds \int_0^{e_s} dr\; f(\sigma_{r,s}).$$

We now state our main result on the invariance principle.

Theorem 3.1 *Almost surely, for all $f \in \mathcal{C}^0((0,1])$ such that $\lim_{x\downarrow 0} x^{\frac{1}{2}-\varepsilon} f(x) = 0$ for some $0 < \varepsilon < \frac{1}{2}$, we have:*

$$\lim_{n\to+\infty} A_n(f) = \Phi(f).$$

Proof: We only present the main ideas of the proof, as the detailed proofs will be given in a forthcoming paper. Let f be a smooth enough function defined on $[0,1]$. We first notice that $A_n(f)$ is well approximated by:

$$\begin{aligned} A_{n,1}(f) &= |T_n|^{-\frac{3}{2}} \sum_{v\in T_n} (|T_v|+1) f\left(\frac{|T_v|}{|T_n|}\right) \\ &= 2|T_n|^{-\frac{3}{2}} \sum_{u\in\mathcal{L}(T_n)} \sum_{v\in T_n,\, v\leq u} f\left(\frac{|T_v|}{|T_n|}\right), \end{aligned}$$

where we recall that $\mathcal{L}(T)$ denotes the leaves of the tree T and $|T| = 2|\mathcal{L}(T)| - 1$ in a full binary tree. The precise distribution of the heights $(h_1, \ldots, h_{2n+1})$ given in Lemma 2.1 and the fact that $L_n/\sqrt{n}$

converges a.s. towards $1/\sqrt{\alpha}$ gives that h_v is close to $L_n/(2n+1)$ that is of $1/(2\sqrt{\alpha n})$. In the spirit of the law of the large number, using L^2 computations, we obtain that $A_{n,1}(f)$ is well approximated by:

$$\begin{aligned} A_{n,2}(f) &= 4\sqrt{\alpha n}\,|T_n|^{-\frac{3}{2}} \sum_{u\in\mathcal{L}(T_n)} \sum_{v\in T_n,\, v\le u} f\left(\frac{|T_v|}{|T_n|}\right) h_v \\ &= 4\sqrt{\alpha n}\,|T_n|^{-\frac{3}{2}} \sum_{k=1}^{n+1} \int_0^{e_{U_k}} dr\, f\left(\frac{2X_{r,k}+1}{|T_n|}\right), \end{aligned}$$

where $X_{r,k}+1$ denotes the number of integers $i \in \{1,\dots,n+1\}$ such that the random variable U_i belongs to the same excursion interval of e above level r as U_k, that is $m(e_{U_i}, e_{U_k}) > r$. Conditionally on e, the random variable $X_{r,k}$ is binomial with parameter (n, σ_{r,U_k}). In particular, for large n, $2X_{r,k}+1$ is close to $2n\sigma_{r,U_k}$ and thus $(2X_{r,k}+1)/|T_n|$ is close to σ_{r,U_k}. Using $|T_n| = 2n+1$, the smoothness of f and L^2 computations, we get that $A_{n,2}(f)$ is well approximated by:

$$A_{n,3}(f) = \sqrt{2\alpha}\,\frac{1}{n+1} \sum_{k=1}^{n+1} \int_0^{e_{U_k}} dr\, f\left(\sigma_{r,U_k}\right).$$

Then use the law of large number (conditionally on e) to get that a.s.

$$\lim_{n\to+\infty} A_{n,3}(f) = \sqrt{2\alpha} \int_0^1 ds \int_0^{e_s} dr\, f\left(\sigma_{r,s}\right).$$

We deduce that for all (smooth enough) functions, we have a.s. $\lim_{n\to+\infty} A_n(f) = \Phi(f)$. Since the considered family of smooth functions is convergence determining, this implies that a.s. $(A_n, n \in \mathbb{N}^*)$ converges towards Φ (for the weak convergence or vague convergence of finite measure on $[0,1]$). This in turns gives that a.s. for all continuous functions, $\lim_{n\to+\infty} A_n(f) = \Phi(f)$. More work is required to extend this result to the class of functions considered in the Theorem. □

According to Lemma 2.2, the random variable Z_β defined by (5) is a.s. finite (resp. infinite) if $\beta > 1/2$ (resp. $0 < \beta \le 1/2$). Considering the function $f(x) = x^{\beta-1}$ for $\beta > 0$, we easily deduce from Theorem 3.1 the following convergence. For $n \in \mathbb{N}^*$, we set:

$$Z_\beta^{(n)} = \frac{1}{\sqrt{2\alpha}}\,|T_n|^{-(\beta+\frac{1}{2})} \sum_{v\in T_n} |T_v|^\beta.$$

Corollary 3.2 *We have almost surely, for all $\beta > 0$,*

$$\lim_{n\to+\infty} Z_\beta^{(n)} = Z_\beta.$$

Remark 3.3 *Corollary 3.2 gives directly that $(|T_n|^{-3/2}\sum_{v\in T_n}|T_v|, |T_n|^{-5/2}\sum_{v\in T_n}|T_v|^2)$ is asymptotically distributed as $\sqrt{2\alpha}\,(Z_1, Z_2)$. Since, according to [6] or [18], the quantity $\sum_{v\in T_n}\min(|L_v|,|R_v|)$ is of smaller magnitude than $|T_n|^{3/2}$, we can directly recover the joint asymptotic distribution of the total length path, the Wiener, Sackin, Colless and cophenetic indexes defined in Section 1.1 for the Catalan model. More precisely, we have the following a.s. convergence as n goes to infinity:*

$$\left(\frac{P(T_n)}{|T_n|^{3/2}}, \frac{W(T_n)}{|T_n|^{5/2}}, \frac{S(T_n)}{|T_n|^{3/2}}, \frac{C(T_n)}{|T_n|^{3/2}}, \frac{\mathrm{Co}(T_n)}{|T_n|^{5/2}}\right) \to \sqrt{2\alpha}\left(Z_1, 2(Z_1-Z_2), \frac{Z_1}{2}, \frac{Z_1}{2}, \frac{Z_2}{4}\right).$$

The next proposition gives the fluctuations corresponding to the invariance principles of Corollary 3.2 when $\beta \geq 1$. Notice the speed of convergence in the invariance principle is of order $|T_n|^{-1/4}$ and the limiting variance is (up to a multiplicative constant) given by $Z_{\beta'}$ with $\beta' = 2\beta$.

Proposition 3.4 *For all $\beta \geq 1$, we have the following convergence in distribution as n goes to infinity:*

$$\left(|T_n|^{1/4}(Z_\beta^{(n)} - Z_\beta),\, Z_\beta^{(n)}\right) \to \left((2\alpha)^{-1/4}\sqrt{Z_{2\beta}}\; G, Z_\beta\right),$$

where G is a centered reduced Gaussian random variable independent of the excursion e.

The contribution to the fluctuations is given by the error of approximation of $A_{n,1}(f)$ by $A_{n,2}(f)$ with $f(x) = x^{\beta-1}$, see notations from the proof of Theorem 3.1. This corresponds to the fluctuations coming from the approximation of the branch lengths $(h_v, v \in T_n)$ by their mean, which relies on the explicit representation on their joint distribution given in Lemma 2.1. In particular, there is no other contribution to the fluctuations from the approximation of the continuum tree $\mathcal{T}_e$ by the sub-tree $\mathcal{T}_n$.

References

[1] R. Abraham and J.-F. Delmas. Record process on the continuum random tree. *ALEA Lat. Am. J. Probab. Math. Stat.*, 10(1):225–251, 2013.

[2] D. Aldous. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.*, 1(2):228–266, 1991.

[3] D. Aldous. The continuum random tree. II. An overview. In *Stochastic analysis (Durham, 1990)*, volume 167 of *London Math. Soc. Lecture Note Ser.*, pages 23–70. Cambridge Univ. Press, Cambridge, 1991.

[4] D. Aldous. The continuum random tree. III. *Ann. Probab.*, 21(1):248–289, 1993.

[5] P. Biane. Relations entre pont et excursion du mouvement brownien réel. *Ann. Inst. H. Poincaré Probab. Statist.*, 22(1):1–7, 1986.

[6] M. Blum, O. François, and S. Janson. The mean, variance and limiting distriution of two statistics sensitive to phylogenetic tree balance. *Ann. Appl. Probab.*, 16(4):2195–2214, 2006.

[7] L. Devroye. Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.*, 32(1):152–171, 2002.

[8] R. P. Doborw and J. A. Fill. Total path length for random recursive trees. *Combin. Probab. Comput.*, 8(4):317–333, 1999.

[9] T. Duquesne. A limit theorem for the contour process of conditioned Galton-Watson trees. *Ann. Probab.*, 31(2):996–1027, 2003.

[10] T. Duquesne and J.-F. Le Gall. *Random trees, Lévy processes and spatial branching processes*, volume 281 of *Astérisque*. 2002.

[11] T. Duquesne and J.-F. Le Gall. Probabilistic and fractal aspects of Lévy trees. *Probab. Th. and Rel. Fields*, 131:553–603, 2005.

[12] S. Evans. *Probability and real trees*, volume 1920 of *Lecture Notes in Math.* Springer, 2008.

[13] J. A. Fill, P. Flajolet, and N. Kapur. Singularity analysis, Hadamard products, and tree recurrences. *J. Comput. Appl. Math.*, 174(2):271–313, 2005.

[14] J. A. Fill and S. Janson. Precise logarithmic asymptotics for the right tails of some limit random variables for random trees. *Ann. Comb.*, 12(4):403–416, 2009.

[15] J. A. Fill and N. Kapur. A repertoire for additive functionals of uniformly distributed m-ary search trees (extended abstract). In *2005 International Conference on Analysis of Algorithms*, Discrete Math. Theor. Comput. Sci. Proc., AD, pages 105–114 (electronic).

[16] J. A. Fill and N. Kapur. Limiting distributions for additive functionals on Catalan trees. *Theoret. Comput. Sci.*, 326(1-3):69–102, 2004.

[17] J. A. Fill and N. Kapur. Transfer theorems and asymptotic distributional results for m-ary search trees. *Random Structures Algorithms*, 26(4):359–391, 2005.

[18] D. J. Ford. Probabilities on cladograms: introduction to the alpha model. *ArXiv preprint*, 2005.

[19] C. Holmgren and S. Janson. Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Elect. J. Probab.*, 20:1–51, 2005.

[20] H.-K. Hwang and R. Neininger. Phase change of limit laws in the "Quicksort" recurrence under varying toll functions. *SIAM J. Comput.*, 31(6):1687–1722, 2002.

[21] S. Janson. The Wiener index of simply generated random trees. *Random Struct. Algo.*, 22(4):337–358, 2003.

[22] S. Janson. Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton-Watson trees. *Random Struct. Algo.*, To appear.

[23] S. Janson and P. Chassaing. The center of mass of the ISE and the Wiener index of trees. *Electron. Comm. Probab.*, pages 178–187, 2004.

[24] N. Kapur. *Additive functionals on random search trees*. ProQuest LLC, Ann Arbor, MI, 2003. Thesis (Ph.D.)–The Johns Hopkins University.

[25] J.-F. Le Gall. Random trees and applications. *Probab. Surv.*, 2:245–311, 2005.

[26] H. M. Mahmoud. Limiting distribution for path lengths in recursive trees. *Probab. Engin. Inform. Sci.*, 5(1):53–59, 1991.

[27] A. Mir, F. Rossellò, and L. Rotger. A new balance index for phylogenetic trees. *Mathematical Biosciences*, 241(1):125 – 136, 2013.

[28] R. Neininger. On binary search tree recursions with monomials as toll functions. *J. Comput. Appl. Math.*, 142(1):185–196, 2002.

[29] R. Neininger. The Wiener index of random trees. *Combin., Probab. Comput.*, 11(6):587–597, 2002.

[30] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Math*. Springer, 2002.

[31] M. Régnier. A limiting distribution for "Quicksort". *RAIRO Theoret. Inform. and Appl.*, 23:335–343, 1989.

[32] U. Rösler. A limit theorem for "Quicksort". *RAIRO Theoret. Inform. and Appl.*, 25:85–100, 1991.

[33] U. Rösler and L. Rüschendorf. The contraction method for recursive algorithms. *Algorithmica*, 29(1-2):3–33, 2001.

[34] L. Takàcs. On the total heights of random rooted binary trees. *J. Combin. Theory Ser. B*, 61(2):155–166, 1994.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Asymmetric Rényi Problem and PATRICIA Tries

Michael Drmota[1†], Abram Magner[2], Wojciech Szpankowski[3‡§]

[1] *Institute for Discrete Mathematics and Geometry, TU Wien, A-1040 Wien, Austria (email: michael.drmota@tuwien.ac.at).*
[2] *Coordinated Science Lab, UIUC, Champaign, IL, USA (email: anmagner@illinois.edu)*
[3] *Department of Computer Science, Purdue University, IN 47907, USA (email: spa@cs.purdue.edu)*

Abstract: In 1960 Rényi asked for the number of random queries necessary to recover a hidden bijective labeling of n distinct objects. In each query one selects a random subset of labels and asks, what is the set of objects that have these labels? We consider here an asymmetric version of the problem in which in every query an object is chosen with probability $p > 1/2$ and we ignore "inconclusive" queries. We study the number of queries needed to recover the labeling in its entirety (the *height*), to recover at least one single element (the *fillup level*), and to recover a randomly chosen element (the *typical depth*). This problem exhibits several remarkable behaviors: the depth D_n converges in probability but not almost surely and while it satisfies the central limit theorem its local limit theorem doesn't hold; the height H_n and the fillup level F_n exhibit phase transitions with respect to p in the second term. To obtain these results, we take a unified approach via the analysis of the *external profile* defined at level k as the number of elements recovered by the kth query. We first establish new precise asymptotic results for the average and variance, and a central limit law, for the external profile in the regime where it grows polynomially with n. We then extend the external profile results to the boundaries of the central region, leading to the solution of our problem for the height and fillup level. As a bonus, our analysis implies novel results for random PATRICIA tries, as it turns out that the problem is probabilistically equivalent to the analysis of the height, fillup level, typical depth, and external profile of a PATRICIA trie built from n independent binary sequences generated by a biased(p) memoryless source.

Keywords: Rényi problem, PATRICIA trie, profile, height, fillup level, analytic combinatorics, Mellin transform, depoissonization

1 Introduction

In his lectures in the summer of 1960 at Michigan State University, Alfred Rényi discussed several problems related to random sets [21]. Among them there was a problem regarding recovering a labeling of

[†]M. Drmota was supported by the Austrian Science Foundation FWF Grant No. S9604.
[‡]A. Magner and W. Szpankowski were supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, NIH Grant 1U01CA198941-01, and the NCN grant, grant UMO-2013/09/B/ST6/02258.
[§]W. Szpankowski is also with the Faculty of Electronics, Telecommunications, and Informatics, Gdańsk University of Technology, Poland.

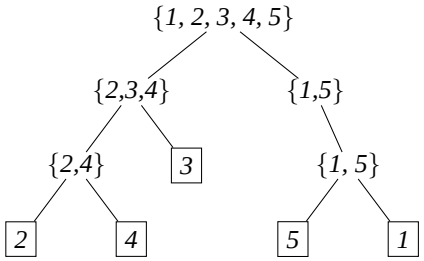
a set X of n distinct objects by asking random subset questions of the form "which objects correspond to the labels in the (random) set B?" For a given method of randomly selecting queries, Rényi's original problem asks for the typical behavior of the number of queries necessary to recover the hidden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

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

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

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

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

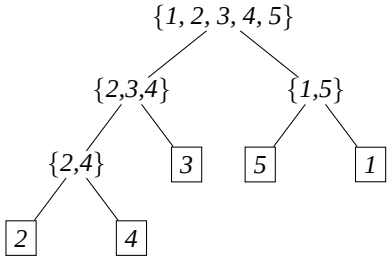


Each level $j \geq 0$ *of the tree depicts the partition* $\mathfrak{P}_j$*, where a right child node corresponds to the subset of objects in the parent set which are contained in the response to the* j*th query. Singletons are only explicitly depicted in the first level in which they appear.* □

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

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

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



□

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

The symmetric version (i.e., $p = 1/2$) of the problem (with a variation) was discussed by Pittel and Rubin in [19], where they analyzed the typical value of H_n. In their model, a query is constructed by deciding whether or not to include each label from A independently with probability $p = 1/2$. To make the problem interesting, they added a constraint similar to ours: namely, a query is, as in our model, admissible if and only if it splits every nontrivial element of the current partition. In contrast with our model, however, Pittel and Rubin completely discard inconclusive queries (rather than modifying their inconclusive subsets as we do). Despite this difference, the model considered in [19] is probabilistically equivalent to ours for the symmetric case. Our primary contribution is the analysis of the problem in the asymmetric case ($p > 1/2$), but our methods of proof allow us to recover the results of Pittel and Rubin.

The question asked by Rényi brings some surprises. For the symmetric model ($p = 1/2$) Pittel and Rubin [19] were able to prove that the number of necessary queries is with high probability (whp) (see Theorem 1)

$$H_n = \log_2 n + \sqrt{2 \log_2 n} + o(\sqrt{\log n}). \tag{1}$$

In this paper, we re-establish this result using a different approach *and* prove that for $p > 1/2$ the number of queries grows whp as

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

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

$$F_n = \begin{cases} \log_{1/q} n - \log_{1/q} \log \log n + o(\log \log \log n) & p > q \\ \log_2 n - \log_2 \log n + o(\log \log n) & p = q = 1/2. \end{cases} \tag{3}$$

We then prove in Theorem 2 some interesting probabilistic behaviors of D_n. We have $D_n/\log n \to 1/h(p)$ (in probability) where $h(p) := -p\log p - q\log q$, but we do not have almost sure convergence. Moreover, D_n appropriately normalized satisfies a central limit result, but not a local limit theorem due to some oscillations discussed below.

We establish these results in a novel way by considering first the *external profile* $B_{n,k}$, whose analysis was, until recently, an open problem of its own (the second and third authors gave a precise analysis of the external profile in an important range of parameters in [13, 15], but the present paper requires nontrivial extensions). The external profile at level k is the number of bijection elements revealed by the kth query (one may also define the *internal* profile at level k as the number of non-singleton elements of the partition immediately after the kth query). Its study is motivated by the fact that many other parameters, including all of those that we mention here, can be written in terms of it. Indeed, $\Pr[D_n = k] = \mathbb{E}[B_{n,k}]/n$, $H_n = \max\{k :\ B_{n,k} > 0\}$, and $F_n = \min\{k :\ B_{n,k} > 0\} - 1$.

We now discuss our new results concerning the probabilistic behavior of the external profile. We establish in [15, 13] precise asymptotic expressions for the expected value and variance of $B_{n,k}$ in the *central range*, that is, with $k \sim \alpha\log n$, where, for any fixed $\epsilon > 0$, $\alpha \in (1/\log(1/q)+\epsilon, 1/\log(1/p)-\epsilon)$ (the left and right endpoints of this interval are associated with F_n and H_n, respectively). Specifically, we show that both the mean and the variance are of the same (explicit) polynomial order of growth (with respect to n) (see Theorem 3). More precisely, we show that both expected value and variance grow for $k \sim \alpha\log n$ as

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

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

In the present paper, we exploit the expected value analysis of $B_{n,k}$ in the central range to give precise distributional information about D_n via the identity $\Pr[D_n = k] = \mathbb{E}[B_{n,k}]/n$. Note that the oscillations in $\mathbb{E}[B_{n,k}]$ are the source of the peculiar behavior of D_n.

In order to establish the most interesting results claimed in the present paper for H_n and F_n, the analysis sketched above does not suffice: we need to estimate the mean and the variance of the external profile *beyond* the range $\alpha \in (1/\log(1/q)+\epsilon, 1/\log(1/p)-\epsilon)$; in particular, for F_n and H_n we need expansions at the left and right side, respectively, of this range. This, it turns out, requires a novel approach and analysis, as discussed in detail in our forthcoming journal paper [5], leading to the announced results on the Rényi problem in (2) and (3).

Having described most of our main results, we mention an important equivalence pointed out by Pittel and Rubin [19]. They observed that their version of the Rényi process resembles the construction of a digital tree known as a PATRICIA trie[1] [12, 23]. In fact, the authors of [19] show that H_n is probabilistically equivalent to the height (longest path) of a PATRICIA trie built from n binary sequences generated independently by a memoryless source with bias $p = 1/2$ (that is, with a "1" generated with probability p; this is often called the *Bernoulli model with bias* p); the equivalence is true more generally, for $p \ge 1/2$. It is easy to see that F_n is equivalent to the fillup level (depth of the deepest full level), D_n to the typical

[1] We recall that a PATRICIA trie is a trie in which non-branching paths are *compressed*; that is, there are no unary paths.

depth (depth of a randomly chosen leaf), and $B_{n,k}$ to the external profile of the tree (the number of leaves at level k; the internal profile at level k is similarly defined as the number of non-leaf nodes at that level). We spell out this equivalence in the following simple claim.

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

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

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

Clearly, this scheme induces the correct distribution on queries. Furthermore, the resulting partition refinement tree (ignoring inconclusive queries) is easily seen to be isomorphic to T. Since the trees are isomorphic, the parameters of interest are equal in each case. □

Thus, our results on these parameters for the Rényi problem directly lead to novel results on PATRICIA tries, and vice versa. In addition to their use as data structures, PATRICIA tries also arise as combinatorial structures which capture the behavior of various processes of interest in computer science and information theory (e.g., in leader election processes without trivial splits [9] and in the solution to Rényi's problem which we study here [19, 2]).

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

Now we briefly review known facts about PATRICIA tries and other digital trees when built over n independent strings generated by a memoryless source. Profiles of tries in both the asymmetric and symmetric cases were studied extensively in [16]. The expected profiles of digital search trees in both cases were analyzed in [6], and the variance for the asymmetric case was treated in [10]. Some aspects of trie and PATRICIA trie profiles (in particular, the concentration of their distributions) were studied using probabilistic methods in [4, 3]. The depth in PATRICIA for the symmetric model was analyzed in [2, 12] while for the asymmetric model in [22]. The leading asymptotics for the PATRICIA height for the symmetric Bernoulli model was first analyzed by Pittel [17] (see also [23] for suffix trees). The two-term expression for the height of PATRICIA for the symmetric model was first presented in [19] as discussed above (see also [2]). Finally, in [13, 15], the second two authors of the present paper presented a precise analysis of the external profile (including its mean, variance, and limiting distribution) in the asymmetric case, for the range in which the profile grows polynomially. The present work relies on this

previous analysis, but the analyses for H_n and F_n involve a significant extension, since they rely on precise asymptotics for the external profile outside this central range.

Regarding methodology, the basic framework (which we use here) for analysis of digital tree recurrences by applying the Poisson transform to derive a functional equation, converting this to an algebraic equation using the Mellin transform, and then inverting using the saddle point method/singularity analysis followed by depoissonization, was worked out in [6] and followed in [16]. While this basic chain is common, the challenges of applying it vary dramatically between the different digital trees, and this is the case here. As we discuss later (see (7) and the surrounding text), this variation starts with the quite different forms of the Poisson functional equations, which lead to unique analytic challenges.

The plan for the paper is as follows. In the next section we formulate more precisely our problem and present our main results regarding the external profile, height, fillup level, and depth. Sketches of proofs are provided in the last section (the full proofs are provided in the journal version of this paper).

2 Main Results

In this section, we formulate precisely Rényi's problem and present our main results. Our goal is to provide precise asymptotics for three natural parameters of the Rényi problem on n objects with each label in a given query being included with probability $p \geq 1/2$: the number F_n of queries needed to identify at least one single element of the bijection, the number H_n needed to recover the bijection in its entirety, and the number D_n needed to recover an element of the bijection chosen uniformly at random from the n objects. If one wishes to determine the label for a particular object, these quantities correspond to the best, worst, and average case performance, respectively, of the random subset strategy proposed by Rényi. We call these parameters, the fillup level F_n, the height H_n, and the depth D_n, respectively (these names come from the corresponding quantities in random digital trees). One more parameter is relevant: we can present a unified analysis of our main three parameters $F_n, H_n,$ and D_n via the *external profile* $B_{n,k}$, which is the number of elements of the bijection on n items identified by the kth query.

Our analysis reveals several remarkable behaviors: the depth D_n converges in probability but not almost surely and while it satisfies the central limit theorem its local limit theorem doesn't hold. Perhaps most interestingly, the height H_n and the fillup level F_n exhibit phase transitions with respect to p in the second term.

To begin, we recall the relations of F_n, H_n, and D_n to $B_{n,k}$:

$$F_n = \min\{k : \ B_{n,k} > 0\} - 1 \qquad H_n = \max\{k : \ B_{n,k} > 0\} \qquad \Pr[D_n = k] = \mathbb{E}[B_{n,k}]/n.$$

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

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

and

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

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

In the next section, we show that the fillup level F_n and the height H_n have the following precise asymptotic expansions. Both exhibit a phase transition with respect to p in the second term. A complete proof can be found in our journal version of this paper [5].

Theorem 1 (Asymptotics for F_n and H_n)**.** *With high probability,*

$$H_n = \begin{cases} \log_{1/p} n + \frac{1}{2}\log_{p/q}\log n + o(\log\log n) & p > q \\ \log_2 n + \sqrt{2\log_2 n} + o(\sqrt{\log n}) & p = q \end{cases} \tag{4}$$

and

$$F_n = \begin{cases} \log_{1/q} n - \log_{1/q}\log\log n + o(\log\log\log n) & p > q \\ \log_2 n - \log_2\log n + o(\log\log n) & p = q \end{cases} \tag{5}$$

for large n.

While the behavior of the fillup level F_n could be anticipated [18] (by comparing it to the corresponding result in the version of Rényi's problem allowing inconclusive queries), the behavior of the height H_n is rather more unusual. It is difficult to compare the height result to the analogous quantity for tries or digital search trees, because only the first term is given for $p > 1/2$ in the literature: for tries, it is $\frac{2}{\log(1/(p^2+q^2))}\log n$, while for digital search trees it is $\log_{1/p} n$, as in PATRICIA tries.

Focusing on the second term of each expression given in the theorem, this result says that the deviation of the typical height from $\log_{1/p} n$ is asymptotically larger when $p = 1/2$ than when $p > 1/2$. That is, the height of the tallest fringe subtree (i.e., a subtree rooted near $\log_{1/p} n$) is asymptotically larger in the symmetric case. A complete explanation of this phenomenon would likely require consideration of the number of such subtrees (i.e., the internal profile at level $\log_{1/p} n$) and the number of strings participating in each of them. In the language of the Rényi problem, this latter parameter is the number of objects that remain unidentified after approximately $\log_{1/p} n$ queries.

Moving to the number of questions D_n needed to identify a random element of the bijection, we have the following theorem (note that due to the evolution process of the random PATRICIA trie, all random variables can be defined on the same probability space).

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

$$\liminf_{n\to\infty} D_n/\log n = 1/\log(1/q) \quad \textit{(a.s)} \quad \limsup_{n\to\infty} D_n/\log n = 1/\log(1/p).$$

Furthermore, D_n satisfies a central limit theorem; that is, $(D_n - \mathbb{E}[D_n])/\sqrt{\mathrm{Var}[D_n]} \to \mathcal{N}(0,1)$, where $\mathbb{E}[D_n] \sim \frac{1}{h(p)}\log n$ and $\mathrm{Var}[D_n] \sim c\log n$ where c is an explicit constant. A local *limit theorem does not hold: for $x = O(1)$ and $k = \frac{1}{h}(\log n + x\sqrt{\kappa_*(-1)\log n/h})$, where $\kappa_*(-1)$ is some explicit constant and $h = h(p)$, we obtain*

$$\Pr[D_n = k] \sim H\left(-1; \log_{p/q} p^k n\right) \frac{e^{-x^2/2}}{\sqrt{2\pi C\log n}}$$

for an oscillating function $H(-1; \log_{p/q} p^k n)$ (see Figure 1) defined in Theorem 3 below and an explicitly known constant C.

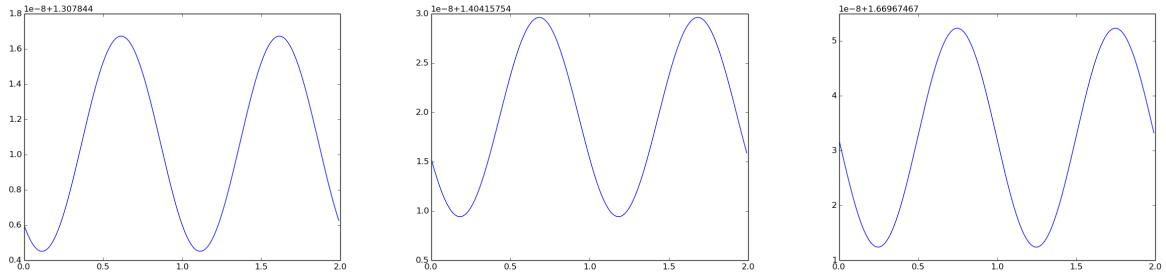


Fig. 1: Plots of $H(\rho, x)$ for $\rho = -0.5, 0, 0.5$.

Again, the depth exhibits a phase transition: for $p = 1/2$ we have $D_n/\log n \to 1/\log 2$ almost surely, which doesn't hold for $p > 1/2$. We note that some of the results on the depth (namely, the convergence in probability and the central limit theorem) are already known (see [20]), but our contribution is a novel derivation of these facts via the profile analysis. Qualitatively, the oscillatory behavior of the external profile that is responsible for the lack of local limit theorem for the depth occurs also in both tries and digital search trees.

We now explain our approach to the analysis of the moments of $B_{n,k}$ in appropriate ranges (we follow [13, 15]). For this, we take an analytic approach [8, 23]. We first explain it for the analysis relevant to D_n, and then show how to extend it for H_n and F_n. More details can be found in the next section.

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

$$\mu_{n,k} = (p^n + q^n)\mu_{n,k} + \sum_{j=1}^{n-1} \binom{n}{j} p^j q^{n-j} (\mu_{j,k-1} + \mu_{n-j,k-1}) \tag{6}$$

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

The tools that we use to solve this recurrence (for details see [13, 15]) are similar to those of the analyses for digital trees [23] such as tries and digital search trees (though the analytical details differ significantly). We first derive a functional equation for the Poisson transform $\tilde{G}_k(z) = \sum_{m\geq 0} \mu_{m,k} \frac{z^m}{m!} e^{-z}$ of $\mu_{n,k}$, which gives

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

This we write as

$$\tilde{G}_k(z) = \tilde{G}_{k-1}(pz) + \tilde{G}_{k-1}(qz) + \tilde{W}_{k,G}(z), \tag{7}$$

We contrast this functional equation with those for tries [16] and for digital search trees [6]: in tries, the expression $\tilde{W}_{k,G}(z)$ does not appear, which significantly simplifies the analysis in that case. In digital search trees, the functional equation is a differential equation, and the analysis is consequently quite different.

At this point the goal is to determine asymptotics for $\tilde{G}_k(z)$ as $z \to \infty$ in a cone around the positive real axis. When solving (7), $\tilde{W}_{k,G}(z)$ complicates the analysis because it has no closed-form Mellin transform (see below); we handle it via its Taylor series. Finally, depoissonization [23] will allow us to transfer the asymptotic expansion for $\tilde{G}_k(z)$ back to one for $\mu_{n,k}$:

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

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

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

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

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

where $A_k(s)$ (see (14) below) is an infinite series arising from the contributions coming from the function $\tilde{W}_{k,G}(z)$:

$$A_k(s) = \sum_{j=0}^{k} T(s)^{-j} \sum_{m=j}^{\infty} T(-m)(\mu_{m,j} - \mu_{m,j-1})\frac{\Gamma(m+s)}{\Gamma(m+1)\Gamma(s+1)}, \tag{8}$$

where we define $\mu_{m,-1} = 0$ for all m. Note that it involves $\mu_{m,j} - \mu_{m,j-1}$ for various m and j (see [13, 14]). Locating and characterizing the singularities of $G_k^*(s)$ then becomes important. We find that, for any k, $A_k(s)$ is entire, with zeros at $s \in \mathbb{Z} \cap [-k, -1]$, so that $G_k^*(s)$ is meromorphic, with possible simple poles at the negative integers less than $-k$. The fundamental strip of $\tilde{G}_k(z)$ then contains $(-k-1, \infty)$. It turns out that the main asymptotic contribution comes from an infinite number of saddle points (see (10) below) defined by the kernel $T(s) = p^{-s} + q^{-s}$.

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

$$\tilde{G}_k(z) = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} z^{-s} G_k^*(s)\, \mathrm{d}s = \frac{1}{2\pi i}\int_{\rho-i\infty}^{\rho+i\infty} z^{-s}\Gamma(s+1)A_k(s)T(s)^k\, \mathrm{d}s, \tag{9}$$

where ρ is any real number inside the fundamental strip associated with $\tilde{G}_k(z)$. For k in the range in which the profile grows polynomially (that coincides with the range of interest in our analysis of D_n), we

evaluate this integral via the saddle point method [8]. Examining $z^{-s}T(s)^k$ and solving the associated saddle point equation

$$\frac{\mathrm{d}}{\mathrm{d}s}[k\log T(s) - s\log z] = 0,$$

we find an explicit formula (12) below for $\rho(\alpha)$, the real-valued saddle point of our integrand. The multivaluedness of the complex logarithm then implies that there are *infinitely many* regularly spaced saddle points s_j, $j \in \mathbb{Z}$, on this vertical line:

$$s_j = \rho(\alpha) + i\frac{2\pi j}{\log(p/q)}. \tag{10}$$

These lead directly to oscillations in the $\Theta(1)$ factor in the final asymptotics for $\mu_{n,k}$). The main challenge in completing the saddle point analysis is then to elucidate the behavior of $\Gamma(s+1)A_k(s)$ for $s \to \infty$ along vertical lines: it turns out that this function inherits the exponential decay of $\Gamma(s+1)$ along vertical lines, and we prove it by splitting the sum defining $A_k(s)$ into two pieces, which decay exponentially for different reasons (the first sum decays as a result of the superexponential decay of $\mu_{m,j}$ for $m = \Theta(j)$, which is outside the main range of interest). We end up with an asymptotic expansion for $\tilde{G}_k(z)$ as $z \to \infty$ in terms of $A_k(s)$.

Finally, we must analyze the convergence properties of $A_k(s)$ as $k \to \infty$. We find that it converges uniformly on compact sets to a function $A(s)$ (see (14)) which is, because of the uniformity, entire. We then apply Lebesgue's dominated convergence theorem to conclude that we can replace $A_k(s)$ with $A(s)$ in the final asymptotic expansion of $\tilde{G}_k(z)$. All of this yields the following theorem which is proved in [13, 15].

Theorem 3 (Moments and limiting distribution for $B_{n,k}$ for k in the central region)**.** *Let $\epsilon > 0$ be independent of n and k, and fix $\alpha \in \left(\frac{1}{\log(1/q)} + \epsilon, \frac{1}{\log(1/p)} - \epsilon\right)$. Then for $k = k_{\alpha,n} \sim \alpha \log n$:*
(i) *The expected external profile becomes*

$$\mathbb{E}[B_{n,k}] = H(\rho(\alpha), \log_{p/q}(p^k n)) \cdot \frac{n^{\beta(\alpha)}}{\sqrt{2\pi\kappa_*(\rho(\alpha))\alpha\log n}} \left(1 + O(\sqrt{\log n})\right), \tag{11}$$

where

$$\rho(\alpha) = -\frac{1}{\log(p/q)} \log\left(\frac{\alpha\log(1/q) - 1}{1 - \alpha\log(1/p)}\right), \qquad \beta(\alpha) = \alpha\log(T(\rho(\alpha))) - \rho(\alpha), \tag{12}$$

and $\kappa_*(\rho)$ is an explicitly known function of ρ. Furthermore, $H(\rho, x)$ (see Figure 1) is a non-zero periodic function with period 1 in x given by

$$H(\rho, x) = \sum_{j\in\mathbb{Z}} A(\rho + it_j)\Gamma(\rho + 1 + it_j)e^{-2j\pi i x}, \tag{13}$$

where $t_j = 2\pi j/\log(p/q)$, and

$$A(s) = \sum_{j=0}^{\infty} T(s)^{-j} \sum_{n=j}^{\infty} T(-n)(\mu_{n,j} - \mu_{n,j-1})\frac{\phi_n(s)}{n!}, \tag{14}$$

where $\phi_n(s) = \prod_{j=1}^{n-1}(s+j)$ *for* $n > 1$ *and* $\phi_n(s) = 1$ *for* $n \le 1$. *We recall that* $T(s) = p^{-s} + q^{-s}$. *Here,* $A(s)$ *is an entire function which is zero at the negative integers.*
(ii) *The variance of the profile is* $\mathrm{Var}[B_{n,k}] = \Theta(\mathbb{E}[B_{n,k}])$.
(iii) *The limiting distribution of the normalized profile is Gaussian; that is,*

$$\frac{B_{n,k} - \mu_{n,k}}{\sqrt{\mathrm{Var}[B_{n,k}]}} \xrightarrow{D} \mathcal{N}(0,1)$$

where $\mathcal{N}(0,1)$ *is the standard normal distribution.*

We should point out that the unusual behavior of D_n in Rényi's problem is a direct consequence of the oscillatory behavior of the profile, which disappears for the symmetric case. Furthermore, for the height and fillup level analyses we need to extend Theorem 3 beyond its original central range for α, as discussed in the next section.

3 Proof sketches

Now we give sketches of the proofs of Theorems 1 and 2 with more details regarding the proof of Theorem 1 in the forthcoming journal version [5]. In particular, in this conference version, we only sketch derivations for H_n and for F_n by upper and lower bounding, respectively. As stated earlier, the proof of Theorem 3 can be found in [13, 15].

3.1 Sketch of the proof of Theorem 1

To prove our results for H_n and F_n, we extend the analysis of $B_{n,k}$ to the boundaries of the central region (i.e., $k \sim \log_{1/p} n$ and $k \sim \log_{1/q} n$).

Derivation of H_n. Fixing any $\epsilon > 0$, we write, for the lower bound on the height,

$$k_L = \log_{1/p} n + (1-\epsilon)\psi(n)$$

and, for the upper bound,

$$k_U = \log_{1/p} n + (1+\epsilon)\psi(n),$$

for a function $\psi(n) = o(\log n)$ which we are to determine. In order for the first and second moment methods to work, we require $\mu_{n,k_L} \xrightarrow{n\to\infty} \infty$ and $\mu_{n,k_U} \xrightarrow{n\to\infty} 0$. (We additionally need that $\mathrm{Var}[B_{n,k_L}] = o(\mu^2_{n,k_L})$, but this is not too hard to show by induction using the recurrence for $\tilde{V}_k(z)$, the Poisson variance of $B_{n,k}$.) In order to identify the $\psi(n)$ at which this transition occurs, we define $k = \log_{1/p} n + \psi(n)$, and the plan is to estimate $\mathbb{E}[B_{n,k}]$ via the integral representation (9) for its Poisson transform. Specifically, we consider the inverse Mellin integrand for some $s = \rho \in \mathbb{Z}^- + 1/2$ to be set later. This is sufficient for the upper bound, since, by the exponential decay of the Γ function, the entire integral is at most of the same order of growth as the integrand on the real axis. We expand the integrand in (9), that is,

$$J_k(n,s) := \sum_{j=0}^{k} n^{-s} T(s)^{k-j} \sum_{m \ge j} T(-m)(\mu_{m,j} - \mu_{m,j-1}) \frac{\Gamma(m+s)}{\Gamma(m+1)}, \tag{15}$$

and apply a simple extension of Theorem 2.2, part (iii) of [14] to approximate $\mu_{m,j} - \mu_{m,j-1}$ when $j \to \infty$ and is close enough to m:

Lemma 2 (Precise asymptotics for $\mu_{n,k}$, $k \to \infty$ and n near k)**.** *Let $p \geq q$. For $n \to \infty$ with $1 \leq k < n$ and $\log^2(n-k) = o(k)$,*

$$\mu_{n,k} \sim (n-k)^{3/2+\frac{\log q}{\log p}} \frac{n!}{(n-k)!} p^{k^2/2+k/2} q^k \cdot \exp\left(-\frac{\log^2(n-k)}{2\log(1/p)}\right) \Theta(1). \tag{16}$$

Moreover, for $n \to \infty$ and $k < n$, for some constant $C > 0$,

$$\mu_{n,k} \leq C \frac{n!}{(n-k-1)!} p^{k^2/2+k/2+O(log(n-k)^2)} q^k.$$

Now, we continue with the evaluation of (15). The jth term of (15) is then of order $p^{\nu_j(n,s)}$, where we set

$$\begin{aligned}\nu_j(n,s) = \; & (j-\psi(n))^2/2 + (j-\psi(n))(s + \log_{1/p}(1+(p/q)^s) + \psi(n) + 1) \\ & - \log_{1/p} n \log_{1/p}(1+(p/q)^s) + \psi(n)^2/2 + o(\psi(n)^2).\end{aligned}$$

The factor $T(s)^{k-j}$ ensures that the bounded j terms are negligible.

Our next goal is to find the j which gives the dominant contribution to the sum in (15); that is, the j for which the contributions $p^{\nu_j(n,s)}$ dominate. By elementary calculus, we can find the j term which minimizes $\nu_j(n,s)$:

$$j = -(s + \log_{1/p}(1+(p/q)^s) + 1).$$

Then $\nu_j(n,s)$ for this value of j becomes

$$\begin{aligned}\nu_j(n,s) = & -\frac{(s + \log_{1/p}(1+(p/q)^s) + \psi(n) + 1)^2}{2} \\ & - \log_{1/p} n \log_{1/p}(1+(p/q)^s) + \psi(n)^2/2 + o(\psi(n)^2).\end{aligned} \tag{17}$$

We then minimize over all s, which requires us to split into the symmetric and asymmetric cases.

Symmetric case: When $p = q = 1/2$, we have $\log_{1/p}(1+(p/q)^s) = \log_2(2) = 1$, so that the expression for $\nu_j(n,s)$ simplifies, and we get $s = -\psi(n) + O(1)$. The optimal value for $\nu_j(n,s)$ then becomes

$$\nu_j(n,s) = -\log_2 n + \psi(n)^2/2 + o(\psi(n)^2). \tag{18}$$

We have thus succeeded in finding a likely candidate for the range of j terms that contribute maximally, as well as an upper bound on their contribution. This gives a tight upper bound on $J_k(n,s)$ and, hence, on $\tilde{G}_k(n)$, of $\Theta(2^{-\nu_j(n,s)})$.

Now, to find $\psi(n)$ for which there is a phase transition in this bound from tending to ∞ to tending to 0, we set the exponent in the above expression equal to zero and solve for $\psi(n)$. This gives

$$-\log_2 n + \psi(n)^2/2(1+o(1)) = 0 \implies \psi(n) \sim \sqrt{2\log_2 n},$$

as expected.

Asymmetric case: On the other hand, when $p > 1/2$, the equation that we need to solve to find the minimizing value of s for (17) is a bit more complicated, owing to the fact that $\log_{1/p}(1 + (p/q)^s)$ now depends on s: taking a derivative with respect to s in (17) and setting this equal to 0, after some algebra, we must solve

$$-\frac{(p/q)^s \log(p/q)}{\log(1/p)} \log_{1/p} n - \psi(n)(1 + O((p/q)^s)) - s(1 + O((p/q)^s)) + O((p/q)^s) = 0 \tag{19}$$

for s. Here, we note that we used the approximation

$$\log_{1/p}(1 + (p/q)^s) = \frac{(p/q)^s}{\log(1/p)} + O((p/q)^2),$$

which is valid since we are looking for $s \to -\infty$.

To find a solution to (19), we first note that it implies that $s < -\psi(n)$ (since the first term involving $\log n$ is negative), and, if $\psi(n) > 0$, this implies that

$$-\psi(n) - s = -O(s). \tag{20}$$

The plan, then, is to use this to guess a solution s for (19), which we can then verify. The equality (20) suggests that we replace $-\psi(n) - s + O((p/q)^s)$ with $-C \cdot s$ in (19), for some constant $C > 0$. Then the equation becomes

$$-Cs - \frac{(p/q)^s \log(p/q)}{\log(1/p)} \log_{1/p} n = 0.$$

After some trivial rearrangement and multiplication of both sides by $\log(p/q)$, we get

$$-s \log(p/q) \cdot e^{-s \log(p/q)} = \Theta(\log n).$$

Setting $W = -s \log(p/q)$ brings us to an expression of the form that defines the Lambert W function [1] (i.e., a function $W(z)$ satisfying $W(z)e^{W(z)} = z$).

Using the asymptotics of the W function for large z [1], we thus find that

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

Note that $s \to -\infty$, as required. This may be plugged into (17) to see that it is indeed a solution to the equation.

Now, to find the correct choice of $\psi(n)$ for which there is a phase transition, we plug this choice of s into (17), set it equal to 0, and solve for $\psi(n)$. This gives

$$\psi(n) = -\frac{s}{2} = \frac{1}{2} \log_{p/q} \log n + O(\log \log \log n), \tag{21}$$

as desired.

Note that replacing $\psi(n)$ in (17) with $(1 + \epsilon)\psi(n)$ yields a maximum contribution to the inverse Mellin integral of

$$J_{k_U}(n, s) = O(p^{\frac{\epsilon}{2}(\log_{p/q} \log n)^2 + o((\log \log n)^2)}) \to 0. \tag{22}$$

When we replace $\psi(n)$ with $(1-\epsilon)\psi(n)$, we get

$$J_{k_L}(n,s) = O(p^{-\frac{\epsilon}{2}(\log_{p/q}\log n)^2 + o((\log\log n)^2)}), \tag{23}$$

so that the upper bound tends to infinity (in [5], we prove a matching lower bound).

The above analysis gives asymptotic estimates for $\tilde{G}_k(n)$. We then apply analytic depoissonization [23] to get

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

(where the second term can be handled in the same way as the first). This gives the claimed result.

Derivation of F_n. We now set $k = \log_{1/q} n + \psi(n)$ and

$$k_L = \log_{1/q} n + (1+\epsilon)\psi(n), \qquad k_U = \log_{1/q} n + (1-\epsilon)\psi(n). \tag{24}$$

Here, $\psi(n) = o(\log n)$ is to be determined so as to satisfy $\mu_{n,k_L} \to 0$ and $\mu_{n,k_U} \to \infty$. We use a technique similar to that used in the height proof to determine $\psi(n)$, except now the Γ function asymptotics play a role, since we will choose $\rho \in \mathbb{R}$ tending to ∞. Our first task is to upper bound (as tightly as possible), for each j, the magnitude of the jth term of (15). First, we upper bound

$$T(-m)(\mu_{m,j} - \mu_{m,j-1}) \le 2p^m \mu_{m,j} \le 2p^m m, \tag{25}$$

using the boundary conditions on $\mu_{m,j}$. Next, we apply Stirling's formula to get

$$\frac{\Gamma(m+\rho)}{\Gamma(m+1)} \sim \sqrt{1+\rho/m}\left(\frac{m+\rho}{e}\right)^{m+\rho}\left(\frac{m+1}{e}\right)^{-(m+1)} \tag{26}$$

$$= e^{(m+\rho)\log(m+\rho) - (m+\rho) + m + 1 - (m+1)\log(m+1) + O(\log\rho)} \tag{27}$$

$$= \exp((m+\rho)\log(m+\rho) - (m+1)\log(m+1) + O(\rho)) \tag{28}$$

$$= \exp(m\log(m(1+\rho/m)) + \rho\log(\rho(1+m/\rho)) - m\log m - \log m + O(\rho)) \tag{29}$$

$$= \exp(m\log(1+\rho/m) + \rho\log(\rho) + \rho\log(1+m/\rho) - \log m + O(\rho)). \tag{30}$$

Multiplying (25) and (30), then optimizing over all $m \ge j$, we find that the maximum term of the m sum occurs at $m = \rho p/q$ and has a value of

$$\exp(\rho\log\rho + O(\rho)). \tag{31}$$

Now, observe that when $\log m \gg \log\rho$, the contribution of the mth term is $p^{m+o(m)} = e^{-\Theta(m)}$. Thus, setting $j' = \rho^{\log\rho}$ (note that $\log j' = (\log\rho)^2 \gg \log\rho$), we split the m sum into two parts:

$$\sum_{m\ge j} 2p^m m\frac{\Gamma(m+\rho)}{\Gamma(m+1)} = \sum_{m=j}^{j'} 2p^m m\frac{\Gamma(m+\rho)}{\Gamma(m+1)} + \sum_{m=j'+1}^{\infty} 2p^m m\frac{\Gamma(m+\rho)}{\Gamma(m+1)}.$$

The terms of the initial part can be upper bounded by (31), while those of the final part are upper bounded by $e^{-\Theta(m)}$ (so that the final part is the tail of a geometric series). This gives an upper bound of

$$j'e^{\rho\log\rho + O(\rho)} + e^{-\Theta(j')} = e^{(\log\rho)^2 + \rho\log\rho + O(\rho)} = e^{\rho\log\rho + O(\rho)},$$

which holds for any j.

Multiplying this by $n^{-\rho}T(\rho)^{k-j} = q^{\rho\cdot(j-\psi(n))+(j-\psi(n)-\log_{1/q} n)\log_{1/q}(1+(q/p)^\rho)}$ gives

$$q^{\rho(j-\psi(n))+(j-\psi(n)-\log_{1/q} n)\log_{1/q}(1+(q/p)^\rho)-\rho\log_{1/q}\rho+O(\rho)}. \tag{32}$$

Maximizing over the j terms, we find that the largest contribution comes from $j = 0$. Then, just as in the height upper bound, the behavior with respect to ρ depends on whether or not $p = q$, because $\log_{1/q}(1 + (q/p)^\rho) = 1$ when $p = q$ and is dependent on ρ otherwise. Taking this into account and minimizing over ρ gives that the maximum contribution to the j sum is minimized by setting $\rho = 2^{-\psi(n)-\frac{1}{\log 2}}$ when $p = q$ and $\rho \sim \log_{p/q}\log n$ otherwise. Plugging these choices for ρ into the exponent of (32), setting it equal to 0, and solving for $\psi(n)$ gives $\psi(n) = -\log_2\log n + O(1)$ when $p = q$ and $\psi(n) \sim -\log_{1/q}\log\log n$ when $p > q$. The evaluation of the inverse Mellin integral with $k = k_L$ as defined in (24) and the integration contour given by $\Re(s) = \rho$ proceeds along lines similar to the height proof, and this yields the desired result.

We remark that the lower bound for F_n may also be derived by relating it to the analogous quantity in regular tries: by definition of the fillup level, there are no unary paths above the fillup level in a standard trie. Thus, when converting the corresponding PATRICIA trie, no path compression occurs above this level, which implies that F_n for PATRICIA is lower bounded by that of tries (and the typical value for tries is the same as in our theorem for PATRICIA). We include the lower bound for F_n via the bounding of the inverse Mellin integral because it is similar in flavor to the corresponding proof of the upper bound (for which no short proof seems to exist).

The upper bound for F_n can similarly be handled by an exact evaluation of the inverse Mellin transform.

3.2 Proof of Theorem 2

Using Theorem 3, we can prove Theorem 2.

Convergence in probability: For the typical value of D_n, we show that

$$\Pr[D_n < (1-\epsilon)\frac{1}{h(p)}\log n] \xrightarrow{n\to\infty} 0, \qquad \Pr[D_n > (1+\epsilon)\frac{1}{h(p)}\log n] \xrightarrow{n\to\infty} 0. \tag{33}$$

For the lower bound, we have

$$\Pr[D_n < (1-\epsilon)\frac{1}{h(p)}\log n] = \sum_{k=0}^{\lfloor(1-\epsilon)\frac{1}{h(p)}\log n\rfloor} \Pr[D_n = k] = \sum_{k=0}^{\lfloor(1-\epsilon)\frac{1}{h(p)}\log n\rfloor} \frac{\mu_{n,k}}{n}.$$

We know from Theorem 3 and the analysis of F_n that, in the range of this sum, $\mu_{n,k} = O(n^{1-\epsilon})$. Plugging this in, we get

$$\Pr[D_n < (1-\epsilon)\frac{1}{h(p)}\log n] = \sum_{k=0}^{\lfloor(1-\epsilon)\frac{1}{h(p)}\log n\rfloor} O(n^{-\epsilon}) = O(n^{-\epsilon}\log n) = o(1).$$

The proof for the upper bound is very similar, except that we appeal to the analysis of H_n instead of F_n.

No almost sure convergence: To show that $D_n/\log n$ does not converge almost surely, we show that

$$\liminf_{n\to\infty} D_n/\log n = 1/\log(1/q), \qquad \limsup_{n\to\infty} D_n/\log n = 1/\log(1/p). \tag{34}$$

For this, we first show that, almost surely, $F_n/\log n \xrightarrow{n\to\infty} 1/\log(1/q)$ and $H_n/\log n \xrightarrow{n\to\infty} 1/\log(1/p)$. Knowing this, we consider the following sequences of events: A_n is the event that $D_n = F_n + 1$, and A'_n is the event that $D_n = H_n$. We note that all elements of the sequences are independent, and $\Pr[A_n], \Pr[A'_n] \geq 1/n$. This implies that $\sum_{n=1}^{\infty} \Pr[A_n] = \sum_{n=1}^{\infty} \Pr[A'_n] = \infty$, so that the Borel-Cantelli lemma tells us that both A_n and A'_n occur infinitely often almost surely (moreover, $F_n < D_n \leq H_n$ by definition of the relevant quantities). This proves (34).

To show the claimed almost sure convergence of $F_n/\log n$ and $H_n/\log n$, we cannot apply the Borel-Cantelli lemmas directly, because the relevant sums do not converge. Instead, we apply a trick which was used in [17]. We observe that both (F_n) and (H_n) are non-decreasing sequences. Next, we show that, on some appropriately chosen subsequence, both of these sequences, when divided by $\log n$, converge almost surely to their respective limits. Combining this with the observed monotonicity yields the claimed almost sure convergence, and, hence, the equalities in (34).

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

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

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

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

which implies

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

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

Asymptotics for probability mass function of D_n: The asymptotic formula for $\Pr[D_n = k]$ with k as in the theorem follows directly from the fact that $\Pr[D_n = k] = \mathbb{E}[B_{n,k}]/n$, plugging in the expression of Theorem 3 for $\mathbb{E}[B_{n,k}]$.

References

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

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

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

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

[5] Michael Drmota, Abram Magner, and Wojciech Szpankowski. Asymmetric Rényi problem and PATRICIA tries. In preparation.

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

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

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

[9] Svante Janson and Wojciech Szpankowski. Analysis of an asymmetric leader election algorithm. *Electronic J. Combin*, 4:1–6, 1996.

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

[11] Philippe Jacquet, Charles Knessl, and Wojciech Szpankowski. A note on a problem posed by D. E. Knuth on a satisfiability recurrence. *Combinatorics, Probability, and Computing*, 23, 839-841, 2014.

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

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

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

[15] Abram Magner and Wojciech Szpankowski. Profile of PATRICIA tries. submitted. `http://www.cs.purdue.edu/~spa/papers/patricia2015.pdf`

[16] G. Park, H. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. *SIAM Journal on Computing*, 38(5):1821–1880, 2009.

[17] B. Pittel. Asymptotic growth of a class of random trees. *Ann. Probab.*, 18:414–427, 1985.

[18] B. Pittel, Paths in a random digital tree: limiting distributions, *Adv. in Applied Probability*, 18, 139–155, 1986.

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

[20] B. Rais, P. Jacquet, and W. Szpankowski. Limiting distribution for the depth in PATRICIA tries. *SIAM Journal on Discrete Mathematics*, 6(2):197–213, 1993.

[21] A. Rényi, On Random Subsets of a Finite Set, *Mathematica*, 3, 355-362, 1961.

[22] Wojciech Szpankowski. Patricia tries again revisited. *J. ACM*, 37(4):691–711, October 1990.

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

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Solutions of First Order Linear Partial Differential Equations Related to Urn Models and Central Limit Theorems

Michael Drmota[1]†, Mehri Javanian

[1]*TU Wien, Institute of Discrete Mathematics and Geometry, Wiedner Hauptstrasse 8–10, A-1040 Wien, Austria*
[2]*University of Zanjan, Department of Statistics, University Blvd., 45371-38791, Zanjan, Iran*

Abstract. We study first order linear partial differential equations that appear, for example, in the analysis of dimishing urn models with the help of the method of characteristics and formulate sufficient conditions for a central limit theorem.

Keywords: urn models, first order partial differential equations, central limit theorem, singularity analysis

1 Introduction and Main result

The purpose of this paper is to study solutions $H(z,w)$ of special first order linear partial differential equations that appear in the analysis of dimishing urn models. In particular we follow the work of Kuba and Panholzer (2007).

More precisely, we consider a Pólya-Eggenberger urn model with two kinds of balls and transition matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. The process runs as follows. Suppose that the urn contains m balls of the first kind and n balls of the second kind - we can interprete this state as the point (m,n) on the integer lattice. Then with probability $m/(n+m)$ we add a balls of the first kind and b balls of the second kind, whereas with probability $n/(n+m)$ we add c balls of the first kind and d balls of the second kind. (Of course, adding a negative number of balls means taking away this number of balls.) An absorbing state $\mathcal{S}$ is a subset $\mathcal{S} \subset \mathbb{N} \times \mathbb{N}$, where the process stops when we arrive in $\mathcal{S}$. In what follows we will only consider (special) dimishing urn-models, where the number of balls of the first kind eventually reaches zero, so that the y-axis $\mathcal{S} = \{(0,n) : n \geq 0\}$ is a natural absorbing state.

Suppose now that the process starts at $(m,n) \in \mathbb{N} \times \mathbb{N}$ with $m \geq 1$ and let $h_{n,m}(v) = \mathbb{E}[v^{X_{n,m}}]$ denote the probability generating function of the random variable $X_{n,m}$ that describes the position $(0,n_0)$ of the absorbing state in $\mathcal{S}$ when the process starts at (m,n).

†Partially supported by the Austrian Science Fund FWF, Project SFB F50-02

By definition the probability generating functions $h_{n,m}(v)$ satisfy the recurrence

$$h_{n,m}(v) = \frac{n}{n+m} h_{n+a,m+c}(v) + \frac{m}{n+m} h_{n+c,m+d}(v). \tag{1}$$

for $(m,n) \notin \mathcal{S}$. The boundary values at an absorbing state $(m,n) \in \mathcal{S}$ is $h_{n,m}(v) = v^n$.

By setting

$$\overline{H}(z,w;v) = \sum_{n\geq 0, m\geq 1} h_{n,m}(v) z^n w^m$$

it follows that this generating function $\overline{H}(z,w;v)$ satisfies the partial differential equation

$$z(1-z^{-a}w^{-b})\overline{H}_z + w(1-z^{-c}w^{-d})\overline{H}_w + (az^{-a}w^{-b} + dz^{-c}w^{-d})\overline{H} = F(z,w) \tag{2}$$

with some inhomogeneous part $F(z,w)$ that is given by the boundary values which are partly unknown (for example $\overline{H}(0,w,v)$, see Kuba and Panholzer (2007).

We want to mention that first order linear partial differential equations related to urn models were first systematically discussed by Flajolet et al. (2005), see also Morcrette (2012), where a special case is detailly treated. On the other hand, it is possible to describe the probabilistic behavior of the development of urn models very precisely, sse Janson (2004, 2006), even with absorbing states. Nevertheless the analysis of dimishing urns with the y-axis as the absorbing state is still quite special. Here we also refer to Kuba (2011); Kuba and Panholzer (2012), where the analysis is based directly on the recurrence (1). Another interesting paper that is related to dimishing urn models and lines as absorbing states is Kuba et al. (2009). There the authors observe several different kinds of limiting behaviors (with five phase changes).

It turns out that there are some special cases, where it is more convenient to study the generating function

$$H(z,w;v) = \sum_{n\geq 0, m\geq 1} \binom{n+m}{m} h_{n,m}(v) z^n w^m \tag{3}$$

that (also) satisfies a first order linear partial differential equation of the form

$$A(z,w)H_z + B(z,w)H_w - C(z,w)H = D(z,w;v), \tag{4}$$

with analytic functions $A(z,w), B(z,w), C(z,w), D(z,w;v)$. (In the examples below $A(z,w), B(z,w)$, and $C(z,w)$ are polynomials.) For these particular cases it turns out that the unknown boundary conditions are not needed since they cancel in the equation. Nevertheless the methods that we are developing below are – although we do not work out the general case – suitable to deal with equations of the form (2).

Note that by definition

$$H(z,0;v) = 0. \tag{5}$$

Furthermore, if $v = 1$ then $h_{n,m}(1) = 1$ so that

$$H(z,w;1) = \frac{1}{1-z-w} - \frac{1}{1-z}.$$

This means that $D(z,w;1)$ is determined by

$$D(z,w;1) = \frac{A(z,w) + B(z,w) - (1-z-w)C(z,w)}{(1-z-w)^2} - \frac{A(z,w) - (1-z)C(z,w)}{(1-z)^2}.$$

In the present context it is convenient to assume that the function

$$H_0(z,w) = \sum_{n,m\geq 0} \binom{n+m}{m} z^n w^m = 1/(1-z-w)$$

is a solution of the homogeneous differential equation $A(z,w)H_z + B(z,w)H_w - C(z,w)H = 0$ so that

$$A(z,w) + B(z,w) = (1-z-w)C(z,w) \tag{6}$$

and, thus,

$$D(z,w;1) = -\frac{A(z,w) - (1-z)C(z,w)}{(1-z)^2} \tag{7}$$

We first state the following three examples from Kuba and Panholzer (2007) (that we present in a slightly modified way).

Example 1 The pill's problem (see Brennan and Prodinger (2003); Knuth and Mccarthy (1991)) has transition matrix $M = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix}$ and absorbing state $\mathcal{S} = \{(0,n) : n \geq 0\}$, and the corresponding differential equation is given by

$$(z - z^2 - w)H_z + w(1-z)H_w - zH = \frac{wv}{(1-vz)^2}.$$

Here it follows that $h_{n,m}(v)$ is given by

$$h_{n,m}(v) = mv \int_0^1 (1 + (v-1)q)^n (1 - q - (v-1)q\log q)^{m-1} dq.$$

Finally the corresponding random variable $X_{n,m}$ has limiting distribution

$$\frac{X_{n,m}}{\frac{n}{m} + \log m} \to X \qquad (m \to \infty),$$

where X has density e^{-x}, $x \geq 0$, or

$$\frac{X_{n,m}}{n} \to \mathrm{Beta}(1,m) \qquad (\text{fixed } m \geq 1, n \to \infty),$$

where (the beta distribution) $\mathrm{Beta}(1,m)$ has density $m(1-x)^{m-1}$, $0 \leq x \leq 1$.

Example 2 A variant of the pill's problem has transition matrix $M = \begin{pmatrix} -1 & 0 \\ 1 & -2 \end{pmatrix}$ and absorbing state $\mathcal{S} = \{(0,n) : n \geq 0\} \cup \{(1,n) : n \geq 0\}$. Due to the parity condition in m (that is, only even m occur), it is convenient to consider the generating function

$$H(z,w;v) = \sum_{n\geq 0, m\geq 1} \binom{n+2m}{n} h_{n,2m}(v) z^n w^m$$

that satisfies the differential equation

$$-wH_z + 2w(1-z)H_w - (1-z)H = \frac{wv}{(1-vz)^2}.$$

Here we obtain an explicit solution of the form

$$\begin{aligned} H(z,w) &= \frac{w}{v\left((1-z)^2 - w - ((v-1)/v))^2\right)(1-z-(v-1/v))} \\ &+ \frac{(v-1)\sqrt{w}}{v^2\left((1-z)^2 - w - ((v-1)/v))^2\right)^{3/2}} \arctan\left(\frac{\sqrt{w}\sqrt{(1-z)^2 - w - ((v-1)/v))^2}}{(1-z)^2 - w - (1-z)(v-1)/v)}\right). \end{aligned}$$

which leads to the limiting behavior:

$$\frac{X_{n,2m}}{\frac{n}{\sqrt{m}} + 2\sqrt{m}} \to R \qquad (m \to \infty),$$

where R has density $2xe^{-x^2}$, $x \geq 0$, or

$$\frac{X_{n,2m}}{n} \to \sqrt{\mathrm{Beta}(1,m)}, \qquad (m \geq 1 \text{ fixed}, n \to \infty).$$

Example 3 The cannibal urn (see Pittel (1987); Kuba (2011)) has transition matrix $M = \begin{pmatrix} 0 & -1 \\ 1 & -2 \end{pmatrix}$ and absorbing state $\mathcal{S} = \{(0,n) : n \geq 0\} \cup \{(1,n) : n \geq 0\}$ and the generating function

$$H(z,w;v) = \sum_{n\geq 0, m\geq 1} \binom{n+m}{n} h_{n+1,m}(v)$$

satisfies the differential equation

$$-(z+w)H_z + H_w - H = \frac{(1+wv)v}{(1-vz)^2}.$$

The solution is explicitly given by

$$H(z,w;v) = \frac{ve^w}{1-(1-e^w(1-z-w))v} - \frac{v}{1-vz}$$

and we have a central limit theorem of the form

$$\frac{X_{n,m} - \mathbb{E}\, X_{n,m}}{\sqrt{\mathbb{V}\mathrm{ar} X_{n,m}}} \to N(0,1) \qquad (m+n \to \infty).$$

These three examples show that although the linear differential equations look very similar the limiting behavior of the encoded random variable $X_{n,m}$ seems to be far from being universal. The main purpose of the present paper is to shed some light on this phenomenon. In particular we detect a sufficient condition that ensures a central limit theorem.

Theorem 1 *Suppose that $X_{n,m}$, $n \geq 0$, $m \geq 1$ are non-negative discrete random variables with probabilty generating function $h_{n,m}(v) = \mathbb{E}[v^{X_{n,m}}]$ such that the generating function $H(z, w; v)$, given by (3) satisfies a first order linear differential equation of the form (4), where the coefficient functions A, B, C as well as the ratios $A(z,w)/B(z,w)$, $C(z,w)/B(z,w)$ are analytic in an open set that contains z, w with $|z| + |w| \leq 1$ such that the ratio $A(z,w)/B(z,w)$ is negative for non-negative z, w. Furthermore we assume that (6) is satisfied (which also implies (7)) and that $D(z, w; v)$ can be represented as*

$$D(z, w; v) = \frac{a(z, w; v)}{(1 - b(z, w; v))^2},$$

where the functions a, b are also in an open set that contains z, w with $|z| + |w| \leq 1$. In particular in accordance with (6) we have $a(z, w; 1) = -A(z, w) + (1 - z)C(z, w)$ and $b(z, w; 1) = z$.

Let $f(c, s)$ be the solution of the differential equation $\frac{\partial f}{\partial s} = A(f, s)/B(f, s)$ with $f(c, 0) = c$ and let $Q(z, w)$ denote the function that satisfies $f(Q(z, w), w) = z$. We further assume that the function $f(Q(z, w), s)$ is analytic in an open set that contains s, z, w with $|z| + |w| \leq 1$ and $|z| + |s| \leq 1$ and non-decreasing for positive and real z and w,

Let $z_0(\rho; v)$ and $w_0(\rho; v)$ denote the solutions of the system of equations

$$b(f(Q(z, w), 0), 0; v) = 1, \qquad z\frac{\partial}{\partial z} b(f(Q(z, w), 0), 0; v) = \rho w \frac{\partial}{\partial w} b(f(Q(z, w), 0), 0; v)$$

with $z_0(\rho; 1) = \rho/(1 + \rho)$ and $w_0(\rho; 1) = 1/(1 + \rho)$. Furthermore set $h(\rho; v) = -\log z_0(\rho; v) - \rho \log w_0(\rho; v)$, $\mu(\rho) = \frac{\partial}{\partial v} h(\rho; v)\big|_{v=1}$ and $\sigma^2(\rho) = \frac{\partial^2}{\partial v^2} h(\rho; v)\Big|_{v=1} + \mu$. If

$$\mu(\rho) > 0 \qquad \text{for } \rho \in [\alpha, \beta]$$

for some positive α, β then $X_{n,m}$ satisfies a central limit theorem of the form

$$\frac{X_{n,m} - \mathbb{E}\, X_{n,m}}{\sqrt{n}} \to N(0, \sigma^2(m/n))$$

uniformly for $m + n \to \infty$, $m/n \in [\alpha, \beta]$, where

$$\mathbb{E}\, X_{n,m} \sim \mu(m/n)\, n \qquad \text{and} \quad \mathbb{V}\mathrm{ar} X_{n,m} \sim \sigma^2(m/n)\, n.$$

This theorem does not provide a full answer to the problem. However, it is a first step that covers at least a part, where we obtain a central limit theorem. In future work we will provide a more complete picture, also covering the cases, where there is no central limit theorem. For example it is not clear whether it is possible to formulate conditions that refer directly to the entries of the transition matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. In particular it is an open question whether it is possible to adapt Theorem 1 so that all cases of (Kuba (2011)) are covered.

Nevertheless, we will discuss the three examples (from above) and another one in the next section. We also present a (short version of the) proof of Theorem 1 in the remaining parts of the paper.

2 Discussion of Examples

We do not work out the details here but in **Examples 1** and **2** several conditions of Theorem 1 are not satisfied, in particular we have $\mu(\rho) = 0$.

The most interesting example is **Example 3**. Here we have $A(z,w) = -z - w$, $B(z,w) = C(z,w) = 1$, and $D(z,w;v) = (1+wv)v/(1-vz)^2$, that is $a(z,w;v) = (1+wv)v$ and $b(z,w;v) = vz$. It is easy to check that the conditions of Theorem 1 are satisfied.

In particular it follows that $f(c,s) = 1 - s - e^{-s}(1-c)$, $Q(z,w) = 1 - e^w(1-z-w)$, and $f(Q(z,w),s) = 1-s-e^{w-s}(1-z-w)$. From that we obtain $b(f(Q(z,w,0),0;v) = (1-e^w(1-z-w))v$. Hence the functions $z = z_0(\rho;v)$ and $w = w_0(\rho;v)$ satisfy the system of equations

$$(1 - e^w(1-z-w))v = 1, \qquad z = \rho w(z+w)$$

from which we obtain (by implicit differentiation)

$$\mu(\rho) = -\frac{z_{0,v}(\rho;1)}{z_0(\rho;1)} - \rho\frac{w_{0,v}(\rho;1)}{w_0(\rho;1)} = 2e^{-1/(1+\rho)} > 0.$$

Thus, the central limit theorem follows automatically.

We add a new example in order to demonstrate the applicabilty of Theorem 1 (even if this example is not related to an urn model). By the way this example can be easily generalized. Suppose that $H(z,w;z)$ satisfies the differential equation

$$-(z+2w)H_z + (1+w)H_w - H = \frac{(1+2w)v}{(1-vz)^2}.$$

Then again all assumptions of Theorem 1 are satisfied. Here we have $A(z,w) = -z - 2w$, $B(z,w) = 1 + w$, $C(z,w) = 1$, and $D(z,w;v) = (1+2w)v/(1-vz)^2$, that is $a(z,w;v) = (1+2w)v$ and $b(z,w;v) = vz$.

From this it follows that

$$f(c,s) = \frac{c}{1+s} - \frac{s^2}{1+s} \quad \text{and} \quad Q(z,w) = (1+w)z + w^2$$

and consequently

$$f(Q(z,w),s) = \frac{(1+w)z + w^2 - s^2}{1+s}.$$

The functions $z = z_0(\rho;v)$ and $w = w_0(\rho;v)$ satisfy the system of equations

$$((1+w)z + w^2)v = 1, \qquad z(1+w) = \rho w(z+2w)$$

from which we obtain (by implicit differentiation)

$$\mu(\rho) = -\frac{z_{0,v}(\rho;1)}{z_0(\rho;1)} - \rho\frac{w_{0,v}(\rho;1)}{w_0(\rho;1)} = 2\frac{(1+\rho)^2}{(2+\rho)^2} > 0.$$

Thus, the central limit theorem follows (again) automatically.

3 The method of characteristics

The first step of the proof is to use the theory of characteristics to provide an integral representation (13) of the solution of the partial differential equation (4).

We start with the inhomogeneous differential equation (4), where v is considered as a parameter. It is a standard procedure to transform (4) into a homogeneous equation. Let $Q = Q(z, w, H; v)$ denote the solution of the linear differential equation

$$A(z,w)Q_z + B(z,w)Q_w + (C(z,w)H + D(z,w;v))Q_H = 0. \tag{8}$$

Then the solution $H(z, w; v)$ of the original equation (4) satisfies the implicit equation

$$Q(z, w, H(z, w; v); v) = const. \tag{9}$$

Thus, if we can solve (8) then we also get the solution of (4). The advantage of the equation (8) is that it can be handled with the method of characteristics (see Hellwig (1977)).

First we translate (8) into a system of first order ordinary differential equations:

$$\frac{dz}{dt} = A(z,w), \quad \frac{dw}{dt} = B(z,w), \quad \frac{dH}{dt} = C(z,w)H + D(z,w;v), \tag{10}$$

where $z = z(t)$, $w = w(t)$, $H = H(t)$ are functions in t. A characteristic of (10) is a function $F(z, w, H)$ for which we have $Q(z(t), w(t), H(t)) = const.$ Clearly, every characteristic Q is a solution of (8). It is well known that a system of three equations has two independent characteristics Q_1, Q_2 as a basis and every characteristic Q can be expressed as $Q = F(Q_1, Q_2)$ for an arbitrary (differentiable) function F. In the present case we have to solve the equation (9) which simplifies the situation. More precisely we can rewrite (9) to an equation of the form

$$Q_2(z, w, H) = \tilde{F}(Q_1(z, w, H)), \tag{11}$$

where $\tilde{F}$ is an arbitrary (differentiable) function.

In order to calculate two independent characteristics it is convenient to *eliminate* t from the system (10) which gives rise to a simpler system of differential equation:

$$\frac{dz}{dw} = \frac{A(z,w)}{B(z,w)}, \quad \frac{dH}{dw} = \frac{C(z,w)}{B(z,w)}H + \frac{D(z,w;v)}{B(z,w)}, \tag{12}$$

where $z = z(w)$ and $H = H(w)$ are now considered as functions is w.

Let $z = f(c_1, w)$ be a one-parametric solution of the differential equation $\frac{dz}{dw} = \frac{A(z,w)}{B(z,w)}$, where c_1 is, for example, the initial value $c_1 = z(0)$. If we express c_1 from the expression $z = f(c_1, w)$, that is, $c_1 = Q_1(z, w)$ then Q_1 is a characteristic of the system (10). Note that Q_1 does not depend on H and also not on v. Actually Q_1 just solves the equation $A(z,w)Q_z + B(z,w)Q_w = 0$. Nevertheless it is a non-trivial characteristic of (10).

In order to obtain a second characteristic we have to solve the second equation of (12) which is a first order linear differential equation. Note that we can substitute $z = f(c_1, w)$ and obtain as a solution

$$H = \exp\left(\int_0^w \frac{C(f(c_1,s),s)}{B(f(c_1,s),s)}\,ds\right)\left(\int_0^w \frac{D(f(c_1,s),s;v)}{B(f(c_1,s),s)} \exp\left(-\int_0^s \frac{C(f(c_1,t),t)}{B(f(c_1,t),t)}\,dt\right)ds + c_2\right),$$

where c_2 is some constant. Again if we express c_2 explicitly (and eliminate c_1 with the help of $c_1 = Q_1(z,w)$) we get another characteristic:

$$\begin{aligned} c_2 &= Q_2(z,w,H) \\ &= H\exp\left(-\int_0^w \frac{C(f(Q_1(z,w),s),s)}{B(f(Q_1(z,w),s),s)}\,ds\right) \\ &- \int_0^w \frac{D(f(Q_1(z,w),s),s;v)}{B(f(Q_1(z,w),s),s)}\exp\left(-\int_0^s \frac{C(f(Q_1(z,w),t),t)}{B(f(Q_1(z,w),t),t)}\,dt\right)ds. \end{aligned}$$

Now if we apply (11) we obtain the following representation for H:

$$\begin{aligned} H &= \exp\left(\int_0^w \frac{C(f(Q_1(z,w),s),s)}{B(f(Q_1(z,w),s),s)}\,ds\right) \\ &\times\left(\int_0^w \frac{D(f(Q_1(z,w),s),s;v)}{B(f(Q_1(z,w),s),s)}\exp\left(-\int_0^s \frac{C(f(Q_1(z,w),t),t)}{B(f(Q_1(z,w),t),t)}\,dt\right)ds + \tilde{F}(Q_1(z,w))\right). \end{aligned}$$

In our context we will assume that (5) holds, that is, $H(z,0;v) = 0$, which implies that $\tilde{F}(x) = 0$. Consequently we have

$$\begin{aligned} H(z,w;v) &= \exp\left(\int_0^w \frac{C(f(Q_1(z,w),s),s)}{B(f(Q_1(z,w),s),s)}\,ds\right) \\ &\times\left(\int_0^w \frac{D(f(Q_1(z,w),s),s;v)}{B(f(Q_1(z,w),s),s)}\exp\left(-\int_0^s \frac{C(f(Q_1(z,w),t),t)}{B(f(Q_1(z,w),t),t)}\,dt\right)ds\right). \end{aligned} \tag{13}$$

4 Singularity analysis

Next we assume that the assumptions of Theorem 1 are satisfied so that we can analyze the analytic properties of the solution function $H(z,w;v)$ that is given by (13). Actually we will show that if v is close to 1 that the dominant singularity comes from a curve that is a pertubation of the curve $z+w=1$.

First we note that by assumption the function $f(Q_1(z,w),s)$ is regular as well as the fraction $C(z,w)/B(z,w)$. Consequently the function

$$(z,w) \mapsto K(z,w) = \int_0^w \frac{C(f(Q_1(z,w),s),s)}{B(f(Q_1(z,w),s),s)}\,ds$$

is analytic, too. Thus, it remains to consider the integral

$$\begin{aligned} &\int_0^w \frac{D(f(Q_1(z,w),s),s;v)}{B(f(Q_1(z,w),s),s)}\exp\left(-K(z,s)\right)ds \\ &= \int_0^w \frac{a(f(Q_1(z,w),s),s;v)\exp\left(-K(z,s)\right)/B(f(Q_1(z,w),s),s)}{(1-b(f(Q_1(z,w),s),s;v))^2}\,ds. \end{aligned}$$

First let us assume that $v=1$. In this case we know by assumption that $H(z,w;1) = 1/(1-z-w) - 1/(1-z)$. Furthermore we have $b(z,w;1) = z$. Thus the above integral simplifies to

$$\int_0^w \frac{L(z,w,s)}{(1-f(Q_1(z,w),s))^2}\,ds,$$

where $L(z,w,s)$ is a non-zero regular function. As long as $f(Q_1(z,w),s) \neq 1$ for $0 \leq s \leq w$ then the integral represents a regular function in z and w. Hence, we have to detect s for which $f(Q_1(z,w),s) = 1$. Let us first assume that z and w are real and positive. We also recall that by assumption $\frac{\partial f}{\partial s} = A(f,s)/B(f,s) < 0$. Thus, if we start with z,w close to zero and increase them we observe that the first critical instance occurs when $f(Q_1(z,w),0) = 1$. Of course this has to coincide with the condition $z+w=1$ and we have to recover the (known) singular behaviour $1/(1-z-w)$.

Actually we can use the following easy lemma (which follows from partial integration).

Lemma 1 *Suppose that $N(s)$ and $D(s)$ are three times continuously differentiable functions such that $D(s) \neq 0$ and $D'(s) \neq 0$ Then we have*

$$\begin{aligned}\int \frac{N(s)}{D(s)^2}\,ds &= -\frac{N(s)}{D(s)D'(s)} + \frac{\log D(s)}{D'(s)}\left(\frac{N(s)}{D'(s)}\right)' \\ &- \int \log D(s)\left(\frac{1}{D'(s)}\left(\frac{N(s)}{D'(s)}\right)'\right)'\,ds.\end{aligned}$$

If we apply this lemma in our context it follows that

$$\int_0^w \frac{L(z,w,s)}{(1-f(Q_1(z,w),s))^2}\,ds = \frac{\tilde{L}_1(z,w)}{1-f(Q_1(z,w),0)} + O\left(\log|1-f(Q_1(z,w),0)|\right)$$

for positive real z,w with $z+w \to 1$ (and a proper non-zero analytic function $\tilde{L}_1(z,w)$). Summing up we obtain for positive real z,w with $z+w \to 1$ the asymptotic representation

$$H(z,w;1) = \frac{\tilde{L}_2(z,w)}{1-f(Q_1(z,w),0)} + O\left(\log|1-f(Q_1(z,w),0)|\right)$$

for some non-zero analytic function $\tilde{L}_2(z,w)$. In particular it follows that $1-f(Q_1(z,w),0)$ can be written as

$$1-f(Q_1(z,w),0) = \tilde{L}_2(z,w)(1-z-w).$$

Of course the same kind of analysis applies if z and w are complex numbers close to the positive real line. Furthermore we observe that the integral representation for $H(z,w;1)$ will not get singular if $z+w \neq 1$. By continuity this also holds if v is close to 1 and $|1-z-w| \geq \delta$ for some $\delta > 0$.

Finally if v is close (but different) to 1 and z and w satisfy $|1-z-w| < \delta$ then we just have to modify the above analysis slightly and observe that $H(z,w;v)$ can be represented as

$$H(z,w;v) = \frac{\tilde{L}_2(z,w;v)}{1-b(f(Q_1(z,w),0),0;v)} + O\left(\log|1-b(f(Q_1(z,w),0),0;v)|\right).$$

Thus, the equation

$$b(f(Q_1(z,w),0),0;v) = 1 \tag{14}$$

determines the dominant singularity of $H(z,w;v)$. By the implicit function theorem it follows that there exists a solution of (14) of the form $z = z_0(w;v)$ with $z_0(w;1) = 1-w$ (if w is close to the positive real line segment $[0,1]$).

5 A central limit theorem

We start with a lemma on bivariate asymptotics for generating functions in two variables which is a slight generalization of the smooth case in Pemantle and Wilson's book Pemantle and Wilson (2013).

Lemma 2 *Suppose that $f(z,w)$ is a generating function in two variables that can be written in the form*

$$f(z,w)=\frac{N(z,w)}{D(z,w)},$$

where N and D are regular functions such that the system of equations

$$D(z,w)=0,\quad wD_w(z,w)=\rho zD_z(z,w) \tag{15}$$

has a unique positive and analytic solution $z=z_0(\rho)$, $w=w_0(\rho)$ for ρ in a positive interval $[\alpha,\beta]$ such that $D_w(z_0(\rho),w_0(\rho))\neq 0$ in this range and that $D(z,w)=0$ has no other solutions for $|z|\leq z_0(\rho)$, $|w|\leq w_0(\rho)$. Furthermore we assume that $N(z_0(\rho),w_0(\rho))\neq 0$.

Then we have uniformly for $m/n\in[\alpha,\beta]$

$$[z^nw^m]f(z,w)\sim\frac{N(z_0(m/n),w_0(m/n))}{-z_0(m/n)w_0(m/n)D_z(z_0(m/n),w_0(m/n))}\frac{z_0(m/n)^{-n}w_0(m/n)^{-m}}{\sqrt{2\pi n\Delta(m/n)}}, \tag{16}$$

where

$$\Delta(\rho)=\left.\frac{D_{zz}D_w^2-2D_{zw}D_zD_w+D_{ww}D_z^2}{zD_z^3}+\frac{D_w^2}{z^2D_z^2}+\frac{D_w}{zwD_z}\right|_{z=z_0(\rho),w=w_0(\rho)}.$$

Proof: By assumption the map $z\mapsto f(z,w)$ has a unique polar singularity at $z=z(w)$, where $z(w)$ is determined by $D(z(w),w)=0$ (for w close to the real interval $[w_0(a),w_0(b)]$) which implies

$$[z^n]f(z,w)\sim\frac{N(z(w),w)}{-z(w)D_z(z(w),w)}z(w)^{-n}.$$

Finally we fix the ratio $m/n=\rho$ and a direct application of the saddle point method on the Cauchy integral evaluating

$$[w^mz^n]f(z,w)=\frac{1}{2\pi i}\int_{|w|=w_0(\rho)}\left([z^n]f(z,w)\right)w^{-m-1}\,dw$$

leads to the result. Note that the saddle point $w=w_0(\rho)$ that comes from the power $z(w)^{-n}w^{-\rho n}$ has to satisfy (15). □

We now apply this procedure to a slightly more general situation, namely when there is a further parameter v (that is assumed to be close to 1):

$$f(z,w;v)=\frac{N(z,w;v)}{D(z,w;v)}.$$

In our context we have to identify $f(z,w;v)$ with $H(z,w;v)$ and $D(z,w;v)$ with $1-b(f(Q_1(z,w),0),0;v)$. Of course we have to formulate proper assumptions (similar to the above which are actually satisfied for $H(z,w;v)$) and, hence, by (16) we obtain an asymptotic expansion of the form

$$[z^n w^m]H(z,w;v) \sim \frac{C(m/n;v)}{\sqrt{2\pi n}} z_0(m/n;v)^{-n} w_0(m/n;v)^{-m}$$

that is uniform in v (for v sufficiently close to 1).

If we fix the ratio $\rho = m/n$ the leading asymptotics is then just a power in n:

$$z_0(\rho;v)^{-n} w_0(\rho;v)^{-\rho n} = e^{h(\rho;v)n}$$

with $h(\rho;v) = -\log z_0(\rho;v) - \rho \log w_0(\rho;v)$. Actually we have a so-called *quasi-power*, where we can expect that (after proper normalization) a central limit theorem should hold.

In our context we obtain

$$\mathbb{E}[v^{X_{\rho n,n}}] = \frac{[z^n w^{\rho n}]H(z,w;v)}{[z^n w^{\rho n}]H(z,w;1)} \sim \frac{C(\rho;v)}{C(\rho;1)} \left(\frac{z_0(\rho;1)w_0(\rho;1)^\rho}{z_0(\rho;v)w_0(\rho;v)^\rho} \right)^n .$$

And this is precisely the assumption that is needed in order to apply Hwang's *Quasi-Power Theorem* Hwang (1994).

Lemma 3 *Let X_n be a random variable with the property that*

$$\mathbb{E}\, v^{X_n} = e^{\lambda_n \cdot A(v) + B(v)} \left(1 + O\left(\frac{1}{\varphi_n} \right) \right) \tag{17}$$

holds uniformly in a complex neighbourhood of $v = 1$, where λ_n and φ_n are sequences of positive real numbers with $\lambda_n \to \infty$ and $\varphi_n \to \infty$, and $A(v)$ and $B(v)$ are analytic functions in this neighbourhood of $v = 1$ with $A(1) = B(1) = 0$. Then X_n satisfies a central limit theorem of the form

$$\frac{1}{\sqrt{\lambda_n}} \left(X_n - \mathbb{E}\, X_n \right) \to N\left(0, \sigma^2\right) \tag{18}$$

and we have

$$\mathbb{E}\, X_n = \lambda_n \mu + O\left(1 + \lambda_n/\varphi_n\right)$$

and

$$\mathbb{V}\mathrm{ar}\, X_n = \lambda_n \sigma^2 + O\left(\left(1 + \lambda_n/\varphi_n\right)^2 \right),$$

where

$$\mu = A'(1)$$

and

$$\sigma^2 = A''(1) + A'(1).$$

Recall that $A(v) = h(\rho;v) = -\log z_0(\rho;v) - \rho \log w_0(\rho;v)$ so that

$$\mu = \mu(\rho) = -\frac{z_{0,v}(\rho;1)}{z_0(\rho;1)} - \rho \frac{w_{0,v}(\rho;1)}{w_0(\rho;1)}.$$

Since we have assumed that $X_{n,m}$ are non-negative random variables we can only expect a central limit theorem if $\mu > 0$, since for $\mu = 0$ it would follow that $X_{n,m}$ is negative with probability $1/2$.

Finally we mention that since the convergence is uniform in $\rho \in [a,b]$ we also get a central limit theorem for $n, m \to \infty$ if $m/n \in [a,b]$. This completes the proof of our main Theorem 1.

Acknowledgements

The authors are grateful to three anonymous referees for their careful reading and their valuable comments.

References

C. A. Brennan and H. Prodinger. The pills problem revisited. *Quaest. Math.*, 26(4):427–439, 2003.

P. Flajolet, J. Gabarró, and H. Pekari. Analytic urns. *Ann. Probab.*, 33(3):1200–1233, 2005.

G. Hellwig. Partial differential equations. An introduction. 2nd ed. Mathematische Leitfäden. Stuttgart: B. G. Teubner. XI, 259 p., 35 figs. DM 48.00 (1977)., 1977.

H.-K. Hwang. Théoremes limites pour les structures combinatoires et les fonctions arithmetiques. PhD Thesis, École Polytechnique, 1994.

S. Janson. Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Processes Appl.*, 110(2):171–245, 2004.

S. Janson. Limit theorems for triangular urn schemes. *Probab. Theory Relat. Fields*, 134(3):417–452, 2006.

D. E. Knuth and J. Mccarthy. Problem E3429: Big Pills and Little Pills. *Amer. math. Monthly*, 98:264, 1991.

M. Kuba. Analysis of a class of cannibal urns. *Electron. Commun. Probab.*, 16:583–599, 2011.

M. Kuba and A. Panholzer. Limit laws for a class of diminishing urn models. In *2007 Conference on analysis of algorithms, AofA 07. Papers from the 13th Conference held in Juan-les-Pins, France, June 17–22, 2007.*, pages 341–352. Nancy: The Association. Discrete Mathematics & Theoretical Computer Science (DMTCS), 2007.

M. Kuba and A. Panholzer. On death processes and urn models. In *Proceeding of the 23rd international meeting on probabilistic, combinatorial, and asymptotic methods in the analysis of algorithms (AofA'12), Montreal, Canada, June 18–22, 2012*, pages 29–42. Nancy: The Association. Discrete Mathematics & Theoretical Computer Science (DMTCS), 2012.

M. Kuba, A. Panholzer, and H. Prodinger. Lattice paths, sampling without replacement, and limiting distributions. *Electron. J. Comb.*, 16(1):research paper r67, 12, 2009.

B. Morcrette. Fully analyzing an algebraic Pólya urn model. In *LATIN 2012: Theoretical informatics. 10th Latin American symposium, Arequipa, Peru, April 16–20, 2012. Proceedings*, pages 568–581. Berlin: Springer, 2012.

R. Pemantle and M. C. Wilson. *Analytic combinatorics in several variables.* Cambridge: Cambridge University Press, 2013.

B. Pittel. An urn model for cannibal behavior. *J. Appl. Probab.*, 24:522–526, 1987.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Analytic approach for reflected Brownian motion in the quadrant

Sandro Franceschi[1,2], Irina Kurkova[1], Kilian Raschel[2]

[1]*Laboratoire de Probabilités et Modèles Aléatoires, Université Pierre et Marie Curie, 4 Place Jussieu, 75252 Paris Cedex 05, France*
[2]*Laboratoire de Mathématiques et Physique Théorique, Université de Tours, Parc de Grandmont, 37200 Tours, France*

Abstract. Random walks in the quarter plane are an important object both of combinatorics and probability theory. Of particular interest for their study, there is an analytic approach initiated by Fayolle, Iasnogorodski and Malyšev, and further developed by the last two authors of this note. The outcomes of this method are explicit expressions for the generating functions of interest, asymptotic analysis of their coefficients, etc. Although there is an important literature on reflected Brownian motion in the quarter plane (the continuous counterpart of quadrant random walks), an analogue of the analytic approach has not been fully developed to that context. The aim of this note is twofold: it is first an extended abstract of two recent articles of the authors of this paper, which propose such an approach; we further compare various aspects of the discrete and continuous analytic approaches.

Keywords: Reflected Brownian motion in the quadrant; (Random) Walks in the quarter plane; Stationary distribution; Laplace transform; Generating function; Boundary value problem; Asymptotic analysis

1 Introduction

1.1 Random walks in the quarter plane

Since the seventies and the pioneered papers Malyšev (1972); Fayolle and Iasnogorodski (1979), random walks in the quarter plane (cf. Figure 1) are extensively studied. They are indeed an important object of probability theory and have been studied for their recurrence/transience, for their links with queueing systems (Fayolle and Iasnogorodski (1979)), representation theory (Biane (1992)), potential theory. Moreover, the state space $\mathbb{N}^2 = \{0, 1, 2, \ldots\}^2$ offers a natural framework for studying any two-dimensional population; accordingly, quadrant walks appear as models in biology and in finance (Cont and de Larrard (2013)). Another interest of random walks in the quarter plane is that in the large class of random processes in cones, they form a family for which remarkable exact formulas exist. Moreover, quadrant walks are popular in combinatorics, see Bousquet-Mélou and Mishna (2010); Bostan and Kauers (2010); Kurkova and Raschel (2012). Indeed, many models of walks are in bijection with other combinatorial objects: maps, permutations, trees, Young tableaux, etc. In combinatorics again, famous models have emerged from quadrant walks, as Kreweras' or Gessel's ones, see Bousquet-Mélou and Mishna (2010); Bostan and Kauers (2010). Finally, walks in the quarter plane are interesting for the numerous tools used

for their analysis: combinatorial (Bousquet-Mélou and Mishna (2010)), from complex analysis (Malyšev (1972); Fayolle and Iasnogorodski (1979); Fayolle et al. (1999); Kurkova and Raschel (2011, 2012); Bernardi et al. (2015)), computer algebra (Bostan and Kauers (2010)), for instance.

1.2 Issues and technicalities of the analytic approach

In the literature (see, e.g., Malyšev (1972); Fayolle and Iasnogorodski (1979); Fayolle et al. (1999); Kurkova and Raschel (2011, 2012)), the analytic approach relies on six key steps:

(i) Finding a functional equation between the generating functions of interest;

(ii) Reducing the functional equation to boundary value problems (BVP);

(iii) Solving the BVP;

(iv) Introducing the group of the walk;

(v) Defining the Riemann surface naturally associated with the model, continuing meromorphically the generating functions and finding the conformal gluing function;

(vi) Deriving the asymptotics of the (multivariate) coefficients.

Before commenting these different steps, let us note that altogether, they allow for studying the following three main problems:

(P1) Explicit expression for the generating functions of interest (needs (i), (ii), (iii) and (v));

(P2) Algebraic nature of these functions (needs (iv) and (v));

(P3) Asymptotics of their coefficients in various regimes (needs (iii), (v) and (vi)).

The point (i) reflects the inherent properties of the model and is easily obtained. Point (ii), first shown in Fayolle and Iasnogorodski (1979), is now standard (see Fayolle et al. (1999)) and follows from algebraic manipulations of the functional equation of (i). Item (iii) uses specific literature devoted to BVP (our main reference for BVP is the book of Litvinchuk (2000)). The idea of introducing the group of the model (iv) was proposed in Malyšev (1972), and brought up to light in the combinatorial context in Bousquet-Mélou and Mishna (2010). Point (v) is the most technical a priori; it is however absolutely crucial, as it allows to access key quantities (as a certain conformal gluing function which appears in the exact formulation of (iii)). Finally, (vi) uses a double refinement of the classical saddle point method: the uniform steepest descent method.

1.3 Reflected Brownian motion in the quarter plane

There is a large literature on reflected Brownian motion in quadrants (and also in orthants, generalization to higher dimension of the quadrant), to be rigorously introduced in Section 3. First, it serves as an approximation of large queuing networks (see Foddy (1984); Baccelli and Fayolle (1987)); this was the initial motivation for its study. In the same vein, it is the continuous counterpart of (random) walks in the quarter plane. In other directions, it is studied for its Lyapunov functions in Dupuis and Williams (1994), cone points of Brownian motion in Le Gall (1987), intertwining relations and crossing probabilities in Dubédat (2004), and of particular interest for us, for its recurrence/transience in Hobson and

Rogers (1993). The asymptotic behavior of the stationary distribution (when it exists) is now well known, see Harrison and Hasenbein (2009); Dai and Miyazawa (2011); Franceschi and Kurkova (2016). There exist, however, very few results giving exact expressions for the stationary distribution. Let us mention Foddy (1984) (dealing with the particular case of a Brownian motion with identity covariance matrix), Baccelli and Fayolle (1987) (on a diffusion having a quite special behavior on the boundary), Harrison and Williams (1987b); Dieker and Moriarty (2009) (on the special case when stationary densities are exponential) and Franceschi and Raschel (2016) (on the particular case of an orthogonal reflection). We also refer to Burdzy et al. (2015) for the analysis of reflected Brownian motion in bounded planar domains by complex analysis techniques.

1.4 Main results and plan

This note is an extended abstract of the papers Franceschi and Kurkova (2016); Franceschi and Raschel (2016), whose main contributions are precisely to export the analytic method for reflected Brownian motion in the quarter plane. Our study constitutes one of the first attempts to apply these techniques to the continuous setting, after Foddy (1984); Baccelli and Fayolle (1987). In addition of reporting about the works Franceschi and Kurkova (2016); Franceschi and Raschel (2016), we also propose a comparative study of the discrete/continuous cases.

Our paper is organized as follows: Section 2 concerns random walks and Section 3 Brownian motion. For clarity of exposition we have given the same structure to Sections 2 and 3: in Section 2.1/3.1 we first state the key functional equation (a kernel equation), which is the starting point of our entire analysis. We study the kernel (a second degree polynomial in two variables). In Section 2.2/3.2 we state and solve the BVP satisfied by the generating functions. We then move to asymptotic results (Section 2.3/3.3). In Section 2.4/3.4 we introduce the Riemann surface of the model and some important related facts.

2 Random walks in the quarter plane

This section is devoted to the discrete case and is based mainly on Fayolle et al. (1999).

2.1 Functional equation

One considers a piecewise homogeneous random walk with sample paths in $\mathbb{N}^2$. There are four domains of spatial homogeneity (the interior of $\mathbb{N}^2$, the horizontal and vertical axes, the origin), inside of which the transition probabilities (of unit size) are denoted by $p_{i,j}$, $p'_{i,j}$, $p''_{i,j}$ and $p^0_{i,j}$, respectively. See Figure 1. The inventory polynomial of the inner domain is called the kernel and equals

$$K(x,y) = xy\{\textstyle\sum_{-1\leqslant i,j\leqslant 1} p_{i,j}x^iy^j - 1\}. \tag{1}$$

The inventory polynomials associated to the other homogeneity domains are

$$k(x,y) = x\{\textstyle\sum p'_{i,j}x^iy^j - 1\}, \quad \widetilde{k}(x,y) = y\{\textstyle\sum p''_{i;j}x^iy^j - 1\}, \quad k_0(x,y) = \{\textstyle\sum p^0_{i,j}x^iy^j - 1\}.$$

Assuming the random walk ergodic (we refer to (Fayolle et al., 1999, Theorem 1.2.1) for necessary and sufficient conditions), we denote the invariant measure by $\{\pi_{i,j}\}_{i,j\geqslant 0}$ and introduce the generating functions

$$\pi(x,y) = \textstyle\sum_{i,j\geqslant 1}\pi_{i,j}x^{i-1}y^{j-1}, \quad \pi(x) = \sum_{i\geqslant 1}\pi_{i,0}x^{i-1}, \quad \widetilde{\pi}(y) = \sum_{j\geqslant 1}\pi_{0,j}y^{j-1}.$$

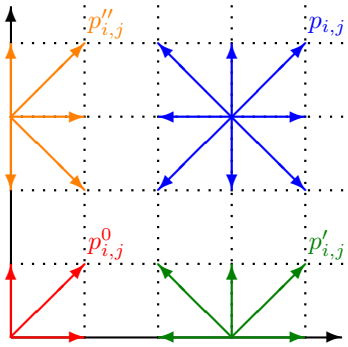


Fig. 1: Transition probabilities of the reflected random walk in the quarter plane, with four domains of spatial homogeneity

Writing the balance equations at the generating function level, we have (see (Fayolle et al., 1999, Equation (1.3.6)) for the original statement):

Lemma 1 *The fundamental functional equation holds*

$$-K(x,y)\pi(x,y) = k(x,y)\pi(x) + \widetilde{k}(x,y)\widetilde{\pi}(y) + k_0(x,y)\pi_{0,0}. \tag{2}$$

Equation (2) holds a priori in the region $\{(x,y) \in \mathbb{C}^2 : |x| \leqslant 1, |y| \leqslant 1\}$. Indeed, the $\pi_{i,j}$ sum up to 1, so that the generating functions $\pi(x,y)$, $\pi(x)$ and $\widetilde{\pi}(y)$ are well defined on the (bi)disc. The identity (2) is a kernel equation, and a crucial role will be played by the kernel (1). This polynomial K is of second order in both x and y; its roots $X(y)$ and $Y(x)$ defined by

$$K(X(y),y) = K(x,Y(x)) = 0 \tag{3}$$

are thus algebraic of degree 2. Writing the kernel as $K(x,y) = a(y)x^2 + b(y)x + c(y)$ and defining its discriminant $d(y) = b(y)^2 - 4a(y)c(y)$, one has obviously

$$X(y) = \frac{-b(y) \pm \sqrt{d(y)}}{2a(y)}.$$

The polynomial d has three or four roots, and exactly two of them are located in the unit disc, see (Fayolle et al., 1999, Lemma 2.3.8). They are named y_1, y_2, cf. Figure 2. On (y_1, y_2) one has $d(y) < 0$, so that the two values (or branches) of $X(y)$ (that we shall call $X_0(y)$ and $X_1(y)$) are complex conjugate of one another. In particular, the set

$$\mathcal{M} = X([y_1,y_2]) = \{x \in \mathbb{C} : K(x,y) = 0 \text{ and } y \in [y_1,y_2]\}$$

is symmetrical w.r.t. the real axis (Figure 2). This curve will be used to set a boundary condition for the unknown function π (Lemma 2).

2.2 Statement and resolution of the BVP

The analytic approach of Malyšev (1972); Fayolle and Iasnogorodski (1979); Fayolle et al. (1999) proposes a way for solving the functional equation (2), by reduction to a BVP. Generally speaking, a BVP consists of a regularity condition and a boundary condition.

Lemma 2 *The function π satisfies the following BVP:*

- *π is meromorphic in the bounded domain delimitated by $\mathcal{M}$ and has there identified poles;*
- *for any $x \in \mathcal{M}$,*

$$\frac{k(x,Y_0(x))}{\widetilde{k}(x,Y_0(x))}\pi(x) - \frac{k(\overline{x},Y_0(\overline{x}))}{\widetilde{k}(\overline{x},Y_0(\overline{x}))}\pi(\overline{x}) = \frac{k_0(\overline{x},Y_0(\overline{x}))}{\widetilde{k}(\overline{x},Y_0(\overline{x}))}\pi_{0,0} - \frac{k_0(x,Y_0(x))}{\widetilde{k}(x,Y_0(x))}\pi_{0,0}.$$

Proof: The regularity condition follows from Theorem 5, which provides a (maximal) meromorphic continuation of the function π. We now turn to the boundary condition. For $i \in \{0,1\}$, we evaluate the functional equation (2) at $(X_i(y),y)$ and divide by $\widetilde{k}(X_i(y),y)$. We then make the difference of the identities corresponding to $i=0$ and $i=1$. Finally, we substitute $X_0(y)=x$ and $X_1(y)=\overline{x}$, noting that when $y \in [y_1,y_2]$, $x \in \mathcal{M}$ by construction. Notice that we have chosen the segment $[y_1,y_2]$ connecting the points inside of the unit disc (Figure 2), in which we know that the generating function $\widetilde{\pi}$ is well defined. □

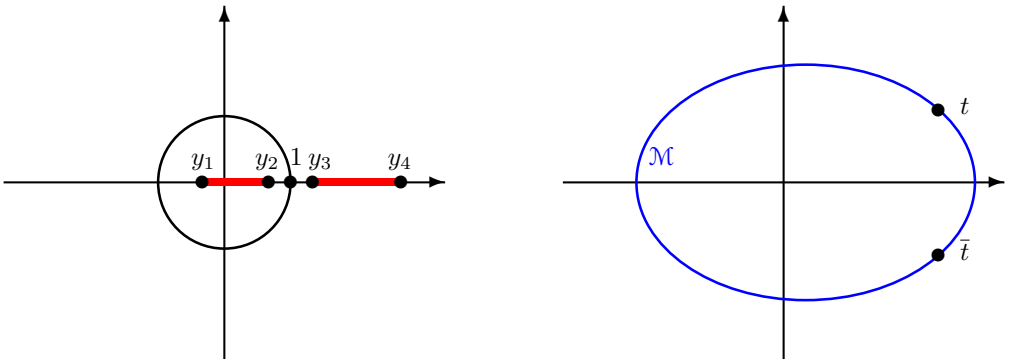


Fig. 2: Left: the polynomial d has three or four roots, denoted by y_1, y_2, y_3, y_4; exactly two of them are inside the unit disc. Right: the curve $\mathcal{M} = X([y_1,y_2])$ is symmetrical w.r.t. the horizontal axis

Although Lemma 2 could be written more precisely (by giving the number and the location of the poles of π), we shall prefer the above version, since we focus in this note on the main ideas of the analytic approach.

Lemma 2 happens to characterize the generating functions, as it eventually leads to an explicit expression for π, see Theorem 3. Before stating this central result (borrowed from (Fayolle et al., 1999, Theorem 5.4.3)), we need to introduce a function w called a conformal gluing function. By definition it satisfies $w(x) = w(\overline{x})$ for $x \in \mathcal{M}$ and is one-to-one inside of $\mathcal{M}$. This function will be constructed in Theorem 6 of Section 2.4.

Theorem 3 *There exist two functions f and g, constructed from k, $\widetilde{k}$, k_0 and w, such that the following integral formulation for π holds:*

$$\pi(x) = f(x) \int_{\mathcal{M}} g(u) \frac{w'(u)}{w(u) - w(x)} \mathrm{d}u.$$

A similar contour integral representation exists for $\widetilde{\pi}$, and eventually the functional equation (2) provides us with an explicit expression for the bivariate function $\pi(x, y)$.

2.3 Asymptotics of the stationary probabilities

The asymptotics of coefficients $\pi_{i,j}$ of unknown generating functions satisfying the functional equation (2) has been obtained by Malyšev (1973) via analytic arguments. He computed the asymptotics of the stationary probabilities $\pi_{i,j}$ as $i, j \to \infty$ and $j/i = \tan\alpha$, for any given $\alpha \in (0, \pi/2)$. Let us briefly present these results. It is assumed in Malyšev (1973) that the random walk is simple, meaning that

$$p_{-1,1} = p_{1,1} = p_{-1,-1} = p_{1,-1} = 0. \tag{4}$$

It is also assumed that both coordinates of the interior drift vector are negative (as in Figure 3). For $\alpha \in (0, \pi/2)$, we define the point $(x(\alpha), y(\alpha))$ as follows. Introducing as in Kurkova and Raschel (2011) the function $P(u, v) = \sum_{i,j} p_{i,j} e^{iu} e^{jv}$ on $\mathbb{R}^2$, the mapping

$$(u, v) \mapsto \frac{\nabla P(u, v)}{|P(u, v)|}$$

is a homeomorphism between $\{(u, v) \in \mathbb{R}^2 : P(u, v) = 1\}$ and the unit circle. The point $(u(\alpha), v(\alpha))$ is the unique solution to $\frac{\nabla P(u,v)}{|P(u,v)|} = (\cos\alpha, \sin\alpha)$. Finally, $(x(\alpha), y(\alpha)) = (e^{u(\alpha)}, e^{v(\alpha)})$.

Following Malyšev (1973), we introduce the sets of parameters

$$\begin{aligned} \mathcal{P}_{--} &= \big\{(\{p_{i,j}\}, \alpha) : k(\psi(x(\alpha), y(\alpha))) \leqslant 0 \text{ and } \widetilde{k}(\phi(x(\alpha), y(\alpha))) \leqslant 0\big\}, \\ \mathcal{P}_{+-} &= \big\{(\{p_{i,j}\}, \alpha) : k(\psi(x(\alpha), y(\alpha))) > 0 \text{ and } \widetilde{k}(\phi(x(\alpha), y(\alpha))) \leqslant 0\big\}, \end{aligned}$$

and $\mathcal{P}_{-+}$ and $\mathcal{P}_{++}$ accordingly. The automorphisms ψ and ϕ are defined in Section 2.4 by (8). The following theorem is proven in Malyšev (1973).

Theorem 4 *Let $(i, j) = (r\cos\alpha, r\sin\alpha)$ with $\alpha \in (0, \pi/2)$. Then as $r \to \infty$ we have*

$$\pi_{i,j} = (1 + o(1)) \cdot \begin{cases} \frac{C_0(\alpha)}{\sqrt{r}} x^{-i}(\alpha) y^{-j}(\alpha) & \text{in } \mathcal{P}_{--}, \\ C_1 p_1^{-i} q_1^{-j} & \text{in } \mathcal{P}_{-+}, \\ C_2 p_2^{-i} q_2^{-j} & \text{in } \mathcal{P}_{+-}, \\ C_1 p_1^{-i} q_1^{-j} + C_2 p_2^{-i} q_2^{-j} & \text{in } \mathcal{P}_{++}, \end{cases} \tag{5}$$

where C_0, C_1 and C_2 are constants that can be expressed in terms of the functions π and $\widetilde{\pi}$. The point (p_1, q_1) is a solution of the system $\{K(x, y) = 0,\ k(\psi(x, y)) = 0\}$ and similarly, (p_2, q_2) is a solution of $\{K(x, y) = 0,\ \widetilde{k}(\phi(x, y)) = 0\}$.

Proof: The stationary probabilities $\pi_{i,j}$ are first written as two-dimensional Cauchy integrals, then reduced via the residue theorem to one-dimensional integrals along some contours. The asymptotics of these integrals is characterized either by the saddle point $(x(\alpha), y(\alpha))$ in the case of the set of parameters $\mathcal{P}_{--}$ or by a pole (p_1, q_1) or (p_2, q_2) that is encountered when moving the integration contour to the saddle point; this happens for the sets of parameters $\mathcal{P}_{-+}$, $\mathcal{P}_{+-}$ and $\mathcal{P}_{++}$. □

This approach has been applied for the analysis of the join the shortest queue problem in Kurkova and Suhov (2003), and for computing the asymptotics of Green functions for transient random walks in the quarter plane reflected at the axes (see Kurkova and Malyshev (1998)) or killed at the axes (cf. Kurkova and Raschel (2011)). Moreover, as illustrated in Kurkova and Raschel (2011); Kurkova and Suhov (2003), the assumption (4) is not crucial for the applicability of the method. The limiting cases $\alpha = 0$ and $\alpha = \pi/2$ can also be treated via this approach, with some additional technical details (the saddle point then coincides with a branch point of the integrand), it is done in Kurkova and Raschel (2011).

2.4 Riemann surface and related facts

In Section 2.1 the set

$$\mathcal{K} = \{(x, y) \in \mathbb{C}^2 : K(x, y) = 0\} = \{(x, y) \in \mathbb{C}^2 : \sum p_{i,j} x^i y^j = 1\}$$

has appeared very naturally, since in order to state the BVP (our Lemma 2), we introduced the functions $X(y)$ and $Y(x)$, which by construction cancel the kernel, see (3).

In this section the central idea is to consider the (global) complex structure of $\mathcal{K}$. The set $\mathcal{K}$ turns out to be a Riemann surface of genus 1, i.e., a torus. This simply comes from the reformulation of the identity $K(x, y) = 0$ as

$$\{2a(y)x + b(y)\}^2 = d(y).$$

Moreover, the Riemann surface of the square root of a polynomial of degree 3 or 4 is classically a torus (with this terminology, the roots of the discriminant are branch points).

This new point of view on $\mathcal{K}$ brings powerful tools. Of particular interest is a parametrization of $\mathcal{K}$ in terms of Weierstrass elliptic functions:

$$\mathcal{K} = \{(x(\omega), y(\omega)) : \omega \in \mathbb{C}/(\omega_1\mathbb{Z} + \omega_2\mathbb{Z})\}. \tag{6}$$

This parametrization is totally explicit: $x(\omega)$ and $y(\omega)$ are rational functions in the $\wp$-Weierstrass function and its derivative $\wp'$ (see (Fayolle et al., 1999, Lemma 3.3.1)); the periods ω_1 and ω_2 admit expressions as elliptic integrals in terms of $\{p_{i,j}\}$ (cf. (Fayolle et al., 1999, Lemma 3.3.2)), etc. Moreover, as any functions of x and/or y, the functions $\pi(x)$ and $\widetilde{\pi}(y)$ can be lifted on $\mathcal{K}$ by setting

$$\Pi(\omega) = \pi(x(\omega)), \qquad \widetilde{\Pi}(\omega) = \widetilde{\pi}(y(\omega)). \tag{7}$$

Group of the walk

Introduced in Malyšev (1972) in a probabilistic context and further used in Fayolle et al. (1999); Bousquet-Mélou and Mishna (2010), the group of the walk is a dihedral group generated by

$$\zeta(x, y) = \left(x, \frac{\sum_i p_{i,-1} x^i}{\sum_i p_{i,+1} x^i} \frac{1}{y}\right), \qquad \eta(x, y) = \left(\frac{\sum_j p_{-1,j} y^j}{\sum_j p_{+1,j} y^j} \frac{1}{x}, y\right). \tag{8}$$

(One easily verifies that these generators are idempotent: $\zeta^2 = \eta^2 = 1$.) The group $\langle \zeta, \eta \rangle$ can be finite or infinite, according to the order of the element $\zeta \circ \eta$. The generator ζ (resp. η) exchanges the roots in y (resp. in x) of $K(x, y) = 0$. Viewed as a group of birational transformations in Bousquet-Mélou and Mishna (2010), we shall rather see it as a group of automorphisms of the Riemann surface $\mathcal{K}$.

This group has many applications. First, it allows for a continuation of the functions $\pi(x)$ and $\widetilde{\pi}(y)$ (Theorem 5 below). It further connects the algebraic nature of the generating functions to the (in)finiteness of the group (Theorem 7). Finally, in the finite group case, elementary algebraic manipulations of the functional equations can be performed (typically, via the computation of certain orbit-sums) so as to eventually obtain D-finite expressions for the unknowns, see (Fayolle et al., 1999, Chapter 4) and Bousquet-Mélou and Mishna (2010).

Using the structure of the automorphisms of a torus, the lifted versions of ζ and η admit simple expressions (Fayolle et al., 1999, Section 3.1.2):

$$\zeta(\omega) = -\omega + \omega_1 + \omega_2, \qquad \eta(\omega) = -\omega + \omega_1 + \omega_2 + \omega_3, \tag{9}$$

where, as the periods ω_1 and ω_2, $\omega_3 \in (0, \omega_2)$ is an elliptic integral (Fayolle et al., 1999, Lemma 3.3.3). Accordingly, the group is finite if and only if $\omega_2/\omega_3 \in \mathbb{Q}$, which provides a nice criterion in terms of elliptic integrals.

Continuation

While the generating function $\pi(x)$ is defined through its power series in the unit disc, it is a priori unclear how to continue it to a larger domain. This is however crucial, since the curve $\mathcal{M}$ on which it satisfies a BVP (Lemma 2) is not included in the unit disc in general.

Theorem 5 *The function π can be continued as a meromorphic function to $\mathbb{C} \setminus [x_3, x_4]$.*

We notice that $\mathcal{M}$ does not intersect $[x_3, x_4]$ by (Fayolle et al., 1999, Theorem 5.3.3), so that Theorem 5 indeed provides a continuation of the generating function in the domain delimitated by $\mathcal{M}$.

Proof: This result, stated as Theorem 3.2.3 in Fayolle et al. (1999), is a consequence of a continuation of the lifted generating functions (7) on the Riemann surface (or, better, on its universal covering — but we shall not go into these details here). The continuation on $\mathcal{K}$ uses the (lifted) functional equation (2) and the group of the walk $\langle \zeta, \eta \rangle$. □

Conformal mapping

In the integral expression of Theorem 3, the conformal gluing function w is all-present, as it appears in the integrand and in f and g as well. The introduction of the Riemann surface $\mathcal{K}$ allows to derive an expression for this function (this is another major interest of introducing $\mathcal{K}$). Let us recall that ω_1 and ω_2 are the periods of the elliptic functions of the parametrization (6), while ω_3 comes out in the lifted expression (9) of the automorphisms.

Theorem 6 *The conformal gluing function w admits the expression:*

$$w(x) = \wp(\wp^{-1}(x; \omega_1, \omega_2); \omega_1, \omega_3),$$

where for $i \in \{2, 3\}$, $\wp(\,\cdot\,; \omega_1, \omega_i)$ is the $\wp$-Weierstrass elliptic function with periods ω_1 and ω_i.

Proof: While on the complex plane, $\mathcal{M}$ is a quartic curve, it becomes on $\mathcal{K}$ much simpler (typically, a segment). This remark (which again illustrates all the benefit of having introduced the Riemann surface) is used in (Fayolle et al., 1999, Section 5.5.2) so as to obtain the above expression for w. □

Algebraic nature of the generating functions

Recall that a function of one variable is D-finite if it satisfies a linear differential equation with polynomial coefficients.

Theorem 7 *If the group is finite, the generating functions $\pi(x)$ and $\widetilde{\pi}(y)$ are D-finite.*

Proof: This follows from manipulations on the Riemann surface, see (Fayolle et al., 1999, Chapter 4). The D-finiteness is proved on $\mathbb{R}$; refined results (in the combinatorial context of the enumeration of paths) can be found in Bousquet-Mélou and Mishna (2010), where the D-finiteness is proved on $\mathbb{Q}$. □

The converse of Theorem 7 is not shown in full generality. It is true in combinatorics, see Kurkova and Raschel (2012).

3 Reflected Brownian motion in the quadrant

Defining reflected Brownian motion in the quadrant

The object of study here is the reflected Brownian motion with drift in the quarter plane $\mathbb{R}_+^2$

$$Z(t) = Z_0 + W(t) + \mu t + RL(t), \qquad \forall t \geqslant 0, \tag{10}$$

associated to the triplet (Σ, μ, R), composed of a non-singular covariance matrix, a drift and a reflection matrix, see Figure 3:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{21} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad R = (R^1, R^2) = \begin{pmatrix} r_{11} & r_{21} \\ r_{12} & r_{22} \end{pmatrix}.$$

In Equation (10), Z_0 is any initial point in $\mathbb{R}_+^2$, the process $(W(t))_{t\geqslant 0}$ is an unconstrained planar Brownian motion starting from the origin, and for $i \in \{1,2\}$, $L^i(t)$ is a continuous non-decreasing process, that increases only at time t such that $Z^i(t) = 0$, namely $\int_0^t \mathbb{1}_{\{Z^i(s)\neq 0\}} \mathrm{d}L^i(s) = 0$. The columns R_1 and R_2 represent the directions in which the Brownian motion is pushed when the axes are reached.

The reflected Brownian motion $(Z(t))_{t\geqslant 0}$ associated with (Σ, μ, R) is well defined, see for instance Williams (1995). Its stationary distribution exists and is unique if and only if the following (geometric flavored) conditions are satisfied (see, e.g., Harrison and Williams (1987a); Hobson and Rogers (1993))

$$r_{11} > 0, \;\; r_{22} > 0, \;\; r_{11}r_{22} - r_{12}r_{21} > 0, \;\; r_{22}\mu_1 - r_{12}\mu_2 < 0, \;\; r_{11}\mu_2 - r_{21}\mu_1 < 0. \tag{11}$$

More that the Brownian motion in the quadrant, all results presented below concern the Brownian motion in two-dimensional cones (by a simple linear transformation of the cones). This is a major difference and interest of the continuous case, which also illustrates that the analytic approach is very well suited to that context.

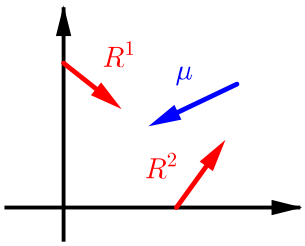


Fig. 3: Drift μ and reflection vectors R^1 and R^2

3.1 Functional equation

Laplace transform of the stationary distribution

The continuous analogues of the generating functions are the Laplace transforms. As their discrete counterparts, they characterize the stationary distribution. Under assumption (11), that we shall do throughout the manuscript, the stationary distribution is absolutely continuous w.r.t. the Lebesgue measure, see Harrison and Williams (1987a); Dai (1990). We denote its density by $\pi(x) = \pi(x_1, x_2)$. Let the Laplace transform of π be defined by

$$\varphi(\theta) = \mathbb{E}_\pi[e^{\langle\theta|Z\rangle}] = \iint_{\mathbb{R}_+^2} e^{\langle\theta|x\rangle}\pi(x)\mathrm{d}x.$$

We further define two finite boundary measures ν_1 and ν_2 with support on the axes, by mean of the formula

$$\nu_i(B) = \mathbb{E}_\pi\left[\int_0^1 \mathbb{1}_{\{Z(t)\in B\}}\mathrm{d}L^i(t)\right].$$

The measures ν_i are continuous w.r.t. the Lebesgue measure by Harrison and Williams (1987a), and may be viewed as boundary invariant measures. We define their moment Laplace transform by

$$\varphi_2(\theta_1) = \int_{\mathbb{R}_+} e^{\theta_1 x_1}\nu_2(x_1)\mathrm{d}x_1, \qquad \varphi_1(\theta_2) = \int_{\mathbb{R}_+} e^{\theta_2 x_2}\nu_1(x_2)\mathrm{d}x_2.$$

Functional equation

There is a functional equation between the Laplace transforms φ, φ_1 and φ_2, see (12), which is reminiscent of the discrete functional equation (2).

Lemma 8 *The following key functional equation between the Laplace transforms holds*

$$-\gamma(\theta)\varphi(\theta) = \gamma_1(\theta)\varphi_1(\theta_2) + \gamma_2(\theta)\varphi_2(\theta_1), \tag{12}$$

where

$$\begin{cases} \gamma(\theta) = \frac{1}{2}\langle\theta|\sigma\theta\rangle + \langle\theta|\mu\rangle = \frac{1}{2}(\sigma_{11}\theta_1^2 + \sigma_{22}\theta_2^2 + 2\sigma_{12}\theta_1\theta_2) + \mu_1\theta_1 + \mu_2\theta_2, \\ \gamma_1(\theta) = \langle R^1|\theta\rangle = r_{11}\theta_1 + r_{21}\theta_2, \\ \gamma_2(\theta) = \langle R^2|\theta\rangle = r_{12}\theta_1 + r_{22}\theta_2. \end{cases}$$

By definition of the Laplace transforms, this equation holds at least for any $\theta = (\theta_1, \theta_2)$ with $\Re\,\theta_1 \leqslant 0$ and $\Re\,\theta_2 \leqslant 0$. The polynomial γ in (12) is the kernel and is the continuous analogue of the kernel (1) in the discrete case. Polynomials γ_1 and γ_2 are the counterparts of k and $\widetilde{k}$.

Proof: To show (12), the main idea is to use an identity called a basic adjoint relationship (first proved in Harrison and Williams (1987a) in some particular cases, then extended in Dai and Harrison (1992)), which characterizes the stationary distribution. (It is the continuous analogue of the well-known equation $\pi Q = 0$, where π is the stationary distribution of a recurrent continuous-time Markov chain with infinitesimal generator Q.) This basic adjoint relationship connects the stationary distribution π and the corresponding boundary measures ν_1 and ν_2. We refer to Foddy (1984); Dai and Miyazawa (2011) for the details. □

Elementary properties of the kernel

The kernel γ in (12) can be alternatively written as

$$\gamma(\theta_1, \theta_2) = \widetilde{a}(\theta_2)\theta_1^2 + \widetilde{b}(\theta_2)\theta_1 + \widetilde{c}(\theta_2) = a(\theta_1)\theta_2^2 + b(\theta_1)\theta_2 + c(\theta_1). \tag{13}$$

The equation $\gamma(\theta_1, \theta_2) = 0$ defines a two-valued algebraic function $\Theta_1(\theta_2)$ by $\gamma(\Theta_1(\theta_2), \theta_2) = 0$, and similarly $\Theta_2(\theta_1)$ such that $\gamma(\theta_1, \Theta_2(\theta_1)) = 0$. Expressions of their branches are given by

$$\Theta_2^{\pm}(\theta_1) = \frac{-b(\theta_1) \pm \sqrt{d(\theta_1)}}{2a(\theta_1)},$$

where $d(\theta_1) = b^2(\theta_1) - 4a(\theta_1)c(\theta_1)$ is the discriminant. The polynomial d has two zeros, real and of opposite signs; they are denoted by $\theta_1^{\pm}$ and are branch points of the algebraic function Θ_2. In the same way we define $\Theta_1^{\pm}$ and its branch points $\theta_2^{\pm}$.

Finally, notice that d is negative on $\mathbb{R} \setminus [\theta_1^-, \theta_1^+]$. Accordingly, the branches $\Theta_2^{\pm}$ take complex conjugate values on this set.

3.2 Statement and resolution of the BVP

An important hyperbola

For further use, we need to introduce the curve

$$\mathcal{R} = \{\theta_2 \in \mathbb{C} : \gamma(\theta_1, \theta_2) = 0 \text{ and } \theta_1 \in (-\infty, \theta_1^-)\} = \Theta_2^{\pm}((-\infty, \theta_1^-)). \tag{14}$$

It is the analogue of the curve $\mathcal{M}$ in Section 2.1. The curve $\mathcal{R}$ is symmetrical w.r.t. the real axis, see Figure 4 (this is a consequence of d being negative on $(-\infty, \theta_1^-)$, see above). Furthermore, it is a (branch of a) hyperbola by Baccelli and Fayolle (1987). We shall denote by $\mathcal{G}_{\mathcal{R}}$ the open domain of $\mathbb{C}$ bounded by $\mathcal{R}$ and containing 0, see Figure 4. Obviously $\overline{\mathcal{G}_{\mathcal{R}}}$, the closure of $\mathcal{G}_{\mathcal{R}}$, is equal to $\mathcal{G}_{\mathcal{R}} \cup \mathcal{R}$.

BVP for orthogonal reflections

In the case of orthogonal reflections (see Figure 3), R is the identity matrix in (10), and we have $\gamma_1(\theta_1, \theta_2) = \theta_1$ and $\gamma_2(\theta_1, \theta_2) = \theta_2$. We set

$$\psi_1(\theta_2) = \frac{1}{\theta_2}\varphi_1(\theta_2), \qquad \psi_2(\theta_1) = \frac{1}{\theta_1}\varphi_2(\theta_1). \tag{15}$$

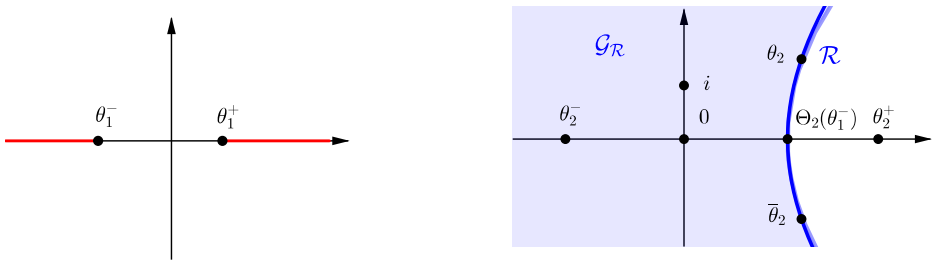


Fig. 4: Left: the discriminant d has two roots θ_1^- and θ_1^+ of opposite signs. Right: the curve $\mathcal{R}$ in (14) is symmetric w.r.t. the horizontal axis and $\mathcal{G}_\mathcal{R}$ is the domain in blue

Lemma 9 *The function ψ_1 in* (15) *satisfies the following BVP:*

(i) *ψ_1 is meromorphic on $\mathcal{G}_\mathcal{R}$ with a single pole at 0, of order 1 and residue $\varphi_1(0)$, and vanishes at infinity;*

(ii) *ψ_1 is continuous on $\overline{\mathcal{G}_\mathcal{R}} \setminus \{0\}$ and*

$$\psi_1(\overline{\theta_2}) = \psi_1(\theta_2), \qquad \forall \theta_2 \in \mathcal{R}. \tag{16}$$

Proof: The regularity condition of point (i) follows from Theorem 15, which provides a (maximal) meromorphic continuation of the function. Let us now consider (ii). Evaluating the (continued) functional equation (12) at $(\theta_1, \Theta_2^\pm(\theta_1))$, we obtain $\psi_1(\Theta_2^\pm(\theta_1)) + \psi_2(\theta_1) = 0$, which immediately implies that

$$\psi_1(\Theta_2^+(\theta_1)) = \psi_1(\Theta_2^-(\theta_1)). \tag{17}$$

Choosing $\theta_1 \in (-\infty, \theta_1^-)$, the two quantities $\Theta_2^+(\theta_1)$ and $\Theta_2^-(\theta_1)$ are complex conjugate the one of the other, see Section 3.1. Equation (17) can then be reformulated as (16), using the definition (14) of the curve $\mathcal{R}$. □

The BVP stated in Lemma 9 is called a homogeneous BVP with shift (the shift stands here for the complex conjugation, but the theory applies to more general shifts, see Litvinchuk (2000)). It has a simpler form than the BVP in Lemma 2 for the discrete case, because there is no inhomogeneous term (as $\pi_{0,0}$) and also because in the coefficients in front of the unknowns there is no algebraic function (as Y_0) involved. Due to its particularly simple form, we can solve it in an explicit way, using the two following steps:

- Using a certain conformal mapping w (to be introduced below), we can construct a particular solution to the BVP of Lemma 9.
- The solution to the BVP of Lemma 9 is unique (see the invariant Lemma 2 in (Litvinchuk, 2000, Section 10.2)). In other words, two different solutions must coincide, and the explicit solution constructed above must be the function ψ_1.

In Franceschi and Raschel (2016) it is explained that the above method may be viewed as a variation of Tutte's invariant approach, first introduced by Tutte for solving a functional equation arising in the enumeration of properly colored triangulations, see Tutte (1995).

The function w glues together the upper and lower parts of the hyperbola $\mathcal{R}$. There are at least two ways to find such a w. First, it turns out that in the literature there exist expressions for conformal gluing functions for relatively simple curves as hyperbolas, see (Baccelli and Fayolle, 1987, Equation (4.6)). Here (based on Franceschi and Raschel (2016)), we use instead the Riemann sphere $\mathcal{S}$, as we will see in Section 3.4. Indeed, many technical aspects (and in particular finding the conformal mapping) happen to be quite simpler on that surface.

We will deduce from Section 3.4 that function w can be expressed in terms of the generalized Chebyshev polynomial

$$T_a(x) = \cos(a \arccos(x)) = \frac{1}{2}\Big\{\big(x + \sqrt{x^2-1}\big)^a + \big(x - \sqrt{x^2-1}\big)^a\Big\}$$

as follows:

$$w(\theta_2) = T_{\frac{\pi}{\beta}}\bigg(-\frac{2\theta_2 - (\theta_2^+ + \theta_2^-)}{\theta_2^+ - \theta_2^-}\bigg), \tag{18}$$

where we have noted

$$\beta = \arccos -\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}. \tag{19}$$

In the case of orthogonal reflection, this methods leads to the main result of Franceschi and Raschel (2016), which is:

Theorem 10 *Let R be the identity matrix in* (10)*. The Laplace transform φ_1 is equal to*

$$\varphi_1(\theta_2) = \frac{-\mu_1 w'(0)}{w(\theta_2) - w(0)}\theta_2.$$

Statement of the BVP in the general case

We would like to close Section 3.2 by stating the BVP in the case of arbitrary reflections (non-necessarily orthogonal). Let us define for $\theta_2 \in \mathcal{R}$

$$G(\theta_2) = \frac{\gamma_1}{\gamma_2}(\Theta_1^-(\theta_2), \theta_2)\frac{\gamma_2}{\gamma_1}(\Theta_1^-(\theta_2), \overline{\theta_2}).$$

Similarly to Lemma 9, there is the following result:

Lemma 11 *The function φ_1 satisfies the following BVP:*

(i) *φ_1 is meromorphic on $\mathcal{G}_{\mathcal{R}}$ with at most one pole p of order* 1 *and is bounded at infinity;*

(ii) *φ_1 is continuous on $\overline{\mathcal{G}_{\mathcal{R}}} \setminus \{p\}$ and*

$$\varphi_1(\overline{\theta_2}) = G(\theta_2)\varphi_1(\theta_2), \qquad \forall \theta_2 \in \mathcal{R}. \tag{20}$$

Due to the presence of the function $G \neq 1$ in (20), this BVP (still homogeneous with shift) is more complicated than the one encountered in Lemma 9 and cannot be solved thanks to an invariant lemma. Instead, the resolution is less combinatorial and far more technical, and the solution should be expressed in terms of both Cauchy integrals and the conformal mapping w of Theorem 10. This will be achieved in a future work.

3.3 Asymptotics of the stationary probabilities

Overview

Let Π be a random vector that has the stationary distribution of the reflected Brownian motion. Dai and Miyazawa (2011) obtain the following asymptotic result: for a given directional vector $c \in \mathbb{R}_+^2$ they find (up to a multiplicative constant) a function $f_c(x)$ such that

$$\lim_{x\to\infty} \frac{\mathbb{P}[\langle c|\Pi\rangle \geqslant x]}{f_c(x)} = 1.$$

In Franceschi and Kurkova (2016) we solve a harder problem arisen in (Dai and Miyazawa, 2011, §8), namely computing the asymptotics of $\mathbb{P}[\Pi \in xc + B]$ as $x \to \infty$, where $c \in \mathbb{R}_+^2$ is any directional vector and $B \subset \mathbb{R}_+^2$ any compact subset. Furthermore, we are able to find the full asymptotic expansion of the density $\pi(x_1, x_2)$ of Π as $x_1, x_2 \to \infty$ and $x_2/x_1 \to \tan(\alpha)$, for any given angle $\alpha \in (0, \pi/2)$.

Main results

First we need to introduce some notations. The equation $\gamma(\theta) = 0$ determines an ellipse $\mathcal{E}$ on $\mathbb{R}^2$ passing through the origin, see Figure 5. Here we restrict ourselves to the case $\mu_1 < 0$ and $\mu_2 < 0$, although our methods can be applied without additional difficulty to other cases. For a given angle $\alpha \in [0, \pi/2]$, let us

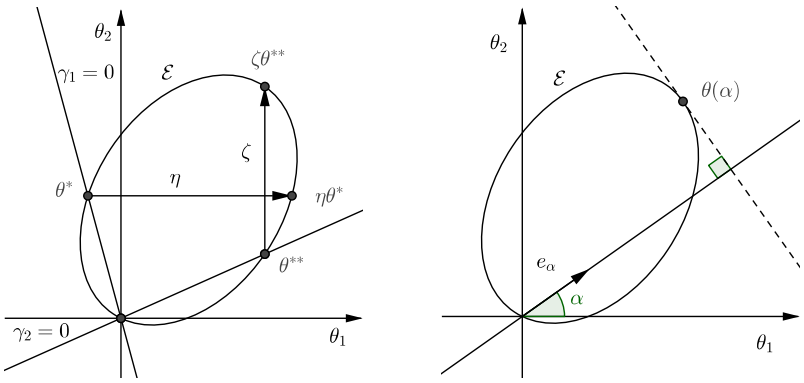


Fig. 5: Left: representation of the ellipse $\mathcal{E}$, straight lines $\{\gamma_1(\theta) = 0\}$, $\{\gamma_2(\theta) = 0\}$, and points θ^*, θ^{**}, $\eta\theta^*$ and $\zeta\theta^{**}$. Right: geometric interpretation of the point $\theta(\alpha)$ in (21) on $\mathcal{E}$

define the point $\theta(\alpha)$ on the ellipse $\mathcal{E}$ by

$$\theta(\alpha) = \mathrm{argmax}_{\theta\in\mathcal{E}} \langle\theta|e_\alpha\rangle, \qquad \text{where } e_\alpha = (\cos\alpha, \sin\alpha). \tag{21}$$

The coordinates of $\theta(\alpha)$ can be given explicitly. One can also construct $\theta(\alpha)$ geometrically as on Figure 5.

Secondly, consider the straight lines $\{\gamma_1(\theta) = 0\}$ and $\{\gamma_2(\theta) = 0\}$, depending on the reflection matrix R only. They cross the ellipse $\mathcal{E}$ at the origin. The line $\{\gamma_1(\theta) = 0\}$ (resp. $\{\gamma_2(\theta) = 0\}$) intersects the

ellipse at a second point called θ^* (resp. θ^{**}). To present our results, we need to define the images on $\mathcal{E}$ of these points via the so-called Galois automorphisms ζ and η, to be introduced in Section 3.4. Namely, for the point $\theta^* = (\theta_1^*, \theta_2^*) \in \mathcal{E}$ there exists a unique point $\eta\theta^* = (\eta\theta_1^*, \theta_2^*) \in \mathcal{E}$ with the same second coordinate. Likewise, there exists a unique point $\zeta\theta^{**} = (\theta_1^{**}, \zeta\theta_2^{**}) \in \mathcal{E}$ with the same first coordinate as $\theta^{**} = (\theta_1^{**}, \theta_2^{**}) \in \mathcal{E}$. Points θ^*, θ^{**}, $\eta\theta^*$ and $\zeta\theta^{**}$ are pictured on Figure 5. Their coordinates can be made explicit.

Similarly to the discrete case, we introduce the set of parameters

$$\mathfrak{Q}_{--} = \big\{((\Sigma, \mu, R), \alpha) : \gamma_1(\eta\theta(\alpha)) < 0 \text{ and } \gamma_2(\zeta\theta(\alpha)) < 0\big\}$$

and $\mathfrak{Q}_{+-}$, $\mathfrak{Q}_{-+}$ and $\mathfrak{Q}_{++}$ accordingly. The following theorem provides the main term in the asymptotic expansion of $\pi(r\cos\alpha, r\sin\alpha)$.

Theorem 12 *Let $(x, y) = (r\cos\alpha, r\sin\alpha)$ with $\alpha \in (0, \pi/2)$. We assume that $\theta(\alpha) \in \mathbb{R}^2_+$. Then as $r \to \infty$ we have*

$$\pi(re_\alpha) = (1 + o(1)) \cdot \begin{cases} \frac{C_0}{\sqrt{r}} e^{-r\langle e_\alpha | \theta(\alpha)\rangle} & \text{in } \mathfrak{Q}_{--}, \\ C_1 e^{-r\langle e_\alpha | \eta\theta^*\rangle} & \text{in } \mathfrak{Q}_{+-}, \\ C_2 e^{-r\langle e_\alpha | \zeta\theta^{**}\rangle} & \text{in } \mathfrak{Q}_{-+}, \\ C_1 e^{-r\langle e_\alpha | \eta\theta^*\rangle} + C_2 e^{-r\langle e_\alpha | \zeta\theta^{**}\rangle} & \text{in } \mathfrak{Q}_{++}, \end{cases} \tag{22}$$

where C_0, C_1 and C_2 are constants that can be expressed in terms of functions φ_1 and φ_2 and the parameters.

In Franceschi and Kurkova (2016) the constants mentioned in Theorem 12 are specified in terms of functions φ_1 and φ_2. But these functions are for now unknown. As we explained in Section 3.2, in a next work we are going to obtain φ_1 and φ_2 as solutions of BVP, thereby determining the constants in Theorem 12.

Proof of the key step of Theorem 12: Theorem 12 is proven in Franceschi and Kurkova (2016). The first step consists in continuing meromorphically the functions φ_1 and φ_2 on $\mathbb{C} \setminus [\theta_2^+, \infty)$ or on the Riemann surface $\mathcal{S}$, see Section 3.4. Then by the functional equation (12) and the inversion formula of Laplace transform (we refer to Doetsch (1974) and Brychkov et al. (1992)), the density $\pi(x_1, x_2)$ can be represented as a double integral. Using standard computations from complex analysis, we are able to reduce it to a sum of single integrals. We obtain the following (with the notation (13)):

$$\begin{aligned} \pi(x_1, x_2) &= \frac{-1}{(2\pi i)^2} \int_{-i\infty}^{i\infty} \int_{-i\infty}^{i\infty} e^{-x_1\theta_1 - x_2\theta_2} \frac{\gamma_1(\theta)\varphi_1(\theta_2) + \gamma_2(\theta)\varphi_2(\theta_1)}{\gamma(\theta)} \mathrm{d}\theta_1 \mathrm{d}\theta_2 \\ &= \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \varphi_2(\theta_1)\gamma_2(\theta_1, \Theta_2^+(\theta_1)) e^{-x_1\theta_1 - x_2\Theta_2^+(\theta_1)} \frac{\mathrm{d}\theta_1}{\sqrt{d(\theta_1)}} \\ &+ \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \varphi_1(\theta_2)\gamma_1(\Theta_1^+(\theta_2), \theta_2) e^{-x_1\Theta_1^+(\theta_2) - x_2\theta_2} \frac{\mathrm{d}\theta_2}{\sqrt{\widetilde{d}(\theta_2)}}. \end{aligned}$$

These integrals are typical to apply the saddle point method, see Fedoryuk (1986). The coordinates of the saddle point are the critical points of the functions

$$\cos(\alpha)\theta_1 + \sin(\alpha)\Theta_2^+(\theta_1) \quad \text{and} \quad \cos(\alpha)\Theta_1^+(\theta_2) + \sin(\alpha)\theta_2.$$

It is the point $\theta(\alpha)$. Then we have to shift the integration contour up to new contours which coincide with the steepest-descent contour near the saddle point. When we shift the contours we have to take into account the poles of the integrands and their residues. The asymptotics will be determined by the pole if we cross a pole when we shift the contour and by the saddle point otherwise. □

3.4 Riemann surface and related facts

Riemann surface

The Riemann surface

$$\mathcal{S} = \{(\theta_1, \theta_2) \in \mathbb{C}^2 : \gamma(\theta_1, \theta_2) = 0\}$$

may be viewed as the set of zeros of the kernel (equivalently, it is the Riemann surface of the algebraic functions Θ_2 and Θ_1). Due to the degree of γ, the surface $\mathcal{S}$ has genus 0 and is a Riemann sphere, i.e., homeomorphic to $\mathbb{C} \cup \{\infty\}$, see Franceschi and Kurkova (2016). It admits a very useful rational parametrization, given by

$$\theta_1(s) = \frac{\theta_1^+ + \theta_1^-}{2} + \frac{\theta_1^+ - \theta_1^-}{4}\left(s + \frac{1}{s}\right), \qquad \theta_2(s) = \frac{\theta_2^+ + \theta_2^-}{2} + \frac{\theta_2^+ - \theta_2^-}{4}\left(\frac{s}{e^{i\beta}} + \frac{e^{i\beta}}{s}\right), \quad (23)$$

with β as in (19). The equation $\gamma(\theta_1(s), \theta_2(s)) = 0$ holds and $\mathcal{S} = \{(\theta_1(s), \theta_2(s)) : s \in \mathbb{C} \cup \{\infty\}\}$.

Group of the process

We finally introduce the notion of group of the model, similar to the notion of group of the walk in the discrete setting (see Malyšev (1972); Fayolle et al. (1999); Bousquet-Mélou and Mishna (2010)). This group $\langle \zeta, \eta \rangle$ is generated by ζ and η, given by (with the notation (13))

$$\zeta(\theta_1, \theta_2) = \left(\theta_1, \frac{c(\theta_1)}{a(\theta_1)}\frac{1}{\theta_2}\right), \qquad \eta(\theta_1, \theta_2) = \left(\frac{\widetilde{c}(\theta_2)}{\widetilde{a}(\theta_2)}\frac{1}{\theta_1}, \theta_2\right).$$

By construction, the generators satisfy $\gamma(\zeta(\theta_1, \theta_2)) = \gamma(\eta(\theta_1, \theta_2)) = 0$ as soon as $\gamma(\theta_1, \theta_2) = 0$. In other words, there are (covering) automorphisms of the surface $\mathcal{S}$. Since $\zeta^2 = \eta^2 = 1$, the group $\langle \zeta, \eta \rangle$ is a dihedral group, which is finite if and only if the element $\zeta \circ \eta$ (or equivalently $\eta \circ \zeta$) has finite order.

Algebraic nature of the Laplace transforms

With the above definition, it is not clear how to see if the group is finite, nor to see it its finiteness would have any implication on the problem. In fact, we have, with β defined in (19):

Lemma 13 *The group $\langle \zeta, \eta \rangle$ is finite if and only if $\pi/\beta \in \mathbb{Q}$.*

The proof of Lemma 13 is simple, once the elements ζ and η have been lifted and reformulated on the sphere $\mathcal{S}$:

$$\zeta(s) = \frac{1}{s}, \qquad \eta(s) = \frac{e^{2i\beta}}{s}.$$

These transformations leave invariant $\theta_1(s)$ and $\theta_2(s)$, respectively, see (23). In particular, we have the following result (see Franceschi and Raschel (2016)), which connects the nature of the solution of the BVP to the finiteness of the group. Such a result holds for discrete walks, see our Theorem 7 and Bousquet-Mélou and Mishna (2010); Bernardi et al. (2015).

Theorem 14 *The solution φ_1 given in Theorem 10 and the conformal gluing function w in* (18) *are algebraic if and only if the group $\langle \zeta, \eta \rangle$ is finite.*

Conformal mapping

The conformal gluing function w introduced in Section 3.2 may be lifted on $\mathcal{S}$. In fact its expression is even simpler using the parametrization of $\mathcal{S}$. We show in Franceschi and Raschel (2016) that

$$w(\theta_2(s)) = -\frac{i}{2}\big\{(-s)^{\frac{\pi}{\beta}} + (-s)^{-\frac{\pi}{\beta}}\big\} = -\frac{i}{2}\big\{e^{\frac{\pi}{\beta}\log(-s)} + e^{-\frac{\pi}{\beta}\log(-s)}\big\}, \tag{24}$$

where we make use of the principal determination of the logarithm.

Continuation of the Laplace transforms

To establish the BVP, we have stated a boundary condition for the functions φ_1 and φ_2, on curves which lie outside their natural domains of definition (the half-plane with negative real-part), see Figure 4. In the same way, in the asymptotic study we use the steepest descent method on some curves outside of the initial domain of definition. We therefore need to extend the domain of definition of the Laplace transforms.

Theorem 15 *The function φ_1 can be continued meromorphically on the cut plane $\mathbb{C} \setminus [\theta_2^+, \infty)$.*

Proof: The first step is to continue meromorphically $\varphi_1(\theta_2)$ to the open and simply connected set $\{\theta_2 \in \mathbb{C} : \Re\,\theta_2 \leqslant 0 \text{ or } \Re\,\Theta_1^-(\theta_2) < 0\}$, by setting

$$\varphi_1(\theta_2) = \frac{\gamma_2}{\gamma_1}(\Theta_1^-(\theta_2), \theta_2)\varphi_2(\Theta_1^-(\theta_2)).$$

This is immediate (see Franceschi and Kurkova (2016) for the details). It is then possible to pursue the extension to the whole $\mathcal{S}$ using the invariance properties by the automorphisms ζ and η satisfied by the lifted Laplace transforms on $\mathcal{S}$. □

Acknowledgements

We acknowledge support from the projet MADACA of the Région Centre and from Simon Fraser University (British Columbia, Canada). We thank our three AofA referees for useful comments.

References

Baccelli, F. and Fayolle, G. (1987). Analysis of models reducible to a class of diffusion processes in the positive quarter plane. *SIAM J. Appl. Math.*, 47(6):1367–1385.

Bernardi, O., Bousquet-Mélou, M., and Raschel, K. (2015). Counting quadrant walks via Tutte's invariant method. *Preprint arXiv:1511.04298 (to appear in the proceedings of FPSAC 2016)*, pages 1–13.

Biane, P. (1992). Minuscule weights and random walks on lattices. In *Quantum probability & related topics*, QP-PQ, VII, pages 51–65. World Sci. Publ., River Edge, NJ.

Bostan, A. and Kauers, M. (2010). The complete generating function for Gessel walks is algebraic. *Proc. Amer. Math. Soc.*, 138(9):3063–3078. With an appendix by Mark van Hoeij.

Bousquet-Mélou, M. and Mishna, M. (2010). Walks with small steps in the quarter plane. In *Algorithmic probability and combinatorics*, volume 520 of *Contemp. Math.*, pages 1–39. Amer. Math. Soc., Providence, RI.

Brychkov, Y., Glaeske, H.-J., Prudnikov, A., and Tuan, V. K. (1992). *Multidimensional Integral Transformations*. CRC Press.

Burdzy, K., Chen, Z.-Q., Marshall, D., and Ramanan, K. (2015). Obliquely reflected Brownian motion in non-smooth planar domains. *Preprint arXiv:1512.02323*, pages 1–60.

Cont, R. and de Larrard, A. (2013). Price dynamics in a Markovian limit order market. *SIAM J. Financial Math.*, 4(1):1–25.

Dai, J. (1990). *Steady-state analysis of reflected Brownian motions: Characterization, numerical methods and queueing applications*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University.

Dai, J. and Harrison, J. (1992). Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *Ann. Appl. Probab.*, 2(1):65–86.

Dai, J. and Miyazawa, M. (2011). Reflecting Brownian motion in two dimensions: Exact asymptotics for the stationary distribution. *Stoch. Syst.*, 1(1):146–208.

Dieker, A. and Moriarty, J. (2009). Reflected Brownian motion in a wedge: sum-of-exponential stationary densities. *Electron. Commun. Probab.*, 14:1–16.

Doetsch, G. (1974). *Introduction to the Theory and Application of the Laplace Transformation*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Dubédat, J. (2004). Reflected planar Brownian motions, intertwining relations and crossing probabilities. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(5):539–552.

Dupuis, P. and Williams, R. (1994). Lyapunov functions for semimartingale reflecting Brownian motions. *Ann. Probab.*, 22(2):680–702.

Fayolle, G. and Iasnogorodski, R. (1979). Two coupled processors: the reduction to a Riemann-Hilbert problem. *Z. Wahrsch. Verw. Gebiete*, 47(3):325–351.

Fayolle, G., Iasnogorodski, R., and Malyshev, V. (1999). *Random Walks in the Quarter-Plane*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Fedoryuk, M. (1986). Asymptotic methods in analysis. In *Current problems of mathematics. Fundamental directions, Vol. 13 (Russian)*, Itogi Nauki i Tekhniki, pages 93–210. Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Inform., Moscow.

Foddy, M. (1984). *Analysis of Brownian motion with drift, confined to a quadrant by oblique reflection (diffusions, Riemann-Hilbert problem)*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Stanford University.

Franceschi, S. and Kurkova, I. (2016). Asymptotic expansion of the stationary distribution for reflected Brownian motion in the quarter plane via analytic approach. *Preprint arXiv:1604.02918*, pages 1–45.

Franceschi, S. and Raschel, K. (2016). Tutte's invariant approach for Brownian motion reflected in the quadrant. *Preprint arXiv:1602.03054*, pages 1–14.

Harrison, J. and Hasenbein, J. (2009). Reflected Brownian motion in the quadrant: tail behavior of the stationary distribution. *Queueing Syst.*, 61(2-3):113–138.

Harrison, J. M. and Williams, R. J. (1987a). Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, 22(2):77–115.

Harrison, J. M. and Williams, R. J. (1987b). Multidimensional reflected Brownian motions having exponential stationary distributions. *The Annals of Probability*, 15(1):115–137.

Hobson, D. and Rogers, L. (1993). Recurrence and transience of reflecting Brownian motion in the quadrant. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 113, pages 387–399. Cambridge Univ Press.

Kurkova, I. and Malyshev, V. (1998). Martin boundary and elliptic curves. *Markov Process. Related Fields*, 4(2):203–272.

Kurkova, I. and Raschel, K. (2011). Random walks in $(\mathbb{Z}_+)^2$ with non-zero drift absorbed at the axes. *Bull. Soc. Math. France*, 139(3):341–387.

Kurkova, I. and Raschel, K. (2012). On the functions counting walks with small steps in the quarter plane. *Publ. Math. Inst. Hautes Études Sci.*, 116:69–114.

Kurkova, I. and Suhov, Y. (2003). Malyshev's theory and JS-queues. Asymptotics of stationary probabilities. *Ann. Appl. Probab.*, 13(4):1313–1354.

Le Gall, J.-F. (1987). Mouvement brownien, cônes et processus stables. *Probab. Theory Related Fields*, 76(4):587–627.

Litvinchuk, G. (2000). *Solvability Theory of Boundary Value Problems and Singular Integral Equations with Shift*. Springer Netherlands, Dordrecht.

Malyšev, V. (1972). An analytic method in the theory of two-dimensional positive random walks. *Sibirsk. Mat. Ž.*, 13:1314–1329, 1421.

Malyšev, V. (1973). Asymptotic behavior of the stationary probabilities for two-dimensional positive random walks. *Sibirsk. Mat. Ž.*, 14:156–169, 238.

Tutte, W. (1995). Chromatic sums revisited. *Aequationes Math.*, 50(1-2):95–134.

Williams, R. (1995). Semimartingale reflecting Brownian motions in the orthant. In *Stochastic networks*, volume 71 of *IMA Vol. Math. Appl.*, pages 125–137. Springer, New York.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Dependence between External Path-Length and Size in Random Tries

M. Fuchs[1†] and H.-K. Hwang[2]

[1] *Department of Applied Mathematics, National Chiao Tung University, Hsinchu, 300, Taiwan*
[2] *Institute of Statistical Science, Academia Sinica, Taipei, 115, Taiwan*

Abstract.

We study the size and the external path length of random tries and show that they are asymptotically independent in the asymmetric case but strongly dependent with small periodic fluctuations in the symmetric case. Such an unexpected behavior is in sharp contrast to the previously known results that the internal path length is totally positively correlated to the size and that both tend to the same normal limit law. These two examples provide concrete instances of bivariate normal distributions (as limit laws) whose correlation is 0, 1 and periodically oscillating.

Keywords: Random tries, Pearson's correlation coefficient, asymptotic normality, Poissonization/de-Poissonization, Mellin transform, contraction method

1 Introduction

Tries are one of the most fundamental tree-type data structures in computer algorithms. Their general efficiency depends on several shape parameters, the principal ones including the depth, the height, the size, the internal path-length (IPL), and the external path-length (EPL); see below for a more precise description of those studied in this paper. While most of these measures have been extensively investigated in the literature, we are concerned here with the question: *how does the EPL depend on the size in a random trie?* Surprisingly, while the IPL and the size are known to have asymptotic correlation coefficient tending to one and to have the same normal limit law after each being properly normalized (see [4, 6]), this paper aims to show that the EPL exhibits a completely different behavior depending on the parameter of the underlying random bits being biased or unbiased. This is a companion paper to [1].

Given a sequence of binary strings (or keys), one can construct a (binary) trie as follows. If $n = 1$, then the trie consists of a single root-node holding the sole string; otherwise, the root is used to direct the strings into the corresponding subtree: if the first bit of the input string is 0 (or 1), then the string goes to the left (or right) subtree; strings going to the same subtree are then constructed recursively in the same manner but instead of splitting according to the first bit, the second bit of each string is then used. In this way, a binary dictionary-type tree with two types of nodes is constructed: external nodes for storing strings and internal nodes for splitting the strings; see Figure 1 for a trie of seven strings.

[†] Partially supported by MOST under the grant MOST-104-2923-M-009-006-MY3

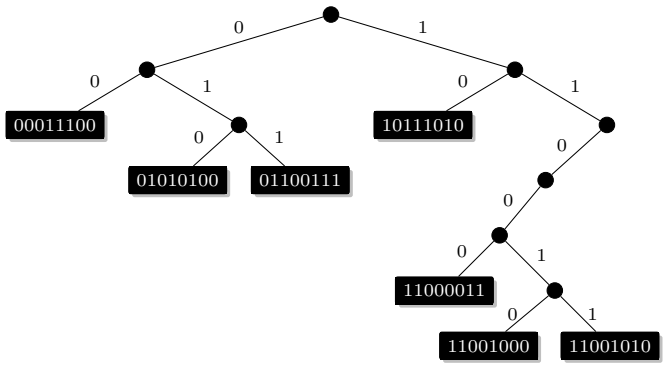


Fig. 1: *A trie with $n = 7$ records: the (filled) circles represent internal nodes and rectangles holding the binary strings are external nodes. In this example, $S_n = 8$, $K_n = 27$, and $N_n = 18$.*

The random trie model we consider here assumes that each of the n binary keys is an infinite sequence consisting of independent Bernoulli bits each with success probability $0 < p < 1$. Then the trie constructed from this sequence is a random trie. We define three shape parameters in a random trie of n strings:

- size S_n: the total number of internal nodes used;
- IPL (or node path-length, NPL) N_n: the sum of the distances between the root to each internal node;
- EPL (or key path-length, KPL) K_n: the sum of the distances between the root to each external node.

We will use mostly NPL in place of IPL, and KPL in place of EPL, the reason being an easier comparison with the corresponding results in random m-ary search trees in the companion paper [1]; see below for more details.

By the recursive definition, we have the following recurrence relations

$$\begin{cases} S_n \stackrel{d}{=} S_{B_n} + S^*_{n-B_n} + 1, \\ K_n \stackrel{d}{=} K_{B_n} + K^*_{n-B_n} + n, \\ N_n \stackrel{d}{=} N_{B_n} + N^*_{n-B_n} + S_{B_n} + S^*_{n-B_n}, \end{cases} \qquad (n \geq 2), \tag{1}$$

where $B_n = \text{Binom}(n, p)$ and $S_0 = S_1 = K_0 = K_1 = N_0 = N_1 = 0$. Here $(S^*_n), (K^*_n)$, and (N^*_n) are independent copies of $(S_n), (K_n)$ and (N_n), respectively. While many stochastic properties of these random variables are known (see [4] and the references therein), much less attention has been paid to their correlation and dependence structure.

The asymptotic behaviors of the moments of random variables defined on tries typically depend on the ratio $\frac{\log p}{\log q}$ being rational or irrational, where $q = 1 - p$. So we introduce, similar to [4], the notation

$$\mathscr{F}[g](z) = \begin{cases} \sum_{k\in\mathbb{Z}} g_k z^{-\chi_k}, & \text{if } \frac{\log p}{\log q} \in \mathbb{Q}; \\ g_0, & \text{if } \frac{\log p}{\log q} \notin \mathbb{Q}, \end{cases} \tag{2}$$

where g_k represents a sequence of coefficients and $\chi_k = \frac{2rk\pi i}{\log p}$ when $\frac{\log p}{\log q} = \frac{r}{l}$ with r and l coprime. In simpler words, $\mathscr{F}[g](z)$ is a periodic function in the rational case, and a constant in the irrational case. We also use $\mathscr{F}[\cdot](z)$ as a generic symbol if the exact form of underlying sequence matters less, and in this case each occurrence may not represent the same function.

With this notation, the asymptotics of the mean and the variance of the above three shape parameters are summarized in the following table; see [4] and the references therein for more information.

Shape parameters	$\frac{1}{n}$(mean) $\sim$	$\frac{1}{n}$(variance) $\sim$
Size S_n	$\mathscr{F}[\cdot](n)$	$\mathscr{F}[g^{(1)}](n)$
NPL N_n	$\frac{\mathbb{E}(S_n)}{n} \cdot \frac{\log n}{h}$	$\frac{\mathbb{V}(S_n)}{n} \cdot \frac{(\log n)^2}{h^2}$
KPL K_n	$\frac{\log n}{h} + \mathscr{F}[\cdot](n)$	$\frac{pq \log^2 \frac{p}{q}}{h^2} \cdot \frac{\log n}{h} + \mathscr{F}[g^{(3)}](n)$
Depth D_n	$\mathbb{E}(D_n) = \frac{\mathbb{E}(K_n)}{n}$	$\mathbb{V}(D_n) = \frac{\mathbb{V}(K_n)}{n} + O(1)$

Tab. 1: *Asymptotic patterns of the means and the variances of the shape parameters discussed in this paper. Here $\mathscr{F}[\cdot](n)$ differs from one occurrence to another and $h = -p\log p - q\log q$ denotes the entropy. Expressions for $g_k^{(1)}$ and $g_k^{(3)}$ will be given below. Asymptotic normality holds for all three random variables S_n, N_n, K_n.*

Note specially that the leading constant

$$\lambda = \lambda_p := \frac{pq\log^2 \frac{p}{q}}{h^3} = \frac{(p\log^2 p + q\log^2 q) - h^2}{h^3}$$

in the asymptotic approximation to $\mathbb{V}(K_n)$ equals zero when $p = q$, implying that $\mathbb{V}(K_n)$ is not of order $n\log n$ but of linear order in the symmetric case. *This change of order can be regarded as the source property distinguishing the dependence and independence of K_n on S_n.*

On the other hand, if we denote by D_n the depth, which is defined to be the distance between the root and a randomly chosen external node (each with the same probability), then we have not only the relation $\mathbb{E}(D_n)n = \mathbb{E}(K_n)$, but also the asymptotic equivalent $\mathbb{V}(D_n)n \sim \mathbb{V}(K_n)$ when $p \neq 1/2$ (or $\lambda > 0$), and a central limit theorem holds; see Devroye [2].

From Table 1, we see roughly that each internal node contributes $\frac{\log n}{h}$ to N_n, namely, that $N_n \approx S_n \cdot \frac{\log n}{h}$. Indeed, it was proved in [4] that the correlation coefficient of S_n and N_n satisfies

$$\rho(S_n, N_n) \sim 1 \qquad (0 < p < 1). \tag{3}$$

Such a linear correlation was further strengthened in [6], where it was proved that both random variables tend to the *same* normal limit law $\mathcal{N}_1$ (with zero mean and unit variance)

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}}, \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathbb{V}(N_n)}}\right) \xrightarrow{d} (\mathcal{N}_1, \mathcal{N}_1),$$

where $\xrightarrow{d}$ denotes convergence in distribution. In terms of the bivariate normal law $\mathcal{N}_2$ (see Tong [16]), we can write

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}}, \frac{N_n - \mathbb{E}(N_n)}{\sqrt{\mathbb{V}(N_n)}}\right)^{\intercal} \xrightarrow{d} \mathcal{N}_2(0, E_2),$$

where $E_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is a singular matrix and $\mathbf{A}^\intercal$ denotes the transpose of matrix $\mathbf{A}$.

We show that the correlation and dependence of K_n on S_n are drastically different. We start with their correlation coefficient.

Theorem A *The covariance of the number of internal nodes and KPL in a random trie of n strings satisfies*

$$\operatorname{Cov}(S_n, K_n) \sim n\mathscr{F}[g^{(2)}](n),$$

where $g_k^{(2)}$ is given in Proposition A below, and their correlation coefficient satisfies

$$\rho(S_n, K_n) \sim \begin{cases} 0, & \text{if } p \neq \frac{1}{2} \\ F(n), & \text{if } p = \frac{1}{2}. \end{cases} \tag{4}$$

Here $F(n) = \frac{\mathscr{F}[g^{(2)}](n)}{\sqrt{\mathscr{F}[g^{(1)}](n)\mathscr{F}[g^{(3)}](n)}}$ is a periodic function with average value $0.927\cdots$.

The result (4) is to be compared with (3) (which holds for all $p \in (0,1)$): *the surprising difference here comes not only from the (common) distinction between $p = \frac{1}{2}$ and $p \neq \frac{1}{2}$ but also from the (less expected) intrinsic asymptotic nature.*

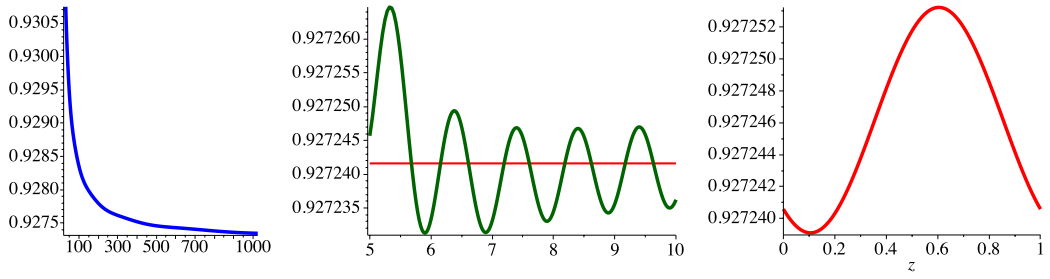


Fig. 2: *$p = \frac{1}{2}$: periodic fluctuations of (i) $\rho(S_n, K_n)$ (left) for $n = 32, \ldots, 1024$, (ii) $\frac{\operatorname{Cov}(S_n,K_n)}{\sqrt{\mathbb{V}(S_n)(\mathbb{V}(K_n)+1.046)}}$ (middle) in logarithmic scale, and (iii) $F(n)$ by its Fourier series expansion (right). Note that the fluctuations are only visible by proper corrections either in the denominator or in the numerator because the amplitude of F is very small: $|F(\cdot)| \leq 1.5 \times 10^{-5}$.*

Furthermore, we show that this different behavior cannot be ascribed to the weak measurability of nonlinear dependence of Pearson's correlation coefficient ρ since the same dependence is also present in the limiting distribution. (For the univariate central limit theorems implied by the result below, see Jacquet and Régnier [8] where such results were first established.)

Theorem B *(i) For $p \neq \frac{1}{2}$, we have*

$$\left(\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\mathbb{V}(S_n)}}, \frac{K_n - \mathbb{E}(K_n)}{\sqrt{\mathbb{V}(K_n)}} \right)^\intercal \xrightarrow{d} \mathcal{N}_2(0, I_2),$$

where I_2 denotes the 2×2 identity matrix.

(ii) For $p = \frac{1}{2}$, we have

$$\Sigma_n^{-\frac{1}{2}} \begin{pmatrix} S_n - \mathbb{E}(S_n) \\ K_n - \mathbb{E}(K_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, I_2),$$

where Σ_n denotes the (asymptotic) covariance matrix of S_n and K_n:

$$\Sigma_n := n \begin{pmatrix} \mathscr{F}[g^{(1)}](n) & \mathscr{F}[g^{(2)}](n) \\ \mathscr{F}[g^{(2)}](n) & \mathscr{F}[g^{(3)}](n) \end{pmatrix}.$$

Alternatively, we may define $\Sigma_n := n \begin{pmatrix} \mathscr{F}[g^{(1)}](n) & \mathscr{F}[g^{(2)}](n) \\ \mathscr{F}[g^{(2)}](n) & \lambda \log n + \mathscr{F}[g^{(3)}](n) \end{pmatrix}$. Then both cases can be stated in one as $\Sigma_n^{-\frac{1}{2}} \begin{pmatrix} S_n - \mathbb{E}(S_n) \\ K_n - \mathbb{E}(K_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, I_2)$. On the other hand, since for bivariate normal distribution, zero correlation implies independence (see [16]), it is more transparent to split the statement into two cases. See Figure 3 for 3D-plots of the joint distributions of (S_n, K_n) when $n = 10^7$.

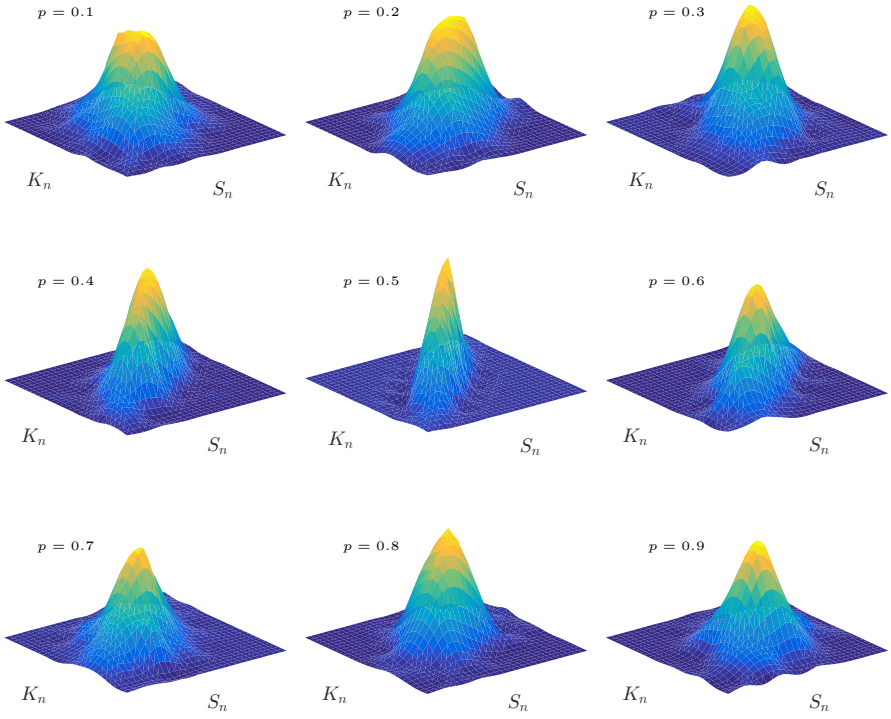


Fig. 3: *Joint distributions of (S_n, K_n) by Monte-Carlo simulations for $n = 10^7$ and varying p: the case $p = 0.5$ is seen to have stronger dependence than the others.*

These results are to be compared with the corresponding ones for random m-ary search trees [1], and the differences for correlation coefficients are summarized in Table 2. Furthermore, the joint distribution for

trees	$\rho(S_n, K_n)$	$\rho(S_n, N_n)$
tries	$\begin{cases} p \neq q :\to 0 \\ p = q : \text{periodic} \end{cases}$	~ 1
m-ary search trees	$\begin{cases} 3 \leq m \leq 26 :\to 0 \\ m \geq 27 : \text{periodic} \end{cases}$	

Tab. 2: *A comparison of the correlation coefficients for random tries and random m-ary search trees: the size of m-ary search trees corresponds to the space requirement, and the KPL and NPL are defined similarly as in tries.*

m-ary search trees undergoes a phase change at $m = 26$: if the branching factor m satisfies $3 \leq m \leq 26$, then the space requirement is asymptotically independent from KPL and NPL, while for $m \geq 27$, their limiting joint distributions contain periodic fluctuations and are dependent; see [1] for more information.

Finally, similar results as those in this paper also hold for other digital families of trees, but for simplicity we focus on tries in this paper; see [7, 4] for more references.

2 Covariance and Correlation Coefficient

In this section, we sketch the main ideas leading to the proof of Theorem A on the asymptotics of the covariance and correlation coefficient of S_n and K_n. For the latter, we also need the variances of S_n and K_n which have been known for a long time; see Jacquet and Régnier [8], Kirschenhofer and Prodinger [10], Kirschenhofer et al. [11], Régnier and Jacquet [14] or the recent paper [4]. (See also Table 1 for a summary of these results.)

Our method of proof is based on the by-now standard two-stage approach relying on the theory of analytic de-Poissonization and Mellin transform whose origin can be traced back to Jacquet and Régnier [8]. See Flajolet et al. [3] for a survey on Mellin transform, and Jacquet and Szpankowski [9] for a survey on analytic de-Poissonization. For the computation of the covariance, the manipulation can be largely simplified by the additional notions of Poissonized variance and admissible functions further developed in our previous papers [4, 7].

The starting point of our analysis is the recurrence satisfied by S_n and K_n in (1). A standard means in the computation of moments of S_n and K_n is the Poisson generating function, which corresponds to the moments of S_n and K_n with n replaced by a Poisson random variable with parameter z (this step is called *Poissonization*).

More precisely, define the Poisson generating function of $\mathbb{E}(S_n)$ and that of $\mathbb{E}(K_n)$: $\tilde{f}_{1,0}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(S_n)\frac{z^n}{n!}$ and $\tilde{f}_{0,1}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(K_n)\frac{z^n}{n!}$. Then the recurrences (1) lead to the functional equations

$$\begin{cases} \tilde{f}_{1,0}(z) = \tilde{f}_{1,0}(pz) + \tilde{f}_{1,0}(qz) + 1 - (1+z)e^{-z}, \\ \tilde{f}_{0,1}(z) = \tilde{f}_{0,1}(pz) + \tilde{f}_{0,1}(qz) + z(1-e^{-z}). \end{cases} \tag{5}$$

From these equations, we obtain, by Mellin transform techniques [3],

$$\tilde{f}_{1,0}(z) \sim z\mathscr{F}[\cdot](z), \qquad \text{and} \qquad \tilde{f}_{0,1}(z) \sim h^{-1}z\log z + z\mathscr{F}[\cdot](z), \tag{6}$$

for large $|z|$ in the half-plane $\Re(z) \geq \varepsilon > 0$, where h denotes the entropy of Bernoulli(p). Then, by Cauchy's integral representation and analytic de-Poissonization techniques [9], we obtain precise asymptotic approximations to $\mathbb{E}(S_n)$ and to $\mathbb{E}(K_n)$.

Similarly, for the variances $\mathbb{V}(S_n)$ and $\mathbb{V}(K_n)$, we introduce the Poisson generating functions of the second moments: $\tilde{f}_{2,0}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(S_n^2)\frac{z^n}{n!}$ and $\tilde{f}_{0,2}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(K_n^2)\frac{z^n}{n!}$, which then satisfy, by (1), the same type of functional equations as in (5) but with different non-homogeneous parts. Instead of computing directly asymptotic approximations to the second moments, it proves computational more advantageous to consider the Poissonized variances

$$\begin{cases}\tilde{V}_S(z) := \tilde{f}_{2,0}(z) - \tilde{f}_{1,0}(z)^2 - z\tilde{f}'_{1,0}(z)^2,\\ \tilde{V}_K(z) := \tilde{f}_{0,2}(z) - \tilde{f}_{0,1}(z)^2 - z\tilde{f}'_{0,1}(z)^2,\end{cases} \tag{7}$$

and then following the same Mellin-de-Poissonization approach (as for the means) to derive the first and the third asymptotic estimate in the second column of Table 1. It remains to derive the claimed estimate for the covariance. For that purpose, we then introduce the Poisson generating function $\tilde{f}_{1,1}(z) := e^{-z}\sum_{n\geq 0}\mathbb{E}(S_nK_n)\frac{z^n}{n!}$, which satisfies, again by (1),

$$\begin{aligned}\tilde{f}_{1,1}(z) = \tilde{f}_{1,1}(pz) + \tilde{f}_{1,1}(qz) + \tilde{f}_{1,0}(pz)\big(\tilde{f}_{0,1}(qz)+z\big) + \tilde{f}_{1,0}(qz)\big(\tilde{f}_{0,1}(pz)+z\big)\\ + pz\tilde{f}'_{1,0}(pz) + qz\tilde{f}'_{1,0}(qz) + \tilde{f}_{0,1}(pz) + \tilde{f}_{0,1}(qz) + z(1-e^{-z}).\end{aligned}$$

To compute the covariance, it is beneficial to introduce now the *Poissonized covariance* (see (7) or [4] for similar details)

$$\tilde{C}(z) = \tilde{f}_{1,1}(z) - \tilde{f}_{1,0}(z)\tilde{f}_{0,1}(z) - z\tilde{f}'_{1,0}(z)\tilde{f}'_{0,1}(z),$$

which satisfies

$$\tilde{C}(z) = \tilde{C}(pz) + \tilde{C}(qz) + \tilde{h}_1(z) + \tilde{h}_2(z), \tag{8}$$

where

$$\tilde{h}_1(z) = pqz\big(\tilde{f}'_{1,0}(pz) - \tilde{f}'_{1,0}(qz)\big)\big(\tilde{f}'_{0,1}(pz) - \tilde{f}'_{0,1}(qz)\big),$$

and

$$\begin{aligned}\tilde{h}_2(z) = ze^{-z}\big(\tilde{f}_{1,0}(pz) + \tilde{f}_{1,0}(qz) + p(1-z)\tilde{f}'_{1,0}(pz) + q(1-z)\tilde{f}'_{1,0}(qz)\big)\\ + e^{-z}\big((1+z)\tilde{f}_{0,1}(pz) + (1+z)\tilde{f}_{0,1}(qz) - pz^2\tilde{f}'_{0,1}(pz) - qz^2\tilde{f}'_{0,1}(qz)\big)\\ + ze^{-z}\big(1-(1+z^2)e^{-z}\big).\end{aligned}$$

Note that $\tilde{h}_1$ is zero when $p=\frac{1}{2}$. Furthermore, from (6) (which can be differentiated since they hold in a sector $\mathscr{S} = \{z\in\mathbb{C} : \Re(z)\geq\epsilon, |\mathrm{Arg}(z)|\leq\theta_0\}$ with $0<\theta_0<\pi/2$ in the complex plane), we obtain that $\tilde{h}_1(z) = O(|z|)$ and $\tilde{h}_2(z)$ is exponentially small for large $|z|$ in $\Re(z)>0$. Also $\tilde{h}_1(z)+\tilde{h}_2(z) = O(|z|^2)$ as $z\to 0$. Thus the Mellin transform of $\tilde{h}_1(z)+\tilde{h}_2(z)$ exists in the strip $\langle -2, 0\rangle$, and we have then the inverse Mellin integral representation

$$\tilde{C}(z) = \frac{1}{2\pi i}\int_{-\frac{3}{2}-i\infty}^{-\frac{3}{2}+i\infty}\frac{\mathscr{M}[\tilde{h}_1(z)+\tilde{h}_2(z);s]}{1-p^{-s}-q^{-s}}\,z^{-s}\mathrm{d}s,$$

where $\mathscr{M}[\phi(z);s] := \int_0^\infty \phi(z)z^{s-1}\mathrm{d}z$ denotes the Mellin transform of ϕ.

We then show that $\mathscr{M}[\tilde{h}_1(z);s]$ can be analytically continued to the vertical line $\Re(s)=-1$ and has no singularities there. This is the most complicated part of the proof because $\tilde{h}_1(z)$ contains the product of the two terms $\tilde{f}'_{1,0}(pz)-\tilde{f}'_{1,0}(qz)$ and $\tilde{f}'_{0,1}(pz)-\tilde{f}'_{0,1}(qz)$ and thus $\mathscr{M}[\tilde{h}_1(z);s]$ becomes a Mellin convolution integral. In [4], a general procedure was given for the simplification of such integrals (see [4, p. 24 *et seq.*]). This simplification procedure and a direct application of the theory of admissible functions of analytic de-Poissonization now yield

Proposition A *The covariance of S_n and K_n is asymptotically linear:*

$$\operatorname{Cov}(S_n,K_n)\sim n\mathscr{F}[g^{(2)}](n).$$

Here

$$\begin{aligned}g_k^{(2)}&=\frac{\Gamma(\chi_k)}{h}\Big(1-\frac{\chi_k+2}{2^{\chi_k+1}}\Big)-\frac{1}{h^2}\sum_{j\in\mathbb{Z}\setminus\{0\}}\Gamma(\chi_{k-j}+1)(\chi_j-1)\Gamma(\chi_j)\\&\quad-\frac{\Gamma(\chi_k+1)}{h^2}\Big(\gamma+1+\psi(\chi_k+1)-\frac{p\log^2p+q\log^2q}{2h}\Big)\\&\quad+\frac{1}{h}\sum_{\ell\ge2}\frac{(-1)^\ell(p^\ell+q^\ell)}{\ell!(1-p^\ell-q^\ell)}\Gamma(\chi_k+\ell-1)(2\ell^2-2\ell+1+\chi_k(2\ell-1)),\end{aligned}\tag{9}$$

where γ denotes Euler's constant, $\psi(z)$ is the digamma function and χ_k is defined in (2)*.*

Remark 1 *If $\frac{\log p}{\log q}\notin\mathbb{Q}$, then only $k=0$ is relevant and the second term (the sum over j) on the right-hand side of* (9) *has to be dropped. Also the first term here $\frac{\Gamma(\chi_k)}{h}\big(1-\frac{\chi_k+2}{2^{\chi_k+1}}\big)$ is taken to be its limit $\frac{1}{h}(\log 2+\frac12)$ as $\chi_k\to0$ when $k=0$.*

The asymptotic estimate for the correlation coefficient in Theorem A now follows from this and the results for the the variances of S_n and K_n (see Table 1), where expressions for $g_k^{(1)}$ and $g_k^{(3)}$ can be found, e.g., in [4]. For convenience, we give below the expressions in the unbiased case. Note that both $\mathscr{F}[g^{(1)}](n)$ and $\mathscr{F}[g^{(3)}](n)$ are strictly positive; see Schachinger [15] for details.

When $p=\frac12$, an alternative expression to (9) (avoiding the convolution of two Fourier series) is

$$g_k^{(2)}=\frac{\Gamma(\chi_k)\Big(1-\frac{\chi_k^2+\chi_k+4}{2^{\chi_k+2}}\Big)}{\log 2}+\frac{1}{\log 2}\sum_{\ell\ge1}\frac{(-1)^\ell\Gamma(\chi_k+\ell)\big(\ell(2\ell+1)(\chi_k+\ell)-(\ell+1)^2\big)}{(\ell+1)!(2^\ell-1)};$$

see the discussion of the size of tries in [4], where a similar alternative expression was given for $g_k^{(1)}$, which reads

$$g_k^{(1)}=-\frac{\Gamma(\chi_k-1)\chi_k(\chi_k+1)^2}{4\log 2}+\frac{2}{\log 2}\sum_{\ell\ge1}\frac{(-1)^\ell\Gamma(\chi_k+\ell)\ell\big(\ell(\chi_k+\ell)-1\big)}{(\ell+1)!(2^\ell-1)}.$$

Moreover, also in [4], the following expression for $g_k^{(3)}$ can be found

$$g_k^{(3)}=\frac{\Gamma(\chi_k)\Big(1-\frac{\chi_k^2-\chi_k+4}{2^{\chi_k+2}}\Big)}{\log 2}+\frac{2}{\log 2}\sum_{\ell\ge1}\frac{(-1)^\ell\Gamma(\chi_k+\ell)(\ell(\chi_k+\ell-1)-1)}{\ell!(2^\ell-1)}.$$

Note that $\chi_k = \frac{2k\pi i}{\log 2}$ and $2^{\chi_k} = 1$, and the reason of retaining 2^{χ_k+2} in the denominator is to give a uniform expression for all k (notably $k = 0$). These provide an explicit expression for the periodic function $F(n)$ in Theorem A. Also, since all the periodic functions have very small amplitude, the average value of the periodic function $F(z)$ can be well-approximated by

$$\frac{g_0^{(2)}}{\sqrt{g_0^{(1)} g_0^{(3)}}} \approx 0.9272416035\cdots .$$

3 Limit Law

In this section, we prove Theorem B, part (i); the proof of part (ii) is similar and skipped here. The key tool of the proof is the multivariate version of the contraction method; see Neininger and Rüschendorf [13]. More precisely, we will use Theorem 3.1 in [13].

We first recall the expression for the square-root of a positive-definite 2×2 matrix $M = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$. It is well-known that such a matrix has exactly one positive-definite square root which is given by

$$M^{\frac{1}{2}} = \frac{1}{\sqrt{a + c + 2\sqrt{ac - b^2}}} \begin{pmatrix} a + \sqrt{ac - b^2} & b \\ b & c + \sqrt{ac - b^2} \end{pmatrix},$$

with the inverse

$$M^{-\frac{1}{2}} = \frac{1}{\sqrt{(ac - b^2)\big(a + c + 2\sqrt{ac - b^2}\big)}} \begin{pmatrix} c + \sqrt{ac - b^2} & -b \\ -b & a + \sqrt{ac - b^2} \end{pmatrix}. \tag{10}$$

Now we sketch the proof of Theorem B, Part (i).

Proof of Theorem B, Part (i). First note that

$$\begin{pmatrix} S_n \\ K_n \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S_{B_n} \\ K_{B_n} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S^*_{n-B_n} \\ K^*_{n-B_n} \end{pmatrix} + \begin{pmatrix} 1 \\ n \end{pmatrix},$$

where the notation is as in Section 1. The contraction method was specially developed for obtaining limiting distribution results for such recurrences; see [13].

We need some notation. First, define

$$\widehat{\Sigma}_n := \begin{pmatrix} \mathbb{V}(S_n) & \mathrm{Cov}(S_n, K_n) \\ \mathrm{Cov}(S_n, K_n) & \mathbb{V}(K_n) \end{pmatrix}. \tag{11}$$

This matrix is clearly positive-definite for all n sufficiently large. Next define

$$M_n^{(1)} := \widehat{\Sigma}_n^{-\frac{1}{2}} \widehat{\Sigma}_{B_n}^{\frac{1}{2}}, \qquad M_n^{(2)} := \widehat{\Sigma}_n^{-\frac{1}{2}} \widehat{\Sigma}_{n-B_n}^{\frac{1}{2}}$$

and

$$\begin{pmatrix} b_n^{(1)} \\ b_n^{(2)} \end{pmatrix} = \widehat{\Sigma}_n^{-\frac{1}{2}} \begin{pmatrix} 1 - \mu(n) + \mu(B_n) + \mu(n - B_n) \\ n - \nu(n) + \nu(B_n) + \nu(n - B_n) \end{pmatrix},$$

where $\mu(n) = \mathbb{E}(S_n)$ and $\nu(n) = \mathbb{E}(K_n)$.

Now to apply the contraction method in [13], it suffices to show that the following conditions hold

$$b_n^{(i)} \xrightarrow{L_3} 0, \qquad M_n^{(i)} \xrightarrow{L_3} M_i, \tag{12}$$

$$\mathbb{E}\big(\|M_1\|_{\text{op}}^3 + \|M_2\|_{\text{op}}^3\big) < 1, \qquad \mathbb{E}\big(\|M_n^{(i)}\|_{\text{op}}^3 \chi_{\{B_n^{(i)} \le j\} \cup \{B_n^{(i)} = n\}}\big) \longrightarrow 0 \tag{13}$$

for $i = 1, 2$ and $j \in \mathbb{N}$, where $\xrightarrow{L_3}$ denotes convergence in the L_3-norm, $\|\cdot\|_{\text{op}}$ is the operator norm, χ_S denotes the characteristic function of set S, $B_n^{(1)} = B_n$, $B_n^{(2)} = n - B_n$ and

$$M_1 = \begin{pmatrix} \sqrt{p} & 0 \\ 0 & \sqrt{p} \end{pmatrix}, \qquad M_2 = \begin{pmatrix} \sqrt{q} & 0 \\ 0 & \sqrt{q} \end{pmatrix}.$$

Then the contraction method in [13] guarantees that (S_n, K_n) (centralized and normalized) converges in distribution to the unique fixed-point with mean 0, covariance matrix the unity matrix and finite L_3-norm of

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} \sqrt{p} & 0 \\ 0 & \sqrt{p} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} \sqrt{q} & 0 \\ 0 & \sqrt{q} \end{pmatrix} \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix},$$

where (X_1^*, X_2^*) is an independent copy of (X_1, X_2). Obviously, the bivariate normal distribution is the solution. All this is summarized as follows.

Proposition B *The following convergence in distribution holds:*

$$\widehat{\Sigma}_n^{-\frac{1}{2}} \begin{pmatrix} S_n - \mathbb{E}(S_n) \\ K_n - \mathbb{E}(K_n) \end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, I_2).$$

Proof: We only check (12) because the second condition of (13) follows along similar lines and the first condition of (13) follows from (12) in view of

$$\|M_1\|_{\text{op}} = \sqrt{p} \qquad \text{and} \qquad \|M_2\|_{\text{op}} = \sqrt{q}.$$

We start with proving the claimed property for $b_n^{(i)}$ for which we use the notations

$$\Omega_1(n) = \mathbb{V}(S_n), \quad \Omega_2(n) = \text{Cov}(S_n, K_n), \quad \Omega_3(n) = \mathbb{V}(K_n)$$

and

$$D(n) = \Omega_1(n)\Omega_3(n) - \Omega_2(n)^2.$$

Also define

$$R(n) = \Omega_1(n) + \Omega_3(n) + 2\sqrt{D(n)}.$$

Then, by (10), we see that

$$\begin{aligned} b_n^{(1)} = (1 - \mu(n) + \mu(B_n) + \mu(n - B_n))&\frac{\Omega_3(n) + \sqrt{D(n)}}{\sqrt{D(n)R(n)}} \\ - (n - \nu(n) + \nu(B_n) + \nu(n - B_n))&\frac{\Omega_2(n)}{\sqrt{D(n)R(n)}} \end{aligned}$$

and a similar expression for $b_n^{(2)}$ holds. From the normality of both S_n and K_n (proved for S_n via the contraction method in [5] and a similar method of proof also applies to K_n), we have

$$\frac{1-\mu(n)+\mu(B_n)+\mu(n-B_n)}{\sqrt{n}} \xrightarrow{L_3} 0 \qquad \text{and} \qquad \frac{n-\nu(n)+\nu(B_n)+\nu(n-B_n)}{\sqrt{n\log n}} \xrightarrow{L_3} 0.$$

Moreover, we have

$$\sqrt{n}\,\frac{\Omega_3(n)+\sqrt{D(n)}}{\sqrt{D(n)R(n)}} \sim \frac{1}{\sqrt{\mathscr{F}[g^{(1)}](n)}},$$

and

$$\sqrt{n\log n}\,\frac{\Omega_2(n)}{\sqrt{D(n)R(n)}} \sim \frac{\mathscr{F}[g^{(2)}](n)}{\lambda\sqrt{\log n\mathscr{F}[g^{(1)}](n)}},$$

where $g^{(1)}, g^{(2)}$ and λ are as above. Thus, both sequences are bounded and, consequently, we obtain the claimed result with L_3-convergence above. Similarly, one proves the claimed result for $b_n^{(2)}$.

Next, we consider $M_n^{(i)}$. Here, we only show the claim for the $(1,1)$ entry of $M_n^{(1)}$ (denoted by $M_n^{(1)}(1,1)$) all other cases being treated similarly. First, observe that by definition and matrix square-root, we have

$$M_n^{(1)}(1,1) = \frac{\sqrt{R(n)}}{\sqrt{R(B_n)}} \cdot \frac{(\Omega_3(n)+\sqrt{D(n)})(\Omega_1(B_n)+\sqrt{D(B_n)})-\Omega_2(n)\Omega_2(B_n)}{\sqrt{D(n)R(n)}}.$$

Now, from the strong law of large numbers for the binomial distribution

$$\frac{B_n}{n} \xrightarrow{\text{a.s.}} p$$

and from Taylor series expansion (note that all periodic functions are infinitely differentiable), we have

$$\frac{\sqrt{R(n)}}{\sqrt{R(B_n)}} \xrightarrow{\text{a.s.}} \frac{1}{\sqrt{p}},$$

and

$$\frac{(\Omega_3(n)+\sqrt{D(n)})(\Omega_1(B_n)+\sqrt{D(B_n)})-\Omega_2(n)\Omega_2(B_n)}{\sqrt{D(n)R(n)}} \xrightarrow{\text{a.s.}} p.$$

Thus, $M_n^{(1)}(1,1) \xrightarrow{\text{a.s.}} \sqrt{p}$ from which the claim follows by the dominated convergence theorem. □

Next, set

$$\widetilde{\Sigma}_n := \begin{pmatrix} n\mathscr{F}[g^{(1)}](n) & 0 \\ 0 & \lambda n\log n \end{pmatrix}.$$

Then, we have the following simple lemma.

Lemma 1 *We have, as $n \to \infty$,*

$$\widehat{\Sigma}_n^{-\frac{1}{2}}\widetilde{\Sigma}_n^{\frac{1}{2}} \to I_2.$$

Proof: This follows by a straightforward computation using the expressions of the matrix square-root and its inverse from above. □

From this lemma and Proposition B our claimed result now follows.

Acknowledgments

We thank the referees for helpful comments; in particular, the detailed comments by one referee on the proof of Theorem B led to considerable improvements in Section 3.

References

[1] H.-H. Chern, M. Fuchs, H.-K. Hwang and R. Neininger. Dependence and phase changes in random m-ary search trees, *Random Struct. Algor.*, to appear.

[2] L. Devroye (1999). Universal limit laws for depths in random trees, *SIAM J. Comput.*, **28**, 409–432.

[3] P. Flajolet, X. Gourdon and P. Dumas (1995). Mellin transforms and asymptotics: harmonic sums, *Theoret. Comput. Sci.*, **144**, 3–58.

[4] M. Fuchs, H.-K. Hwang and V. Zacharovas (2014). An analytic approach to the asymptotic variance of trie statistics and related structures, *Theor. Comput. Sci.*, **527**, 1–36.

[5] M. Fuchs and C.-K. Lee (2014). A general central limit theorem for shape parameters of m-ary tries and PATRICIA tries, *Electron. J. Combin.*, **21:1**, 26 pages.

[6] M. Fuchs and C.-K. Lee (2015). The Wiener index of random digital trees, *SIAM J. Discrete Math.*, **29**, 586–614.

[7] H.-K. Hwang, M. Fuchs and V. Zacharovas (2010). Asymptotic variance of random symmetric digital search trees, *Discrete Math. Theor. Comput. Sci.*, **12**, 103–166.

[8] P. Jacquet and M. Régnier (1986). Trie partitioning process: limiting distributions. In *CAAP '86* (Nice, 1986), vol. **214** of *Lecture Notes in Comput. Sci.*, Springer, Berlin, 196–210.

[9] P. Jacquet and W. Szpankowski (1998). Analytical de-Poissonization and its applications, *Theoret. Comput. Sci.*, **201**, 1–62.

[10] P. Kirschenhofer and H. Prodinger (1991). On some applications of formulae of Ramanujan in the analysis of algorithms, *Mathematika*, **38**, 14–33.

[11] P. Kirschenhofer, H. Prodinger and W. Szpankowski (1989). On the variance of the external path length in a symmetric digital trie, *Discrete Appl. Math.*, **25**, 129–143.

[12] R. Neininger and L. Rüschendorf (2004). A general limit theorem for recursive algorithms and combinatorial structures, *Ann. Appl. Probab.*, **14:1**, 378–418.

[13] R. Neininger and L. Rüschendorf (2006). A survey of multivariate aspects of the contraction method, *Discrete Math. Theor. Comput. Sci.*, **8**, 31–56.

[14] M. Régnier and P. Jacquet (1989). New results on the size of tries, *IEEE Trans. Inform. Theory*, **35**, 203–205.

[15] W. Schachinger (1995). On the variance of a class of inductive valuations of data structures for digital search, *Theoret. Comput. Sci.*, **144**, 251–275.

[16] Y. L. Tong, *The Multivariate Normal Distribution*. Springer-Verlag, New York, 1990.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Variance of the Internal Profile in Suffix Trees

Jeffrey Gaither[1], Mark Daniel Ward[2†]

[1]*Mathematical Biosciences Institute of The Ohio State University, Columbus, OH 43210, USA*
[2]*Department of Statistics, Purdue University, West Lafayette, IN 47907, USA*

The precise analysis of the variance of the profile of a suffix tree has been a longstanding open problem. We analyze three regimes of the asymptotic growth of the variance of the profile of a suffix tree built from a randomly generated binary string, in the nonuniform case. We utilize combinatorics on words, singularity analysis, and the Mellin transform.

Keywords: suffix tree, asymptotic analysis, combinatorics on words, singularity analysis, Mellin transform

1 Introduction

One open problem about suffix trees is how to characterize the number of internal nodes on the kth level of a suffix tree that has n leaves. Park et al. [PHNS09] precisely analyzed the profile of retrieval tries in 2009. Ward has been working on the analogous problem in suffix trees for a decade; see, e.g., [NW11, War07]. While the mean profile of retrieval trees and suffix trees are the same (asymptotically, up to first order, in the main range of interest of the parameters), the variances of the profiles of these two classes of trees are different. The goal of this paper is to precisely analyze the variance of the profile of suffix trees.

In retrieval trees, the strings inserted into the tree structure are often considered to be independent; such was the case in [PHNS09]. In contrast to this, in suffix trees, the strings inserted into the tree are suffixes of a common string, so these strings are overlapping. The overlaps make the corresponding analysis much trickier, as compared to [PHNS09].

We analyze a suffix tree built from the suffixes of a common string $S = S_1S_2S_3\ldots$, where the S_j's are randomly generated, independent, and identically distributed. We view each S_j as a letter from the alphabet $\mathcal{A} = \{a, b\}$, where $P(S_j = a) = p$ and $P(S_j = b) = q$. (Without loss of generality, we assume throughout that $p > q$.) We use $\mathcal{A}^\ell$ to denote the set of words of length ℓ. For a word u that consists of i occurrences of letter a and j occurrences of letter b, we use $\mathbb{P}(u)$ to denote the probability that a randomly chosen word of length $|u|$ is exactly equal to u, i.e., $\mathbb{P}(u) := p^iq^j$.

The jth string to be inserted into the suffix tree is $S^{(j)} := S_jS_{j+1}S_{j+2}\ldots$. We consider a randomly generated suffix tree $\mathcal{T}_n$ built over the first n suffixes of S, i.e., built from the suffixes $S^{(1)}$ through $S^{(n)}$. Briefly, all n of these suffixes can be viewed as initially being placed at the root of the suffix tree. The n suffixes are then filtered down to the left or right children of the root, making the classification of the

†M. D. Ward's work is supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370; his work is also supported by NSF DMS-124681 and NSF DMS-1560332.

suffixes according to whether the first letter of each suffix is "a" or "b", respectively. The filtering continues down through the tree, with splitting at the jth level according to the jth letter in the corresponding suffixes in that portion of the tree.

For each word $u \in \mathcal{A}^k$, the suffix tree $\mathcal{T}_n$ will contain the internal node corresponding to u if and only if the base-string S contains at least two copies of the word u within its first $n+k-1$ characters. (Equivalently, $\mathcal{T}_n$ contains the internal node corresponding to u if and only if at least two of the suffixes $S^{(1)}$ through $S^{(n)}$ have u as a prefix.) For this reason, we define $I_{n,u} := 1$ if u appears at least twice in $S_1 S_2 \dots S_{n+k-1}$, or $I_{n,u} := 0$ otherwise. We use $X_{n,k}$ to denote the number of internal nodes in $\mathcal{T}_n$ at level k. With the above notation in place, we observe that $X_{n,k} = \sum_{u \in \mathcal{A}^k} I_{n,u}$. This decomposition will be crucial to our proofs, which start in Section 3.

Finally, following the lead of [PHNS09], we assume that the limit $\alpha := \lim_{n\to\infty} k/\log(n)$ exists.

2 Main Results

The value of $\mathrm{Var}(X_{n,k})$ depends qualitatively on the quantity α, which describes the relationship between n and k via the relation $k/\log(n) \to \alpha$. It turns out that there are two particular alpha-values of importance,

$$\alpha_1 = -\frac{1}{\log(q)}, \qquad \alpha_2 = -\frac{p^2+q^2}{p^2\log(p)+q^2\log(q)}.$$

We do not attempt, as Park et al. did in [PHNS09], to analyze the cases where α is exactly equal to one of these α_i, but instead assume that both $|\alpha - \alpha_i|$ are strictly positive. Given this restriction, it is permissible to take the approximation $k = \alpha\log(n)$, which we do henceforth without comment.

The variance obeys different laws depending on where the value of α falls in the ranges defined by these α_i. The range of most interest is (perhaps) the range in which $\alpha_1 < \alpha < \alpha_2$; we discuss this case in Theorem 2. (The case $\alpha < \alpha_1$ is discussed in Theorem 1; and the case $\alpha_2 < \alpha$ is handled in Theorem 3.)

When α is small, we have an easy and very strong bound on the decay of $\mathrm{Var}(X_{n,k})$.

Theorem 1 *When $\alpha < \alpha_1$, there exists $B > 0$ such that*

$$\mathrm{Var}(X_{n,k}) = O(e^{-n^B}).$$

The proof of Theorem 1 follows from lemmas that mimic the techniques of [War05]; we omit it from this shortened version. The intuitive meaning behind Theorem 1 is that level k of the suffix tree is extremely likely to be completely filled (meaning the variance will be extremely small) if $\log(n)$ is sufficiently large in comparison to k.

Our main results deal with the less trivial case when $\alpha > \alpha_1$. We first introduce the functions involved in our main estimates, and provide a word on how we obtain them.

2.1 Functions Involved in Main Results; Methodology

Our basic device for computing the variance of the internal profile is to write $X_{n,k}$ as a sum of indicator variables $I_{n,u}$, and then evaluate

$$\mathrm{Var}(X_{n,k}) = \mathrm{Var}\Big(\sum_{u\in\mathcal{A}^k} I_{n,u}\Big) = \sum_{u\in\mathcal{A}^k} \mathrm{Var}(I_{n,u}) + \sum_{\substack{u,v\in\mathcal{A}^k \\ u\neq v}} \mathrm{Cov}(I_{n,u}, I_{n,v}). \tag{1}$$

Our final analysis of the sum of the $\mathrm{Var}(I_{n,u})$ will be fairly simple: we will ultimately just have to evaluate the inverse Mellin integral

$$\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} n^{-s} f(s) \sum_{u\in\mathcal{A}^k} \mathbb{P}(u)^{-s}\, ds = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} f(s)\, n^{h(s)}\, ds, \tag{2}$$

where the function $h(s)$ will be given by

$$h(s) := -s + \alpha \log(p^{-s} + q^{-s}).$$

(See [FGD95] for more details about the Mellin transform.) The function $h(s)$ is the same as analyzed in [PHNS09], and their arguments extend seamlessly to our case.

On the other hand, the terms $\mathrm{Cov}(I_{n,u}, I_{n,v})$ for $u \neq v$ will be novel and much more interesting. To deal with them, we will consider all possible overlapping decompositions $(\sigma w, w\theta)$ of (u, v). To accomplish this, we observe that

$$n^{-s}\sum_{\ell=1}^{k-1} \sum_{\substack{w\in\mathcal{A}^{k-\ell}\\ \sigma,\theta\in\mathcal{A}^{\ell}}} \mathbb{P}(w)^{-s}(\mathbb{P}(\sigma)+\mathbb{P}(\theta))^{-s} = \sum_{\ell=1}^{k-1}\sum_{i,j=0}^{\ell} \binom{\ell}{i}\binom{\ell}{j} n^{H(s,\,(k-\ell)/k,\,i/\ell,\,j/\ell)}, \tag{3}$$

where $H(s, r, c, d)$ is defined as

$$H(s,r,c,d) := -s + \alpha(1-r)\log(p^{-s}+q^{-s}) - s\Big(\frac{\alpha}{k}\Big)\log((p^c q^{1-c})^{kr} + (p^d q^{1-d})^{kr}).$$

Note: For ease of the (already cumbersome) notation, we have not written α nor k as a parameter of H. We will substitute the right hand side of (3) for $n^{-s}\sum_{u\in\mathcal{A}^k}\mathbb{P}(u)^{-s}$ into equation (2). We will use a technique for H similar to that used for h, namely, summing over all possible values $p^i q^{\ell-i}$ and $p^j q^{\ell-j}$ of $\mathbb{P}(\sigma)$ and $\mathbb{P}(\theta)$ respectively, and summing $\mathbb{P}(w)$ into a closed form, as was done at (2).

The dominant contribution to (3) comes from terms with small r. Since $\lim_{r\to 0} H(s,r,c,d) = h(s)$, this implies that $\sum_{u,v}\mathrm{Cov}(I_{n,u}, I_{n,v})$ and $\sum_u \mathrm{Var}(I_{n,u})$ have the same first-order asymptotic growth, as functions of n.

We will evaluate the inverse Mellin integral at (2) (and the analogous integral for H) by using either the saddle point method or by taking the residue of the pole of $\Gamma(s+2)$ at $s=-2$; which device we use will depend on the value of α. Before giving our main results, we list the saddle points of the functions $h(s)$ and $H(s,r,c,d)$, which are

$$\begin{aligned}
\rho &:= \frac{\Big(-\dfrac{\alpha\log(p)+1}{\alpha\log(q)+1}\Big)}{\log(p/q)},\\
\rho_{r,c,d} &:= \frac{\Big(-\dfrac{\alpha(1-r)\log(p)+1+(\alpha/k)\log((p^cq^{1-c})^{kr}+(p^dq^{1-d})^{kr})}{\alpha(1-r)\log(q)+1+(\alpha/k)\log((p^cq^{1-c})^{kr}+(p^dq^{1-d})^{kr})}\Big)}{\log(p/q)}.
\end{aligned} \tag{4}$$

It is also easy to verify that for any $y\in\mathbb{Z}$, the value $s = \rho + 2\pi i y/\log(p/q)$ is also a saddle point of h, and similarly, $s = \rho_{r,c,d} + 2\pi i y/\log(p/q)$ is a saddle point of H.

These saddle points will (at last) allow us to express an asymptotic value for $\mathrm{Var}(X_{n,k})$ in the case where $\alpha_1 < \alpha < \alpha_2$.

2.2 Behavior in the main regime

Theorem 2 *Assume α satisfies $\alpha_1 < \alpha < \alpha_2$. Let ρ and $\rho_{r,c,d}$ be as in (4). Then we have*

$$\mathrm{Var}(X_{n,k}) = \frac{n^{h(\rho)}(C_1(n) + 2C_2(n))}{\sqrt{\log(n)}} \times \big(1 + O(\log(n)^{-1})\big).$$

The $C_1(n)$ is given by

$$C_1(n) = \sum_{y\in\mathbb{Z}} \frac{n^{i\Im(h(\rho+iyK))} f_1(\rho+iyK)\Gamma(\rho+iyK+1)}{\sqrt{2\pi h''(\rho+iyK)}},$$

where $K := 2\pi/\log(p/q)$ and where $f_1(s) := 1 - 2^{-s} - s2^{-s-2}$. Regarding $C_2(n)$, we define $r = \frac{\ell}{k}$, $c = \frac{i}{\ell}$, $d = \frac{j}{\ell}$, and then $C_2(n)$ is given by

$$C_2(n) = \sum_{\substack{0<\ell<k\\ 0\le i,j\le \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(\rho_{r,c,d},r,c,d)}}{n^{h(\rho)}} \sum_{y\in\mathbb{Z}} \frac{n^{i\Im(H(\rho_{r,c,d}+iyK,r,c,d))} f_2(\rho_{r,c,d}+iyK,\ell,i,j)\Gamma(\rho_{r,c,d}+iyK+2)}{\sqrt{2\pi\frac{\partial H}{\partial s}(\rho_{r,c,d}+iyK,r,c,d)}}$$
$$\times (1 + O(\log(n)^{-1})).$$

with the function $f_2(s,\ell,i,j)$ given by

$$f_2(s,\ell,i,j) = \sum_{m\ge 2} \Big(\frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}}\Big)^{m-1} \frac{\Gamma(s+m)}{\Gamma(s+2)m!} L_m\Big(\frac{p^i q^{\ell-i} p^j q^{\ell-j}}{p^i q^{\ell-i} + p^j q^{\ell-j}},\ \frac{p^i q^{\ell-i} p^j q^{\ell-j}}{(p^i q^{\ell-i} + p^j q^{\ell-j})^2},\ s+m\Big),$$

with

$$L_m(a,b,x) = a(m-1)^2 + m(2-m) + bmx.$$

Furthermore, the outer sum in $C_2(n)$ satisfies the decay condition that for any positive integer ℓ_0, the sum over all $\ell > \ell_0$ and $1 \le i,j \le \ell$ is $O(n^{-(\ell_0/k)\times\beta})$ for a fixed $\beta > 0$.

2.3 Behavior in the polar regime

In the final α-regime, where $\alpha > \alpha_2$, the asymptotics arise from the pole at $s = -2$, as the following theorem states.

Theorem 3 *Assume the parameter α satisfies $\alpha > \alpha_2$. Then for some $\epsilon > 0$, we have*

$$\mathrm{Var}(X_{n,k}) = n^{h(-2)}(C_1(n) + 2C_2(n)) \times (1 + O(n^{-\epsilon}))$$

with f_1, f_2 as defined in Theorem 2, and $C_1(n)$, $C_2(n)$ are given by

$$C_1(n) = f_1(-2), \qquad C_2(n) = f_2(-2) \sum_{\substack{0<\ell<k\\ 0\le i,j\le \ell}} \binom{\ell}{i}\binom{\ell}{j} \frac{n^{H(-2,r,c,d)}}{n^{h(-2)}}$$

with the decay of $C_2(n)$ as in Theorem 2.

Having stated our main results, we now proceed to the proof of Theorems 2 and 3, which will occupy the remainder of the paper.

3 An Expression for the Variance

Our first task in proving Theorems 2 and 3 is to obtain an exact expression for the variance of the internal profile $X_{n,k}$. Recalling equation (1), we need to derive the values of $\mathrm{Var}(I_{n,u})$ and $\mathrm{Cov}(I_{n,u}, I_{n,v})$, so we let U_n denote the number of occurrences of u in the first n characters of S, and we define V_n analogously. Then inclusions-exclusion yields the representations

$$\begin{aligned}\mathrm{Var}(I_{n,u}) &= \Big(1 - \sum_{i=0}^{1} \mathbb{P}(U_{n+k-1} = i)\Big) - \Big(1 - \sum_{i=0}^{1} \mathbb{P}(U_{n+k-1} = i)\Big)^2 \\ \mathrm{Cov}(I_{n,u}, I_{n,v}) &= \sum_{0 \le i,j \le 1} \Big(\mathbb{P}(U_{n+k-1} = i,\ V_{n+k-1} = j) - \mathbb{P}(U_{n+k-1} = i) \times \mathbb{P}(V_{n+k-1} = j)\Big) \quad (5)\end{aligned}$$

where we require u and v to be distinct. Thus, to obtain an expression for $\mathrm{Var}(X_{n,k})$, we just have to evaluate all the probabilities in (5).

4 Explicit Expressions for Word-Occurrence Probabilities

To estimate the probabilities in (5), we use generating functions, and complex analysis. Motivated by [BCN12], we define

$$\psi(z) = C_{u,u}(z)C_{v,v}(z) - C_{u,v}(z)C_{v,u}(z), \qquad \text{and} \qquad \phi_u(z) = C_{v,v}(z) - C_{u,v}(z), \qquad (6)$$

where the functions $C_{x,y}(z)$ are *correlation polynomials*, the fundamental device for dealing with the phenomenon of word-overlaps. With these functions in hand, we can define generating-functions for all the probabilities in (5). We summarize the result in the following proposition.

Proposition 1 *Let $\psi(z)$ and $\phi_u(z)$ be as defined at (6), and define the functions*

$$\begin{aligned}&D_u(z) = (1-z)C_{u,u}(z) + z^k\mathbb{P}(u), \qquad \delta_{u,v}(z) = (1-z)\psi(z) + z^k(\phi_u(z)\mathbb{P}(u) + \phi_v(z)\mathbb{P}(v)), \\ &G_0^{(u)}(z) = C_{u,u}(z),\ \ G_1^{(u)}(z) = \mathbb{P}(u)z^k, G_{0,0}^{(u,v)}(z) = \psi(z),\ \ G_{1,0}^{(u,v)}(z) = \delta_{u,v}(z)C_{v,v}(z) - \psi(z)D_v(z), \\ &G_{1,1}^{(u,v)}(z) = \delta_{u,v}(z)^2 - \delta_{u,v}(z)\big(C_{v,v}(z)D_u(z) + C_{u,u}(z)D_v(z) + (1-z)\psi(z)\big) + 2\psi(z)D_u(z)D_v(z), \qquad (7)\end{aligned}$$

with all v-counting functions defined in a manner analogous to the u-counting functions. Then we have the closed-form power series expressions

$$\frac{G_i^{(u)}(z)}{D_u(z)^{i+1}} = \sum_{n\ge 0} z^n \mathbb{P}(U_n = i), \quad \textit{and} \quad \frac{G_{i,j}^{(u,v)}(z)}{\delta_{u,v}(z)^{i+j+1}} = \sum_{n\ge 0} z^n \mathbb{P}(U_n = i,\ V_n = j), \quad 0 \le i,j \le 1. \qquad (8)$$

Now we must derive the $(n+k-1)$st coefficients of these generating functions. To do this, we use Cauchy's Integral Formula, following a standard argument in combinatorics on words. Our specific methodology will rely on a vital fact about the denominators $D_u(z)$, $D_v(z)$ and $\delta_{u,v}(z)$ of the probability generating functions in (8).

Lemma 1 *There exist $K, \rho > 0$ such that for all $k > K$ and all $u, v \in \mathcal{A}^k$, each of the polynomials $D_u(z)$, $D_v(z)$, and $\delta_{u,v}(z)$ has a unique root (defined respectively as R_u, R_v and $R_{u,v}$) in the disc $|z| \le \rho$.*

The proof for $D_u(z)$ and $D_v(z)$ is given in [JS05]; spatial constraints prevent us from giving the proof for the $\delta_{u,v}(z)$ portion.

Armed with Lemma 1, we can estimate the word-counting coefficients of our generating functions to within a factor of $O(\rho^{-n})$ by applying Cauchy's Theorem to the contour $z = |\rho|$. The following theorem gives the resultant estimates.

Theorem 4 *Let the polynomials $D_u, D_v, \delta_{u,v}$ and $G_0^{(u)}, G_1^{(u)}$, etc. be as in (7) and (8). If we define*

$$c_{0,0}^{(u)} = -\frac{C_{u,u}(R_u)}{D_u'(R_u)}, \qquad c_{1,0}^{(u)} = \frac{\mathbb{P}(u) D_u''(R_u)}{D_u'(R_u)^3}, \qquad c_{1,1}^{(u)} = \frac{\mathbb{P}(u)}{D_u'(R_u)^2},$$

then we have the following estimates

$$\mathbb{P}(U_{n+k-1} = 0) \approx c_{0,0}^{(u)} \frac{1}{R_u^{n+k}}, \qquad \textit{and} \qquad \mathbb{P}(U_{n+k-1} = 1) \approx c_{1,0}^{(u)} \frac{1}{R_u^{n}} + c_{1,1}^{(u)} \frac{n}{R_u^{n+1}},$$

and the error in each case is $O(\rho^{-n})$.

Similarly, for the joint events $(U_{n+k-1} = i,\ V_{n+k-1} = j)$, and

$$a_{0,0}^{(u,v)} = -\frac{\psi'(R_{u,v})}{\delta_{u,v}'(R_{u,v})}, \qquad a_{1,0,u}^{(u,v)} = -\frac{G_{1,0}^{(u,v)}(R_{u,v})\delta_{u,v}''(R_{u,v})}{\delta'(R_{u,v})^3}, \qquad a_{1,1,u}^{(u,v)} = \frac{G_{1,0}^{(u,v)}(R_{u,v})}{\delta'(R_{u,v})^2},$$

$$\begin{aligned} a_{2,0}^{(u,v)} = &-\frac{G_{1,1}^{(u,v)\prime\prime}(R_{u,v})}{2\delta_{u,v}'(R_{u,v})^3} + \frac{3G_{1,1}^{(u,v)\prime}(R_{u,v})\delta_{u,v}{}''(R_{u,v})}{2\delta_{u,v}'(R_{u,v})^4} \\ &- \frac{G_{1,1}^{(u,v)}(R_{u,v})(-\delta_{u,v}'(R_{u,v})\delta_{u,v}'''(R_{u,v}) + 3\delta_{u,v}''(R_{u,v})^2)}{2\delta_{u,v}'(R_{u,v})^5}, \end{aligned}$$

$$a_{2,1}^{(u,v)} = \frac{G_{1,1}^{(u,v)\prime}(R_{u,v})}{\delta_{u,v}{}'(R_{u,v})^3} - \frac{3G_{0,0}^{(u,v)}(R_{u,v})\delta_{u,v}{}''(R_{u,v})}{2\delta_{u,v}{}'(R_{u,v})^4}, \qquad a_{2,2}^{(u,v)} = -\frac{G_{1,1}^{(u,v)}(R_{u,v})}{2\delta_{u,v}'(R_{u,v})^3},$$

with $G_{i,j}^{(u,v)}(z)$ as in (8), we also obtain these estimates, where again, the error in each case is $O(\rho^{-n})$:

$$\begin{aligned} \mathbb{P}(U_{n+k-1} = 0, V_{n+k-1} = 0) &\approx a_{0,0}^{(u,v)} \frac{1}{R_{u,v}^{n+k}}, \\ \mathbb{P}(U_{n+k-1} = 1, V_{n+k-1} = 0) &\approx a_{1,0,u}^{(u,v)} \frac{1}{R_{u,v}^{n+k}} + a_{1,1,u}^{(u,v)} + \frac{(n+k)}{R_{u,v}^{n+k+1}}, \\ \mathbb{P}(U_{n+k-1} = 1, V_{n+k-1} = 1) &\approx a_{2,0}^{(u,v)} \frac{1}{R_{u,v}^{n+k}} + a_{2,1}^{(u,v)} \frac{(n+k)}{R_{u,v}^{n+k+1}} + a_{2,2}^{(u,v)} \frac{(n+k)(n+k+1)}{R_{u,v}^{n+k+2}}. \end{aligned}$$

Using these expressions, we can evaluate the expressions for $\mathrm{Var}(I_{n,u})$ and $\mathrm{Cov}(I_{n,u}, I_{n,v})$ at (5) to within a factor of $O(\rho^{-n})$. In doing this, however, it will be helpful to break up our estimates from

Theorem 4 so that terms of common order in n are denoted under a single variable. We therefore define the upper-case constants (we suppress the dependence on u and v in the notation)

$$C_0 = \frac{c_{0,0}^{(u)} + c_{1,0}^{(u)}}{R_u^k} + \frac{k c_{1,1}^{(u)}}{R_u^{k+1}}, \qquad C_1 = \frac{c_{1,1}^{(u)}}{R_u^{k+1}},$$
$$A_0 = \frac{a_{0,0}^{(u,v)} + a_{1,0,u}^{(u,v)} + a_{1,0,v}^{(u,v)} + a_{2,0}^{(u,v)}}{R_{u,v}^k} + \frac{(a_{1,1,u}^{(u,v)} + a_{1,1,v}^{(u,v)})k}{R_{u,v}^{k+1}} + \frac{a_{1,1}^{(u,v)} k(k+1)}{R_{u,v}^{k+2}},$$
$$A_1 = \frac{a_{1,1,u}^{(u,v)} + a_{1,1,v}^{(u,v)} + a_{2,1}^{(u,v)}}{R_{u,v}^{k+1}} + \frac{a_{2,2}^{(u,v)}(2k+1)}{R_{u,v}^{k+2}}, \qquad A_2 = \frac{a_{2,2}}{R_{u,v}^{k+2}}, \qquad B_0 = \frac{c_{0,0}^{(v)} c_{0,0}^{(u)}}{(R_u R_v)^k},$$
$$B_1 = \Big(c_{1,0}^{(u)} + \frac{c_{1,1}^{(u)}}{R_u}\Big)\frac{c_{0,0}^{(u)}}{R_v^k} + \Big(c_{1,0}^{(v)} + \frac{c_{1,1}^{(v)}}{R_v}\Big)\frac{c_{0,0}^{(u)}}{R_u^k}, \qquad B_2 = \Big(c_{1,0}^{(u)} + \frac{c_{1,1}^{(u)}}{R_u}\Big)\Big(c_{1,0}^{(v)} + \frac{c_{1,1}^{(v)}}{R_v}\Big). \tag{9}$$

Returning to the expression $\mathrm{Var}(X_{n,k}) = \sum_{u\in\mathcal{A}^k} \mathrm{Var}(I_{n,u}) + \sum_{\substack{u,v\in\mathcal{A}^k \\ u\neq v}} \mathrm{Cov}(I_{n,v})$, we obtain an expression for our ultimate desired quantity.

Corollary 1 *Let A_i, B_i, C_i be as defined in (9). With A_i, B_i and C_i as in (9), we have the estimate*

$$\mathrm{Var}(X_{n,k}) = \sum_{u\in\mathcal{A}^k} \Big(1 - \frac{C_0 + nC_1}{R_u^n}\Big) - \Big(1 - \frac{C_0 + nC_1}{R_u^n}\Big)^2 + \sum_{\substack{u,v\in\mathcal{A}^k \\ u\neq v}} \sum_{i=0}^{2} \Big(\frac{A_i}{R_{u,v}^n} - \frac{B_i}{(R_u R_v)^n}\Big) n^i + O(\rho^{-n}).$$

4.1 High-Probability Approximations

Our task is now to approximate the expression from Corollary 1. To achieve this, we follow the usual suffix-tree strategy: we compare the terms to simpler ones which will be accurate with very high probability, and use Mellin transforms to show that sum of the the differences between the old terms and the new ones is negligible. Our two main tools for demonstrating this negligibility are bounds provided by the following lemma.

Lemma 2 *We have the bounds*

$$\sum_{u\in\mathcal{A}^k} \mathbb{P}(u)(C_{u,u}(1) - 1) = O(p^{k/2}), \qquad \sum_{\substack{u,v\in\mathcal{A}^k \\ u\neq v}} \mathbb{P}(u) C_{u,v}(1) C_{v,u}(1) = O(p^{k/2})$$

The first portion of Lemma 2 is proved in [JS05]; spatial constraints prevent us from proving the second portion here. However, by rigorously expanding on the heuristic $C_{u,u}(1) \approx 1$ and $C_{u,v}(1)C_{v,u}(1) \approx 0$, we obtain the following theorem which is one of the major steps of the proof.

Theorem 5 *We define the terms $\boldsymbol{P}_{u,v} := \mathbb{P}(u) + \mathbb{P}(v)$,, $\Theta_{u,v} := \mathbb{P}(u)C_{u,v}(1) + \mathbb{P}(v)C_{v,u}(1)$, and $K_{u,v} = (2k-1)\mathbb{P}(u)\mathbb{P}(v)$, and the expressions*

$$V_1(n) := \sum_{u\in\mathcal{A}^k} 1-(1+n\mathbb{P}(u))e^{-n\mathbb{P}(u)} - \big(1-(1+n\mathbb{P}(u))e^{-n\mathbb{P}(u)}\big)^2,$$

$$V_2(n) := \sum_{\substack{u,v\in\mathcal{A}^k\\ u\neq v}} n^3\mathbb{P}(u)\mathbb{P}(v)K_{u,v}e^{-n(\boldsymbol{P}_{u,v}-\Theta_{u,v})},$$

$$V_3(n) := \sum_{\substack{u,v\in\mathcal{A}^k\\ u\neq v}} e^{-n\boldsymbol{P}_{u,v}}(e^{n\Theta_{u,v}}-1)\big(1+n\boldsymbol{P}_{u,v}+n^2\mathbb{P}(u)\mathbb{P}(v)\big) - e^{-n(\boldsymbol{P}_{u,v}-\Theta_{u,v})}n\Theta_{u,v}\big(1+n(\boldsymbol{P}_{u,v}-\Theta_{u,v})\big).$$

Then, for every $\epsilon>0$, we have the estimate

$$\mathrm{Var}(X_{n,k}) = V_1(n) - V_2(n) + 2V_3(n) + O\big(n^{1+(\alpha/2)\log(p)+\epsilon}\big).$$

We mention that the term $V_1(n)$ has already been analyzed in Park [PHNS09]. It gives the asymptotic variance of the internal profile in a *trie*. The term $V_2(n)$ is negligible. Thus, after proving Theorem 5, all that will remain will be to analyze $V_3(n)$.

5 Distilling Essence of Estimate

We must now analyze the estimate from Theorem 5, which consists of the terms $V_1(n)$, $V_2(n)$ and $V_3(n)$. We can deal with the first two of these terms in two quick theorems. Theorem 6 was proven in [PHNS09]. Theorem 7 has a short proof, which we omit in this concise version.

Theorem 6 *An asymptotic expression for $V_1(n)$ is given by the $C_1(n)$ portions from Theorems 2 and 3.*

Theorem 7 *The term $V_2(n)$ from Theorem 8 satisfies $V_2(n) = \mathrm{Var}(X_{n,k})O(n^{-\epsilon})$ for some $\epsilon>0$.*

For the rest of the paper, then, we concentrate on the portion $V_3(n)$, which contains the term $\Theta_{u,v} = \mathbb{P}(u)C_{u,v}(1) + \mathbb{P}(v)C_{v,u}(1)$ and constitutes the really novel part of the whole enterprise. We deal with $\Theta_{u,v}$ by nothing that, by Lemma 2, the quantities $C_{u,v}(1)$ and $C_{v,u}(1)$ are unlikely to simultaneously be large, so the approximation $\Theta_{u,v}\approx\mathbb{P}(u)C_{u,v}(1)$ is reasonable. From here, we note that for $\Theta_{u,v}$ to be nonzero we must have $C_{u,v}(1)>0$, in which case there exists some maximal suffix of u which is also a prefix of v. If we call this word w, and then have the precise equality $\mathbb{P}(u)C_{u,v}(1) = \mathbb{P}(\sigma)\mathbb{P}(w)\mathbb{P}(\theta)C_{w,w}(1)$. where $\sigma,\theta\in\mathcal{A}^{k-|w|}$ are such that $u=\sigma w$ and $v=w\theta$. Then we employ the estimate $C_{w,w}(1)\approx 1$, again as suggested by Lemma 2. We thus have the central estimate $\Theta_{u,v}\approx\mathbb{P}(\sigma)\mathbb{P}(w)\mathbb{P}(\theta)$. Our strategy, then, is to make the substitutions $u=\sigma w$, $v=w\theta$, and $\Theta_{u,v}=\mathbb{P}(\sigma)\mathbb{P}(w)\mathbb{P}(\theta)$ in the summand of $V_3(n)$, and then sum over all possible such decompositions. In the proof and final result it will be helpful to have the shorthand $\mathbf{Q}_{\sigma,\theta} := \mathbb{P}(\sigma)+\mathbb{P}(\theta)$ and $\mathbf{T}_{\sigma,\theta} := \mathbb{P}(\sigma)\mathbb{P}(\theta)$, The following theorem states that this heuristic can be rigorously justified.

Theorem 8 *Let $\boldsymbol{Q}_{\sigma,\theta}, \boldsymbol{T}_{\sigma,\theta}$ be as defined above, and define the functions*

$$\begin{aligned} g_{w,\sigma,\theta}(n) = {} & e^{-n\mathbb{P}(w)\boldsymbol{Q}_{\sigma,\theta}}(e^{x\mathbb{P}(w)\boldsymbol{T}_{\sigma,\theta}}-1)\big(1+x\mathbb{P}(w)\boldsymbol{Q}_{\sigma,\theta}+n^2\mathbb{P}(w)^2\boldsymbol{T}_{\sigma,\theta}\big)\\ & - e^{-x\mathbb{P}(w)(\boldsymbol{Q}_{\sigma,\theta}-\boldsymbol{T}_{\sigma,\theta})}x\mathbb{P}(w)\boldsymbol{T}_{\sigma,\theta}\big(1+x\mathbb{P}(w)(\boldsymbol{Q}_{\sigma,\theta}-\boldsymbol{T}_{\sigma,\theta})\big)\end{aligned}$$

and $\widetilde{V}_3(n) := \sum_{\ell=1}^{k-1} \sum_{\substack{w\in\mathcal{A}^{k-\ell} \\ \sigma,\theta\in\mathcal{A}^{\ell}}} g_{w,\sigma,\theta}(n)$. *Then for* $V_3(n)$ *as given in Theorem 5, we have the estimate*

$$V_3(n) = 2\widetilde{V}_3(n) + O\big(n^{1+(\alpha/2)\log(p)+\epsilon}\big).$$

One proves Theorem 8 by making the substitutions $\mathbb{P}(w)\mathbf{Q}_{\sigma,\theta} \approx \mathbf{P}_{u,v}$ and $\mathbb{P}(w)\mathbf{T}_{\sigma,\theta} \approx \Theta_{u,v}$, and then using Mellin transforms and Lemma 2 to show that the derived error-bound is satisfied.

6 Derivation of Asymptotics

To complete the main proof, it remains only to analyze $\widetilde{V}_3(n)$. We present the key results in this process in a series of subsections.

6.1 Partitioning the Sum

Our first step is to partition the sum which comprises $\widetilde{V}_3(n)$. into subsets which share a common value for the ordered pair $(\mathbb{P}(\sigma), \mathbb{P}(\theta))$. We can rewrite the function $g_{w,\sigma,\theta}(n)$ from Thereom 8 as an infinite sum,

$$g_{w,\sigma,\theta}(x) = e^{-x\mathbb{P}(w)\mathbf{Q}_{\sigma,\theta}} \sum_{m\geq 2} \frac{(x\mathbb{P}(w))^m \mathbf{T}_{\sigma,\theta}^{m-1}\mathbf{Q}_{\sigma,\theta}}{m!} L_m\Big(\frac{\mathbf{T}_{\sigma,\theta}}{\mathbf{Q}_{\sigma,\theta}},\ \frac{\mathbb{P}(w)\mathbf{T}_{\sigma,\theta}}{\mathbf{Q}_{\sigma,\theta}},\ x\Big).$$

with the function L_m given by $L_m(a,b,x) := a(m-1)^2 + m(2-m) + bmx$. The terms $\mathbf{Q}_{\sigma,\theta}$ and $\mathbf{T}_{\sigma,\theta}$ only depend on the *probabilities* of σ and θ; their internal composition does not matter. This allows a great reduction in the number of terms to handle. With some abuse of notation, we define the terms

$$\begin{aligned}\mathbf{Q}_{r,c,d}^{(k)} &:= \mathbf{Q}_{a^{krc}b^{kr(1-c)},a^{krd}b^{kr(1-d)}} = p^{krc}q^{kr(1-c)} + p^{krd}q^{kr(1-d)},\\ \mathbf{T}_{r,c,d}^{(k)} &:= \mathbf{T}_{a^{krc}b^{kr(1-c)},a^{krd}b^{kr(1-d)}} = p^{krc}q^{kr(1-c)} \times p^{krd}q^{kr(1-d)}\end{aligned}$$

and then define the atom of all our remaining analysis, which is

$$g(x,r,c,d) = \sum_{w\in\mathcal{A}^{k(1-r)}} e^{-x\mathbb{P}(w)\mathbf{Q}_{r,c,d}^{(k)}} \sum_{m\geq 2} \frac{(x\mathbb{P}(w))^m \mathbf{T}_{r,c,d}^{(k)\,m-1}\mathbf{Q}_{r,c,d}^{(k)}}{m!} L_m\Big(\frac{\mathbf{T}_{r,c,d}^{(k)}}{\mathbf{Q}_{r,c,d}^{(k)}},\ \frac{\mathbb{P}(w)\mathbf{T}_{r,c,d}^{(k)}}{\mathbf{Q}_{r,c,d}^{(k)}},\ x\Big). \tag{10}$$

With this notation, we have the following proposition.

Proposition 2 *Let* $g(x,r,c,d)$ *be as in (10). Then* $\widetilde{V}_3(n)$ *from Theorem 8 admits the representation*

$$\widetilde{V}_3(n) = \sum_{\substack{0<\ell<k \\ 0\leq i,j\leq \ell}} \binom{\ell}{i}\binom{\ell}{j} g(n, \tfrac{\ell}{k}, \tfrac{i}{\ell}, \tfrac{j}{\ell}). \tag{11}$$

Now we analyze g.

6.2 Analysis of $g(n,r,c,d)$

All our final estimates rest on our analysis of the function g given in Proposition 2. To begin that analysis, we take the Mellin transform of g and, specifying the bounded portion

$$W(s,r,c,d) = \sum_{m\geq 2} \Big(\frac{\mathbf{T}^{(k)}_{r,c,d}}{\mathbf{Q}^{(k)}_{r,c,d}}\Big)^{m-1} \frac{\Gamma(s+m)}{\Gamma(s+2)m!} L_m\Big(\frac{\mathbf{T}^{(k)}_{r,c,d}}{\mathbf{Q}^{(k)}_{r,c,d}},\ \frac{\mathbf{T}^{(k)}_{r,c,d}}{{\mathbf{Q}^{(k)}_{r,c,d}}^2},\ s+m\Big),$$

we obtain

$$\begin{aligned} g^*(s,r,c,d) &= \Gamma(s+2)W(s,r,c,d){\mathbf{Q}^{(k)}_{r,c,d}}^{-s} \sum_{w\in\mathcal{A}^{k(1-r)}} \mathbb{P}(w)^{-s} \\ &= \Gamma(s+2)W(s,r,c,d){\mathbf{Q}^{(k)}_{r,c,d}}^{-s}(p^{-s}+q^{-s})^{k(1-r)}. \end{aligned}$$

We then consider the value of $n^{-s}g^*(s,r,c,d)$, which will be the integrand of our inverse Mellin integral. Using the relation $k=\alpha\log(n)$, we can write $n^{-s}g^*(s,r,c,d) = \Gamma(s+2)W(s,r,c,d)n^{H(s,r,c,d)}$, where the function H is as defined in Section 2.1. From here, we can recover the value of $g(n,r,c,d)$ via an inverse Mellin transform. We summarize the results in the following theorem.

Theorem 9 *Define the discriminant*

$$A(r,c,d) = \frac{\alpha(1-r)}{(\alpha/k)\log(\mathbf{Q}^{(k)}_{r,c,d})+1}.$$

Then the function $g(n,r,c,d)$ defined in (10) obeys the following asymptotic scheme.
If $A(r,c,d)<\alpha_1$, then $g(n,r,c,d)=O(n^{-M})$ for every $M>0$.
If $\alpha_1 < A(r,c,d) < \alpha_2$, then

$$\begin{aligned} g(n,r,c,d) = {} & \frac{n^{H(\rho_{r,c,d},r,c,d)}}{\sqrt{\log(n)}} \sum_{y\in\mathbb{Z}} \frac{n^{i\Im(H(\rho_{r,c,d}+iyK,r,c,d))}W(\rho_{r,c,d}+iyK,r,c,d)\Gamma(\rho_{r,c,d}+iyK+2)}{\sqrt{2\pi\frac{\partial H}{\partial s}(\rho_{r,c,d}+iyK,r,c,d)}} \\ & \times(1+O(\log(n)^{-1/2})). \end{aligned}$$

If $A(r,c,d)>\alpha_2$, then $g(n,r,c,d) = n^{H(-2,r,c,d)}W(-2,r,c,d)(1+O(n^{-\epsilon}))$ for some $\epsilon>0$.

The estimates of Theorem 9 can be derived using techniques that are standard (albeit pretty technical) in the analysis of tree structures. In the first regime, one can show that $H(s,r,c,d)$ is always decreasing in s, so integrating along $\Re(s)=s_0$ for $H(s_0)=-M$ gives the desired bound. In the second regime we use the saddle-point method, and in the final regime, we derive the asymptotics by taking the residue from the pole of $\Gamma(s+2)$ at $s=-2$.

Theorem 9, though certainly essential, is not in itself sufficient for our purposes, since we have to sum $g(n,\frac{\ell}{k},\frac{i}{\ell},\frac{j}{\ell})$ over a set of triplets (ℓ,i,j) that will grow unboundedly large as $n\to\infty$. The next lemma gives the needed statement about uniform convergence.

Lemma 3 *Suppose $\alpha_1<\alpha<\alpha_2$. Then there exists $r_0>0$ such that for all triplets (r,c,d) in the rectangle $R_0=[0,r_0]\times[0,1]^2$, we have $\alpha_1<A(r,c,d)<\alpha_2$, and the saddle-point estimate of Theorem 9 holds uniformly. Furthermore, the analogous result holds in the polar case, when $\alpha>\alpha_2$.*

The claims about $A(r,c,d)$ lying in particular ranges follow easily from the definition of $A(r,c,d)$. To show uniformity in the saddle point case, we use bounds from [Olv70], which are uniform on the compact set R_0. In the polar regime, we again use the compactness of R_0 to show that the s-partial of $H(s,r,c,d)$ at $s=0$ is bounded below by a positive constant, meaning that for some $\epsilon > 0$, we can uniformly take the left-hand side our Mellin box to be $\Re(s) = -2-\epsilon$, thereby obtaining an error that is $O(n^{H(-2-\epsilon,r,c,d)})$, with the (r,c,d) portion controlled by compactness.

7 Bounding the Tail

Theorem 9 justifies the content of $C_2(n)$ in the main Theorems 2 and 3. However, we still have to justify the uniform $(1+O(\cdot))$ error-bounds given in the leading equations of those theorems (which amounts to showing that our estimates for $g(n,r,c,d)$ are uniform outside the compact rectangle R_0) as well as prove our claim about the decay of the outer sum in $C_2(n)$.

We can accomplish both these tasks using the same argument. First, we unify the s-arguments for H in the polar and saddle-point cases into a single term,

$$\hat{\rho}_{r,c,d} := \begin{cases} \rho_{r,c,d} & : \alpha_1 < \alpha < \alpha_2 \\ -1 & : \alpha > \alpha_2. \end{cases} \tag{12}$$

Then we note that if we define

$$G(r,c,d) = \alpha r(-c\log(c) - (1-c)\log(1-c) - d\log(d) - (1-d)\log(1-d)) + H(\hat{\rho}_{r,c,d}, r, c, d), \tag{13}$$

then by Stirling's Formula we have

$$\binom{kr}{krc}\binom{kr}{krd} g(n,r,c,d) = n^{G(r,c,d)} \times Y(\log(n)),$$

where the function $Y(\log(n))$ is unimportant except for the fact that its growth/decay are in $\log(n)$. We now state an important and somewhat surprising result about the function G.

Lemma 4 *Let the function $G(r,c,d)$ be as in (13), and $A(r,c,d)$ the discriminant from Theorem 9. Then for any fixed r such that the set $\Omega_r := \{(c,d) \ : \ A(r,c,d) > \alpha_1\}$ is nonempty, the map $(c,d) \to G(r,c,d)$ attains its maximum at a unique ordered pair $(c_m(r), c_m(r))$ on the diagonal of Ω_r.*

The proof of Lemma 4, although not exceedingly difficult or technical, is rather long and (to us) not very intuitive. We therefore omit it. Lemma 4 allows us to define the function

$$F(r) = G(r, c_m(r), c_m(r)) \tag{14}$$

for every r on which the set Ω_r defined in Lemma 4 is nonempty. We now state two vital facts about this F, which are exactly the results needed complete the proof.

Lemma 5 *The function $F(r)$ defined at (14) is concave, and moreover $lim_{r\to 0} F'(r) < 0$.*

The statements in Theorems 2 and 3 about the decay of $C_2(n)$ immediately follow from Lemma 5, since we have $n^{F(0)-(\ell/k)F'(0)} \geq n^{F(\ell/k)} \geq \binom{\ell}{i}\binom{\ell}{j} n^{H(\hat{\rho}_{r,c,d},r,c,d)}$, and one readily verifies that $F(0) = h(\rho)$ in the saddle-point case and $h(0)$ in the polar case. It remains only to justify the global O-bounds at the beginning of Theorems 2 and 3 for those (r,c,d) outside the rectangle R_0 given in Lemma 3, which the following achieves.

Lemma 6 *With F as defined at (14) and g as at (10), for all sufficiently small r_0 there exists C such that*

$$\binom{kr}{krc}\binom{kr}{krd} g(n,r,c,d) \le C n^{F(0)-(r_0/2)F'(0)}$$

for all $r > r_0$ and all $(c,d) \in [0,1]$.

The main tool in proving Lemma 6 is Lemma 5, although some work is required in proving uniformity in (for example) cases where the saddle point $\hat{\rho}_{r,c,d}$ is very close to the pole at $s = -2$.

Acknowledgements

Ward thanks the authors of [BCN12] for their hospitality during several visits to Paris, which enabled him to work on this problem. Both authors also thank the authors of [PHNS09] for their pioneering work in this area. Moreover, we both acknowledge W. Szpankowski as a constant source of encouragement during the years in which we worked toward a method of solution. Finally, we thank the three anonymous referees for their very insightful comments, corrections, and suggestions.

References

[BCN12] F. Bassino, J. Clément, and P. Nicodème. Counting occurrences for a finite set of words: combinatorial methods. *ACM Transactions on Algorithms*, 8(3):31:1–28, July 2012.

[FGD95] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[FHZ14] M. Fuchs, H.-K. Hwang, and V. Zacharovas. An analytic approach to the asymptotic variance of trie statistics and related structures. *Theoretical Computer Science*, 527:1–36, 2014.

[JS05] P. Jacquet and W. Szpankowski. Analytic approach to pattern matching. In M. Lothaire, editor, *Applied Combinatorics on Words*, chapter 7. Cambridge, 2005.

[NW11] P. Nicodème and M.-D. Ward. *Towards the variance of the profile of suffix trees*, volume 23/2011, pages 1269–1272. Mathematisches Forschungsinstitut Oberwolfach, 2011.

[Olv70] F. W. J. Olver. Why steepest descents? *SIAM Review*, 12:228–247, 1970.

[PHNS09] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. *SIAM Journal on Computing*, 38:1821–1880, 2009.

[Szp01] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Addison-Wesley, Wiley, 2001.

[War05] M. D. Ward. *Analysis of the Multiplicity Matching Parameter in Suffix Trees*. PhD thesis, Purdue University, West Lafayette, IN, May 2005.

[War07] M. D. Ward. The average profile of suffix trees. In *The Fourth Workshop on Analytic Algorithmics and Combinatorics (ANALCO 07)*, pages 183–193, 2007.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Locally Restricted Sequential Structures and Runs of a Subcomposition in Integer Compositions

Edward A. Bender[1], Zhicheng Gao[2†]

[1]*Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112*
[2]*School of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S5B6*

We study part sizes of supercritical locally restricted sequential structures. This extends previous results about locally restricted integer compositions and part sizes in smooth supercritical compositional structures. Applications are given for runs of subcompositions. The problems are formulated as enumerating directed walks in sized infinite digraphs and the proofs depend heavily on earlier results by Bender and Canfield about infinite transfer matrices.

Keywords: sequential structure, local restriction, infinite transfer matrix, composition, run

1 Introduction

In [4] part sizes of compositional structures were studied. It was shown that if the composition is smooth supercritical then the numbers of parts of large sizes are asymptotically Poisson and an asymptotic expression was obtained for the expected value of the maximum part size which is accurate up to $o(1)$. In [3], part sizes of locally restricted nearly free integer compositions were studied. We will extend some of the major results of [3, 4] to locally restricted supercritical sequential structures. Runs of a single letter/part in words/compositions have been studied extensively; see e.g. [4, 6, 12]. In locally restricted structures, such as Carlitz compositions and Smirnov words, runs of a single part may not be allowed, and hence it is natural to consider runs of a substructure. Our main results on large part size distributions will be applied to maximum run length of a subcomposition in several classes of locally restricted compositions. Our approach follows that in [2, 3] using directed walks in sized infinite digraphs and properties of the corresponding infinite transfer matrices. Our proofs rely heavily on results in [2, 3] about infinite transfer matrices and large part size distributions.

2 Definitions

Definition 1 (Sequential structures) *Let $\mathcal{P}$ be a class of combinatorial structures, called* parts.

†Research supported by NSERC

- *Each part p has a positive integer* size*, denoted $|p|$.*
- *We use $\mathcal{P}_n$ to denote the set of parts of size n and assume $P_n := |\mathcal{P}_n| < \infty$ for each $n \in \mathbb{N}$.*
- *For each integer $k \geq 0$, we denote the class consisting of sequences of k parts by*

$$\mathrm{SEQ}_k(\mathcal{P}) := \{p_1 p_2 \dots p_k : p_j \in \mathcal{P}\}.$$

  Thus $\mathrm{SEQ}_0(\mathcal{P})$ contains only the empty sequence which has size and length 0. **Throughout this paper** *ε denotes the empty sequence.*
- *Let $\mathrm{SEQ}_{<k}(\mathcal{P})$ denote the class of sequences of at most $k-1$ parts.*
  Let $\mathrm{SEQ}(\mathcal{P}) := \cup_{k\geq 0}\mathrm{SEQ}_k(\mathcal{P})$, the set of all sequences.
- *If $\mathbf{a} = p_1 p_2 \dots p_k \in \mathrm{SEQ}(\mathcal{P})$, then the* length *of $\mathbf{a}$ is k and we write $\mathrm{len}(\mathbf{a}) = k$. The* size *of $\mathbf{a}$ is*

$$|\mathbf{a}| = |p_1| + |p_2| + \cdots + |p_k|$$

  and the distance *from p_i to p_j is $|i-j|$.*

Example 1 (Generalized compositions) *When studying $\mathrm{SEQ}(\mathcal{P})$, only the values of P_n are important. Thus we could think of $\mathcal{P}$ as parts in a generalized composition where the part n comes in P_n "colors", or whatever you choose to call them. If $P_n = 0$, there are no parts of size n. Thus, ordinary compositions correspond to $P_n = 1$ for all n and words on a k-letter alphabet correspond to $P_1 = k$ and $P_n = 0$ for $n > 1$. When parts of size n occur in n colors, $P_n = n$ and we have n-colored compositions, which were studied in [1, 9, 11]. We may also consider colored compositions where a part of size n corresponds to a multiset of n colored balls with N colors available. Here we have $P_n = \binom{n+N-1}{N-1}$.* □

Let $\mathcal{A} \subseteq \mathrm{SEQ}(\mathcal{P})$. It is called locally restricted if the parts of a structure in $\mathcal{A}$ within a fixed distance satisfy certain restrictions. Locally restricted integer compositions were studied in [2], where local restrictions are defined in terms of local restriction functions. The function was then used to construct a digraph. In this paper, we define local restrictions directly in terms of the digraph. Readers wishing to see the connection between local restriction functions and the digraph should consult [2].

Definition 2 (Locally restricted structures) *Let $m \in \mathbb{N}$, $\mathcal{S}, \mathcal{F} \subseteq \mathrm{SEQ}_{<m}(\mathcal{P})$ and $\mathcal{R} \subseteq \mathrm{SEQ}_m(\mathcal{P})$. The integer m is called the* span *of the locally restricted class of structures associated with the digraph D which has vertex set $V(D) = \mathcal{S} \cup \mathcal{R} \cup \mathcal{F}$. If $p \in \mathcal{S} \cap \mathcal{F}$, we allow two copies as separate vertices in the digraph. For convenience, we introduce two copies of the empty sequence, denoted by ε_s and ε_f, such that $\varepsilon_s \in \mathcal{S}$ and $\varepsilon_f \in \mathcal{F}$. Suppose D satisfies the following conditions.*

(a) There is an arc from ε_s to every other vertex in $\mathcal{S}$, and at least one arc from $\mathcal{S}$ to $\mathcal{R}$.

(b) There is an arc to ε_f from every other vertex in $\mathcal{F}$, and at least one arc from $\mathcal{R}$ to $\mathcal{F}$.

(c) The sub-digraph $D_{\mathcal{R}}$ of D induced by $\mathcal{R}$ is strongly connected, $|\mathcal{R}| \geq 2$, and the sub-digraph of D induced by $\mathcal{S} \cup \mathcal{F}$ contains no directed cycle.

The vertices of $\mathcal{R}$ will be called the recurrent vertices*, and the vertices of $\mathcal{S}$ and $\mathcal{F}$ will be called the* start vertices *and* finish vertices*, respectively. If a part $p \in \mathcal{P}$ appears in some recurrent vertex, then p is called a* recurrent part.

Let $\mathcal{W}$ denote the set of all directed walks in D from ε_s to ε_f. We use $\mathrm{SEQ}(\mathcal{P}; D)$ to denote the class of all structures of the concatenation form $\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_j$, where $\varepsilon_s\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_j\varepsilon_f \in \mathcal{W}$. If each element in $\mathrm{SEQ}(\mathcal{P}; D)$ arises from **only one** *walk in $\mathcal{W}$, we say that $\mathrm{SEQ}(\mathcal{P}; D)$ is a* locally restricted class *associated with D.*

The "only one" condition is required so that an ogf built from D will count each element in $\mathrm{SEQ}(\mathcal{P}; D)$ just once. The span is actually associated with the digraph. One can easily construct a digraph D' with span km for any $k \in \mathbb{N}$ such that $\mathrm{SEQ}(\mathcal{P}; D') = \mathrm{SEQ}(\mathcal{P}; D)$.

Definition 3 (Regular class) *A class $\mathrm{SEQ}(\mathcal{P}; D)$ of locally restricted structures will be called* regular *if it satisfies the following conditions.*

- *The gcd of the lengths of all directed cycles in $D_{\mathcal{R}}$ is equal to 1.*
- *There is a positive integer k and vertices $\mathbf{v}_0, \mathbf{v}_k \in \mathcal{R}$ such that $\gcd\{m - n : m, n \in S\} = 1$, where $S = \{n : n = |\mathbf{v}_0| + \cdots + |\mathbf{v}_k|$ for some directed walk $\mathbf{v}_0\cdots\mathbf{v}_k$ of length k in $D_{\mathcal{R}}\}$.*

Example 2 (Pattern avoidance) *Let B be a finite set of structures in $\mathrm{SEQ}(\mathcal{P})$, and $\mathcal{A}$ be the class of structures in $\mathrm{SEQ}(\mathcal{P})$ which don't contain any structure in B. We say that the structures in $\mathcal{A}$* avoid *the structures in B (or simply, avoid B). We may construct the class $\mathcal{A}$ using the following digraph. Let $m+1$ be the maximum length of the structures in B. Let $\mathcal{S} = \{\varepsilon_s\}$, $\mathcal{F} \subset \mathrm{SEQ}_{<m}(\mathcal{P})$ and $\mathcal{R} \subset \mathrm{SEQ}_m(\mathcal{P})$ be consisting of structures which avoid B. There is an arc from vertex $\mathbf{a}$ to a vertex $\mathbf{b}$ if and only if $\mathbf{ab}$ avoids B. There is also an arc from ε_s to every vertex in $\mathcal{R} \cup \mathcal{F}$. Then $\mathcal{A} = \mathrm{SEQ}(\mathcal{P}; D)$. We note that the second gcd condition in Definition 3 is not satisfied here because each recurrent vertex has size m.* □

Remark: Words over a finite alphabet which avoid certain patterns have been studied extensively (see, e.g., [8]). Here the size of a word is the length of the word. Since $\mathcal{P}$ is finite here, one may use the simpler transfer matrix $T(z)$ such that $T_{i,j}(z) = z^{|r_j|}$ for recurrent vertex $\mathbf{r}_j$ (See Definition 6 on page 8 for the transfer matrix). Consequently the ogf $F_R(z) = \mathbf{s}^t(I - T(z))^{-1}\mathbf{f}$ is a rational function of z. We note that in this case, each recurrent vertex has size m and so $T(z)$ is a function of z^m, and hence we may apply [2, Theorem 1] to T with $x = z^m$.

Definition 4 (Generating functions and supercritical structures) *Let $\mathcal{A} \subseteq \mathrm{SEQ}(\mathcal{P})$.*

- *Define the ordinary generating function (ogf) $P(z) := \sum_{p\in\mathcal{P}} z^{|p|} = \sum_{n\geq 1} P_n z^n$.*
- **Throughout this paper** *ρ is the radius of convergence of $P(z)$.*
- *If the radius of convergence of the ogf $A(z) := \sum_{\mathbf{a}\in\mathcal{A}} z^{|\mathbf{a}|}$ is less than ρ, we call $\mathcal{A}$* supercritical.

Example 3 (Alternating compositions) *Alternating, or up-down, compositions are compositions in which a part is alternately greater and less than the preceding part. We set $m = 2$, $\mathcal{P} = \mathbb{N}$, $\mathcal{S} = \mathrm{SEQ}_{<2}(\mathcal{P})$, $\mathcal{F} = \mathrm{SEQ}_{<2}(\mathcal{P})\backslash\{1\}$, and $\mathcal{R} = \{i, j \mid i > j\}$. The empty sequence in $\mathcal{S}$ (resp. $\mathcal{F}$) connects to every element in $\mathcal{R}$. There is an arc from $i \in \mathcal{S}$ to $j \in \mathcal{F}$ whenever $i < j$ or i is the empty sequence. The arcs having at least one end in $\mathcal{R}$ should be fairly easy to see. Every alternating composition, including the empty one, is associated with a unique directed walk from ε_s to ε_f. It is known [2] that the class of alternating compositions is regular and supercritical.* □

Definition 5 (Asymptotically free set) *Let $\mathcal{A} = \mathrm{SEQ}(\mathcal{P}; D)$ be a class of locally restricted structures with $|\mathcal{P}| = \infty$ and let $\mathcal{L}$ be an infinite set of recurrent parts. If $\mathbf{a} = p_1 \dots p_k$ and $t = \min(k, m-1)$, define $s(\mathbf{a}) := p_1 \dots p_t$ and $f(\mathbf{a}) := p_{k-t+1} \dots p_k$. Suppose there is a function $g : \mathrm{SEQ}_{<m}(\mathcal{P}) \times \mathrm{SEQ}_{<m}(\mathcal{P}) \to \mathbb{N}$ such that the following holds. If $\mathbf{a}l\mathbf{z} \in \mathrm{SEQ}(\mathcal{P}; D)$ with $l \in \mathcal{L}$ and $|l| \geq g(f(\mathbf{a}), s(\mathbf{z}))$, then $\mathbf{a}l'\mathbf{z} \in \mathrm{SEQ}(\mathcal{P}; D)$ whenever $l' \in \mathcal{L}$ and $|l'| \geq g(f(\mathbf{a}), s(\mathbf{z}))$. Then we call $\mathcal{L}$* asymptotically free.

What this says is that if l is large enough as determined by parts closer than distance m, then it can be replaced by any large enough part in $\mathcal{L}$.

Example 4 (k-Carlitz compositions) *A k-Carlitz composition is a composition in which each part is different from each of the preceding m parts, i.e. $p_i \neq p_j$ whenever $|i-j| \leq m$. (Carlitz compositions, for which adjacent parts differ, is the case $k = 1$. They have been studied extensively. See for example [10].) They are regular supercritical structures and a digraph D can be constructed with any $m \geq k$. Let $\mathcal{P} = \mathbb{N}$ and $\mathcal{S} = \{\varepsilon_s\}$. Let $\mathcal{R}$ (respectively $\mathcal{F}$) be all compositions in $\mathrm{SEQ}_m(\mathcal{P})$ (respectively $\mathrm{SEQ}_{<m}(\mathcal{P})$) which are k-Carlitz. There is an arc from ε_s to all vertices in $\mathcal{R} \cup \mathcal{F}$. There is an arc from $\mathbf{a} \in \mathcal{R}$ to $\mathbf{b} \in \mathcal{R} \cup \mathcal{F}$ whenever $\mathbf{ab}$ is an k-Carlitz composition. Since a part that is larger than all parts within distance k can be replaced by any larger part, the set $\mathbb{N}$ is asymptotically free in k-Carlitz compositions.* □

3 Main Results

Let r be the radius of convergence of the ogf for $\mathrm{SEQ}(\mathcal{P}; D)$. In the following, all logarithms will be to the base $1/r$. Our main results are the following.

Theorem 1 *Let $\mathcal{A} = \mathrm{SEQ}(\mathcal{P}; D)$ be a regular supercritical class of locally restricted structures, and $\mathcal{L} = \{l_1, l_2, \dots\}$ be an asymptotically free set. Assume $|l_1| < |l_2| < \cdots$. Select a structure $\mathbf{a}$ uniformly at random from $\mathcal{A}_n$. Let $\zeta_k(n)$ be the number of occurrences of l_k in $\mathbf{a}$. The following are true.*

(a) $|\mathcal{A}_n| = Ar^{-n}\left(1 + O\left(e^{-\delta n}\right)\right)$ *for some positive constants r, A and δ.*

(b) The distribution of $\zeta_k(n)$ is asymptotically normal with mean and variance asymptotically proportional to n.

(c) The limit

$$v_k = \lim_{n\to\infty} \frac{\mathbb{E}(\zeta_k(n))}{n} \tag{1}$$

exists, and $v_k \sim Cr^{|l_k|}$ as $k \to \infty$, for some positive constant C.

(d) Suppose there is a function $\omega_1(n) \to \infty$ such that $\{|l_k| : k \geq 1\} \cap [\log n - \omega_1(n), n]$ is not empty for all sufficiently large n. Then there is a function $\omega_2(n) \to \infty$ such that the random variables $\{\zeta_k(n) : \log n - \omega_2(n) \leq |l_k| \leq n\}$ are asymptotically independent Poisson random variables with means $\mu_k = Cnr^{|l_k|}$.

The distribution of large part sizes in general supercritical compositional structures has been studied extensively. Some latest results can be found in [4]. We will convert runs of a given subcomposition into free parts in a related class of locally restricted structures and apply Theorem 1(d) to derive the following.

Theorem 2 *Let* $\mathbf{c}$ *be a given composition such that* $\mathbf{c}$ *cannot be written in the form* $\mathbf{c} = \mathbf{xyx}$. *Let* $\mathcal{A} = \mathrm{SEQ}(\mathbb{N}; D)$ *be a regular supercritical class of locally restricted compositions with span* $m = \mathrm{len}(\mathbf{c})$. *Assume that* $\mathbf{c}$ *is a recurrent vertex and the digraph* D *contains the arc from* $\mathbf{c}$ *to itself. Let* C, r, $\mathbf{a}$ *and* ζ_k *be as defined in Theorem 1,* R_n *be the maximum run length of* $\mathbf{c}$ *in* $\mathbf{a}$, *and* $g_n(k)$ *be the probability that* $\mathbf{a}$ *contains exactly* k *runs of* $\mathbf{c}$ *of length* R_n. *Let* $\gamma \doteq 0.577216$ *be Euler's constant and define*

$$P_k(x) \;=\; \frac{\log e}{|\mathbf{c}|} \sum_{\ell \neq 0} \Gamma\left(k + \frac{2i\pi\ell \log e}{|\mathbf{c}|}\right) \exp\left(\frac{-2i\ell\pi \log x}{|\mathbf{c}|}\right). \tag{2}$$

Then

(a) $\mathbb{P}(R_n < k) \sim \exp\left(-\dfrac{Cn}{1-r^{|\mathbf{c}|}} r^{k|\mathbf{c}|}\right)$.

(b) $\mathbb{E}(R_n) = \dfrac{1}{|\mathbf{c}|} \log \dfrac{Cn}{1-r^{|\mathbf{c}|}} + \dfrac{\gamma \log e}{|\mathbf{c}|} - \dfrac{1}{2} - P_0\left(\dfrac{Cn}{1-r^{|\mathbf{c}|}}\right) + o(1)$.

(c) $g_n(k) = \dfrac{(1-r^{|\mathbf{c}|})^k}{k!} P_k\left(\dfrac{Cn}{1-r^{|\mathbf{c}|}}\right) + \dfrac{(1-r^{|\mathbf{c}|})^k \log e}{k|\mathbf{c}|} + o(1)$.

Corollary 1 *Theorem 2 holds for the following classes of compositions.*

(a) All compositions for any given composition $\mathbf{c}$, *where* $r = 1/2$ *and* $C = \frac{1}{2}(1-2^{-|\mathbf{c}|})^2$. *In particular,* $C = 1/8$ *when* $\mathbf{c} = 1$, *which gives Gafni's result [6].*

(b) Carlitz compositions for any given Carlitz composition $\mathbf{c}$, *where* $r \doteq 0.571350$ *is the smallest positive number satisfying* $\displaystyle\sum_{j\geq 1} \frac{r^j}{1+r^j} = 1$. *In particular, when* $\mathbf{c} = ab$ *with* $a \neq b$, *we have*

$$C = \frac{\left(1-r^{a+b}\right)^2}{(1+r^a)(1+r^b)} \frac{1}{\sum_{j\geq 1} \frac{jr^j}{(1+r^j)^2}}.$$

(c) k*-Carlitz compositions for any given* k*-Carlitz composition* $\mathbf{c}$.

(d) Alternating compositions for any given alternating composition $\mathbf{c}$, *where* $r \doteq 0.6363$.

(e) n*-color compositions (defined in Example 1) for any given* n*-color composition* $\mathbf{c}$, *where* $r = \frac{3-\sqrt{5}}{2}$.

(f) Colored compositions with $P_n = \binom{n+N-1}{N-1}$ *(defined in Example 1) for any given colored composition* $\mathbf{c}$, *where* $r = 1 - 2^{-1/N}$.

4 Converting Runs into Run Parts

The basic idea in the proof of Theorem 2 is to replace a k-long run of a given subcomposition $\mathbf{c}$ with a new part $\bar{k}$ with $|\bar{k}| = k|\mathbf{c}|$, and then apply Theorem 1(d) to a new class $\mathcal{A}'$ of locally restricted structures with parts in $\mathbb{N}$ as well as *run parts* $\bar{k}$ for $k \geq 1$. Let θ denote this replacement operation. For example, if $\mathbf{c} = 12$ and $\mathbf{a} = 12123122121$, then $\theta(\mathbf{a}) = \bar{2}3\bar{1}2\bar{1}1$.

Example 5 (Runs in unrestricted compositions) *Let $\mathcal{A}$ be the class of all compositions, and $\mathbf{c}$ be a given composition with $\mathrm{len}(\mathbf{c}) = m$ such that it cannot be written in the form $\mathbf{c} = \mathbf{xyx}$. The digraph D' for $\mathcal{A}' = \theta(\mathcal{A})$ is defined as follows.*

- $\mathcal{S}(D') = \{\varepsilon_s\}$.
- $\mathbf{a}' \in \mathcal{F}(D')$ *if and only if* $\mathbf{a}' \in \mathrm{SEQ}_{<m}(\mathcal{P})$, $\mathbf{a}'$ *does not contain two consecutive run parts.*
- $\mathbf{r}' \in \mathcal{R}(D')$ *if and only if* $\mathbf{r}' \in \mathrm{SEQ}_{m}(\mathcal{P})$, $\mathbf{r}' \neq \mathbf{c}$, $\mathbf{r}'$ *does not contain two consecutive run parts.*
- *There is an arc from a vertex $\mathbf{a}'$ to a vertex $\mathbf{b}'$ if and only if $\mathbf{a}'\mathbf{b}'$ does not contain $\mathbf{c}$ and does not contain two consecutive run parts.*

□

Let $\mathbf{c}$ be a given composition which cannot be written in the form $\mathbf{c} = \mathbf{xyx}$. It is easy to see that copies of $\mathbf{c}$ in a compositions $\mathbf{a}$ cannot overlap with each other. Define $\mathcal{P} = \mathbb{N} \cup \{\bar{k} : k \geq 1\}$ with $|\bar{k}| = k|\mathbf{c}|$. We have the following

Proposition 1 *Let $\mathcal{A} = \mathrm{SEQ}(\mathcal{P}; D)$ be a given regular supercritical class of locally restricted compositions with span m such that $\mathcal{S}(D) = \{\varepsilon_s\}$. Let $\mathbf{c} \in \mathcal{R}(D)$. Assume that D contains an arc from $\mathbf{c}$ to itself and $\mathbf{c}$ cannot be written in the form $\mathbf{c} = \mathbf{xyx}$. Let $\mathcal{A}' = \theta(\mathcal{A})$. Then*

(a) $\mathcal{A}' = \mathrm{SEQ}(\mathcal{P}; D')$ is a regular supercritical class with $\mathcal{P} = \mathbb{N} \cup \{\bar{k} : k \geq 1\}$ and some digraph D'.

(b) For each $\mathbf{a} \in \mathcal{A}$, $|\theta(\mathbf{a})| = |\mathbf{a}|$, and θ is a bijection between $\mathcal{A}_n$ and $\mathcal{A}'_n$.

(c) For each $\mathbf{a} \in \mathcal{A}$, Let $R(\mathbf{a})$ be the maximum run length of $\mathbf{c}$ in $\mathbf{a}$, and $M(\mathbf{a})$ be the maximum value of k such that $\bar{k}$ appears in $\theta(\mathbf{a})$. We have $R(\mathbf{a}) = M(\mathbf{a})$.

Proof: (a) We define the digraph D' as follows.

- $\mathcal{S}(D') = \{\varepsilon_s\}$.
- $\mathbf{a}' \in \mathcal{F}(D')$ if and only if $\mathbf{a}' \in \mathrm{SEQ}_{<m}(\mathcal{P})$, $\mathbf{a}'$ does not contain two consecutive run parts and $\theta^{-1}(\mathbf{a}')$ appears at the end of a composition in $\mathcal{A}$.
- $\mathbf{r}' \in \mathcal{R}(D')$ if and only if $\mathbf{r}' \in \mathrm{SEQ}_{m}(\mathcal{P})$, $\mathbf{r}' \neq \mathbf{c}$, $\mathbf{r}'$ does not contain two consecutive run parts, and $\theta^{-1}(\mathbf{r}')$ appears in a composition in $\mathcal{A}$.
- There is an arc from a vertex $\mathbf{a}'$ to a vertex $\mathbf{b}'$ if and only if $\mathbf{a}'\mathbf{b}'$ does not contain two consecutive run parts, does not contain $\mathbf{c}$, and $\theta^{-1}(\mathbf{a}'\mathbf{b}')$ appears in a composition in $\mathcal{A}$.

It can be verified that $\mathcal{A}' = \mathrm{SEQ}(\mathcal{P}; D')$ is regular supercritical.
(b) This follows immediately from the definition of θ.
(c) Since the copies of $\mathbf{c}$ in $\mathbf{a}$ cannot overlap with each other, θ induces a bijection between the set of run parts in $\theta(\mathbf{a})$ and the set of runs in $\mathbf{a}$, which implies $R(\mathbf{a}) = M(\mathbf{a})$. □

Example 6 (Runs of a subcomposition in m-Carlitz compositions) *Let $\mathbf{c}$ be a sequence of m distinct integers. The digraph D' for $\mathcal{A}' = \theta(\mathcal{A})$ is defined as follows.*

- $\mathcal{S}(D') = \{\varepsilon_s\}$.
- $\mathbf{a}' \in \mathcal{F}(D')$ *if and only if* $\mathbf{a}' \in \mathrm{SEQ}_{<m}(\mathcal{P})$, $\mathbf{a}'$ *does not contain two consecutive run parts and* $\theta^{-1}(\mathbf{a}')$ *is m-Carlitz.*
- $\mathbf{r}' \in \mathcal{R}(D')$ *if and only if* $\mathbf{r}' \in \mathrm{SEQ}_{m}(\mathcal{P})$, $\mathbf{r}' \neq \mathbf{c}$, $\mathbf{r}'$ *does not contain two consecutive run parts, and* $\theta^{-1}(\mathbf{a}')$ *is m-Carlitz.*
- *There is an arc from a vertex* $\mathbf{a}'$ *to a vertex* $\mathbf{b}'$ *if and only if* $\mathbf{a}'\mathbf{b}'$ *does not contain* $\mathbf{c}$*, does not contain two consecutive run parts, and* $\theta^{-1}(\mathbf{a}'\mathbf{b}')$ *is m-Carlitz.*

□

Example 7 (Runs of a subcomposition in alternating compositions) *Let $m \geq 2$, $\mathbf{c} = c_1 c_2 \cdots c_m$ be an alternating composition such that $\mathbf{c}^2$ is also alternating. The digraph D' for $\mathcal{A}' = \theta(\mathcal{A})$ is defined as follows.*

- $\mathcal{S}(D') = \{\varepsilon_s\}$.
- $\mathbf{a}' \in \mathcal{F}(D')$ *if and only if* $\mathbf{a}' \in \mathrm{SEQ}_{<m}(\mathcal{P})$, $\mathbf{a}'$ *does not contain two consecutive run parts and* $\theta^{-1}(\mathbf{a}')$ *is alternating.*
- $\mathbf{r}' \in \mathcal{R}(D')$ *if and only if* $\mathbf{r}' \in \mathrm{SEQ}_{m}(\mathcal{P})$, $\mathbf{r}' \neq \mathbf{c}$, $\mathbf{r}'$ *does not contain two consecutive run parts, and* $\theta^{-1}(\mathbf{a}')$ *is alternating.*
- *There is an arc from a vertex* $\mathbf{a}'$ *to a vertex* $\mathbf{b}'$ *if and only if* $\mathbf{a}'\mathbf{b}'$ *does not contain* $\mathbf{c}$*, does not contain two consecutive run parts, and* $\theta^{-1}(\mathbf{a}'\mathbf{b}')$ *is alternating.*

□

5 Outline of Proofs

The proofs are essentially the same as those in [2, 3]. In particular, we will make use of the infinite transfer matrix.

Definition 6 (Transfer matrix) *Let $\mathrm{SEQ}(\mathcal{P}; D)$ be a class of locally restricted structures as in Definition 2. Let $\mathbf{r}_1, \mathbf{r}_2, \ldots$ be an ordered list of vertices in $\mathcal{R}$. We define the transfer matrix $T(z)$ such that the (i, j)th entry of $T(z)$ is $T_{i,j}(z) = z^{|\mathbf{r}_i|+|\mathbf{r}_j|}$ if there is an arc in $D_{\mathcal{R}}$ from $\mathbf{r}_i$ to $\mathbf{r}_j$; otherwise $T_{i,j}(z) = 0$.*

The *weight* of an arc $(\mathbf{v}, \mathbf{w})$ in D is $z^{|\mathbf{v}|+|\mathbf{w}|}$. We define the weight of a directed walk in D to be the product of the weights of all the arcs in the walk. Let $F_R(z)$ be the ogf for structures in $\mathrm{SEQ}(\mathcal{P}; D)$ containing at least one recurrent vertex. It is not difficult to see that $F_R(z^2)$ is the sum of weights of all directed walks in $\mathcal{W}$ containing at least one recurrent vertex. We may express $F_R(z)$ in terms of $T(z)$, the *start vector* $\mathbf{s}(z)$ and the *finish vector* $\mathbf{f}(z)$, which are defined as follows. The ith component of $\mathbf{s}(z)$ is the sum of weights of all directed walks from ε_s to $\mathbf{r}_i$, and the jth component of $\mathbf{f}(z)$ is the sum of weights of all directed walks from $\mathbf{r}_j$ to ε_f. Since $T^k_{i,j}(z)$ is the sum of weights of all directed walks of length k from $\mathbf{r}_i$ to $\mathbf{r}_j$, we have

$$F_R(z^2) = \mathbf{s}(z)^t \sum_{k\geq 0} T^k(z)\mathbf{f}(z).$$

The following lemma summarizes the results which are used in the proof of our Theorems 1 and 2. These results are simple extensions of the corresponding results from [2] for locally restricted compositions.

Lemma 1 *Let $\mathcal{A} = \mathrm{SEQ}(\mathcal{P}; D)$ be a regular class of locally restricted structures.*

(a) Let r be the radius of convergence of the generating function $F_R(z)$. Then $r < 1$ and it is a simple pole of $F_R(z)$. Moreover $F_R(z)$ has no other singularity in $|z| \leq 1$.

(b) Let $p \in \mathcal{P}$ be a recurrent part and let X_n be the number of occurrences of p in a random structure in $\mathcal{A}_n$. There are constants $C_i > 0$ such that

$$\Pr(X_n < C_1 n) \ < \ C_2(1 + C_3)^{-n} \ \text{ for all } n.$$

(c) Let $\mathbf{c}$ be a given structure in $\mathcal{A}$. There is a constant B such that the probability that $\mathbf{c}$ occurs in a random structure in $\mathcal{A}_n$ is at most $Bnr^{|\mathbf{c}|}$.

Proof of Lemma 1: Part (a) is lifted from [2, Theorem 2] and its proof remains exactly the same. Parts (b) and (c) are lifted from [2, Lemma 1] and their proofs remain exactly the same. □

Proof of Theorem 1: Theorem 1(a) follows from Lemma 1 as shown in [2]. Now the proof of Theorem 1(b–d) is essentially the same as that of [2, Theorem 1], using Lemma 1 above, and Lemmas 3, 4 and 5 from [2]. □

Lemma 2 *Let $\mathcal{A} = \mathrm{SEQ}(\mathbb{N}; D)$ be a class of locally restricted compositions (or colored compositions defined in Example 1). Then $\mathcal{A}$ is supercritical.*

Proof: Let $G(z)$ be the ogf for all compositions in $\mathcal{A}$ which don't contain any recurrent vertex. Since each composition counted by $G(z)$ has at most K parts for some fixed integer K, we have $G(z) \leq \sum_{j=0}^{K} P(z)^j$ (coefficient-wise). Since the radius of convergence of $P(z)$ is 1, the radius of convergence of $G(z)$ is at least 1. □

Proof of Theorem 2: Let $\mathcal{P} = \mathbb{N} \cup \{\bar{k} : k \geq 1\}$, and $\mathcal{A}' = \mathrm{SEQ}(\mathcal{P}; D')$ be the new class defined in Proposition 1. Since θ is a bijection between $\mathcal{A}_n$ and $\mathcal{A}'_n$, the radius of convergence for $\mathcal{A}'$ is the same as

that for $\mathcal{A}$. It is easy to see that $\mathcal{L} = \{\bar{k} : k \geq 1\}$ is asymptotically free in $\mathcal{A}'$. Applying Theorem 1(d) with $|l_j| = j|\mathbf{c}|$, we obtain

$$\begin{aligned}\mathbb{P}(R_n < k) &= \prod_{k \leq j \leq n/|\mathbf{c}|} \mathbb{P}(\zeta_j(n) = 0) \exp\left(-Cnr^{j|\mathbf{c}|}\right) \\ &\sim \exp\left(-Cn\frac{r^{k|\mathbf{c}|}}{1-r^{|\mathbf{c}|}}\right).\end{aligned}$$

This establishes part (a).
Using $\mathbb{E}(R_n) = \sum_{k\geq 1}(1 - \mathbb{P}(R_n \leq k-1))$ and the same argument as in [3, page 25] (replacing r with $r^{|\mathbf{c}|}$), we obtain part (b). We remark that the $+$ sign before P_0 should be $-$ in the expression of $f(x)$ in [3, page 25]. The same applies to [3, Theorem 1(b,c)]
Part (c) follows from the same argument as in [3, page 27], with r being replaced by $r^{|\mathbf{c}|}$. □

Proof of Corollary 1: This follows immediately from Theorem 2. The values of r and C are computed as follows.
For part (a), it is clear $r = 1/2$. To obtain the value of C, we let $F(z)$ be the ogf of compositions containing a marked k-long run of a subcomposition $\mathbf{c}$. We note that $(1-z)/(1-2z)$ is the ogf of all compositions, and so $z^{|\mathbf{c}|}(1-z)/(1-2z)$ is the ogf of compositions ending with (or starting with) $\mathbf{c}$. Hence we have

$$F(z) = \left(\frac{1-z}{1-2z} - \frac{z^{|\mathbf{c}|}(1-z)}{1-2z}\right)^2 z^{k|\mathbf{c}|} = (1-z)^2(1-z^{|\mathbf{c}|})^2 z^{k|\mathbf{c}|}(1-2z)^{-2}.$$

It follows from the "transfer theorem" [5] that

$$[z^n]F(z) \sim \frac{n}{4}\left(1-2^{-|\mathbf{c}|}\right)^2 2^{n-k|\mathbf{c}|}.$$

Hence $C = \frac{1}{2}\left(1-2^{-|\mathbf{c}|}\right)^2$. In particular, $C = 1/8$ for the runs of 1, which gives Gafni's result.

For part (b), the value of r can be found, for example, in [3]. The expression of C (and r) can be derived as follows. Let $K(z)$ be the ogf of Carlitz compositions, and $F(z)$ be the ogf of Carlitz compositions with a marked k-long run of $\mathbf{c} = ab$, where $a \neq b$. For a given Carlitz composition $\mathbf{v}$, let $K_{\mathbf{v}}(z)$ be the ogf of Carlitz compositions that start with $\mathbf{v}$. Then we have

$$\begin{aligned}K_a(z) &= (K(z) - K_a(z))z^a, \\ K_{ab}(z) &= z^{a+b}(K(z) - K_b(z)), \\ K_{ba}(z) &= z^{b+a}(K(z) - K_a(z)), \\ F(z) &= z^{k(a+b)}(K(z) - K_b(z) - K_{ab}(z))(K(z) - K_a(z) - K_{ba}(z)).\end{aligned} \tag{3}$$

It follows that

$$F(z) = \frac{1}{(1+z^a)(1+z^b)} z^{k(a+b)}(1-z^{a+b})^2 K(z)^2. \tag{4}$$

The expression of $K(z)$ can be obtained from (3), which gives

$$K_a(z) = \frac{z^a}{1+z^a} K(z).$$

Summing over a, and noting $K(z) = 1 + \sum_{a\geq 1} K_a(z)$, we obtain

$$K(z) = \frac{1}{1 - \sum_{j\geq 1} \frac{z^j}{1+z^j}}.$$

Let r be the smallest positive number satisfying $\sum_{a\geq 1} \frac{z^j}{1+z^j} = 1$, it is easy to see that r is a simple pole of $K(z)$ and

$$K(z) = \frac{1}{\sum_{j\geq 1} \frac{jr^j}{(1+r^j)^2}} \frac{1}{1-z/r} + h(z),$$

where $h(z)$ is analytic in $|z| \leq r$. Consequently

$$[z^n]K(z) \sim \frac{1}{\sum_{j\geq 1} \frac{jr^j}{(1+r^j)^2}} r^{-n}.$$

It follows from (4) and the "transfer theorem" that

$$\begin{aligned} [z^n]F(z) &\sim \frac{n}{(1+r^a)(1+r^b)} r^{k(a+b)} (1-r^{a+b})^2 \left(\frac{1}{\sum_{j\geq 1} \frac{jr^j}{(1+r^j)^2}} \right)^2 r^{-n}, \\ \frac{[z^n]F(z)}{[z^n]K(z)} &\sim \frac{(1-r^{a+b})^2}{(1+r^a)(1+r^b)} \frac{1}{\sum_{j\geq 1} \frac{jr^j}{(1+r^j)^2}} n r^{k(a+b)}. \end{aligned}$$

Hence

$$C = \frac{(1-r^{a+b})^2}{(1+r^a)(1+r^b)} \frac{1}{\sum_{j\geq 1} \frac{jr^j}{(1+r^j)^2}}.$$

The value of r for part (d) can be found in [3]. To obtain the value of r for part (e), we note that the corresponding part generating function is $P(z) = z(1-z)^{-2}$. Solving the equation $r(1-r)^{-2} = 1$, we obtain $r = \frac{3-\sqrt{5}}{2}$. To obtain the value of r for part (f), we note that the corresponding part generating function is $P(z) = (1-z)^{-N} - 1$. Solving the equation $(1-r)^{-N} = 2$, we obtain $r = 1 - 2^{-1/N}$. □

6 Discussions

In this paper, we showed how infinite transfer matrix method developed in [2] can be applied to enumerate locally restricted regular supercritical sequential structures. Poisson distribution results were derived for sizes of free parts and applications are given for runs of subcompositions in several classes of compositions. In the upcoming full version of the paper, we plan to extend the results using a more general set up in terms of sized digraphs. Also in Theorem 1, we imposed the condition that the set $\mathcal{L}$ contains parts of distinct sizes. Such condition is unnecessary for unrestricted sequential structures as shown in [4]. It

might be possible to relax this condition for some classes of locally restricted sequential structures. For example, when the restriction is size-based, instead of part-based, we may allow the asymptotically free set $\mathcal{L}$ to contain several parts of the same size. Finally it might be possible to remove the restriction that the subcomposition $\mathbf{c}$ cannot be written in the form $\mathbf{c} = \mathbf{xyx}$ in Theorem 2. The mapping θ in Proposition 1 is still a bijection provided that the replacement is made at the earliest opportunity. However, the copies of $\mathbf{c}$ in a composition may overlap here, and the maximum run length $R(\mathbf{a})$ in a composition $\mathbf{a} \in \mathcal{A}$ may exceed the maximum run part $\bar{k}$ in $\theta(\mathbf{a})$. For example, consider the compositions $\mathbf{c} = 121$ and $\mathbf{a} = 12121121121$. We note that the first two copies of $\mathbf{c}$ overlap at the third position and $\theta(\mathbf{a}) = \bar{1}21\bar{2}$. The maximum run part is $\bar{2}$ here, but the maximum run length of $\mathbf{c}$ in $\mathbf{a}$ is 3.

References

[1] A.K. Agarwal, n-Colour Compositions, *Indian J. Pure Appl. Math.* **31**(11) (2000), 1421–1427.

[2] E.A. Bender and E.R. Canfield, Locally restricted compositions II. General restrictions and infinite matrices, *Elec. J. Combin.* **16** (2009) R108, 35pp.

[3] E.A. Bender, E.R. Canfield and Z. Gao, Locally restricted compositions IV. Nearly free large parts and gap-freeness, *Elec. J. Combin.* **19**:P14, 2012.

[4] E.A. Bender and Z. Gao, Part Sizes of Smooth Supercritical Compositional Structures, *Combin. Prob. Comp.*, **23** (2014) 686–716.

[5] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, 2009.

[6] A. Gafni, Longest run of equal parts in a random integer composition, *Discrete Math.* **338** (2015) 236–247.

[7] P. Grabner, A. Knopfmacher and H. Prodinger, Combinatorics of geometrically distributed random variables: Run statistics, *Theoret. Computer Sci.* **297** (2003) 261–270.

[8] S. Heubach and T. Mansour, *Combinatorics of compositions and words*, CRC / Taylor & Francis 2009.

[9] G. Narang and A.K. Agarwal, n-colour self-inverse compositions, *Proc. Indian Acad. Sci. Math. Sci.* **116**(3) (2006), 257–266.

[10] G. Louchard and H. Prodinger, Probabilistic Analysis of Carlitz Compositions, *Disc. Math and Theor. Comp. Sci.* **5** (2002), 71–96.

[11] G. Narang and A. K. Agarwal, Lattice paths and n-colour compositions, *Discrete Mathematics* **308** (2008), 1732–1740.

[12] H.S. Wilf, The distribution of run lengths in integer compositions, *Elec. J. Combin.* **18** #2 (2011) P23, 5pp.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Full asymptotic expansion for Pólya structures

Antoine Genitrini[1]

[1] *Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris.*
Email: Antoine.Genitrini@lip6.fr.

In order to obtain the full asymptotic expansion for Pólya trees, i.e. rooted unlabelled and non-plane trees, Flajolet and Sedgewick observed that their specification could be seen as a slight disturbance of the functional equation satisfied by the Cayley tree function. Such an approach highlights the complicated formal expressions with some combinatorial explanation. They initiated this process in their book but they spared the technical part by only exhibiting the first-order approximation. In this paper we exhibit the university of the method and obtain the full asymptotic expansions for several varieties of trees. We then focus on three different varieties of rooted, unlabelled and non-plane trees, Pólya trees, rooted identity trees and hierarchies, in order to calculate explicitly their full singular expansions and asymptotic expansions.

Keywords: Unlabelled non-plane trees; Full Puiseux expansion; Full asymptotic expansion; Analytic Combinatorics.

1 Introduction

By using either Darboux's method or singularity analysis, we easily get the dominant coefficients of the asymptotic expansions for the number of some specific Pólya structures; a Pólya structure being decomposable by using some Pólya operators like the multiset MSET or the powerset PSET constructions. For the numbers of hierarchies (a specific class of trees) of size 100 the relative error between the exact number and the first-order approximation is only around 0.01% (note that it is only $10^{-10}\%$ with an 8-order approximation). However for small hierarchies, the first-order approximation is not precise: the relative error for the trees of size 20 is around 0.3% whereas it is only around 0.0004% with the 8-order approximation (cf. Fig. 1).

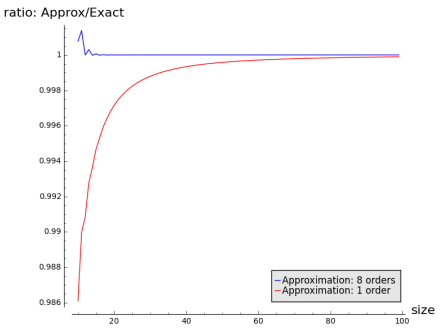


Fig. 1: Ratio between the approximations and the exact numbers of hierarchies

In a technical report [Fin03c], Finch provided recurrence formulas to compute all the coefficients in the asymptotic expansion for Pólya trees. He developed there the classical Darboux's method to derive the recurrences and computed explicitly the five most important coefficients.

According to Finch's report, Flajolet proposed at that time to study the fundamental equation given by the Weierstrass Preparation Theorem as, somehow, a slight disturbance of

the functional equation satisfied by the Cayley tree function. Using this point of view, the procedure to exhibit the full asymptotic expansion is much more highlighted and the complicated formal expressions can be combinatorially understood. Flajolet and Sedgewick initiated this process in their book [FS09, p. 477] in the context of Pólya trees but they spared the technical part of the proof by only exhibiting the first-order approximation.

In this paper, we explain why such an approach is generic to obtain easily the full asymptotic expansions for several varieties of trees. We focus on varieties that can be seen as a disturbance of the Cayley function in the way that they can be described by their generating function $T(z)$ as:

$$T(z) = \zeta(z)\exp(T(z)),$$

for some constrained function $\zeta(z)$. For such classes of trees, we exhibit the full Puiseux (i.e., singular) expansion of the generating series. We then compute the generic full asymptotic expansion of the number of trees. In Section 3, we then focus on three different varieties of rooted, unlabelled and non-plane trees. The first class of trees is the classical set of Pólya trees that already appears in the papers of Cayley [BLW76], Pólya [Pól37] and Otter [Ott48]. The generating function of Pólya trees is easily described with a functional equation using the multiset construction. By replacing the construction by the powerset operator we get the class of rooted identity trees, the second class we are interested in. Such trees are studied, for example, in the work of Harary *et al.* in [HRS75]. Finally we deal with hierarchies, i.e., rooted unlabelled non-plane trees without nodes of arity 1. This class has been introduced by Cayley too, but it is also directly linked to series-parallel networks in the papers of Riordan and Shannon [RS42] and Moon [Moo87]. In the Section 3.4, we give numerical approximations for the first coefficients of the singular and the asymptotic expansions of each specific variety of trees. We conclude the paper (Section 4) by mentioning several other structures where our generic approach could be applied directly.

2 Main results

For each of the varieties under consideration, the fundamental idea consists, from an analytic point of view, at studying its generating function as a disturbance of the classical Cayley tree function (cf. e.g. [FS09, p. 127]). Let $C(z)$ be the Cayley tree function; it satisfies the functional equation

$$C(z) = z \cdot \exp(C(z)). \tag{1}$$

Its dominant singularity is $1/e$ and $C(1/e) = 1$. Recall that the Cayley tree function is closely related to the Lambert W function. Many fundamental results about this classical function are given in the paper of Corless *et al.* [CGH$^+$96].

In order to obtain generically the full asymptotic expansion of the number of the structures of a variety of trees, let us first compute the full Puiseux expansion (i.e., the full singular expansion) of the Cayley tree function and then study how the disturbance induced by a given variety modifies this behaviour. Let us recall the definition of Bell polynomials, extensively studied in Comtet's book [Com74] and denoted by $B_{n,k}(\cdot)$:

$$B_{n,k}\left(x_1,\dots,x_{n-k+1}\right) = \sum_{\substack{c_1,\dots,c_{n-k+1}\geq 0\\ \sum_i c_i = k\\ \sum_i i c_i = n}} \frac{n!}{c_1!\cdots c_{n-k+1}!}\left(\frac{x_1}{1!}\right)^{c_1}\cdots\left(\frac{x_{n-k+1}}{(n-k+1)!}\right)^{c_{n-k+1}}.$$

The Bell polynomials appear naturally in Faà di Bruno's formula [Com74] that states the value of iterated derivatives of the composition of two functions.

Proposition 1 *The full Puiseux expansion of the Cayley tree function is*

$$C(z) \underset{z\to 1/e}{=} 1-\sqrt{2}\sqrt{1-ez}-\sum_{n\geq 2}\left(\sum_{k=1}^{n-1}(-1)^k B_{n-1,k}\left(\frac{1}{3},\frac{1}{4},\ldots,\frac{1}{n-k+2}\right)\prod_{i=0}^{k-1}(n+2i)\right)\frac{2^{n/2}}{n!}\,(1-ez)^{n/2},$$

where the functions $B_{n,k}(\cdot)$ are the Bell polynomials.

The calculation of the first terms of the singular expansion gives

$$C(z) \underset{z\to 1/e}{=} 1-\sqrt{2}\sqrt{1-ez}+\frac{2}{3}(1-ez)-\frac{11}{36}\sqrt{2}(1-ez)^{3/2}+\frac{43}{135}(1-ez)^2-$$
$$\frac{769}{4320}\sqrt{2}(1-ez)^{5/2}+\frac{1768}{8505}(1-ez)^3-\frac{680863}{5443200}\sqrt{2}(1-ez)^{7/2}+\mathcal{O}\left((1-ez)^4\right).$$

Let us recall that the expansion until $\mathcal{O}((1-ez)^{3/2})$ has been derived in [FS09]. We prove the full expansion with their approach but with further precision. Note that, in the formula of Proposition 1, the inner sum of k can be factored in the same way as the classical Ruffini-Horner method for polynomial evaluation. Doing so makes its computations much more efficient.

The second step consists in studying the ordinary generating function $T(z)=\sum_{n\geq 0}T_n z^n$ of the tree variety under consideration as a disturbance of the Cayley tree function. We follow the approach presented in [FS09, p. 477] for Pólya trees. We assume the existence of a function $\zeta(z)$ such that

$$T(z)=\zeta(z)\cdot\exp(T(z)). \tag{2}$$

Theorem 2 *Let $\mathcal{T}$ be a variety of trees whose generating function is $T(z)$, and ρ be its dominant singularity. If the generating function $T(z)$ satisfies the Equation* (2)*, if the dominant singularity of $\zeta(z)$ is strictly larger than ρ and if $\zeta^{(1)}(\rho)\neq 0$, then $T(z)$ satisfies the following full Puiseux expansion*

$$T(z) \underset{z\to\rho}{=} 1+\sum_{n\geq 1}t_n\left(1-\frac{z}{\rho}\right)^{n/2},$$

with $t_1=-\sqrt{2e\rho\zeta^{(1)}(\rho)}$; and, for all $n>1$

$$t_n=-\frac{B(n)}{n!}\left(2e\rho\zeta^{(1)}(\rho)\right)^{n/2}-\sum_{\substack{\ell=1\\ n\equiv\ell \bmod 2}}^{n-1}(-1)^{(n-\ell)/2}\rho^{n/2}\cdot\frac{B(\ell)}{\ell!}\left(2e\zeta^{(1)}(\rho)\right)^{\ell/2}$$
$$\cdot\sum_{r=1}^{\frac{n-\ell}{2}}\binom{\ell/2}{r}\frac{1}{(\zeta^{(1)}(\rho))^r}\sum_{\substack{i_1,\ldots,i_r\geq 1\\ \sum_j i_j=\frac{n-\ell}{2}}}\frac{\zeta^{(i_1+1)}(\rho)}{(i_1+1)!}\cdots\frac{\zeta^{(i_r+1)}(\rho)}{(i_r+1)!},$$

where $\zeta^{(i)}(z)$ stands for the ith derivative of $\zeta(z)$, $B(1)=1$, and for all $\ell>1$,

$$B(\ell)=\sum_{k=1}^{\ell-1}(-1)^k B_{\ell-1,k}\left(\frac{1}{3},\frac{1}{4},\ldots,\frac{1}{\ell-k+2}\right)\prod_{i=0}^{k-1}(\ell+2i).$$

Proof key idea: The complete proof follows the strategy of Flajolet and Sedgewick. The main idea is to compose the Puiseux expansion of $C(z)$ at the singularity $1/e$ and the analytic expansion of $\zeta(z)$ at the dominant singularity of $T(z)$. □

In Theorem 2, the assumption $\zeta^{(1)}(\rho) \neq 0$ could be replaced by a weaker assumption that there exists an integer $r > 0$ such that $\zeta^{(r)}(\rho) \neq 0$. Making this weaker assumption would however make the proof a bit more technical without adding substantial information.

The the first terms of the singular expansion of $T(z)$ are given by

$$\begin{aligned} T(z) \underset{z\to\rho}{=} & 1 - \sqrt{2e\rho\zeta^{(1)}(\rho)}\sqrt{1-\frac{z}{\rho}} + \frac{2e\rho\zeta^{(1)}(\rho)}{3}\left(1-\frac{z}{\rho}\right) - \left(\frac{11\sqrt{2}(e\rho\zeta^{(1)}(\rho))^{3/2}}{36} - \frac{\sqrt{2e}\rho^{3/2}\zeta^{(2)}(\rho)}{4\sqrt{\zeta^{(1)}(\rho)}}\right)\left(1-\frac{z}{\rho}\right)^{3/2} \\ & + \left(\frac{43(e\rho\zeta^{(1)}(\rho))^2}{135} - \frac{e(\rho\zeta^{(2)}(\rho))^2}{3}\right)\left(1-\frac{z}{\rho}\right)^2 - \left(\frac{769\sqrt{2}(e\rho\zeta^{(1)}(\rho))^{5/2}}{4320} - \frac{11\sqrt{2}\rho^{5/2}(e\zeta^{(1)}(\rho))^{3/2}\zeta^{(2)}(\rho)}{48\zeta^{(1)}(\rho)}\right. \\ & \left. - \frac{\sqrt{2}\rho^{5/2}\sqrt{e\zeta^{(1)}(\rho)}}{96}\left(\frac{3(\zeta^{(2)}(\rho))^2}{(\zeta^{(1)}(\rho))^2} - \frac{8\zeta^{(3)}(\rho)}{\zeta^{(1)}(\rho)}\right)\right)\left(1-\frac{z}{\rho}\right)^{5/2} + \mathcal{O}\left(\left(1-\frac{z}{\rho}\right)^3\right). \end{aligned}$$

We are now ready to compute the full asymptotic expansion for the class $\mathcal{T}$.

Theorem 3 *Let $\mathcal{T}$ be a variety of trees whose generating function is $T(z)$, and ρ be its dominant singularity. If the generating function $T(z)$ satisfies the Equation* (2)*, if the dominant singularity of $\zeta(z)$ is strictly larger than ρ and if $\zeta^{(1)}(\rho) \neq 0$, then asymptotically when n tends to infinity,*

$$T_n \underset{n\to\infty}{\sim} \frac{\rho^{-n}}{\sqrt{\pi n^3}} \sum_{\ell\geq 0} \frac{1}{n^\ell} \cdot \left(\sum_{r=1}^{\ell+1} Q_r R_{\ell+1-r}\right),$$

where

$$Q_r = \sum_{j=0}^{r-1} (-1)^{j+1} t_{2j+1} \sum_{\substack{\ell_0,\dots,\ell_j \geq 1 \\ \sum_i \ell_i = r}} \prod_{i=0}^{j} \left(i+\frac{1}{2}\right)^{\ell_j} \qquad \textit{for all } r > 0;$$

with the sequence (t_i) defined in Theorem 2, $R_0 = 1$ and

$$R_\ell = \sum_{\substack{r=1 \\ r\equiv \ell \bmod 2}}^{\ell} \sum_{\substack{k_1,\dots,k_r \geq 1 \\ \sum_j k_j = \frac{\ell+r}{2}}} \prod_{i=1}^{r} \frac{(2^{-2k_i}-1)\sum_{s=0}^{2k_i} \frac{1}{s+1} \sum_{j=0}^{s} (-1)^j \binom{s}{j} j^{2k_i}}{(\ell - 2k_1 - \cdots - 2k_{i-1} + i - 1)k_i} \qquad \textit{for all } \ell > 0.$$

In particular, the first few terms in the asymptotic expansion of T_n are given by

$$\begin{aligned} T_n \underset{n\to\infty}{=} & \frac{\rho^{-n}}{\sqrt{\pi n^3}}\left(-\frac{t_1}{2} - \frac{3(t_1-4t_3)}{16n} - \frac{5(5t_1-72t_3+96t_5)}{256n^2} - \frac{105(t_1-44t_3+160t_5-128t_7)}{2048n^3}\right. \\ & \left. - \frac{21(79t_1-10800t_3+81600t_5-161280t_7+92160t_9)}{65536n^4} + \mathcal{O}\left(\frac{1}{n^5}\right)\right), \end{aligned}$$

where the t_i's are given in the Theorem 2.

3 Different varieties of rooted unlabelled and non-plane trees

In the following three sections, we will show how both Theorems 2 and 3 directly apply to three families of trees, namely the Pólya trees, the rooted identity trees and the hierarchies. In each of these sections, we will use the same notations $\mathcal{T}$, $T(z)$ and $\zeta(z)$ to refer to the family of considered trees.

For each of the three examples, we proceed in two steps. First we focus on efficient recurrences in order to compute the first numbers of the sequence $(T_n)_{n\in\mathbb{N}}$ that encodes for each positive integer n the number of trees of size n. Second, by using the numerical procedure given in [FS09, p. 477], we compute an approximation of the dominant singularity of $T(z)$.

Finally, at the end of the section, we exhibit two Tables 1 and 2 to compare the numerical approximations (according to each class of trees) of the coefficients given in the Theorems 2 and 3. We also exhibit the typical gain in the relative error obtained by using a more precise asymptotic approximation.

3.1 Pólya trees

A Pólya tree is a rooted unlabelled and non-plane tree. Let us denote by $\mathcal{T}$ the set of Pólya trees. It satisfies the following unambiguous specification :

$$\mathcal{T} = \mathcal{Z} \times \textsc{MSet}\,\mathcal{T},$$

because a Pólya tree is by definition a root, specified by $\mathcal{Z}$ (of size 1), followed by a multiset of Pólya trees (we refer the reader to [FS09] for more details). By the *symbolic method* (cf. [FS09]), we get

$$T(z) = z \exp\left(\sum_{i>0} \frac{T(z^i)}{i}\right), \tag{3}$$

with $T(z)$ being the ordinary generating function enumerating $\mathcal{T}$. The latter formula already appears in Pólya's paper [Pól37] and has been sketched by Cayley ([BLW76, p. 67]) as an introduction to the counting theory for unlabelled objects. This method takes into account symmetries of the objects and thus quantifies isomorphisms. We have a classical alternative definition: cf. e.g. [FS09, p. 71].

$$T(z) = z \cdot \prod_{n>0} \frac{1}{(1-z^n)^{T_n}}, \tag{4}$$

with T_n the number of trees of size n in $\mathcal{T}$. Some combinatorial arguments, given in [FS09, p. 27–30], prove that both definitions are equivalent. From the latter Equation (4), we deduce a recurrence for the sequence $(T_n)_{n\in\mathbb{N}}$ for Pólya trees.

Fact 4 *The sequence $(T_n)_{n\in\mathbb{N}}$ enumerating Pólya trees satisfies*

$$T_n = \begin{cases} n & \text{if } n \in \{0,1\} \\ \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n-1} iT_i \left(\sum_{m=1}^{\lfloor \frac{n-1}{i} \rfloor} T_{n-mi}\right) & \text{if } n > 1. \end{cases}$$

This result is given as an exercise by Knuth [Knu97, p. 395]. Furthermore, Otter [Ott48] proved a very similar recurrence for unrooted trees. The first values of the sequence, given in OEIS[(i)] sequence A000081, are

$$0, 1, 1, 2, 4, 9, 20, 48, 115, 286, 719, 1842, 4766, 12486, 32973, 87811, \ldots$$

The number of Pólya trees from each size from 1 to n can be computed in $\mathcal{O}(n^2)$ arithmetic operations (by using memoization).

Proof of Fact 4: Several authors, in particular, Flajolet and Sedgwick obtained such a recurrence by using the logarithmic derivative of $T(z)$: for all $n > 1$

$$z\frac{T'(z)}{T(z)} = 1 + \sum_{n>0} T_n \frac{nz^n}{1-z^n}.$$

We rewrite this equation as

$$zT'(z) = \left(1 + \sum_{n>0} nT_n \frac{z^n}{1-z^n}\right) T(z).$$

Extracting the n-th coefficient of the generating functions gives:

$$nT_n = T_n + \sum_{i=1}^{n-1} \left([z^i] \sum_{m>0} mT_m \frac{z^m}{1-z^m}\right) T_{n-i}.$$

Since $[z^k](1-z^m)^{-1}$ equals 1 if m divides k and 0 otherwise, we get

$$(n-1)T_n = \sum_{i=1}^{n-1} \left(\sum_{m|i} mT_m\right) T_{n-i}.$$

The notation $m|i$ corresponds to the condition that the integer m divides the integer i. The stated formula is obtained by interchanging the two sums. □

By using Flajolet and Sedgewick's numerical procedure (cf. [FS09, p. 477]) with $n = 200$ terms, we get the following 50-digits approximation of ρ:

$$\rho \approx 0.33832185689920769519611262571701705318377460753297\ldots$$

We are now interested in the full Puiseux expansion of the generating function of Pólya trees. In view of Equations (2) and (3), we define have

$$T(z) = \zeta(z) \cdot \exp(T(z)), \qquad \text{where } \zeta(z) = z \cdot \exp\left(\sum_{n\geq 2} \frac{T(z^n)}{n}\right). \tag{5}$$

Fact 5 *The function $\zeta(z)$ defined for Pólya trees satisfies the assumptions of the Theorems 2 and 3.*

[(i)] OEIS: On-line Encyclopedia of Integer Sequences

This fact has already been proved by Cayley as mentioned in [BLW76, p. 67]. We recall here the arguments given in [FS09, p. 477].

Proof: The definition of $\zeta(z)$ given in Equation (**??**) implies that its dominant singularity is $\sqrt{\rho}$, (with the constant ρ being the dominant singularity of $T(z)$). Since $1/e$ is the dominant singularity of the Cayley tree function $C(z)$ and $[z^n]T(z) > [z^n]C(z)$ (by using Equation (3)) for n sufficiently large, we get $\rho \leq 1/e$. Thus $\sqrt{\rho} > \rho$ and we finally infer that the function $\zeta(z)$ is analytic beyond the disc of convergence of $T(z)$. Finally we easily get $\zeta'(\rho) > 0$. □

Theorem 2 and the above approximation for ρ give the first coefficients for the Puiseux expansion of Pólya trees presented in the Table 1. The computations of the numbers t_i's have been done with an approximation of the function $\zeta(z)$, computed with the truncation of the series $T(z)$ after the 100-th first coefficients. Experimentally, it seems that the accuracy is actually much larger than the 20 digits given in Table 1.

Finally the previous approximations and the result of Theorem 3 give

$$T_n \underset{n\to\infty}{=} \frac{\rho^{-n}}{\sqrt{\pi n^3}}\left(0.7797450101873204419\ldots + \frac{0.07828911261061096133\ldots}{n} + \frac{0.3929402676631860168\ldots}{n^2} + \frac{1.537879315978838092\ldots}{n^3} + \frac{8.200844090435596194\ldots}{n^4} + \mathcal{O}\left(\frac{1}{n^5}\right)\right).$$

Note that, from here, it is then easy to get back the first evaluations exhibited by Finch [Fin03c].

3.2 Rooted identity trees

A rooted identity tree is a rooted unlabelled (non-plane) tree for which the only automorphism preserving the root node is the identity. Harary *et al.* studied this class of trees in [HRS75]. In his book [Fin03a], Finch also mentions this class. Intuitively, whereas a Pólya tree can be seen as a root followed by a multiset of Pólya trees, a rooted identity tree can be seen as a root followed by a set of rooted identity trees (i.e., no repetition is allowed). Let us denote by $\mathcal{T}$ the set of rooted identity trees. It satisfies the following unambiguous specification

$$\mathcal{T} = \mathcal{Z} \times \textsc{PSet}\,\mathcal{T}.$$

The symbolic method gives the functional equation

$$T(z) = z\exp\left(\sum_{i>0}(-1)^{i-1}\frac{T(z^i)}{i}\right).$$

An equivalent formula for the function $T(z)$ is

$$T(z) = z\cdot\prod_{n>0}(1+z^n)^{T_n}.$$

In order to obtain an efficient recurrence relation satisfied by the numbers of rooted identity tree, we use the same strategy as above (for Pólya trees), and thus obtain:

Proposition 6 *The sequence $(T_n)_{n\in\mathbb{N}}$ enumerating rooted identity trees satisfies*

$$T_n = \begin{cases} n & \text{if } n \in \{0,1\} \\ \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n-1} iT_i\left(\sum_{m=1}^{\lfloor\frac{n-1}{i}\rfloor}(-1)^{m+1}T_{n-mi}\right) & \text{if } n > 1. \end{cases}$$

The first values of the sequence, see in OEIS A004111, are

$$0, 1, 1, 1, 2, 3, 6, 12, 25, 52, 113, 247, 548, 1226, 2770, 6299, \ldots$$

The number of rooted identity trees from each size from 1 to n can be computed in $\mathcal{O}(n^2)$ arithmetic operations. Once we are able to compute efficiently the first numbers T_n we can estimate the dominant singularity of $T(z)$ to be approximately

$$\rho \approx 0.39721309688424004148565407022739873422987370995276\ldots$$

Obviously this dominant singularity is larger than the one for Pólya trees because there are less rooted identity trees than Pólya trees.

To describe $T(z)$ like in Equation (2), we get $\zeta(z) = z \cdot \exp\left(\sum_{n\geq 2}(-1)^{n-1}\frac{T(z^n)}{n}\right)$.

Proposition 7 *The function $\zeta(z)$ defined in the context of rooted identity trees satisfies the assumptions of the Theorems 2 and 3.*

The approximations of the first coefficients of the Puiseux expansion for rooted identity trees are given in the Table 1. The second Table 2 gives the approximations of the asymptotic expansion of T_n:

$$T_n \underset{n\to\infty}{=} \frac{\rho^{-n}}{\sqrt{\pi n^3}}\left(0.64257907974426947\ldots - \frac{0.18511979777663370\ldots}{n} - \frac{0.42724272900609787\ldots}{n^2} - \frac{2.25545556898721207\ldots}{n^3} - \frac{16.60970953335647846\ldots}{n^4} + \mathcal{O}\left(\frac{1}{n^5}\right)\right).$$

It seems that these numbers do not appear elsewhere in the literature.

3.3 Hierarchies

A hierarchy is a rooted unlabelled and non-plane tree with no node of arity 1. The size notion for hierarchies is the number of leaves. This class already appears in the work of Cayley (cf. [BLW76, p. 43]. Using the notations from [FS09, p. 72] for hierarchies, we have both following specification and functional equation for its generating function

$$\mathcal{T} = \mathcal{Z} + \textsc{MSet}_{\geq 2}\mathcal{T}, \qquad T(z) = \frac{1}{2}\left(z - 1 + \exp\left(\sum_{i>0}\frac{T(z^i)}{i}\right)\right).$$

Again, we obtain a recurrence formula that computes the numbers T_n.

Proposition 8 *The sequence $(T_n)_{n\in\mathbb{N}}$ enumerating hierarchies satisfies*

$$T_n = \begin{cases} n & \text{if } n \in \{0,1\} \\ \dfrac{1}{n}\displaystyle\sum_{\substack{m|n \\ m\neq n}} mT_m + \frac{2}{n}\left(\sum_{i=1}^{n-1} iT_i \sum_{m=1}^{\lfloor\frac{n-1}{i}\rfloor} T_{n-mi} - \frac{1}{2}\delta_{\{n-mi=1\}}\right) & \text{if } n > 1, \end{cases}$$

with the notation $\delta_{\{n-mi=1\}}$ evaluates to 1 if $n - mi = 1$ and to 0 otherwise.

The first values of the sequence, see in OEIS A000669, are given by

$$0, 1, 1, 2, 5, 12, 33, 90, 261, 766, 2312, 7068, 21965, 68954, 218751, 699534, \ldots$$

They are stored (there the sequence is shifted by 1). We note that in this context, we cannot easily simplify the recurrence in order to avoid a sum over the divisors of n (for T_n). However here, the sum is not inside another one, thus the complexity (in the number of arithmetic operations) to compute T_n is quadratic. We estimate the dominant singularity of $T(z)$ to be approximately

$$\rho \approx 0.28083266698420035539318755911632333333736599643391\ldots$$

In order to fall under the framework described by Equation (2), we need to consider the generating function $\tilde{T}(z) = T(z) - \frac{1}{2}(1-z)$. The two generating functions $T(z)$ and $\tilde{T}(z)$ have the same dominant singularity. Thus we get

$$\tilde{T}(z) = \zeta(z) \cdot \exp(\tilde{T}(z)),$$

with

$$\zeta(z) = \frac{1}{2}\exp\left(\frac{1}{2}(1-z) + \sum_{n\geq 2}\frac{T(z^i)}{i}\right).$$

Proposition 9 *The function $\zeta(z)$ defined in the class of objects associated to $\tilde{T}(z)$ satisfies the assumptions of the Theorems 2 and 3.*

It remains to slightly modify the 2 first coefficients in the singular expansion of $\tilde{T}(z)$ to obtain the singular expansion of $T(z)$ and fill both Tables 1. and 2. In particular we get

$$T_n \underset{n\to\infty}{=} \frac{\rho^{-n}}{\sqrt{\pi n^3}}\left(0.3658015862381119375\ldots - \frac{0.2409833212579280352\ldots}{n} - \frac{0.3678657493849431861\ldots}{n^2} - \frac{0.9991064877914853523\ldots}{n^3} - \frac{4.137777553476907813\ldots}{n^4} + \mathcal{O}\left(\frac{1}{n^5}\right)\right).$$

It seems that these numbers do not appear elsewhere in the literature.

Let us conclude this section on hierarchies by mentioning the OEIS sequence A000084, that is directly related. It counts the number of series-parallel networks with n unlabelled edges; both generating functions are essentially the same (up to a simple factor). We thus get the Puiseux expansions and the asymptotic expansion for these objects as a by-product.

3.4 Approximations

In order to obtain the following approximations for the coefficients in the Puiseux expansions or for the asymptotic expansions of the numbers of trees, we have used the open-source mathematics software *Sage* [Dev15] and the Python library *MPmath* [J$^+$14] for some specific high precision calculations.

The first table synthesises the first elements of the sequences $(t_n)_{n\in\mathbb{N}}$ satisfying the Puiseux expansions for the previous Pólya structures:

$$T(z) = \sum_{n\geq 0} t_n \left(1 - \frac{z}{\rho}\right)^{n/2}.$$

Coeff.	Pólya trees	Rooted identity trees	Hierarchies
t_0	1.000000000000000000	1.000000000000000000	0.640416333492100177 7
t_1	−1.559490020374640884	−1.285158159488538943	−0.7316031724762238750
t_2	0.8106697078826992796	0.5505438316333229659	0.03799806716699161541
t_3	−0.2854870216128456058	−0.5681159369076463432	0.1384103018915147449
t_4	0.1653723657120838943	0.4261261857916583247	−0.07387395031732463851
t_5	−0.3424599704021542007	−0.1312888430707878210	−0.05428300802019698042
t_6	0.3174072259465285628	0.1224152517144394163	0.03800381072191918081
t_7	−0.1077788002916310083	−0.3225499663026797778	0.03109684705422999274
t_8	0.06138495705583510410	0.2539454170234272677	−0.02381831461193008886
t_9	−0.1952123835975564636	0.04875363678533678081	−0.02078556533052714092
t_{10}	0.2059848312779074186	−0.00002800001023286558041	0.01666265537126027377
t_{11}	−0.05272470849819056138	−0.3631594631270670335	0.01611178365047090583
t_{12}	0.01702656875495366861	0.2637344037695510765	−0.01295368177079785790
t_{13}	−0.1523706243663253961	0.2617035123807709629	−0.01338408339711046374
t_{14}	0.1737028832998504627	−0.1368754575043169801	0.01075691931570711729
t_{15}	−0.01447370373952704466	−0.5927534134371262366	0.01183388780152404393
t_{16}	−0.02189951761121556237	0.3911340105112945142	−0.009441457380326882677
t_{17}	−0.1445471935709097045	0.6832510269350502136	−0.01084956346194149131
t_{18}	0.1760771088850177779	−0.3902593892984113718	0.008607637481105329431

Tab. 1: Approximation of the Puiseux expansions for Pólya trees, rooted identity trees and hierarchies

The following Table 2 contains the first numbers $(\tau_n)_{n\in\mathbb{N}}$ satisfying the asymptotic expansions for the previous Pólya structures:

$$T_n \underset{n\to\infty}{\sim} \frac{\rho^{-n}}{\sqrt{\pi n^3}} \sum_{i\geq 0} \frac{\tau_i}{n^i}.$$

Coeff.	Pólya trees	Rooted identity trees	Hierarchies
τ_0	0.7797450101873204419	0.6425790797442694714	0.3658015862381119375
τ_1	0.07828911261061096133	−0.1851197977766337056	0.2409833212579280352
τ_2	0.3929402676631860168	−0.4272427290060978745	0.3678657493849431861
τ_3	1.537879315978838092	−2.255455568987212079	0.9991064877914853523
τ_4	8.200840904355961 94	−16.60970953335647846	4.137777553476907813
τ_5	57.29291473494343825	−157.9003693373302727	23.43410248921570084
τ_6	503.0445050262735854	−1840.110517359351172	170.1188811511555370
τ_7	5359.600933884326064	−25387.34869954017854	1514.745295656330186
τ_8	67342.06920114653067	−404610.0663959841556	16007.82637588106931
τ_9	975425.4970695924728	−7.313377058487246593e6	195812.3506172274875
τ_{10}	1.599693249293173348e7	−1.477949138517813328e8	2.719234685827618831e6
τ_{11}	2.928225313353392698e8	−3.301794456762036735e9	4.222444465223140109e7
τ_{12}	5.914523441293936053e9	−8.080229604228356791e10	7.243861962702191648e8
τ_{13}	1.305991927898973201e11	−2.149826267241085239e12	1.359774926415692519e10
τ_{14}	3.128498399789526502e12	−6.179075814699061934e13	2.770908644498957323e11
τ_{15}	8.078305401468914384e13	−1.908151484770832703e15	6.089496262810801422e12
τ_{16}	2.236301680891647428e15	−6.301063280436556255e16	1.435269254893331074e14
τ_{17}	6.605960869699262787e16	−2.215767775919040241e18	3.610881990157578400e15
τ_{18}	2.073828085209932615e18	−8.267080545525264413e19	9.656755540184967275e16

Tab. 2: Asymptotic expansion of the number of Pólya trees, rooted identity trees and hierarchies

It is interesting to note that, in Table 2, for n sufficiently large and due to the sign of the values of the

(τ_i), all truncations after the nth term in the full expansions (for $n = 1 \dots 17$) correspond to lower bounds for the case of Pólya trees and hierarchies and all of them are upper bounds for rooted identity trees.

Size	10	20	50	100	200	500
Order-1 approximation	$1.391 \cdot 10^{-2}$	$2.859 \cdot 10^{-3}$	$4.204 \cdot 10^{-4}$	$1.027 \cdot 10^{-4}$	$2.540 \cdot 10^{-5}$	$4.039 \cdot 10^{-6}$
Order-4 approximation	$1.039 \cdot 10^{-3}$	$3.448 \cdot 10^{-5}$	$2.383 \cdot 10^{-7}$	$6.872 \cdot 10^{-9}$	$2.071 \cdot 10^{-10}$	$2.078 \cdot 10^{-12}$
Order-8 approximation	$7.722 \cdot 10^{-4}$	$3.369 \cdot 10^{-6}$	$3.822 \cdot 10^{-10}$	$6.195 \cdot 10^{-13}$	$1.123 \cdot 10^{-15}$	$2.611 \cdot 10^{-18}$

Tab. 3: Relative error induced by approximations for hierarchies

Finally, by using only 20 digits of precision in our approximations of the values $\zeta^{(r)}(\rho)$'s we cannot hope to obtain a better approximation than the one of order 8 (Table 3) for the number of large trees (i.e. with size larger than 500).

4 Conclusion

The strength of the approach presented here is its universality. We have shown, in full detail, how it applies to Pólya trees, rooted identity trees and hierarchies but many other examples fill in our framework.

1. *Rooted oriented trees and series-reduced planted trees.* The OEIS sequences A000151 and A001678 can be directly studied.

2. *Series-parallel networks.* In the context of [RS42], [Moo87] and [Fin03b] we get back several generating functions (listed in OEIS A058385, A058386 and A058387) that can be studied in the same vein as hierarchies. Let us recall that many links between trees and series-parallel graphs have already been exhibited, thus the fact that the behaviours of their generating series are analogous is not a surprise.

3. *Phylogenetic trees and also total partitions.* The OEIS sequence A000311, counting phylogenetic trees and also total partitions that are labelled objects, can also be analysed with our technique. Note here that the function $\zeta(z)$ does not explicitly depend on $T(z)$ and thus every derivative is explicit. Just put a factor $n!$ in front of T_n to obtain its full asymptotic expansion. We thus exhibit the polynomials whose existence has been stated in [Com74, p. 224].

4. *The unrooted versions of the previous rooted trees.* With some further work, we are able to exhibit the full asymptotic expansion of the unrooted versions of the previous rooted trees we were interested in. In fact their generating functions $P(z)$ satisfy some equation of the form
$$P(z) = T(z) - \frac{1}{2}T^2(z) + \frac{1}{2}T(z^2).$$
Since we have the full Puiseux expansion of the series $T(z)$, we can compute the one of the series $P(z)$. Some examples of such series correspond to the following sequences A000055, A000238, A000014.... An open question would be to be able to write a functional equation for $P(z)$ as a disturbance of the Cayley tree function, and then to use directly an analogous approach as the one studied in Section 2. There, we would get $\zeta^{(1)} = 0$ since we know that these trees are unrooted.

Acknowledgements. The author is very grateful to Cécile Mailler for the carefully reading of this manuscript and to the anonymous referees for the suggested improvements.

References

[BLW76] N. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory, 1736-1936*. Clarendon Press, 1976.

[CGH+96] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the Lambert W Function. In *Advances in Computational Mathematics*, volume 5, pages 329–359, 1996.

[Com74] L. Comtet. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Reidel, 1974.

[Dev15] The Sage Developers. *Sage Mathematics Software (V. 6.10)*, 2015. `http://www.sagemath.org`.

[Fin03a] S. R. Finch. *Mathematical Constants*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2003.

[Fin03b] S. R. Finch. Series-parallel networks. Technical report, Harvard University, 2003.

[Fin03c] S. R. Finch. Two asymptotic series. Technical report, Harvard University, 2003.

[FS09] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[HRS75] F. Harary, R. W. Robinson, and A. J. Schwenk. Twenty-step algorithm for determining the asymptotic number of trees of various species. *Journal of the Australian Mathematical Society (Series A)*, 20:483–503, 11 1975.

[J+14] F. Johansson et al. *mpmath: a Python library for arbitrary-precision floating-point arithmetic (V. 0.19)*, 2014. `http://mpmath.org/`.

[Knu97] D. E. Knuth. *The Art of Computer Programming, Volume 1 (3rd Ed.): Fundamental Algorithms*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1997.

[Moo87] J. W. Moon. Some enumerative results on series-parallel networks. In *Annals of Discrete Mathematics, Proc. of the Int. Conf. on Finite Geometries and Combinatorial Structures*, volume 144 of *North-Holland Mathematics Studies*, pages 199–226, 1987.

[Ott48] R. Otter. On the number of trees. *The Annals of Mathematics*, 49:583–599, 1948.

[Pól37] G. Pólya. Kombinatorische anzahlbestimmungen für gruppen, graphen, und chemische verbindungen. *Acta Mathematica*, 68:145–254, 1937.

[RS42] J. Riordan and C. E. Shannon. The number of two terminal series-parallel networks. *Journal of Mathematical Physics*, 21:83–93, 1942.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Asymptotic Expansions in SageMath (Poster)

Benjamin Hackl[1], Clemens Heuberger[1], Daniel Krenn[1]

[1] *Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Austria,*
benjamin.hackl@aau.at, clemens.heuberger@aau.at, math@danielkrenn.at or daniel.krenn@aau.at

We present a toolbox for everyday life in analytic combinatorics, namely the new asymptotic expansion module which is included in the mathematics software system SageMath. The code of this module was contributed by the authors of this poster.

Keywords: Asymptotic expansions, SageMath, software

SageMath [1] is a free and open-source mathematics software system. Since version 6.10, it is shipped with a module for computations with asymptotic expansions [3]; no workaround or loading of a package is needed. Even better, it is automatically tested with each release to guarantee functionality and repreducibility of the results.

The asymptotic expansion module[(i)] is integrated completely into SageMath's infrastructure and interacts with all of SageMath's other mathematical objects very well. All contributed code and documentation goes through a transparent peer-review process; this ensures that SageMath's quality standards on efficient, readable, and maintainable code are met.

Due to space restrictions, we limit ourselves to the univariate case here. However, the module is designed for multivariate asymptotic expansions as well.

1 Creating an Asymptotic Ring

We use the coefficient ring

```
C = SR.subring(no_variables=True)  # symbolic constants
```

which is a ring of symbolic constants; note that this includes the rationals $\mathbb{Q}$. A univariate asymptotic ring for our calculations is created by

```
A = AsymptoticRing(growth_group='QQ^n * n^QQ * log(n)^QQ',
                   coefficient_ring=C,
                   default_prec=5)
n = A.gen()
```

A typical element of `A` is the asymptotic expansion

$$42\left(\frac{1}{3}\right)^n n^{\frac{5}{2}}\log(n)^2 + O\left(\left(\frac{1}{3}\right)^n n\log(n)^{\frac{1}{2}}\right)$$

as $n \to \infty$.

(i) See `http://doc.sagemath.org/html/en/reference/asymptotic` for the online-documentation.

2 Basic Arithmetic

Beside the very basic arithmetical operations addition, subtraction and multiplications, series expansions are automatically performed for division, the exponential function or the logarithm. For example, typing `log(n + 1)` returns

$$\log(n+1) = \log(n) + n^{-1} - \frac{1}{2}n^{-2} + \frac{1}{3}n^{-3} - \frac{1}{4}n^{-4} + \frac{1}{5}n^{-5} + O\big(n^{-6}\big).$$

More advanced stuff is possible, e.g., `(1 + 1/n)^n` returns

$$\left(1+\frac{1}{n}\right)^n = e - \frac{1}{2}\,en^{-1} + \frac{11}{24}\,en^{-2} - \frac{7}{16}\,en^{-3} + \frac{2447}{5760}\,en^{-4} + O\big(n^{-5}\big).$$

3 Example: Catalan Numbers

There are several possibilities to obtain asymptotics for the Catalan numbers. We use the convenience generator function for $\binom{kn}{n}$ and type

```
binomial_2n_n = asymptotic_expansions.Binomial_kn_over_n(
                    'n', k=2, precision=3)
C_n = binomial_2n_n / (n+1)
```

This results in

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \frac{1}{\sqrt{\pi}}4^n n^{-\frac{3}{2}} - \frac{9}{8\sqrt{\pi}}4^n n^{-\frac{5}{2}} + \frac{145}{128\sqrt{\pi}}4^n n^{-\frac{7}{2}} + O\Big(4^n n^{-\frac{9}{2}}\Big).$$

This could have been achieved by using `factorial()` to build the binomial coefficient manually, as well.

Getting the asymptotic expansions, for example, the harmonic numbers is even easier, since there is a pre-defined generator in SageMath.

4 Singularity Analysis

The Catalan numbers satisfy the generating function

```
def catalan(z):
    return (1 - sqrt(1-4*z)) / (2*z)
```

So, in contrast to the direct calculation of their asymptotic expansion out of the exact formula, we can do a singularity analysis. SageMath assists here as well. We perform

```
C_n = A.coefficients_of_generating_function(
          catalan, singularities=(1/4,), precision=3)
```

to obtain

$$\frac{1}{\sqrt{\pi}}4^n n^{-\frac{3}{2}} - \frac{9}{8\sqrt{\pi}}4^n n^{-\frac{5}{2}} + \frac{145}{128\sqrt{\pi}}4^n n^{-\frac{7}{2}} + O\big(4^n n^{-4}\big)$$

again. We can proceed similarly with the harmonic numbers.

Bootstrapping for finding the dominant singularity is easily possible as well. For example, let us consider longest runs of words over a two letter alphabet, see [2, Example V.4]. The generating function counting runs where one of the two letters has less than n consecutive repetitions is $(1-z^n)/(1-2z+z^{n+1})$. The dominant singularity satisfies the fix-point equation $z = f(z)$ with

```
def f(z):
    return (1 + z^(n+1)) / 2
```

By starting with the approximation $z = \frac{1}{2} + O((\frac{3}{5})^n)$, applying `f` twice yields the known expansion

$$z = \frac{1}{2} + \frac{1}{4}\left(\frac{1}{2}\right)^n + \frac{1}{8}\left(\frac{1}{4}\right)^n n + \frac{1}{8}\left(\frac{1}{4}\right)^n + O\left(\left(\frac{3}{20}\right)^n n^2\right).$$

References

[1] The SageMath Developers, *SageMath Mathematics Software (Version 7.0)*, 2016, `http://www.sagemath.org`.

[2] Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.

[3] Benjamin Hackl, Clemens Heuberger, and Daniel Krenn, *Asymptotic expansions in SageMath*, `http://trac.sagemath.org/17601`, 2015, module in SageMath 6.10.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

The Register Function and Reductions of Binary Trees and Lattice Paths

Benjamin Hackl[1†], Clemens Heuberger[1†], Helmut Prodinger[2‡]

[1] *Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Austria, {benjamin.hackl, clemens.heuberger}@aau.at*
[2] *Department of Mathematical Sciences, Stellenbosch University, South Africa, hproding@sun.ac.za*

Abstract. The register function (or Horton-Strahler number) of a binary tree is a well-known combinatorial parameter. We study a reduction procedure for binary trees which offers a new interpretation for the register function as the maximal number of reductions that can be applied to a given tree. In particular, the precise asymptotic behavior of the number of certain substructures ("branches") that occur when reducing a tree repeatedly is determined.

In the same manner we introduce a reduction for simple two-dimensional lattice paths from which a complexity measure similar to the register function can be derived. We analyze this quantity, as well as the (cumulative) size of an (iteratively) reduced lattice path asymptotically.

Keywords: Register function; binary tree; lattice path; asymptotics

1 Introduction

Binary trees are either a leaf or a root together with a left and a right subtree which are binary trees. It is well-known that the generating function counting these objects with respect to the number of inner nodes is given by

$$B(z) = \frac{1-\sqrt{1-4z}}{2z} = \sum_{n\geq 0} \frac{1}{n+1}\binom{2n}{n} z^n.$$

Thus, the nth Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n}$ counts the number of binary trees with n inner nodes.

By simple algebraic manipulations, it is easy to verify that $B(z)$ fulfills the identity

$$B(z) = 1 + \frac{z}{1-2z} B\Big(\frac{z^2}{(1-2z)^2}\Big).$$

However, as we will see in Section 2, we can justify this identity from a combinatorial point of view as well, and the most important part of this combinatorial interpretation is a reduction procedure for binary trees.

†B. Hackl and C. Heuberger are supported by the Austrian Science Fund (FWF): P 24644-N26 and by the Karl Popper Kolleg "Modeling-Simulation-Optimization" funded by the Alpen-Adria-Universität Klagenfurt and by the Carinthian Economic Promotion Fund (KWF).

‡H. Prodinger is supported by an incentive grant of the National Research Foundation of South Africa.

The aim of this paper is to analyze the binary tree reduction with a focus on the structures that emerge when repeatedly reducing a given tree. After the aforementioned introduction of the reduction in Section 2, we discover an inherent connection to a very well-known branching complexity measure of binary trees: the register function.

Sections 2.1 and 2.2 deal with the analysis of the number of r-branches and the number of all branches within trees of given size, where an r-branch can be thought of a local structure in a binary tree that survives exactly r reductions.

In Section 3, we switch our attention from binary trees to two-dimensional lattice paths. As we will see, the generating function of these objects fulfills a similar functional equation as the generating function for binary trees—and its combinatorial interpretation strongly depends on a reduction process as well. The remainder of Section 3 is devoted to analyzing the lattice path reduction. In particular, Section 3.1 investigates *fringes* of lattice paths, which play a similar role as branches with respect to binary trees.

On a general note, we used the open-source mathematics software system SageMath [16] in order to perform the computationally intensive parts of the asymptotic analysis for each of the quantities investigated in this paper. Furthermore, the proofs and many details are omitted in this extended abstract; they can be found in the full version.

2 Tree Reductions and the Register Function

As mentioned in the introduction, we want to find a combinatorial proof for the following proposition.

Proposition 2.1. *The generating function counting binary trees by the number of inner nodes,* $B(z) = \frac{1-\sqrt{1-4z}}{2z}$*, fulfills the identity*

$$B(z) = 1 + \frac{z}{1-2z} B\Big(\frac{z^2}{(1-2z)^2}\Big). \tag{1}$$

Proof (Sketch): The central idea of this proof is to consider a reduction of a binary tree t, which we write as $\Phi(t)$:

First, all leaves of t are erased. Then, if a node has only one child, these two nodes are merged; this operation will be repeated as long as there are such nodes. Finally, the nodes without children are declared to be leaves.

Observe that this reduction is only defined for trees t that have at least one inner node. The various steps of this operation (which was introduced in [19]) are depicted in Figure 1. The number attached to the nodes will be explained later.

It can be shown that the generating function

$$\frac{z}{1-2z} B\big(\frac{z^2}{(1-2z)^2}\big)$$

counts all binary trees that can be reduced at least once. Thus, the functional equation (1) can be interpreted combinatorially as follows: a binary tree is either just $\square$, or it can be reduced at least once. $\square$

Remark. Note that (1) can be used to find a very simple proof for a well-known identity for Catalan numbers:

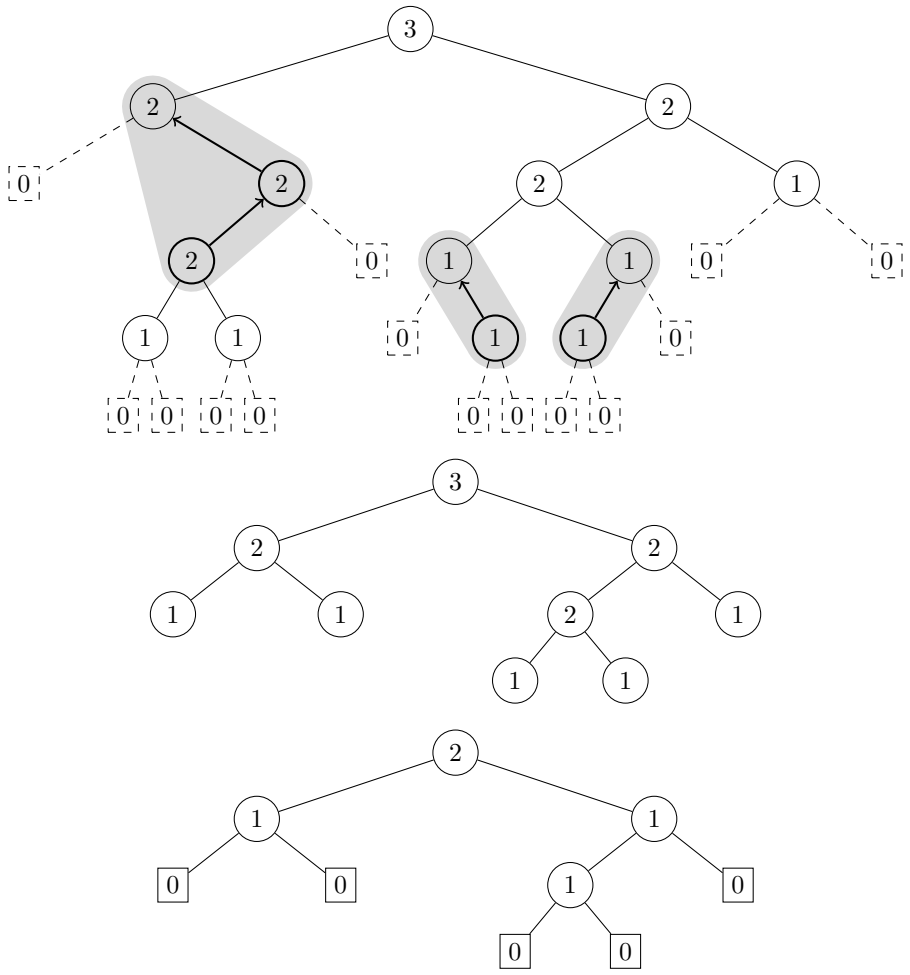


Fig. 1: Illustration of the compactification Φ: in the first tree, the leaves are deleted (dashed nodes) and nodes with exactly one child are merged (gray overlay). The second tree shows the result of these operations. Finally, in the last tree all nodes without children are marked as leaves.

Comparing the coefficients of z^{n+1}, (1) leads to

$$C_{n+1} = [z^{n+1}] \sum_{k\geq 0} C_k \frac{z^{2k+1}}{(1-2z)^{2k+1}} = \sum_{k\geq 0} C_k [z^{n-2k}] \sum_{j\geq 0} 2^j \binom{2k+j}{j} z^j$$
$$= \sum_{0\leq k\leq n/2} C_k 2^{n-2k} \binom{n}{2k},$$

which is known as Touchard's identity [14, 17].

With this interpretation in mind, (1) can also be seen as a recursive process to generate binary trees by repeated substitution of chains. This process can be modeled by the generating functions

$$B_0(z) = 1, \quad B_r(z) = 1 + \frac{z}{1-2z} B_{r-1}\Big(\frac{z^2}{(1-2z)^2}\Big), \quad r \geq 1. \tag{2}$$

By construction, $B_r(z)$ is the generating function of all binary trees that can be constructed from $\square$ with up to r expansions—or, equivalently—all binary trees that can be reduced to $\square$ by applying Φ up to r times.

As it turns out, these generating functions are inherently linked with the *register function* (also known as the Horton-Strahler number) of binary trees. In order to understand this connection, we introduce the register function and prove a simple property regarding the compactification Φ.

The register function is recursively defined: for the binary tree consisting of only a leaf we have $\operatorname{Reg}(\square) = 0$, and if a binary tree t has subtrees t_1 and t_2, then the register function is defined to be

$$\operatorname{Reg}(t) = \begin{cases} \max\{\operatorname{Reg}(t_1), \operatorname{Reg}(t_2)\} & \text{for } \operatorname{Reg}(t_1) \neq \operatorname{Reg}(t_2), \\ \operatorname{Reg}(t_1) + 1 & \text{otherwise.} \end{cases}$$

In particular, the numbers attached to the nodes in Figures 1 and 2 represent the register function of the subtree rooted at the respective node.

Historically, the idea of the register function originated (as the Horton-Strahler numbers) in [8, 15] in the study of the complexity of river networks. However, the very same concept also occurs within a computer science context: arithmetic expressions with binary operators can be expressed as a binary tree with data in the leaves and operators in the internal nodes. Then, the register function of this binary expression tree corresponds to the minimal number of registers needed to evaluate the expression.

There are several publications in which the register function and related concepts are investigated in great detail, for example Flajolet, Raoult, and Vuillemin [5], Kemp [9], Flajolet and Prodinger [4], Louchard and Prodinger [10], Drmota and Prodinger [1], and Viennot [18]. For a detailed survey on the register function and related topics see [13].

We continue by observing that the compactification Φ is a very natural operation regarding the register function:

Proposition 2.2. *Let t be a binary tree with $\operatorname{Reg}(t) = r \geq 1$. Then $\Phi(t)$ is well-defined and the register function of the compactified tree is $\operatorname{Reg}(\Phi(t)) = r - 1$.*

As an immediate consequence of Proposition 2.2 we find that Φ can be applied r times repeatedly to some binary tree t if and only if $\operatorname{Reg}(t) \geq r$ holds. In particular, we obtain

$$\Phi^r(t) = \square \quad \Longleftrightarrow \quad \operatorname{Reg}(t) = r. \tag{3}$$

With (3), the link between the generating functions $B_r(z)$ from above and the register function becomes clear: $B_r(z)$ is exactly the generating function of binary trees with register function $\leq r$.

In order to analyze these recursively defined generating functions an explicit representation is convenient. As it turns out, the substitution $z = \frac{u}{(1+u)^2} =: Z(u)$ is a helpful tool in this context.

In particular, it can be shown that applying $z \mapsto \frac{z^2}{(1-2z)^2}$ corresponds to $u \mapsto u^2$, which helps to find the explicit representation

$$B_r(z) = \frac{1-u^2}{u} \sum_{j=0}^{r} \frac{u^{2^j}}{1-u^{2^{j+1}}}.$$

Note that at this point, we could obtain the generating function for binary trees with register function equal to r simply by computing the difference $B_r(z) - B_{r-1}(z)$ for $r \geq 1$. These functions can be used to study the asymptotic behavior of the average register function value.

However, as these results are well-known (cf. [5]), we will continue in a different direction by studying the number of so-called r-branches.

2.1 r-branches

The register function associates a value to each node (internal nodes as well as leaves), and the value at the root is the value of the register function of the tree. An r-branch is a maximal chain of nodes labeled r. This must be a chain, since the merging of two such chains would already result in the higher value $r+1$. The nodes of the tree are partitioned into such chains, from $r = 0, 1, \ldots$. The goal of this section is the study of the parameter "number of r-branches", in particular, the average number of them, assuming that all binary trees of size n are equally likely.

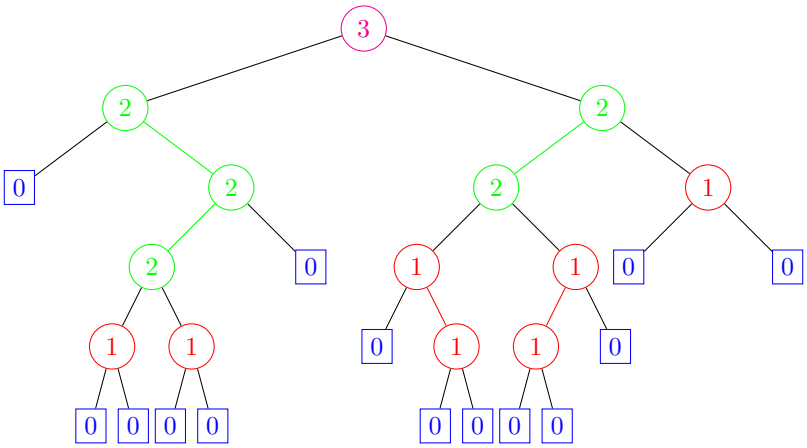


Fig. 2: Binary tree with colored r-branches

This parameter was the main object of the paper [19], and some partial results were given that we are now going to extend. In contrast to this paper, our approach relies heavily on generating functions which, besides allowing us to verify the results in a relatively straightforward way, also enables us to extract explicit formulæ for the expectation (and, in principle, also for higher moments).

A parameter that was not investigated in [19] is the total number of r-branches, for any r, i.e., the sum over $r \geq 0$. Here, asymptotics are trickier, and the basic approach from [19] cannot be applied. However, in this paper we use the Mellin transform, combined with singularity analysis of generating functions, a multi-layer approach that also allowed one of us several years ago to solve a problem by Yekutieli and Mandelbrot, cf. [11]. The origins of singularity analysis can be found in [6], and for a detailed survey see [7].

For reasons of comparisons, let us mention that the value of register function in [19] are one higher than here, and that n generally refers there to the number of leaves, not nodes as here.

According to our previous considerations, after r iterations of Φ, the r-branches become leaves (or, equivalently, 0-branches). The bivariate generating function allowing us to count the leaves of the binary trees is $vB(zv)$.

The proofs of the statements in this section, together with SageMath worksheets containing the corresponding computations, can be found in the full version of this paper.

Theorem 1. *Let $r \in \mathbb{N}_0$ be fixed. The expected number of r-branches in binary trees of size n and the corresponding variance have the following asymptotic expansions:*

$$E_{n;r} = \frac{n}{4^r} + \frac{1}{6}\Big(1 + \frac{5}{4^r}\Big) + \frac{1}{20n}\Big(4^r - \frac{1}{4^r}\Big) + \frac{1}{12n^2}\Big(\frac{5\cdot 16^r}{21} - \frac{7\cdot 4^r}{10} + \frac{97}{210\cdot 4^r}\Big) + O(n^{-3}), \quad (4)$$

$$V_{n;r} = \frac{4^r-1}{3\cdot 16^r}n - \frac{2\cdot 16^r - 25\cdot 4^r + 23}{90\cdot 16^r} - \frac{13\cdot 64^r - 14\cdot 16^r + 7\cdot 4^r - 6}{420\cdot 16^r n} + O(n^{-2}). \quad (5)$$

Of course, the expected number of r-branches can also be computed explicitly by using Cauchy's integral formula. This yields the following result:

Proposition 2.3. *The expected number of r-branches in binary trees of size n is given by the explicit formula*

$$E_{n;r} = \frac{n+1}{\binom{2n}{n}} \sum_{\lambda \geq 1} \lambda \left[\binom{2n}{n+1-\lambda 2^r} - 2\binom{2n}{n-\lambda 2^r} + \binom{2n}{n-1-\lambda 2^r} \right]. \quad (6)$$

2.2 The total number of branches

So far, we were dealing with fixed r, and the number of r-branches in trees of size n, for large n. Now we consider the total number of such branches, i.e., the sum over $r \geq 0$, which was not considered in [19]. First, to get an explicit formula, the results from Proposition 2.3 can be summed.

Corollary 2.4. *The expected number of branches in binary trees of size n, denoted as E_n, is given by the explicit formula*

$$E_n = \frac{n+1}{\binom{2n}{n}} \sum_{k=1}^{n+1} (2 - 2^{-v_2(k)}) k \left[\binom{2n}{n+1-k} - 2\binom{2n}{n-k} + \binom{2n}{n-1-k} \right],$$

where $v_2(k)$ is the dyadic valuation of k, i.e., the highest exponent ν such that 2^ν divides k.

While it is absolutely possible to work out the asymptotic growth from this explicit formula, at it was done in earlier papers [5, 9], we choose a faster method, like in [4]. It works on the level of generating functions and uses the Mellin transform together with singularity analysis of generating functions [7, 12].

The following theorem describes the asymptotic behavior for the expected number of branches in a binary tree.

Theorem 2. *The expected value of the total number of branches in a random binary tree of size n admits the asymptotic expansion*

$$E_n = \frac{4n}{3} + \frac{1}{6}\log_4 n - \frac{2\zeta'(-1)}{\log 2} - \frac{\gamma}{12\log 2} - \frac{1}{6\log 2} + \frac{43}{36} + \delta(\log_4 n) + O\Big(\frac{\log n}{n}\Big),$$

where

$$\delta(x) := \frac{1}{\log 2}\sum_{k\neq 0}\Gamma\Big(\frac{\chi_k}{2}\Big)\zeta(\chi_k - 1)(\chi_k - 1)e^{2\pi i k x}$$

is a 1-periodic function of mean zero, given by its Fourier series expansion.

Remark. Note that the value of the derivative of the zeta function is given by $\zeta'(-1) = -\frac{1}{12} - \log A \approx -0.1654211437$, where A is the Glaisher-Kinkelin constant (cf. [2, Section 2.15]).

Remark. The occurrence of the periodic fluctuation δ where the argument is logarithmic in n is actually not surprising: while this phenomenon is already very common in the context of the register function, fluctuations appear very often in the asymptotic analysis of sums.

While this multi-layer approach enabled us to analyze the expected value of the number of branches in binary trees of size n, the same strategy fails for computing the variance. This is because the random variables modeling the number of r-branches are correlated for different values of r—and thus, the sum of the variances (which we compute by our approach) differs from the variance of the sum.

This concludes our study of the number of branches per binary tree. In the next section, we analyze a quantity that has similar properties as the register function, but is defined on simple two-dimensional lattice paths.

3 A Similar Recursive Scheme Involving Lattice Paths

Recall that the register function describes the number of compactifications of a binary tree required in order to reduce the tree to a leaf. By defining a similar process for simple two-dimensional lattice paths, a function that plays a similar role as the register function is obtained.

Simple two-dimensional lattice paths are sequences of the symbols $\{\uparrow, \rightarrow, \downarrow, \leftarrow\}$. It is easy to see that the generating function counting these paths (without the path of length 0) is

$$L(z) = \frac{4z}{1-4z} = 4z + 16z^2 + 64z^3 + 256z^4 + 1024z^5 + \cdots.$$

Proposition 3.1. *The generating function $L(z) = \frac{4z}{1-4z}$ fulfills the functional equation*

$$L(z) = 4L\Big(\frac{z^2}{(1-2z)^2}\Big) + 4z. \tag{7}$$

Remark. It is easy to verify this result by means of substitution and expansion. However, we want to give a combinatorial proof—this approach also motivates the definition of a recursive generation process for lattice paths, similar to the process for binary trees from above.

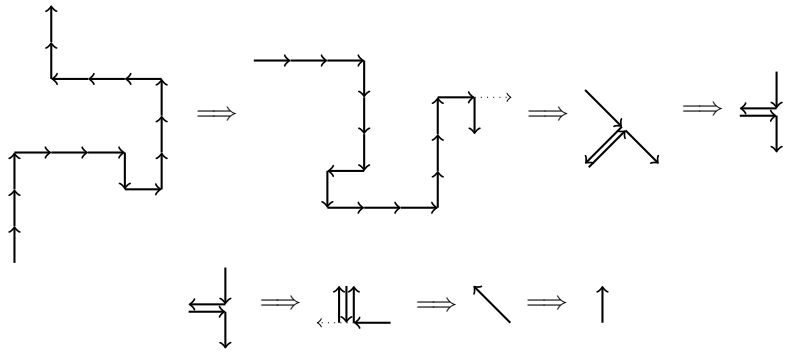


Fig. 3: Repeated application of the reduction Φ_L on a path with compactification degree 2

Proof (Sketch): While we leave the detailed proof to the full version of this paper, we still want to introduce a lattice path reduction which plays an analogous role as the binary tree reduction in the proof of Proposition 2.1.

We consider the reduction Φ_L, which acts on any given lattice path ℓ with length ≥ 2 as follows:

First, the path needs to be modified such that it starts horizontally and ends vertically. This is done by rotation to the right of the entire path and/or the very last step, respectively.

Then, the horizontally starting and vertically ending path is reduced by replacing each pair of successive horizontal-vertical path segments in the following way:

- If a segment starts with $\rightarrow$ and the first vertical step is $\uparrow$, replace it by $\nearrow$,
- if a segment starts with $\rightarrow$ and the first vertical step is $\downarrow$, replace it by $\searrow$,
- if a segment starts with $\leftarrow$ and the first vertical step is $\downarrow$, replace it by $\swarrow$,
- and if a segment starts with $\leftarrow$ and the first vertical step is $\uparrow$, replace it by $\nwarrow$.

Rotating the resulting path by 45° in order to obtain a path with horizontal and vertical steps then yields $\Phi_L(\ell)$. It can be shown that this reduction corresponds to the right-hand side of (7). □

The process described in the proof of Proposition 3.1 allows us to assign a unique number to each lattice path:

Definition. Let ℓ be a simple two-dimensional lattice path consisting of at least one step. We define the *compactification degree* of ℓ, denoted as $\mathrm{cdeg}(\ell)$ as

$$\mathrm{cdeg}(\ell) = n \quad \Longleftrightarrow \quad \Phi_L^n(\ell) \in \{\uparrow, \rightarrow, \downarrow, \leftarrow\}.$$

Remark. The parallels between the compactification degree and the register function are obvious: both count the number of times some given mathematical object can be reduced according to some rules until an atomic form of the respective object is obtained. Therefore, both functions describe, in some sense, the complexity of a given structure.

In the remainder of this section we want to derive some asymptotic results for the compactification degree, namely the expected degree of a lattice path of given length as well as the corresponding variance.

Analogously to our strategy for (1), we want to interpret (7) as a recursive generation process as well and therefore set

$$L_0^{=}(z) = 4z, \quad L_r^{=}(z) = 4L_{r-1}^{=}\Big(\frac{z^2}{(1-2z)^2}\Big), \quad r \geq 1.$$

With the help of the substitution $z = Z(u)$ the generating function can be written explicitly as

$$L_r^{=}(z) = 4^{r+1}\frac{u^{2^r}}{(1+u^{2^r})^2}. \tag{8}$$

The coefficients of this function can be extracted explicitly by applying Cauchy's integral formula.

Proposition 3.2. *The number of two-dimensional simple lattice paths of length n that have compactification degree r is given by*

$$[z^n]L_r^{=}(z) = 4^{r+1}\sum_{\lambda\geq 0}\lambda(-1)^{\lambda-1}\left[\binom{2n-1}{n-\lambda 2^r} - \binom{2n-1}{n-\lambda 2^r-1}\right].$$

In fact, by studying the substitution $z = Z(u)$ closely, the asymptotic behavior of the coefficients of $L_r^{=}(z)$ can be extracted as well.

We turn to the investigation of the expected compactification degree. Let $\mathscr{L}_n$ denote the set of simple two-dimensional lattice paths of size n. Consider the family of random variables $X_n\colon \mathscr{L}_n \to \mathbb{N}_0$ modeling the compactification degree of the lattice paths of length n under the assumption that all paths are equally likely. The following results are immediate consequences of Proposition 3.2.

Corollary 3.3. *The probability that a lattice path of length n has compactification degree r is given by the explicit formula*

$$\mathbb{P}(X_n = r) = \frac{[z^n]L_r^{=}(z)}{4^n} = 4^{r+1-n}\sum_{\lambda\geq 0}\lambda(-1)^{\lambda-1}\left[\binom{2n-1}{n-\lambda 2^r} - \binom{2n-1}{n-\lambda 2^r-1}\right],$$

and the expected compactification degree for paths of length n is given by

$$\mathbb{E}X_n = \sum_{k\geq 1} 8k(2^{v_2(k)}-1)\left[\binom{2n-1}{n-k} - \binom{2n-1}{n-k-1}\right]. \tag{9}$$

Remark. The formula for $\mathbb{P}(X_n = r)$ is very similar to the results for the classical register function obtained by Flajolet (cf. [3]). It is likely that applying the techniques that were used in [10] could be used to determine expansions for arbitrary moments.

The following theorem characterizes the asymptotic behavior of the expected compactification degree and the corresponding variance.

Theorem 3. *The expected compactification degree of simple two-dimensional lattice paths of length n admits the asymptotic expansion*

$$\mathbb{E}X_n = \log_4 n + \frac{\gamma + 2 - 3\log 2}{2\log 2} + \delta_1(\log_4 n) + O(n^{-1}), \tag{10}$$

and for the corresponding variance we have

$$\mathbb{V}X_n = \frac{\pi^2 - 24\log^2\pi - 48\zeta''(0) - 24}{24\log^2 2} - \frac{2\log\pi}{\log 2} - \frac{11}{12} + \delta_2(\log_4 n) - \frac{\gamma + 2 - 3\log 2}{\log 2}\delta_1(\log_4 n) + \delta_1^2(\log_4 n) + O\Big(\frac{1}{\log n}\Big) \quad (11)$$

where $\delta_1(x)$ and $\delta_2(x)$ are 1-periodic fluctuations of mean zero whose Fourier coefficients can be given explicitly.

3.1 Fringes

We define the rth *fringe* of a given lattice path ℓ of length ≥ 1 to be $\Phi_L^r(\ell)$, i.e. the rth fringe is given by the rth reduction of the path. In particular, if ℓ can be reduced r times, we call the length of $\Phi_L^r(\ell)$ the size of the rth fringe. Otherwise, we say that this size is 0.

The rth fringes of positive size can then be enumerated by the bivariate generating function

$$H_r(z,v) = \sum_{\substack{\ell \text{ path} \\ \operatorname{cdeg}(\ell)\geq r}} v^{|\Phi_L^r(\ell)|} z^{|\ell|}$$

where $|\ell|$ denotes the length of a lattice path.

It can be shown that $H_r(z,v)$ fulfills the recursion

$$H_0(z,v) = \frac{4zv}{1-4zv}, \quad H_r(z,v) = 4H_{r-1}\Big(\Big(\frac{z}{1-2z}\Big)^2, v\Big),\ r \geq 1,$$

which can be used to find the explicit representation

$$H_r(z,v) = \frac{4^{r+1}u^{2^r}v}{(1+u^{2^r})^2 - 4u^{2^r}v}.$$

The generating function $H_r(z,v)$ can now be used to derive the asymptotic behavior of the expectation $E_{n;r}^L$ and the variance $V_{n;r}^L$ of the size of the rth fringe, where all paths of length n arise with the same probability.

Theorem 4. *Let $r \in \mathbb{N}_0$ be fixed. The expectation and variance of the rth fringe size of a random path of length n have the asymptotic expansions*

$$E_{n;r}^L = \frac{n}{4^r} + \frac{1-4^{-r}}{3} + O(n^3\theta_r^{-n}) \quad (12)$$

and

$$V_{n;r}^L = \frac{4^r-1}{3\cdot 16^r}n + \frac{-2\cdot 16^r - 5\cdot 4^r + 7}{45\cdot 16^r} + O(n^5\theta_r^{-n}), \quad (13)$$

where $\theta_r = \frac{4}{2+2\cos(2\pi/2^r)} > 1$. If additionally $r > 0$, then for the random variables $Y_{n;r}$ modeling the rth fringe size of lattice paths of length n we have

$$\mathbb{P}\Big(\frac{Y_{n;r} - E_{n;r}}{\sqrt{V_{n;r}}} \leq x\Big) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-w^2/2}\,dw + O(n^{-1/2}),$$

i.e. the random variables $Y_{n;r}$ are asymptotically normally distributed.

As we have the generating function $H_r(z,v)$ in an explicit form, the expected value can also be extracted explicitly by means of Cauchy's integral formula.

Proposition 3.4. *For given $r \in \mathbb{N}_0$, the rth expected fringe size of a random path of length n is given by the explicit formula*

$$E_{n;r}^L = 4^{r+1-n} \sum_{\lambda \geq 1} \frac{2\lambda^3+\lambda}{3} \left[\binom{2n-1}{n-2^r\lambda} - \binom{2n-1}{n-2^r\lambda-1} \right].$$

Analogously to our investigations concerning branches in binary trees, we also study the asymptotic behavior of the expected fringe size, i.e. the sum over the size of the rth fringes for $r \geq 0$. Like the compactification degree, this parameter can also be interpreted as a complexity measure for lattice paths.

Corollary 3.5. *The expected fringe size E_n^L of a random path of length n can be computed as*

$$E_n^L = \frac{1}{12 \cdot 4^n} \sum_{k=1}^{n} \left(2k^3(2-2^{-v_2(k)}) + k(2^{v_2(k)+1}-1)\right) \left[\binom{2n-1}{n-k} - \binom{2n-1}{n-k-1} \right].$$

The following theorem quantifies the asymptotic behavior of $E_n^L := \sum_{r \geq 0} E_{n;r}^L$.

Theorem 5. *Asymptotically, the behavior of the expected fringe size E_n^L for a random path of length n is given by*

$$E_n^L = \frac{4}{3}n + \frac{1}{3}\log_4 n + \frac{5+3\gamma-11\log 2}{18\log 2} + \delta(\log_4 n) + O\Big(\frac{\log n}{n}\Big), \tag{14}$$

where $\delta(x)$ is a 1-periodic fluctuation of mean zero with Fourier series expansion

$$\delta(x) = \sum_{k \neq 0} \frac{2}{3\sqrt{\pi}\log 2} \Gamma\Big(\frac{3+\chi_k}{2}\Big) \big(2\zeta(\chi_k-1)+\zeta(\chi_k+1)\big) e^{2k\pi i x}.$$

References

[1] M. Drmota and H. Prodinger, *The register function for t-ary trees*, ACM Trans. Algorithms **2** (2006), no. 3, 318–334.

[2] S. R. Finch, *Mathematical constants*, Encyclopedia of Mathematics and its Applications, vol. 94, Cambridge University Press, Cambridge, 2003. 2003519 (2004i:00001)

[3] P. Flajolet, *Analyse d'algorithmes de manipulation d'arbres et de fichiers*, Cahiers du Bureau Universitaire de Recherche Oprationnelle **34/35** (1981), 1–209.

[4] P. Flajolet and H. Prodinger, *Register allocation for unary-binary trees*, SIAM J. Comput. **15** (1986), 629–640.

[5] P. Flajolet, J.-C. Raoult, and J. Vuillemin, *The number of registers required for evaluating arithmetic expressions*, Theoret. Comput. Sci. **9** (1979), no. 1, 99–125.

[6] P. Flajolet and A. Odlyzko, *Singularity analysis of generating functions*, SIAM J. Discrete Math. **3** (1990), 216–240.

[7] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.

[8] R. E. Horton, *Erosioned development of systems and their drainage basins*, Geol. Soc. Am. Bull. **56** (1945), 275–370.

[9] R. Kemp, *The average number of registers needed to evaluate a binary tree optimally*, Acta Inform. **11** (1978/79), no. 4, 363–372.

[10] G. Louchard and H. Prodinger, *The register function for lattice paths*, Fifth Colloquium on Mathematics and Computer Science, Discrete Math. Theor. Comput. Sci. Proc., AI, Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2008, pp. 135–148. 2508783

[11] H. Prodinger, *On a problem of Yekutieli and Mandelbrot about the bifurcation ratio of binary trees*, Theoret. Comput. Sci. **181** (1997), no. 1, 181–194.

[12] H. Prodinger, *Analytic methods*, Handbook of Enumerative Combinatorics (M. Bóna, ed.), CRC Press Series on Discrete Mathematics and its Applications, Chapman & Hall/CRC, Boca Raton, FL, 2015, pp. 173–252.

[13] ———, *Introduction to Philippe Flajolet's work on the register function and related topics*, Philippe Flajolet's Collected Papers (M. D. Ward, ed.), vol. V, to appear.

[14] L. W. Shapiro, *A short proof of an identity of Touchard's concerning Catalan numbers*, J. Combinatorial Theory Ser. A **20** (1976), no. 3, 375–376. 0406819 (53 #10605)

[15] A. N. Strahler, *Hypsomic analysis of erosional topography*, Geol. Soc. Am. Bull. **63** (1952), 1117–1142.

[16] The Sage Developers, *Sage Mathematics Software (Version 7.0)*, 2016, `http://www.sagemath.org`.

[17] J. Touchard, *Sur certaines équations fonctionnelles*, Proc. Internat. Math. Congress, vol. I (1928), 1924, pp. 465–472.

[18] X. G. Viennot, *A Strahler bijection between Dyck paths and planar trees*, Discrete Math. **246** (2002), no. 1-3, 317–329, Formal power series and algebraic combinatorics (Barcelona, 1999). 1887493

[19] K. Yamamoto and Y. Yamazaki, *Topological self-similarity on the random binary-tree model*, J. Stat. Phys. **139** (2010), no. 1, 62–71.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

On the Higher Dimensional Quasi-Power Theorem and a Berry–Esseen Inequality

Clemens Heuberger[1†] and Sara Kropf[1,2†]

[1] *Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Austria, {clemens.heuberger, sara.kropf}@aau.at*
[2] *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, sarakropf@stat.sinica.edu.tw*

Hwang's quasi-power theorem asserts that a sequence of random variables whose moment generating functions are approximately given by powers of some analytic function is asymptotically normally distributed. This theorem is generalised to higher dimensional random variables. To obtain this result, a higher dimensional analogue of the Berry–Esseen inequality is proved, generalising a two-dimensional version by Sadikova.

Keywords: Quasi-power theorem, Berry–Esseen inequality, limiting distribution, central limit theorem

1 Introduction

Asymptotic normality is a frequently occurring phenomenon in combinatorics, the classical central limit theorem being the very first example. The first step in the proof is the observation that the moment generating function of the sum of n identically independently distributed random variables is the n-th power of the moment generating function of the distribution underlying the summands. As similar moment generating functions occur in many examples in combinatorics, a general theorem to prove asymptotic normality is desirable. Such a theorem was proved by Hwang [16], usually called the "quasi-power theorem".

Theorem (Hwang [16])**.** *Let $\{\Omega_n\}_{n\geq 1}$ be a sequence of integral random variables. Suppose that the moment generating function satisfies the asymptotic expression*

$$M_n(s) := \mathbb{E}(e^{\Omega_n s}) = e^{W_n(s)}(1 + O(\kappa_n^{-1})), \tag{1.1}$$

the O-term being uniform for $|s| \leq \tau$, $s \in \mathbb{C}$, $\tau > 0$, where

1. *$W_n(s) = u(s)\phi_n + v(s)$, with $u(s)$ and $v(s)$ analytic for $|s| \leq \tau$ and independent of n; and $u''(0) \neq 0$;*
2. *$\lim_{n\to\infty} \phi_n = \infty$;*
3. *$\lim_{n\to\infty} \kappa_n = \infty$.*

†The authors are supported by the Austrian Science Fund (FWF): P 24644-N26.

Then the distribution of Ω_n is asymptotically normal, i.e.,

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\Omega_n-u'(0)\phi_n}{\sqrt{u''(0)\phi_n}}<x\right)-\Phi(x)\right|=O\left(\frac{1}{\sqrt{\phi_n}}+\frac{1}{\kappa_n}\right),$$

where Φ denotes the standard normal distribution

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{1}{2}y^2\right)\,dy.$$

See Hwang's article [16] as well as Flajolet-Sedgewick [8, Sec. IX.5] for many applications of this theorem. A generalisation of the quasi-power theorem to dimension 2 has been provided in [12]. It has been used in [14], [15], [6], [13] and [17]. In [5, Thm. 2.22], an m-dimensional version of the quasi-power theorem is stated without speed of convergence. Also in [2], such an m-dimensional theorem without speed of convergence is proved. There, several multidimensional applications are given, too.

In contrast to many results about the speed of convergence in classical probability theory (see, e.g., [11]), the sequence of random variables is not assumed to be independent. The only assumption is that the moment generating function behaves asymptotically like a large power. This mirrors the fact that the moment generating function of the sum of independent, identically distributed random variables is exactly a large power. The advantage is that the asymptotic expression (1.1) arises naturally in combinatorics by using techniques such as singularity analysis or saddle point approximation (see [8]).

The purpose of this article is to generalise the quasi-power theorem including the speed of convergence to arbitrary dimension m. We first state this main result in Theorem 1 in this section. In Section 2, a new Berry–Esseen inequality (Theorem 2) is presented, which we use to prove the m-dimensional quasi-power theorem. We give sketches of the proofs of these two theorems in Section 4. All details of these proofs can be found in the full version of this extended abstract. In Section 3, we give some applications of the multidimensional quasi-power theorem.

We use the following conventions: vectors are denoted by boldface letters such as $\mathbf{s}$, their components are then denoted by regular letters with indices such as s_j. For a vector $\mathbf{s}$, $\|\mathbf{s}\|$ denotes the maximum norm $\max\{|s_j|\}$. All implicit constants of O-terms may depend on the dimension m as well as on τ which is introduced in Theorem 1.

Our first main result is the following m-dimensional version of Hwang's theorem.

Theorem 1. *Let $\{\mathbf{\Omega}_n\}_{n\geq 1}$ be a sequence of m-dimensional real random vectors. Suppose that the moment generating function satisfies the asymptotic expression*

$$M_n(\mathbf{s}):=\mathbb{E}(e^{\langle\mathbf{\Omega}_n,\mathbf{s}\rangle})=e^{W_n(\mathbf{s})}(1+O(\kappa_n^{-1})),\tag{1.2}$$

the O-term being uniform for $\|\mathbf{s}\|\leq\tau$, $\mathbf{s}\in\mathbb{C}^m$, $\tau>0$, where

1. *$W_n(\mathbf{s})=u(\mathbf{s})\phi_n+v(\mathbf{s})$, with $u(\mathbf{s})$ and $v(\mathbf{s})$ analytic for $\|\mathbf{s}\|\leq\tau$ and independent of n; and the Hessian $H_u(\mathbf{0})$ of u at the origin is non-singular;*
2. *$\lim_{n\to\infty}\phi_n=\infty$;*
3. *$\lim_{n\to\infty}\kappa_n=\infty$.*

Then, the distribution of $\mathbf{\Omega}_n$ is asymptotically normal with speed of convergence $O(\phi_n^{-1/2})$, i.e.,

$$\sup_{\mathbf{x}\in\mathbb{R}^m}\left|\mathbb{P}\left(\frac{\mathbf{\Omega}_n-\operatorname{grad}u(\mathbf{0})\phi_n}{\sqrt{\phi_n}}\le\mathbf{x}\right)-\Phi_{H_u(\mathbf{0})}(\mathbf{x})\right|=O\left(\frac{1}{\sqrt{\phi_n}}\right),\tag{1.3}$$

where Φ_Σ denotes the distribution function of the non-degenerate m-dimensional normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ, i.e.,

$$\Phi_\Sigma(\mathbf{x})=\frac{1}{(2\pi)^{m/2}\sqrt{\det\Sigma}}\int_{\mathbf{y}\le\mathbf{x}}\exp\left(-\frac{1}{2}\mathbf{y}^\top\Sigma^{-1}\mathbf{y}\right)d\mathbf{y},$$

where $\mathbf{y}\le\mathbf{x}$ means $y_\ell\le x_\ell$ for $1\le\ell\le m$.

If $H_u(\mathbf{0})$ is singular, the random variables

$$\frac{\mathbf{\Omega}_n-\operatorname{grad}u(\mathbf{0})\phi_n}{\sqrt{\phi_n}}$$

converge in distribution to a degenerate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $H_u(\mathbf{0})$.

Note that in the case of the singular $H_u(\mathbf{0})$, a uniform speed of convergence cannot be guaranteed. To see this, consider the (constant) sequence of random variables Ω_n which takes values ±1 each with probability $1/2$. Then the moment generating function is $(e^t+e^{-t})/2$, which is of the form (1.2) with $\phi_n=n$, $u(s)=0$, $v(s)=\log(e^t+e^{-t})/2$ and κ_n arbitrary. However, the distribution function of $\Omega_n/\sqrt{n}$ is given by

$$\mathbb{P}\left(\frac{\Omega_n}{\sqrt{n}}\le x\right)=\begin{cases}0 & \text{if } x<-1/\sqrt{n},\\ 1/2 & \text{if } -1/\sqrt{n}\le x<1/\sqrt{n},\\ 1 & \text{if } 1/\sqrt{n}\le x,\end{cases}$$

which does not converge uniformly.

In contrast to the original quasi-power theorem, the error term in our result does not contain the summand $O(1/\kappa_n)$. In fact, this summand could also be omitted in the original proof of the quasi-power theorem by using a better estimate for the error $E_n(\mathbf{s})=M_n(\mathbf{s})e^{-W_n(\mathbf{s})}-1$.

The proof of Theorem 1 relies on an m-dimensional Berry–Esseen inequality (Theorem 2). It is a generalisation of Sadikova's result [22, 23] in dimension 2. The main challenge is to provide a version which leads to bounded integrands around the origin, but still allows to use excellent bounds for the tails of the characteristic functions. To achieve this, linear combinations involving all partitions of the set $\{1,\dots,m\}$ are used.

Note that there are several generalisations of the one-dimensional Berry–Esseen inequality [3, 7] to arbitrary dimension, see, e.g., Gamkrelidze [9, 10] and Prakasa Rao [20]. However, using these results would lead to the less precise error term in (1.3), see the end of Section 2 for more details. For that reason we generalise Sadikova's result, which was already successfully used by the first author in [12] to prove a 2-dimensional quasi-power theorem. Also note that our theorem can deal with discrete random variables, in contrast to [21], where density functions are considered.

For the sake of completeness, we also state the following result about the moments of $\mathbf{\Omega}_n$.

Proposition 1.1. *The cross-moments of $\mathbf{\Omega}_n$ satisfy*

$$\frac{1}{\prod_{\ell=1}^m k_\ell!}\mathbb{E}\Big(\prod_{\ell=1}^m \Omega_{n,\ell}^{k_\ell}\Big) = p_{\mathbf{k}}(\phi_n) + O\Big(\kappa_n^{-1}\phi_n^{k_1+\cdots+k_m}\Big),$$

for k_ℓ nonnegative integers, where $p_{\mathbf{k}}$ is a polynomial of degree $\sum_{\ell=1}^m k_\ell$ defined by

$$p_{\mathbf{k}}(X) = [s_1^{k_1}\cdots s_m^{k_m}]e^{u(\mathbf{s})X+v(\mathbf{s})}.$$

In particular, the mean and the variance-covariance matrix are

$$\begin{aligned}\mathbb{E}(\mathbf{\Omega}_n) &= \operatorname{grad} u(\mathbf{0})\phi_n + \operatorname{grad} v(\mathbf{0}) + O(\kappa_n^{-1}),\\ \operatorname{Cov}(\mathbf{\Omega}_n) &= H_u(\mathbf{0})\phi_n + H_v(\mathbf{0}) + O(\kappa_n^{-1}),\end{aligned}$$

respectively.

2 A Berry–Esseen Inequality

This section is devoted to a generalisation of Sadikova's Berry–Esseen inequality [22, 23] in dimension 2 to dimension m. Before stating the theorem, we introduce our notation.

Let $L = \{1,\dots,m\}$. For $K \subseteq L$, we write $\mathbf{s}_K = (s_k)_{k\in K}$ for the projection of $\mathbf{s} \in \mathbb{C}^L$ to $\mathbb{C}^K$. For $J \subseteq K \subseteq L$, let $\chi_{J,K}\colon \mathbb{C}^J \to \mathbb{C}^K$, $(s_j)_{j\in J} \mapsto (s_k[k\in J])_{k\in K}$ be an injection from $\mathbb{C}^J$ into $\mathbb{C}^K$. Similarly, let $\psi_{J,K}\colon \mathbb{C}^K \to \mathbb{C}^K$, $(s_k)_{k\in K} \mapsto (s_k[k\in J])_{k\in K}$ be the projection which sets all coordinates corresponding to $K\setminus J$ to 0.

We denote the set of all partitions of K by Π_K. We consider a partition as a set $\alpha = \{J_1,\dots,J_k\}$. Thus $|\alpha|$ denotes the number of parts of the partition α. Furthermore, $J \in \alpha$ means that J is a part of the partition α.

Now, we can define an operator which we later use to state our Berry–Esseen inequality. The motivation behind this definition is explained at the end of this section.

Definition 2.1. Let $K \subseteq L$ and $h\colon \mathbb{C}^K \to \mathbb{C}$. We define the non-linear operator

$$\Lambda_K(h) := \sum_{\alpha\in\Pi_K} \mu_\alpha \prod_{J\in\alpha} h\circ\psi_{J,K}$$

where

$$\mu_\alpha = (-1)^{|\alpha|-1}(|\alpha|-1)!\,.$$

We denote Λ_L briefly by Λ.

For any random variable $\mathbf{Z}$, we denote its cumulative distribution function by $F_{\mathbf{Z}}$ and its characteristic function by $\varphi_{\mathbf{Z}}$.

With these definitions, we are able to state our second main result, an m-dimensional version of the Berry–Esseen inequality.

Theorem 2. *Let $m \geq 1$ and $\mathbf{X}$ and $\mathbf{Y}$ be m-dimensional random variables. Assume that $F_{\mathbf{Y}}$ is differentiable.*

Let

$$\begin{aligned} A_j &= \sup_{\mathbf{y}\in\mathbb{R}^m} \frac{\partial F_{\mathbf{Y}}(\mathbf{y})}{\partial y_j}, \\ B_j &= \sum_{k=1}^{j} \left\{ {j \atop k} \right\} k!\,, \\ C_1 &= \sqrt[3]{\frac{32}{\pi\big(1-\big(\frac{3}{4}\big)^{1/m}\big)}}, \\ C_2 &= \frac{12}{\pi} \end{aligned}$$

for $1 \le j \le m$ *where* $\left\{ {j \atop k} \right\}$ *denotes a Stirling partition number (Stirling number of the second kind).*

Let $T > 0$ *be fixed. Then*

$$\begin{aligned} \sup_{\mathbf{z}\in\mathbb{R}^m} |F_{\mathbf{X}}(\mathbf{z}) - F_{\mathbf{Y}}(\mathbf{z})| \le{}& \frac{2}{(2\pi)^m} \int_{\|\mathbf{t}\|\le T} \Big| \frac{\Lambda(\varphi_{\mathbf{X}})(\mathbf{t}) - \Lambda(\varphi_{\mathbf{Y}})(\mathbf{t})}{\prod_{\ell\in L} t_\ell} \Big|\, d\mathbf{t} \\ &+ 2 \sum_{\emptyset\ne J\subsetneq L} B_{m-|J|} \sup_{\mathbf{z}_J\in\mathbb{R}^J} \big| F_{\mathbf{X}_J}(\mathbf{z}_J) - F_{\mathbf{Y}_J}(\mathbf{z}_J) \big| \\ &+ \frac{2\sum_{j=1}^m A_j}{T} (C_1 + C_2). \end{aligned} \tag{2.1}$$

Existence of $\mathbb{E}(\mathbf{X})$ *and* $\mathbb{E}(\mathbf{Y})$ *is sufficient for the finiteness of the integral in* (2.1)*.*

Let us give two remarks on the distribution functions occurring in this theorem: The distribution function $F_{\mathbf{Y}}$ is non-decreasing in every variable, thus $A_j > 0$ for all j. Furthermore, our general notations imply that $F_{\mathbf{X}_J}$ is a marginal distribution of $\mathbf{X}$.

The numbers B_j are known as "Fubini numbers" or "ordered Bell numbers". They form the sequence A000670 in [18].

Recursive application of (2.1) leads to the following corollary, where we no longer explicitly state the constants depending on the dimension.

Corollary 2.2. *Let* $m \ge 1$ *and* $\mathbf{X}$ *and* $\mathbf{Y}$ *be* m*-dimensional random variables. Assume that* $F_{\mathbf{Y}}$ *is differentiable and let*

$$A_j = \sup_{\mathbf{y}\in\mathbb{R}^m} \frac{\partial F_{\mathbf{Y}}(\mathbf{y})}{\partial y_j}, \qquad 1 \le j \le m.$$

Then

$$\begin{aligned} &\sup_{\mathbf{z}\in\mathbb{R}^m} |F_{\mathbf{X}}(\mathbf{z}) - F_{\mathbf{Y}}(\mathbf{z})| \\ &\quad = O\Big(\sum_{\emptyset\ne K\subseteq L} \int_{\|\mathbf{t}_K\|\le T} \Big| \frac{\Lambda_K(\varphi_{\mathbf{X}}\circ\chi_{K,L})(\mathbf{t}_K) - \Lambda_K(\varphi_{\mathbf{Y}}\circ\chi_{K,L})(\mathbf{t}_K)}{\prod_{k\in K} t_k} \Big|\, d\mathbf{t}_K + \frac{\sum_{j=1}^m A_j}{T} \Big) \end{aligned} \tag{2.2}$$

where the O*-constants only depend on the dimension* m*.*

Existence of $\mathbb{E}(\mathbf{X})$ *and* $\mathbb{E}(\mathbf{Y})$ *is sufficient for the finiteness of the integrals in* (2.2)*.*

In order to explain the choice of the operator Λ, we first state it in dimension 2:

$$\Lambda(h)(s_1, s_2) = h(s_1, s_2) - h(s_1, 0)h(0, s_2). \tag{2.3}$$

This coincides with Sadikova's definition. This also shows that our operator is non-linear as, e.g., $\Lambda(s_1 + s_2)(s_1, s_2) \neq \Lambda(s_1)(s_1, s_2) + \Lambda(s_2)(s_1, s_2)$.

In Theorem 2, we apply Λ to characteristic functions; so we may restrict our attention to functions h with $h(\mathbf{0}) = 1$. From (2.3), we see that $\Lambda(h)(s_1, 0) = \Lambda(h)(0, s_2) = 0$, so that $\Lambda(h)(s_1, s_2)/(s_1 s_2)$ is bounded around the origin. This is essential for the boundedness of the integral in Theorem 2. In general, this property will be guaranteed by our particular choice of coefficients. It is no coincidence that for $\alpha \in \Pi_L$, the coefficient μ_α equals the value $\mu(\alpha, \{L\})$ of the Möbius function in the lattice of partitions: Weisner's theorem (see Stanley [24, Corollary 3.9.3]) is crucial in the proof that $\Lambda(h)(\mathbf{s})/(s_1 \cdots s_m)$ is bounded around the origin.

The second property is that our proof of the quasi-power theorem needs estimates for the tails of the integral in Theorem 2. These estimates have to be exponentially small in every variable, which means that every variable has to occur in every summand. This is trivially fulfilled as every summand in the definition of Λ is formulated in terms of a partition.

Note that Gamkrelidze [10] (and also Prakasa Rao [20]) use a linear operator L mapping h to

$$(s_1, s_2) \mapsto h(s_1, s_2) - h(s_1, 0) - h(0, s_2). \tag{2.4}$$

When taking the difference of two characteristic functions, we may assume that $h(0, 0) = 0$ so that the first crucial property as defined above still holds. However, the tails are no longer exponentially small in every variable: The last summand $h(0, s_2)$ in (2.4) is not exponentially small in s_1 because it is independent of s_1 and nonzero in general. However, the first two summands are exponentially small in s_1 by our assumption (1.2).

For that reason, using the Berry–Esseen inequality by Gamkrelidze [10] to prove a quasi-power theorem leads to a less precise error term $O(\phi_n^{-1/2} \log^{m-1} \phi_n)$ in (1.3). It can be shown that the less precise error term necessarily appears when using Gamkrelidze's result by considering the example of $\mathbf{\Omega}_n$ being the 2-dimensional vector consisting of a normal distribution with mean -1 and variance n and a normal distribution with mean 0 and variance n. This is a consequence of the linearity of the operator L in Gamkrelidze's result.

3 Examples of Multidimensional Central Limit Theorems

In this section, we give two examples from combinatorics where we can apply Theorem 1. Asymptotic normality was already shown in earlier publications [4, 2], but we additionally provide an estimate for the speed of convergence.

3.1 Context-Free Languages

Consider the following example of a context-free grammar G with non-terminal symbols S and T, terminal symbols $\{a, b, c\}$, starting symbol S and the rules

$$P = \{S \to aSbS,\ S \to bT,\ T \to bS,\ T \to cT,\ T \to a\}.$$

The corresponding context-free language $L(G)$ consists of all words which can be generated starting with S using the rules in P to replace all non-terminal symbols. For example, $abcababbba \in L(G)$ because it can be derived as

$$S \to aSbS \to abTbaSbS \to abcTbabTbbT \to abcababbba.$$

Let $\mathbb{P}(\mathbf{\Omega}_n = \mathbf{x})$ be the probability that a word of length n in $L(G)$ consists of x_1 and x_2 terminal symbols a and b, respectively. Thus there are $n - x_1 - x_2$ terminal symbols c. For simplicity, this random variable is only 2-dimensional. But it can be easily extended to higher dimensions.

Following Drmota [4, Sec. 3.2], we obtain that the moment generating function is

$$\mathbb{E}(e^{\langle \mathbf{\Omega}_n, \mathbf{s}\rangle}) = \frac{y_n(e^{\mathbf{s}})}{y_n(\mathbf{1})}$$

with $y_n(\boldsymbol{z})$ defined in [4]. Using [4, Equ. (4.9)], this moment generating function has an asymptotic expansion as in (1.2) with $\phi_n = n$. Thus $\mathbf{\Omega}_n$ is asymptotically normally distributed after standardisation (as was shown in [4]) and additionally the speed of convergence is $O(n^{-1/2})$.

Other context-free languages can be analysed in the same way, either by directly using the results in [4] (if the underlying system is strongly connected) or by similar methods. This has applications, for example, in genetics (see [19]).

3.2 Dissections of Labelled Convex Polygons

Let $S_1 \uplus \cdots \uplus S_{t+1} = \{3, 4, \ldots\}$ be a partition. We dissect a labelled convex n-gon into smaller convex polygons by choosing some non-intersecting diagonals. Each small polygon should be a k-gon with $k \notin S_{t+1}$. Define $a_n(\mathbf{r})$ to be the number of dissections of an n-gon such that it consists of exactly r_i small polygons whose number of vertices is in S_i, for $i = 1, \ldots, t$. For convenience, we use $a_2(\mathbf{r}) = [\mathbf{r} = \mathbf{0}]$. Asymptotic normality was proved in [2, Sec. 3], see also [1, Ex. 7.1] for a one-dimensional version. We additionally provide an estimate for the speed of convergence.

Let

$$f(z, \mathbf{x}) = \sum_{\substack{n \ge 2 \\ \mathbf{r} \ge 0}} a_n(\mathbf{r}) \mathbf{x}^{\mathbf{r}} z^{n-1}.$$

Then choosing a k-gon with $k \in S_1 \uplus \cdots \uplus S_t$ and gluing dissected polygons to $k-1$ of its sides translates into the equation

$$f = z + \sum_{i=1}^{t} x_i \sum_{k \in S_i} f^{k-1}.$$

Following [1], this equation can be used to obtain an asymptotic expression for the moment generating function as in (1.2) with $\phi_n = n$. The asymptotic normal distribution follows after suitable standardisation with speed of convergence $O(n^{-1/2})$.

4 Sketch of the Proofs

We now sketch the main ideas of the proofs of Theorems 2 and 1. All details can be found in the full version of this extended abstract.

Sketch of the proof of Theorem 2: As in [23, 10, 20], our proof of the Berry–Esseen inequality proceeds via adding a continuous random variable $\mathbf{Q}$ to our random variables $\mathbf{X}$ and $\mathbf{Y}$. The characteristic function of $\mathbf{Q}$ vanishes outside $[-T, T]^m$. The error resulting from replacing the difference of the distribution functions $|F_{\mathbf{X}} - F_{\mathbf{Y}}|$ by $|F_{\mathbf{X}+\mathbf{Q}} - F_{\mathbf{Y}+\mathbf{Q}}|$ can be estimated by the final summand in (2.1). In principle, Lévy's theorem then allows to bound the difference of the distribution functions by the difference of the characteristic functions. Instead of only using the difference of the characteristic functions, we use the difference $|\Lambda(\varphi_{\mathbf{X}}) - \Lambda(\varphi_{\mathbf{Y}})|$, which ensures boundedness of the integral in (2.1) at least if the first moments exist. However, we have to compensate Λ by the sum over the differences of the marginal distribution functions, which yields the second summand in (2.1). □

Sketch of the proof of Theorem 1: First, the characteristic function of the standardised random variable $\mathbf{X} = (\mathbf{\Omega}_n - \operatorname{grad} u(\mathbf{0})\phi_n)/\sqrt{\phi_n}$ is

$$\varphi_{\mathbf{X}}(\mathbf{s}) = \exp\Big(-\frac{1}{2}\mathbf{s}^\top \Sigma \mathbf{s} + O\Big(\frac{\|\mathbf{s}\|^3 + \|\mathbf{s}\|}{\sqrt{\phi_n}}\Big)\Big)$$

for $\|\mathbf{s}\| < \tau\sqrt{\phi_n}/2$. Thus, we obtain convergence in distribution as stated in the theorem.

To obtain a bound for the speed of convergence, we use the Berry–Esseen inequality given in Theorem 2 for $\mathbf{Y}$ an m-dimensional normal distribution. We bound the difference of Λ evaluated at the characteristic function of $\mathbf{X}$ and the one of the normal distribution by the exponentially decreasing function

$$|\Lambda(\varphi_{\mathbf{X}})(\mathbf{s}) - \Lambda(\varphi_{\mathbf{Y}})(\mathbf{s})| \le \exp\Big(-\frac{\sigma}{4}\|\mathbf{s}\|^2 + O(\|\mathbf{s}\|)\Big) O\Big(\frac{\|\mathbf{s}\|^3 + \|\mathbf{s}\|}{\sqrt{\phi_n}}\Big)$$

for suitable $\mathbf{s}$ where σ is the smallest eigenvalue of Σ.

We then estimate the integral in (2.1). For the variables in a neighbourhood of zero, we get rid of the denominator by Taylor expansion using the zero of $\Lambda(\varphi_{\mathbf{X}}) - \Lambda(\varphi_{\mathbf{Y}})$ at $\mathbf{0}$. The error term of the Taylor expansion can be estimated by the difference of the characteristic functions using Cauchy's formula. The exponentially small tails are used to bound the contribution of the large variables in the integral in (2.1).

The second summand in (2.1) can be estimated inductively. □

References

[1] Edward A. Bender, *Asymptotic methods in enumeration*, SIAM Rev. **16** (1974), no. 4, 485–515.

[2] Edward A. Bender and L. Bruce Richmond, *Central and local limit theorems applied to asymptotic enumeration II: Multivariate generating functions*, J. Combin. Theory Ser. A **34** (1983), 255–265.

[3] Andrew C. Berry, *The accuracy of the Gaussian approximation to the sum of independent variates*, Trans. Amer. Math. Soc. **49** (1941), 122–136.

[4] Michael Drmota, *Systems of functional equations*, Random Structures Algorithms **10** (1997), no. 1–2, 103–124.

[5] ———, *Random trees*, SpringerWienNewYork, 2009.

[6] Christopher Eagle, Zhicheng Gao, Mohamed Omar, Daniel Panario, and Bruce Richmond, *Distribution of the number of encryptions in revocation schemes for stateless receivers*, Fifth Colloquium on Mathematics and Computer Science, Discrete Math. Theor. Comput. Sci. Proc., AI, 2008, pp. 195–206.

[7] Carl-Gustav Esseen, *Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law*, Acta Math. **77** (1945), 1–125.

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

[9] Nicko G. Gamkrelidze, *A multidimensional generalization of Esseen's inequality for distribution functions*, Teor. Verojatnost. i Primenen. **22** (1977), no. 4, 897–900.

[10] ———, *A multidimensional generalization of Esseen's inequality for distribution functions*, Theory Probab. Appl. **22** (1977), 877–880, English Translation of the paper in Teor. Verojatnost. i Primenen.

[11] Allan Gut, *Probability: A graduate course*, Springer Texts in Statistics, Springer-Verlag New York, 2005.

[12] Clemens Heuberger, *Hwang's quasi-power-theorem in dimension two*, Quaest. Math. **30** (2007), 507–512.

[13] Clemens Heuberger, Sara Kropf, and Stephan Wagner, *Variances and covariances in the central limit theorem for the output of a transducer*, European J. Combin. **49** (2015), 167–187.

[14] Clemens Heuberger and Helmut Prodinger, *Analysis of alternative digit sets for nonadjacent representations*, Monatsh. Math. **147** (2006), 219–248.

[15] ———, *The Hamming weight of the non-adjacent-form under various input statistics*, Period. Math. Hungar. **55** (2007), 81–96.

[16] Hsien-Kuei Hwang, *On convergence rates in the central limit theorems for combinatorial structures*, European J. Combin. **19** (1998), 329–343.

[17] Sara Kropf, *Variance and covariance of several simultaneous outputs of a Markov chain*, arXiv:1508.05754 [math.CO], 2015.

[18] *The On-Line Encyclopedia of Integer Sequences*, `http://oeis.org`, 2015.

[19] Svetlana Poznanović and Christine E. Heitsch, *Asymptotic distribution of motifs in a stochastic context-free grammar model of RNA folding*, J. Math. Biol. **69** (2014), no. 6, 1743–1772.

[20] Bhagavatula Lakshmi Surya Prakasa Rao, *Another Esseen-type inequality for multivariate probability density functions*, Statist. Probab. Lett. **60** (2002), no. 2, 191–199.

[21] George G. Roussas, *An Esseen-type inequality for probability density functions, with an application*, Statist. Probab. Lett. **51** (2001), no. 4, 397–408.

[22] S. M. Sadikova, *On two-dimensional analogs of an inequality of Esseen and their application to the central limit theorem*, Teor. Verojatnost. i Primenen. **11** (1966), 369–380.

[23] ———, *On two-dimensional analogues of an inequality of Esseen and their application to the central limit theorem*, Theory Probab. Appl. **XI** (1966), 325–335, English Translation of the paper in Teor. Verojatnost. i Primenen.

[24] Richard P. Stanley, *Enumerative combinatorics. Volume 1*, second ed., Cambridge Studies in Advanced Mathematics, vol. 49, Cambridge University Press, Cambridge, 2012.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Probabilistic consequences of some polynomial recurrences

Paweł Hitczenko[1†], Amanda Lohss[1]

[1]*Department of Mathematics, Drexel University, Philadelphia, PA 19104, USA*

Abstract. In this paper, we consider sequences of polynomials that satisfy differential–difference recurrences. Our interest is motivated by the fact that polynomials satisfying such recurrences frequently appear as generating polynomials of integer valued random variables that are of interest in discrete mathematics. It is, therefore, of interest to understand the properties of such polynomials and their probabilistic consequences. As an illustration we analyze probabilistic properties of tree–like tableaux, combinatorial objects that are connected to asymmetric exclusion processes. In particular, we show that the number of diagonal boxes in symmetric tree–like tableaux is asymptotically normal and that the number of occupied corners in a random tree–like tableau is asymptotically Poisson. This extends earlier results of Aval, Boussicault, Nadeau, and Laborde Zubieta, respectively.

Keywords: Generating polynomial, recurrence, tree–like tableaux

1 Introduction and motivation

In this paper we will consider a sequence of polynomials

$$P_n(x) = \sum_{k=0}^{m} p_{n,k} x^k, \quad n \geq 0$$

that satisfy a differential–difference recurrence of one of the following forms

$$P_n'(x) = f_n(x) P_{n-1}(x) + g_n(x) P_{n-1}'(x) \tag{1}$$

or

$$P_n(x) = f_n(x) P_{n-1}(x) + g_n(x) P_{n-1}'(x) \tag{2}$$

for some sequences of polynomials (f_n), (g_n) and a given $P_0(x)$.

As a motivation for our interest we give examples of recurrences of these types that we encountered in recent literature. The first two examples appear in the context of tree–like tableaux introduced in Aval et al. (2013).

[†]Partially supported by a Simons Foundation grant #208766

(ABN) Aval et al. (2013):

$$\begin{aligned} B_n(x) &= nx(x+1)B_{n-1}(x) + x(1-x^2)B'_{n-1}(x), \\ B_0(x) &= x. \end{aligned}$$

(LZ) Laborde Zubieta (2015):

$$\begin{aligned} P_n^{'}(x) &= nP_{n-1}(x) + 2(1-x)P_{n-1}^{'}(x), \\ P_0(x) &= 1. \end{aligned}$$

Laborde Zubieta also considered the following version

$$\begin{aligned} Q_n^{'}(x) &= 2nxQ_{n-1}(x) + 2(1-x^2)Q_{n-1}^{'}(x), \\ Q_0(x) &= 1, \end{aligned}$$

where $Q_n(x)$ is a polynomial of degree $2n$ whose odd–numbered coefficients vanish. But this recurrence can be reduced to **(LZ)** by considering $Q_n(x) = P_n(x^2)$.

The following recurrence for fixed parameters a and b was considered in Hitczenko and Janson (2014) (see Sections 2 and 4 there):

(HJ) Hitczenko and Janson (2014):

$$\begin{aligned} P_{n,a,b}(x) &= ((n-1+b)x + a)P_{n-1,a,b}(x) + x(1-x)P'_{n-1,a,b}(x) \\ P_{0,a,b}(x) &= 1. \end{aligned}$$

This is a generaliztion of the classical Eulerian polynomials. Specifically, the choice of parameters $a = 1$ and $b = 0$ gives $P_{n,1,0} = E_n(x)$, where

$$E_n(x) = \sum_{k=0}^{n} \left\langle {n \atop k} \right\rangle x^k,$$

and $\left\langle {n \atop k} \right\rangle$ is the number of permutations of $\{1, \ldots, n\}$ with exactly k ascents. The recurrence for the polynomials $E_n(x)$ is:

$$E_n(x) = ((n-1)x + 1)E_{n-1}(x) + x(1-x)E'_{n-1}(x).$$

A very similar recurrence played a role in Dasse-Hartaut and Hitczenko (2013) although it appeared there only implicitly.

(DHH) Dasse-Hartaut and Hitczenko (2013):

$$\begin{aligned} V_n(x) &= ((2n-1)x + 1)V_{n-1}(x) + 2x(1-x)V'_{n-1}(x) \\ V_0(x) &= 1. \end{aligned}$$

As one more example, the following recurrence was used in (Acan and Hitczenko, 2016, Section 3) in connection with the analysis of a version of a card game called the memory game.

(AH) Acan and Hitczenko (2016):

$$\begin{aligned} A_n(x) &= (2n-1)A_{n-1}(x) + x(x-1)A_{n-1}^{'}(x), \\ A_0(x) &= x. \end{aligned}$$

In the examples above the polynomials are generating polynomials of integer valued random variables and it is of interest to understand what bearing the form of a recurrence has on the probabilistic properties of these random variables. This is, of course, not a new idea and in various forms has been studied for a long time (see, for example, many results and references in Flajolet and Sedgewick (2009)). Still, we believe that there is more work to be done to better understand the probabilistic consequences of the above recurrences.

2 Tree–like tableaux

Although we would like to keep the discussion at a general level, we will use particular objects, namely tree–like tableaux as a primary illustration. Therefore we briefly introduce the definition and their basic properties; we refer the reader to Aval et al. (2013); Laborde Zubieta (2015); Hitczenko and Lohss (2015) for more information and details.

A *Ferrers diagram* is a left–aligned finite set of cells arranged in rows and columns with weakly decreasing number of cells in rows. Its *half–perimeter* is the number of rows plus the number of columns. The *border edges* of a Ferrers diagram are the edges of the southeast border, and the number of border edges is equal to the half–perimeter. A *tree–like tableaux* of size n is a Ferrers diagrams of half-perimeter $n+1$ with some cells (called pointed cells) filled with a point according to the following rules:

1. The cell in the first column and first row is always pointed (this point is known as the root point).
2. Every row and every column contains at least one pointed cell.
3. For every pointed cell, all the cells above are empty or all the cells to the left are empty.

We will also consider *symmetric tree–like tableaux*, a subset of tree–like tableaux which are symmetric about their main diagonal (see (Aval et al., 2013, Section 2.2) for more details). As noticed in Aval et al. (2013), the size of a symmetric tree–like tableaux must be odd. It is known that there are $n!$ tree–like tableaux of size n (see (Aval et al., 2013, Corollary 8)) and $2^n n!$ symmetric tree–like tableaux of size $2n+1$ (see (Aval et al., 2013, Corollary 8)).

Corners of a tree–like tableau (symmetric or not) are the cells in which both the right and bottom edges are border edges. *Occupied corners* are corners that contain a point. Figure 1 shows examples of tree–like tableaux.

3 General setting

Motivated by examples discussed in Section 1 we wish to consider a sequence of polynomials

$$P_n(x) = \sum_{k=0}^{m} p_{n,k} x^k, \quad n \geq 0$$

(i) (ii)

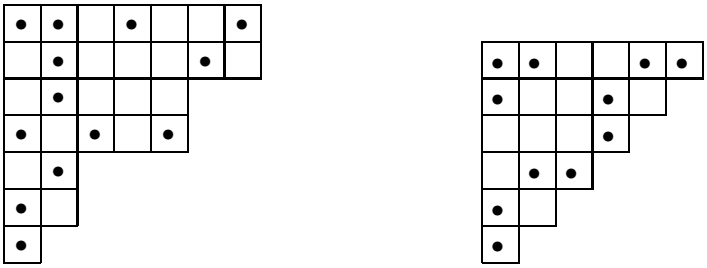


Fig. 1: (i) A tree–like tableaux of size 13 with 4 corners and 2 occupied corners. (ii) A symmetric tree–like tableaux of size 11 with 6 corners, 4 of which are occupied.

that satisfy one of the recurrences (1) or (2) with the given initial polynomial $P_0(x)$. The sequences of polynomials $(f_n(x))$ and $(g_n(x))$ are typically of low degree, but formally this is not required. Similarly, in all of the above examples we have $g_n(1) = 0$ and we will assume that throughout. It should be emphasized, however, that there are natural situations in which the condition $g_n(1) = 0$ fails. For example, Wang (2014) considered a recurrence

$$T_n(x) = (x + c)T_{n-1}(x) + mxT'_{n-1}(x),$$

for fixed numbers c and m. The choice $c = 0$ and $m = 1$ is a classical situation of Bell polynomials (see e. g. a discussion at the end of Section 7.2 in Chapter VII of Comtet (1974)). Furthermore, the choice $c = 1$ and any fixed $m \in \mathbb{N}$ gives polynomials associated with Whitney numbers of Dowling lattices (see Benoumhani (1999)). For polynomials satisfying

$$F_n(x) = (x + 1)F_{n-1}(x) + x(x + m)F'_{n-1}(x)$$

with $m \in \mathbb{N}$ we refer to (Benoumhani, 1997, Section 4) and references therein. So, clearly it is of interest to consider (1) or (2) without the assumption that $g_n(1) = 0$ but as we indicated earlier we will assume this throughout this paper.

Since we are interested in a probabilistic interpretation, we will assume that $p_{n,k} \geq 0$ and that $\sum_k p_{n,k} > 0$ for every n. Then

$$\frac{P_n(x)}{P_n(1)} = \sum_{k\geq 0} \frac{p_{n,k}}{P_n(1)} x^k$$

is the probability generating function of the integer valued random variable X_n whose distribution function is given by

$$\mathbb{P}(X_n = k) = \frac{p_{n,k}}{P_n(1)}, \quad k \geq 0. \tag{3}$$

We note that recurrence (1) defines the polynomials P_n up to an additive constant or, equivalently, up to the value $P_n(1)$. In our context the polynomials arise in the study of discrete combinatorial structures, and thus a natural choice of the normalization is obtained by letting $P_n(1)$ be the cardinality of the structure consisting of all objects of size n. For example, Laborde Zubieta set $P_n(1) = n!$ and $Q_n(1) = 2^n n!$

representing the number of tree–like tableaux of size n and the symmetric tree–like tableaux of size $2n+1$, respectively.

We want to use recurrences (1) and (2) to study the convergence in distribution of the sequences (X_n) associated with these recurrences through (3).

4 Method of moments

One natural approach is to use the method of moments or, more precisely, the method of factorial moments. It is based on the fact that if X is a random variable uniquely determined by its (factorial) moments

$$\mathbb{E}(X)_r = \mathbb{E}X(X-1)\dots(X-(r-1)), \quad r = 1, 2, \dots$$

and (X_n) is a sequence of random variables such that

$$\mathbb{E}(X_n)_r \longrightarrow \mathbb{E}(X)_r, \quad n \to \infty, \quad r = 1, 2, \dots$$

then

$$X_n \xrightarrow{d} X, \quad n \to \infty,$$

where "$\xrightarrow{d}$" denotes the convergence in distribution.

As is well–known, for a random variable X with probability generating function $h(x) = \mathbb{E}x^X$ we have

$$\mathbb{E}(X)_r = h^{(r)}(1),$$

where $h^{(r)}(x)$ is the r^{th} derivative of $h(x)$. Thus, in terms of polynomials $(P_n(x))$ this means

$$\mathbb{E}(X_n)_r = \frac{P_n^{(r)}(1)}{P_n(1)}$$

and consequently, we would be interested in computing $P_n^{(r)}(1)$ and finding the asymptotic of the ratio on the right–hand side above.

For recurrence (1) using Leibniz formula for higher order derivative of the product we obtain

$$\begin{aligned} P_n^{(r)}(x) &= (P_n'(x))^{(r-1)} = (f_n(x)P_{n-1}(x))^{(r-1)} + \left(g_n(x)P_{n-1}^{'}(x)\right)^{(r-1)} \\ &= \sum_{k=0}^{r-1}\binom{r-1}{k} f_n^{(k)}(x)P_{n-1}^{(r-1-k)}(x) + \sum_{k=0}^{r-1}\binom{r-1}{k} g_n^{(k)}(x)P_{n-1}^{(r-k)}(x) \\ &= g_n(x)P_{n-1}^{(r)}(x) + \sum_{k=0}^{r-2}\left(\binom{r-1}{k} f_n^{(k)}(x) + \binom{r-1}{k+1} g_n^{(k+1)}(x)\right) P_{n-1}^{(r-1-k)}(x) \\ &\quad + f_n^{(r-1)}(x)P_{n-1}(x). \end{aligned}$$

The idea now is that if f_n and g_n are low–degree polynomials then one obtains a manageable recurrence for $P_n^{(r)}(1)$. We will illustrate this on Laborde Zubieta's example **(**LZ**)**. In that case $f_n(x)$ and $g_n(x)$ are polynomials of degree zero and one, respectively and thus the above expression reduces to

$$P_n^{(r)}(x) = g_n(x)P_{n-1}^{(r)}(x) + \left(f_n(x) + (r-1)g_n'(x)\right) P_{n-1}^{(r-1)}(x) \tag{4}$$

if $r \geq 2$ (and agrees with (1) if $r = 1$). Laborde Zubieta used this, the specific form of the polynomials $f_n(x)$, $g_n(x)$, and $P_n(1) = n!$ to show that the random variables X_n defined by (3) satisfy

$$\mathbb{E}X_n = 1 \quad \text{and} \quad \text{var}(X_n) = \frac{n-2}{n}.$$

This suggests that the sequence (X_n) converges in distribution to a Poisson random variable with parameter 1. This is, indeed the case, and can be deduced from the recurrence (1) as was shown in Hitczenko and Lohss (2015). Here is a general statement that covers **(LZ)**.

Proposition 1 *Let*

$$P_n(x) = \sum_{k=0}^{m} p_{n,k} x^k$$

be a sequence of polynomials satisfying recurrence (1) where $f_n(x) = f_n$ and $g_n(x) = g_n \cdot (x-1)$ for some sequences of constants (f_n) and (g_n). Assume that $p_{n,k} \geq 0$ and that $\sum_k p_{n,k} > 0$ for every $n \geq 1$, and that $m = m_n$ may depend on n. Consider a sequence of random variables (X_n) defined by (3). If

$$g_n = o(f_n) \quad \textit{and} \quad f_n \frac{P_{n-1}(1)}{P_n(1)} \to c > 0, \quad \textit{as} \quad n \to \infty \tag{5}$$

then

$$X_n \xrightarrow{d} \textit{Pois}(c) \quad \textit{as} \quad n \to \infty,$$

where Pois(c) is a Poisson random variable with parameter c.

As established by Laborde Zubieta (2015), the generating polynomials for the number of occupied corners in tree–like tableaux satisfy recurrence **(LZ)** (that means taking $f_n = n$, $g_n = -2$, and $P_n(1) = n!$ in Proposition 1). Thus, the assumptions (5) are clearly satisfied with $c = 1$ and we obtain the following extension of Laborde Zubieta's result (see Hitczenko and Lohss (2015))

Corollary 2 *As $n \to \infty$, the limiting distribution of the number of occupied corners in a random tree–like tableau of size n is Pois(1).*

A companion result for symmetric tableaux is as follows (see Hitczenko and Lohss (2015) for more details). The expected value and the variance were obtained earlier in Laborde Zubieta (2015).

Corollary 3 *As $n \to \infty$, the limiting distribution of the number of occupied corners in a random symmetric tree–like tableau of size $2n+1$ is $2 \times$ Pois$(1/2)$.*

Proof of Proposition 1: By (Bollobás, 2001, Theorem 20, Chapter 1) it is enough to show that for every $r \geq 1$ the factorial moments

$$\mathbb{E}(X_n)_r = \mathbb{E}X_n(X_n - 1)\dots(X_n - (r-1)),$$

of (X_n) converge to c^r as $n \to \infty$. Using $g_n(1) = 0$ and $g_n'(x) = g_n$ in (4) we obtain

$$P_n^{(r)}(1) \quad = \quad (f_n + (r-1)g_n)\, P_{n-1}^{(r-1)}(1).$$

Consequently,

$$\begin{aligned}\frac{P_n^{(r)}(1)}{P_n(1)} &= (f_n+(r-1)g_n)\frac{P_{n-1}^{(r-1)}(1)}{P_n(1)}\\ &= f_n\frac{P_{n-1}(1)}{P_n(1)}\left(1+(r-1)\frac{g_n}{f_n}\right)\frac{P_{n-1}^{(r-1)}(1)}{P_{n-1}(1)}.\end{aligned}$$

Therefore, upon further iteration,

$$\frac{P_n^{(r)}(1)}{P_n(1)} = \left(\prod_{k=0}^{r-1} f_{n-k}\frac{P_{n-k-1}(1)}{P_{n-k}(1)}\left(1+(r-k-1)\frac{g_{n-k}}{f_{n-k}}\right)\right)\frac{P_{n-r}^{(r-r)}(1)}{P_{n-r}(1)}.$$

Since the last factor is 1, it follows from (5) that for every $r \geq 1$ as $n \to \infty$,

$$\frac{P_n^{(r)}(1)}{P_n(1)} \to c^r$$

as desired. □

Remark 1 *In principle it should be possible to prove a similar result for polynomials of higher degrees than those considered in Proposition 1. However, we have not tried to do that, primarily because we have not encountered instances of such recurrences.*

5 Real–rootedness of $P_n(x)$

The idea we explore in this section is that if all roots of $P_n(x)$ are real then $P_n(x)$ can be written as a product of linear factors. Furthermore, since the coefficients are non–negative the roots are non–positive. Hence, these linear factors may be interpreted as the generating functions of $\{0,1\}$–valued random variables and then knowing that the variance of their sum converges to infinity suffices to conclude that the sum is asymptotically normal. More specifically, assume that

$$-\infty < \gamma_{i,n} \leq 0, i = 1, \ldots, m$$

are roots of $P_n(x)$ and write $\pi_{i,n} = -\gamma_{i,n}$ so that $\pi_{i,n} \geq 0$. Then $P_n(x)$ has a factorization

$$P_n(x) = p_{n,m}\prod_{k=1}^{m}(x+\pi_{k,n}),$$

so that

$$\mathbb{E}x^{X_n} = \frac{P_n(x)}{P_n(1)} = \prod_{k=1}^{m}\frac{x+\pi_{k,n}}{1+\pi_{k,n}} = \prod_{k=1}^{m}\left(\frac{x}{1+\pi_{k,n}}+\frac{\pi_{k,n}}{1+\pi_{k,n}}\right).$$

The factor on the right–hand side is the probability generating function of a random variable $\xi_{k,n}$ such that

$$\mathbb{P}(\xi_{k,n}=1) = \frac{1}{1+\pi_{k,n}} \quad \text{and} \quad \mathbb{P}(\xi_{k,n}=0) = \frac{\pi_{k,n}}{1+\pi_{k,n}}, \quad k = 1, \ldots, m.$$

Moreover, since the product of the probability generating functions corresponds to taking sums of independent random variables we have that

$$X_n = \sum_{k=1}^{n} \xi_{k,n},$$

where $(\xi_{k,n})$ are independent. Therefore, it follows immediately from either Lindeberg or Lyapunov version of the central limit theorem (see e. g. (Billingsley, 1995, Theorem 27.2 or Theorem 27.3)) that

$$\frac{X_n - \mathbb{E}X_n}{\sqrt{\text{var}(X_n)}} \xrightarrow{d} N(0,1),$$

as long as $\text{var}(X_n) \longrightarrow \infty$ as $n \to \infty$. (Here $N(0,1)$ denotes the standard normal random variable.)

Since showing that the variance of X_n tends to infinity is generally not difficult from the recurrences (1) and (2), the main issue is real–rootedness of $P_n(x)$. This is, of course, not a new idea and the problem has a very long history and the questions of real–rootedness for many families of classical polynomials have been settled long time ago. In particular, in the context the present discussion, the proof that all roots of polynomials **(HJ)** are real was a slight modification of the proof for the Eulerian polynomials given by Frobenius (1910) more than hundred years ago. Nonetheless, the techniques seem to be tailored to the particular cases at hand. As far as general criteria for the real–rootedness of a family of recursively defined polynomials, not much seem to have been known until two relatively recent papers Dominici et al. (2011); Liu and Wang (2007). The first concerns recurrence (2) and requires $f_n(x)$ and $g_n(x)$ to have degrees at most one and two, respectively. The second, when specified to generality of (2) does not put any restrictions on the degrees of $f_n(x)$ and $g_n(x)$ but requires that $g_n(x) < 0$ whenever $x \leq 0$. While many of the real–rootedness results for classical polynomials may obtained from one of these criteria (and sometimes from both, e. g. Eulerian or Bell polynomials) some are not covered by them. In particular, neither Dominici et al. (2011) nor Liu and Wang (2007) applies to our first example **(ABN)**. Yet, as it turns out a modification of methods developed in Dominici et al. (2011) may be used to show that the polynomials $B_n(x)$ defined by **(ABN)** do, indeed, have all roots real. We will not prove it in this extended abstract, instead referring the reader to the full version of this paper.

6 Asymptotic normality of the number of diagonal boxes in symmetric tree–like tableaux

In this section we analyze the recurrence **(ABN)**. The polynomials

$$B_n(x) = \sum_{k=1}^{n+1} B(n,k)x^k, \quad n \geq 0,$$

were introduced in (Aval et al., 2013, Section 3.2) and are the generating polynomials for the number of diagonal cells in symmetric tree–like tableaux of size $2n+1$ (that is to say that $B(n,k)$ is the number of symmetric tree–like tableaux of size $2n+1$ with k diagonal cells). As was shown in Aval et al. (2013) $(B_n(x))$ satisfy the recurrence **(ABN)** and it follows readily from that that the expected number of diagonal cells in symmetric tableaux of size $2n+1$ is $3(n+1)/4$ (see (Aval et al., 2013, Proposition 19)). Continuing that work, we find the expression for the variance and show that the number of diagonal cells is asymptotically normal. The precise statement is as follows.

Theorem 4 *Let D_n be the number of diagonal boxes in a random symmetric tree–like tableau of size $2n+1$. Then, as $n \to \infty$*

$$\frac{D_n - 3(n+1)/4}{\sqrt{7(n+1)/48}} \xrightarrow{d} N(0,1).$$

Since (D_n) are random variables defined by

$$\mathbb{P}(D_n = k) = \frac{B(n,k)}{\sum_{k\geq 0} B(n,k)} = \frac{B(n,k)}{B_n(1)},$$

where $(B_n(x))$ satisfy recurrence **(ABN)** it follows form our discussion that theorem will be proved once we show that the variance of D_n grows to infinity with n and that all roots of $B_n(x)$ are real. The precise statements are given it two propositions below.

Proposition 5 *The variance of the number of diagonal cells in a random symmetric tree–like tableaux of size $2n+1$ is,*

$$\mathit{var}(D_n) = \frac{7(n+1)}{48}. \tag{6}$$

Proposition 6 *For all $n \geq 0$, the polynomial $B_n(x)$*

a) has degree $n+1$ with all coefficients non-negative, and

b) all roots real and in the interval $[-1, 0]$.

Because of the space limitation we will include here a proof of Proposition 5 only and we refer the reader to the full version of the paper for the proof of Proposition 6.

Proof of Proposition 5: First we will calculate the second factorial moment of D_n. Differentiating the recurrence **(ABN)** twice and evaluating at $x = 1$ yields

$$B_n''(1) = 2nB_{n-1}(1) + 6(n-1)B'_{n-1}(1) + 2(n-2)B''_{n-1}(1).$$

Furthermore, since

$$B_n(1) = 2nB_{n-1}(1)$$

and

$$\text{var}(D_n) = \mathbb{E}(D_n)_2 - \mathbb{E}^2 D_n + \mathbb{E}D_n \tag{7}$$

we obtain

$$\begin{aligned}
\mathbb{E}(D_n)_2 &= \frac{B_n^{''}(1)}{B_n(1)} = \frac{2nB_{n-1}(1) + 6(n-1)B'_{n-1}(1) + 2(n-2)B''_{n-1}(1)}{2nB_{n-1}(1)} \\
&= 1 + \frac{3(n-1)}{n}\mathbb{E}D_{n-1} + \frac{n-2}{n}\mathbb{E}(D_{n-1})_2 \\
&= 1 + \frac{3(n-1)}{n}\mathbb{E}D_{n-1} + \frac{n-2}{n}\left(\text{var}(D_{n-1}) + \mathbb{E}^2 D_{n-1} - \mathbb{E}D_{n-1}\right) \\
&= 1 + \frac{n-2}{n}\text{var}(D_{n-1}) + \frac{n-2}{n}\mathbb{E}^2 D_{n-1} + \left(\frac{2n-1}{n}\right)\mathbb{E}D_{n-1}.
\end{aligned}$$

Now, using $\mathbb{E}D_n = 3(n+1)/4$ (as computed from **(ABN)** in (Aval et al., 2013, Proposition 19)) and (7) we obtain

$$\begin{aligned}\text{var}(D_n) &= 1 + \frac{n-2}{n}\text{var}(D_{n-1}) + \frac{n-2}{n}\left(\frac{3n}{4}\right)^2 + \frac{2n-1}{n}\frac{3n}{4} \\ &\quad - \left(\frac{3(n+1)}{4}\right)^2 + \frac{3(n+1)}{4} \\ &= \frac{n-2}{n}\text{var}(D_{n-1}) + \frac{7}{16}.\end{aligned}$$

This recurrence is easily solved (see e. g. (Graham et al., 1994, Section 2.2)) and yields (6) completing the proof of Proposition 5 and Theorem 4. □

Remark 2 *The representation of D_n as the sum of independent indicator random variables implies that a local limit theorem holds too. Specifically, using $\mathbb{E}D_n = 3(n+1)/4$ and var$(D_n) = 7(n+1)/48$ we have that*

$$\mathbb{P}(D_n = k) = \frac{2\sqrt{6}}{\sqrt{7\pi(n+1)}}\left(\exp\left(-\frac{24(k-3(n+1)/4)^2}{7(n+1)}\right) + o(1)\right)$$

holds uniformly over k as $n \to \infty$. We refer to (Hitczenko and Janson, 2014, Theorem 2.7 and a discussion of its proof in Section 5) for more detailed explanation and to (Petrov, 1975, Theorem VII.3) for a general statement of a local limit theorem.

7 Conclusion

We have considered recurrences for generating polynomials of sequences of integer valued random variables and tried to use these recurrences to identify the distributional limits of the associated sequences of random variables. Some examples lead to Poisson limits, some other to Gaussian limits. In particular, we established the asymptotic normality for the number of diagonal cells in the random tree–like tableaux by verifying that the generating polynomials have only real roots and that the variance tends to infinity with n. However, there seem to be lack of general criteria that would allow one to find the limiting distribution of the underlying sequence of random variables directly from the recurrences of the form (2) or (1). For example, the limiting distribution of the random variables associated with the recurrence **(AH)** is neither Poisson nor normal. In fact, as have been shown in (Acan and Hitczenko, 2016, Section 3) if (X_n) is a sequence of random variables associated with the recurrence **(AH)** through (3) then

$$\frac{X_n}{2\sqrt{n}} \xrightarrow{d} X,$$

where X is a random variable with the probability density function $2xe^{-x^2}$ if $x \geq 0$ and is 0 otherwise. However, it is not clear how to see it from the recurrence **(AH)**. Factorial moments satisfy

$$\mathbb{E}(X_n)_r = \frac{2n-1+r}{2n-1}\mathbb{E}(X_{n-1})_r + \frac{r(r-1)}{2n-1}\mathbb{E}(X_{n-1})_{r-1}$$

and one can get from there

$$\mathbb{E}X_n = \frac{2n}{2n-1}\mathbb{E}X_{n-1} = \frac{(2n)!!}{(2n-1)!!} = \frac{2^{2n}}{\binom{2n}{n}} \sim \sqrt{\pi n}$$

and

$$\text{var}(X_n) = (4-\pi)n + O(\sqrt{n}).$$

In principle, higher moments can be found too. For example

$$\mathbb{E}(X_n)_3 = 6\left(\frac{\sqrt{\pi}(n+2)n!}{\Gamma(n+1/2)} - 4n - 3\right) \sim 6\sqrt{\pi}n^{3/2}$$

but the computations become increasingly more complicated. Even the asymptotic behavior of the first two moments is not immediately clear from the recurrence **(AH)**.

Thus, it seems worthwhile to further study the recurrences like (1) and (2) to obtain a more comprehensive picture of their probabilistic consequences.

References

H. Acan and P. Hitczenko. On the game of memory and preferential attachment graphs. *Adv. Appl. Probab.*, 49(2), 2016. to appear.

J.-C. Aval, A. Boussicault, and P. Nadeau. Tree-like tableaux. *Electron. J. Combin.*, 20(4):Paper 34, 24, 2013. ISSN 1077-8926.

M. Benoumhani. On some numbers related to Whitney numbers of Dowling lattices. *Adv. in Appl. Math.*, 19(1):106–116, 1997. ISSN 0196-8858.

M. Benoumhani. Log-concavity of Whitney numbers of Dowling lattices. *Adv. in Appl. Math.*, 22(2): 186–189, 1999. ISSN 0196-8858.

P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. ISBN 0-471-00710-2. A Wiley-Interscience Publication.

B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001. ISBN 0-521-80920-7; 0-521-79722-5.

L. Comtet. *Advanced combinatorics*. D. Reidel Publishing Co., Dordrecht, enlarged edition, 1974. ISBN 90-277-0441-4. The art of finite and infinite expansions.

S. Dasse-Hartaut and P. Hitczenko. Greek letters in random staircase tableaux. *Random Structures Algorithms*, 42(1):73–96, 2013. ISSN 1042-9832.

D. Dominici, K. Driver, and K. Jordaan. Polynomial solutions of differential-difference equations. *J. Approx. Theory*, 163(1):41–48, 2011. ISSN 0021-9045.

P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009. ISBN 978-0-521-89806-5.

G. Frobenius. Über die Bernoullischen Zahlen und die Eulerschen Polynome. In *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pages 809–847. Berlin, 1910.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete mathematics*. Addison-Wesley Publishing Company, Reading, MA, second edition, 1994. ISBN 0-201-55802-5. A foundation for computer science.

P. Hitczenko and S. Janson. Weighted random staircase tableaux. *Combin. Probab. Comput.*, 23(6): 1114–1147, 2014. ISSN 0963-5483.

P. Hitczenko and A. Lohss. Corners in tree–like tableaux, 2015. `arXiv:1511.04989`.

P. Laborde Zubieta. Occupied corners in tree–like tableaux. *Séminaire Lotharingien de Combinatoire*, 74, 2015. Article B74b.

L. L. Liu and Y. Wang. A unified approach to polynomial sequences with only real zeros. *Adv. in Appl. Math.*, 38(4):542–560, 2007. ISSN 0196-8858.

V. V. Petrov. *Sums of independent random variables*. Springer-Verlag, New York-Heidelberg, 1975. Translated from the Russian by A. A. Brown, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82.

D. G. L. Wang. On colored set partitions of type B_n. *Cent. Eur. J. Math.*, 12(9):1372–1381, 2014. ISSN 1895-1074.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Using Pólya urns to show normal limit laws for fringe subtrees in m-ary search trees

Cecilia Holmgren [1] [†], Svante Janson [1] [‡], Matas Šileikis[2]

[1] *Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden*
[2] *Department of Applied Mathematics of Faculty of Mathematics and Physics, Charles University in Prague, Malostranské nám. 25, 118 00 Praha, Czech Republic*

We study fringe subtrees of random m-ary search trees, by putting them in the context of generalised Pólya urns. In particular we show that for the random m-ary search tree with $m \leq 26$, the number of fringe subtrees that are isomorphic to an arbitrary fixed tree T converges to a normal distribution; more generally, we also prove multivariate normal distribution results for random vectors of such numbers for different fringe subtrees.

Keywords: Random trees; Fringe trees; Normal limit laws; Pólya urns; m-ary search trees

1 Introduction

The main focus of this paper is to consider fringe subtrees of random m-ary search trees; these random trees are defined in Section 2. Recall that a *fringe subtree* is a subtree consisting of some node and all its descendants, see Aldous [1] for a general theory, and note that fringe subtrees typically are "small" compared to the whole tree.

We will use (generalised) Pólya urns to analyze vectors of the numbers of fringe subtrees of different types in random m-ary search trees. As a result, we prove multivariate normal asymptotic distributions for these random variables, for m-ary search trees when $m \leq 26$. (It is well known that asymptotic normality does not hold for m-ary search trees for $m > 26$, see [2].)

Pólya urns have earlier been used to study the total number of nodes in random m-ary search trees, see [16, 13, 17]. In that case one only needs to consider an urn with $m-1$ different types, describing the nodes holding i keys, where $i \in \{0, 1, \dots, m-2\}$. Recently, in [10] more advanced Pólya urns were used to describe protected nodes in random m-ary search trees, where the types were further divided depending on characteristics of the different fringe subtrees (however, in [10] only the cases $m = 2, 3$ were treated in detail).

In [10] a simpler urn was also used to describe the total number of leaves in random m-ary search trees. In this work we further extend the approach used in [10] for analyzing arbitrary fringe subtrees of a fixed size in random m-ary search trees. This paper is an extended abstract of [12], where we also prove similar results for the general class of linear preferential attachment trees, and also extend the methods used in [10] to analyze the number of protected nodes in m-ary search trees for $m \leq 26$.

2 m-ary search trees

We recall the definition of m-ary search trees, see e.g. [15] or [6]. An m-ary search tree, where $m \geq 2$, is constructed recursively from a sequence of n *keys* (ordered numbers); we assume that the keys are distinct. Each node may contain up to $m-1$ keys. We start with a tree containing just an empty root. The first $m-1$ keys are put in the root, and are placed in increasing order from left to

[†] Partly supported by the Swedish Research Council
[‡] Partly supported by by the Knut and Alice Wallenberg Foundation

right; they divide the set of real numbers into m intervals $J_1, \dots, J_m$. When the root is full (after the first $m-1$ keys are added), it gets m children that are initially empty, and each further key is passed to one of the children depending on which interval it belongs to; a key in J_i is passed to the i'th child. (The binary search tree, i.e., the case $m = 2$, is the simplest case.) The procedure repeats recursively in the subtrees until all keys are added to the tree.

We are primarily interested in the random case when the keys form a uniformly random permutation of $\{1, \dots, n\}$, and we let $\mathcal{T}_n$ denote the random m-ary search tree constructed from such keys. (Only the order of the keys matters, so alternatively, we may assume that the keys are n i.i.d. uniform random numbers in $[0, 1]$.)

Nodes that contain at least one key are called *internal*, while empty nodes are called *external*. We regard the m-ary search tree as consisting only of the internal nodes; the external nodes are places for potential additions, and are useful when discussing the tree but are not really part of the tree. Thus, a *leaf* is an internal node that has no internal children, but it may have external children.

We say that a node with $i \le m-2$ keys has $i+1$ *gaps*, while a full node has no gaps. It is easily seen that an m-ary search tree with n keys has $n+1$ gaps; the gaps correspond to the intervals of real numbers between the keys (and $\pm\infty$), and a new key has the same probability $1/(n+1)$ of being inserted into any of the gaps. Thus, the evolution of the random m-ary search tree may be described by choosing a gap uniformly at random at each step, and inserting a new key there.

Note that the construction above yields the m-ary search tree as an ordered tree. Hence, a nonrandom m-ary search tree is an ordered rooted tree where each node is marked with the number of keys it contains, with this number being in $\{0, \dots, m-1\}$ and such that nodes with $m-1$ keys have exactly m children, and the other nodes are leaves. There is a natural partial order on the set of (isomorphism classes of) nonrandom m-ary search trees, such that $T \preceq T'$ if T' can be obtained from T by adding keys (including the case $T' = T$).

In applications where the order of the children of a node does not matter, we can simplify by ignoring the order and regard the m-ary search tree as an unordered tree. (Then, we can also ignore the external nodes.) A partial order $T \preceq T'$ is defined on the set of (isomorphism classes of) unordered m-ary search trees in the same way as in the ordered case.

3 Main results

In this section we state the main results on fringe subtrees in random m-ary search trees. These results are extensions of results that previously have been shown for the specific case of the random binary search tree with the use of other methods, see e.g., [4, 5, 9].

Remark 3.1 As said in the introduction, m-ary search trees can be regarded as either ordered or unordered trees. The most natural interpretation is perhaps the one as ordered trees, and it implies the corresponding result for unordered trees in, e.g., Theorem 3.2. However, in some applications it is preferable to regard the fringe trees as unordered trees, since this gives fewer types to consider in the Pólya urns that we use, see e.g., Example 5.1.

The following theorem generalises [9, Theorem 1.22], where the specific case of the binary search tree was analyzed.

Let $H_m := \sum_{k=1}^{m} 1/k$ be the m'th harmonic number.

Theorem 3.2 *Assume that $2 \le m \le 26$. Let $T^1, \dots, T^d$ be a fixed sequence of nonrandom m-ary search trees and let $\mathbf{Y}_n = \left(X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d}\right)$, where $X_n^{T^i}$ is the (random) number of fringe subtrees that are isomorphic to T^i in the random m-ary search tree $\mathcal{T}_n$ with n keys. Let k_i be the number of keys of T^i for $i \in \{1, \dots, d\}$. Let*

$$\boldsymbol{\mu}_n := \mathbb{E}\,\mathbf{Y}_n = \left(\mathbb{E}(X_n^{T^1}), \mathbb{E}(X_n^{T^2}), \dots, \mathbb{E}(X_n^{T^d})\right).$$

Then

$$n^{-1/2}(\mathbf{Y}_n - \boldsymbol{\mu}_n) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma), \tag{3.1}$$

where $\Sigma = (\sigma_{ij})_{i,j=1}^{d}$ *is some covariance matrix. Furthermore, in* (3.1)*, the vector* $\boldsymbol{\mu}_n$ *can be replaced by the vector* $\hat{\boldsymbol{\mu}}_n := n\hat{\boldsymbol{\mu}}$*, with*

$$\hat{\boldsymbol{\mu}} := \Big(\frac{\mathbb{P}(\mathcal{T}_{k_1} = T^1)}{(H_m - 1)(k_1 + 1)(k_1 + 2)}, \dots, \frac{\mathbb{P}(\mathcal{T}_{k_d} = T^d)}{(H_m - 1)(k_d + 1)(k_d + 2)} \Big). \tag{3.2}$$

Moreover, if the trees $T^1, \dots, T^d$ *have at least one internal node each, then the covariance matrix* Σ *is non-singular.*

Remark 3.3 That $\boldsymbol{\mu}_n$ can be replaced by the vector $\hat{\boldsymbol{\mu}}_n$ means that

$$\mathbb{E}\big(X_n^{T^i}\big) = \frac{\mathbb{P}(\mathcal{T}_{k_i} = T^i)}{(H_m - 1)(k_i + 1)(k_i + 2)} n + o\big(n^{1/2}\big). \tag{3.3}$$

A weaker version of (3.3) with the error term $o(n)$ follows from the branching process analysis of fringe subtrees in [11], see the proof in Section 6. The vector $\hat{\boldsymbol{\mu}}_n$ can also, using (5.2) below, be calculated from an eigenvector of the intensity matrix of the Pólya urn defined in Section 5, see Theorem 4.1(i). See also [14].

Also the covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1}^{d}$ can be calculated explicitly from the intensity matrix of the Pólya urn, see Theorem 4.1(ii). The results in [14] show also

$$\sigma_{ij} = \lim_{n\to\infty} \frac{1}{n} \operatorname{Cov}\big(X_n^{T^i}, X_n^{T^j}\big). \tag{3.4}$$

The following theorem is an important corollary of Theorem 3.2. It also follows from Fill and Kapur [7, Theorem 5.1]. The special case of the random binary search tree was proved by Devroye [4], and the covariances for $Y_{n,k}$ in that case were given by Dennert and Grübel [3], see also [9, Theorem 1.19 and Proposition 1.10]. For binary search trees also the case when the size k is depending on n has been analyzed; in that case both normal and Poisson limit laws appear, see e.g., Fuchs [8] and [9].

Theorem 3.4 *Assume that* $2 \le m \le 26$*. Let* k *be an arbitrary fixed integer and let* $Y_{n,k}$ *be the (random) number of fringe subtrees with* k *keys in the random* m*-ary search tree* $\mathcal{T}_n$ *with* n *keys. Then, as* $n \to \infty$*,*

$$n^{-1/2}\big(Y_{n,k} - \mathbb{E}\, Y_{n,k}\big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2), \tag{3.5}$$

where σ_k^2 *is some constant with* $\sigma_k^2 > 0$ *except when* $k = 0$ *and* $m = 2$*. Furthermore, we also have*

$$n^{-1/2}\Big(Y_{n,k} - \frac{n}{(H_m - 1)(k + 1)(k + 2)}\Big) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_k^2). \tag{3.6}$$

Remark 3.5 The asymptotic mean $\frac{n}{(H_m-1)(k+1)(k+2)}$ in (3.6) easily follows from (3.3). The constant σ_k^2 can again be calculated explicitly from our proof.

We give one example of Theorem 3.4 in Section 7, where we let $m = 3$ and $k = 4$.

4 Generalised Pólya urns

A (generalised) Pólya urn process is defined as follows, see e.g. [13] or [17]. There are balls of q types (or colours) $1, \dots, q$, and for each n a random vector $\mathcal{X}_n = (X_{n,1}, \dots, X_{n,q})$, where $X_{n,i}$ is the number of balls of type i in the urn at time n. The urn starts with a given vector $\mathcal{X}_0$. For each type i, there is an activity (or weight) $a_i \in \mathbb{R}_{\ge 0}$, and a random vector $\xi_i = (\xi_{i1}, \dots, \xi_{iq})$, where $\xi_{ij} \in \mathbb{Z}_{\ge 0}$ and $\xi_{ii} \in \mathbb{Z}_{\ge -1}$. The urn evolves according to a discrete time Markov process. At each time $n \ge 1$, one ball is drawn at random from the urn, with the probability of any ball proportional to its activity. Thus, the drawn ball has type i with probability $\frac{a_i X_{n-1,i}}{\sum_j a_j X_{n-1,j}}$. If the drawn ball has type i, it is replaced together with $\Delta X_{n,j}^{(i)}$ balls of type j, $j = 1, \dots, q$, where the random vector

$\Delta\mathcal{X}_n^{(i)} = (\Delta X_{n,1}^{(i)}, \dots, \Delta X_{n,q}^{(i)})$ has the same distribution as ξ_i and is independent of everything else that has happened so far. We allow $\Delta X_{n,i}^{(i)} = -1$, which means that the drawn ball is *not* replaced.

The *intensity matrix* of the Pólya urn is the $q \times q$ matrix

$$A := (a_j \, \mathbb{E}\, \xi_{ji})_{i,j=1}^{q}. \tag{4.1}$$

The intensity matrix A with its eigenvalues and eigenvectors is central for proving limit theorems.

We use the basic assumptions (A1)–(A6) on the Pólya urn stated in [13, p. 180] together with the following simplifying assumption, cf. [10]:

(A7) At each time $n \geq 1$, there exists a ball of a dominating type, as defined in [13].

Using the Perron–Frobenius theorem, it is easy to verify all conditions (A1)–(A6) for the Pólya urns used in this paper, and (A7) follows because the urn is irreducible if we ignore balls with activity 0, and there will always be a ball of positive activity, see [13, Lemma 2.1] and the discussion in [12].

Before stating the results that we use, we need some notation. By a vector v we mean a column vector, and we write v' for its transpose (a row vector). More generally, we denote the transpose of a matrix A by A'. By an eigenvector of A we mean a right eigenvector; a left eigenvector is the same as the transpose of an eigenvector of the matrix A'. If u and v are vectors then $u'v$ is a scalar while uv' is a $q \times q$ matrix of rank 1. We also use the notation $u \cdot v$ for $u'v$. We let λ_1 denote the largest real eigenvalue of A. (This exists by our assumptions and the Perron–Frobenius theorem.) Let $a = (a_1, \dots, a_q)$ denote the (column) vector of activities, and let u_1' and v_1 denote left and right eigenvectors of A corresponding to the largest eigenvalue λ_1, i.e., vectors satisfying

$$u_1' A = \lambda_1 u_1', \qquad\qquad A v_1 = \lambda_1 v_1.$$

We assume that v_1 and u_1 are normalised so that

$$a \cdot v_1 = a' v_1 = v_1' a = 1, \qquad\qquad u_1 \cdot v_1 = u_1' v_1 = v_1' u_1 = 1, \tag{4.2}$$

see [13, equations (2.2)–(2.3)]. We write $v_1 = (v_{11}, \dots, v_{1q})$.

We define $P_{\lambda_1} = v_1 u_1'$, and $P_I = I_q - P_{\lambda_1}$, where I_q is the $q \times q$ identity matrix. We define the matrices

$$B_i := \mathbb{E}(\xi_i \xi_i') \tag{4.3}$$

$$B := \sum_{i=1}^{q} v_{1i} a_i B_i \tag{4.4}$$

$$\Sigma_I := \int_0^\infty P_I e^{sA} B e^{sA'} P_I' e^{-\lambda_1 s} ds, \tag{4.5}$$

where we recall that $e^{tA} = \sum_{j=0}^{\infty} t^j A^j / j!$. From [13] it follows that when $\operatorname{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$, the integral Σ_I in (4.5) converges.

Furthermore, it is proved in [13] that, under assumptions (A1)–(A7), $\mathcal{X}_n$ is asymptotically normal if $\operatorname{Re}\lambda \leq \lambda_1/2$ for each eigenvalue $\lambda \neq \lambda_1$. We will apply the following result from [13].

Theorem 4.1 ([13, Theorem 3.22 and Lemma 5.4]) *Assume* (A1)–(A7) *and that we have normalised as in* (4.2)*. Also assume that* $\operatorname{Re}\lambda < \lambda_1/2$*, for each eigenvalue* $\lambda \neq \lambda_1$*.*

(i) *Then, as* $n \to \infty$*,*

$$n^{-1/2}(\mathcal{X}_n - n\mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma), \tag{4.6}$$

with $\mu = \lambda_1 v_1$ *and some covariance matrix* Σ*.*

(ii) *Suppose further that, for some* $c > 0$*,*

$$a \cdot \mathbb{E}(\xi_i) = c, \qquad i = 1, \dots, q. \tag{4.7}$$

Then the covariance matrix in (4.6) *is given by* $\Sigma = c\Sigma_I$*, with* Σ_I *as in* (4.5)*.*

□

Remark 4.2 It is easily seen that (4.7) implies that $\lambda_1 = c$ and $u_1 = a$, see e.g. [13, Lemma 5.4]. There is also an alternative way to evaluate Σ in the case when A is diagonalisable (which is the case at least for many examples of Theorem 3.2 and Theorem 3.4, e.g., the example in Section 7), see [12, Theorem 4.1(iii)] or [13, Lemma 5.3].

Remark 4.3 From (4.6) follows immediately a weak law of large numbers:

$$\mathcal{X}_n/n \xrightarrow{\mathrm{p}} \mu. \tag{4.8}$$

In fact, the corresponding strong law $\mathcal{X}_n/n \xrightarrow{\mathrm{a.s.}} \mu$ holds as well, see [13, Theorem 3.21]. Furthermore, in all applications in the present paper, all ξ_{ij} are bounded and thus each $X_{n,i} \le Cn$ for some deterministic constant; hence (4.8) implies by dominated convergence that also the means converge:

$$\mathbb{E}\,\mathcal{X}_n/n \to \mu. \tag{4.9}$$

5 Pólya urns to count fringe subtrees in random m-ary search trees

In this section we describe the Pólya urns that we will use in the analysis of fringe subtrees to prove Theorem 3.2 and Theorem 3.4 for m-ary search trees. We consider either ordered or unordered trees, see Remark 3.1.

Let $T^1, \dots, T^d$ be a fixed sequence of (nonrandom) m-ary search trees and let as in Theorem 3.2 $\mathbf{Y}_n = (X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d})$, where $X_n^{T^i}$ is the number of fringe subtrees in $\mathcal{T}_n$ that are isomorphic to T^i. We may assume that at least one tree T^i contains at least $m-2$ keys. (Otherwise we simply add one such tree to the sequence.)

Assume that we have a given m-ary search tree $\mathcal{T}_n$ together with its external nodes. Denote the fringe subtree of $\mathcal{T}_n$ rooted at a node v by $\mathcal{T}_n(v)$. We say that a node v is *living* if $\mathcal{T}_n(v) \preceq T^i$ for some $i \in \{1, \dots, d\}$, i.e., if $\mathcal{T}_n(v)$ is isomorphic to some T^i or can be grown to become one of them by adding more keys. Note that this includes all external nodes and all leaves with at most $m-2$ keys (by the assumption that at least one tree T^i contains at least $m-2$ keys). Furthermore, we let all descendants of a living node be living. All other nodes are *dead*.

Now erase all edges from dead nodes to their children. This yields a forest of small trees, where each tree either consists of a single dead node or is living (meaning that all nodes are living) and can be grown to become one of the T^i. We regard these small trees as the balls in our generalised Pólya urn. Hence, the types in this Pólya urn are all (isomorphism types of) nonrandom m-ary search trees T such that $T \preceq T^i$ for some $i \in \{1, \dots, d\}$, plus one dead type. We denote the set of living types by

$$\mathcal{S} := \bigcup_{i=1}^{d} \{T : T \preceq T^i\}, \tag{5.1}$$

and the set of all types by $\mathcal{S}^* := \mathcal{S} \cup \{*\}$, where $*$ is the dead type.

When a key is added to the tree $\mathcal{T}_n$, it is added to a leaf with at most $m-2$ keys or an external node, and thus to one of the living subtrees in the forest just described. If the root of that subtree still is living after the addition, then that subtree becomes a living subtree of a different type; if the root becomes dead, then the subtree is further decomposed into one or several dead nodes and several (at least m) living subtrees. In any case, the transformation does not depend on anything outside the subtree where the key is added. The random evolution of the forest obtained by decomposing $\mathcal{T}_n$ is thus described by a Pólya urn with the types $\mathcal{S}^*$, where each type has activity equal to its number of gaps, and certain transition rules that in general are random, since the way a subtree is decomposed (or perhaps not decomposed) typically depends on where the new key is added.

Note that dead balls have activity 0; hence we can ignore them and consider only the living types (i.e., the types in $\mathcal{S}$) and we will still have a Pólya urn. The number of dead balls can be recovered

from the numbers of balls of other types if it is desired, since the total number of keys is non-random and each dead ball contains $m-1$ keys.

Let $X_{n,T}$ be the number of balls of type T in the Pólya urn, for $T \in \mathcal{S}$. The trees T^i that we want to count correspond to different types in the Pólya urn, but they may also appear as subtrees of larger living trees. Hence, if $n(T,T')$ denotes the number of fringe subtrees in T that are isomorphic to T', then $X_n^{T_i}$ is the linear combination

$$X_n^{T^i} = \sum_{T\in\mathcal{S}} n(T,T^i) X_{n,T}. \tag{5.2}$$

The strategy to prove Theorem 3.2 should now be obvious. We verify that the Pólya urn satisfies the conditions of Theorem 4.1 (this is done in Section 6); then that theorem yields asymptotic normality of the vectors $(X_{n,T})_{T\in\mathcal{S}}$, and then asymptotic normality of $(X_n^{T^1},\dots,X_n^{T^d})$ follows from (5.2).

Example 5.1 (a Pólya urn to count fringe subtrees with k keys) As an important example, let us consider the problem of finding the distribution of the number of fringe subtrees with a given number of keys, as in Theorem 3.4. In this case, the order of children in the tree does not matter so it is easier to regard the trees as unordered.

Thus, fix $k \ge m-2$ and let T^i, $i \in \{1,\dots,d\}$, be the sequence of all m-ary search trees that can be obtained with at most k keys. Hence, (5.1) yields $\mathcal{S} = \{T^i : 1 \le i \le d\}$.

In the decomposition of an m-ary search tree constructed above, a node v is living if and only if the fringe subtree rooted at v has at most k keys. Hence, the decomposition consists of all maximal fringe subtrees with at most k keys, plus dead nodes, which we ignore.

The replacement rules in the Pólya urn are easy to describe. The types are the m-ary search trees with at most k keys. A type T with j keys has $j+1$ gaps, and is thus given activity $j+1$. Let $T_1,\dots,T_{j+1}$ be the trees obtained by adding a key to one of these gaps in T. (Some of these may be equal.) If we draw a ball of type T and $j<k$, then the drawn ball is replaced by one ball of a type randomly chosen among $T_1,\dots,T_{j+1}$ (with probability $1/(j+1)$ each); note that these trees have $j+1 \le k$ keys and are themselves types in the urns. On the other hand, if $j=k$, then each of these trees has $k+1$ keys so its root is dead; the root contains $m-1$ keys so after removing it we are left with m subtrees with together $k+1-(m-1) \le k$ keys, so these subtrees are all living and the decomposition stops there. Consequently, when $j=k$, the drawn ball is replaced by m balls of the types obtained by choosing one of $T_1,\dots,T_{k+1}$ uniformly at random and then removing its root; this leaves m living subtrees and we add balls of the corresponding types.

To find the number of fringe subtrees with k keys, we sum the numbers $X_{n,T}$ of balls of type T in the urn, for all types T with exactly k keys. Note that we similarly, using (5.2), may obtain the number of fringe subtrees with ℓ keys, for any $\ell \le k$, from the same urn. This enables us to obtain joint convergence in Theorem 3.4 for several different k, with asymptotic covariances that can be computed from this urn.

Note that for $k=m-2$, the urn described here consists of $m-1$ types, viz. a single node with $i-1$ keys for $i=1,\dots,m-1$. This urn has earlier been used in [16, 13, 17] to study the number of nodes, and the numbers of nodes with different numbers of keys, in an m-ary search tree.

In Section 7 we give an example with $m=3$ and $k=4$; in that case there are 6 different (living) types in the Pólya urn.

Remark 5.2 The types described by the Pólya urns above all have activities equal to the total number of gaps in the type. Since the total number of gaps increases by 1 in each step, we have $a\cdot\xi_i = 1$ for every i, deterministically; in particular, (4.7) holds with $c=1$. Hence, $\lambda_1 = 1$ by Remark 4.2.

6 Proofs

As said in Section 4, it is easy to see that the Pólya urns constructed in Section 5 satisfy (A1)–(A7), for example with the help of [13, Lemma 2.1]. To apply Theorem 4.1 it remains to show that $\operatorname{Re}\lambda < \lambda_1/2$ for each eigenvalue $\lambda \ne \lambda_1$. We will find the eigenvalues of A by using induction

on the size of $\mathcal{S}$, the set of (living) types. For definiteness we consider the version with ordered unlabelled trees; the version with unordered trees is the same up to minor differences that are left to the reader.

Note that there is exactly one type that has activity j for every $j \in \{1, \dots, m-1\}$. (These correspond to the nodes holding $j-1$ keys.) These types are the $m-1$ smallest in the partial order $\preceq$, and they always belong to the set $\mathcal{S}$ constructed in Section 5.

Let $q := |\mathcal{S}|$ be the number of types in $\mathcal{S}$, and choose a numbering $T_1, \dots, T_q$ of these q types that is compatible with the partial order $\preceq$. For $k \le q$, let

$$\mathcal{S}_k := \{T_1, \dots, T_k\}. \tag{6.1}$$

For $k \ge m-1$, we may thus consider the Pólya urn with the k types in $\mathcal{S}_k$ constructed as in Section 5. Note that this corresponds to decomposing $\mathcal{T}_n$ into a forest with all components in $\mathcal{S}_k \cup \{*\}$. Furthermore, let $\mathcal{X}_n^k := (X_{n,1}^k, \dots, X_{n,k}^k)$, where $X_{n,i}^k$ is the number of balls of type T_i in the urn with types $\mathcal{S}_k$ at time n and let A_k be the intensity matrix of this Pólya urn. Thus $A = A_q$.

First let us take a look at the diagonal values ξ_{ii}. In the result below we assume $m \ge 3$, the case $m = 2$ is similar and we refer to [12] for the corresponding statement and proof in that case.

Proposition 6.1 *Let $m \ge 3$ and $m-1 \le k \le q$. Then $(A_k)_{ii} = -a_i$ for every type $i = 1, \dots, k$. Hence, the trace satisfies*

$$\operatorname{tr}(A_k) = -\sum_{i=1}^{k} a_i. \tag{6.2}$$

Proof: Observe that if we draw a ball of type i with k_i keys, then the ball is replaced either by a single ball of a type with $k_i + 1$ keys or by several different balls obtained by decomposing a tree with $k_i + 1$ keys that has a dead root. In the latter case, $m-1$ of the keys are in the dead root, so each living tree in the decomposition has at most $k_i + 1 - (m-1) = k_i - m + 2$ keys.

Hence, if $m \ge 3$, then in no case will there be a ball with exactly k_i keys among the added balls, and in particular no ball of type i; consequently, $\xi_{ii} = -1$ and $(A_k)_{ii} = -a_i$, see (4.1). □

Theorem 6.2 *Let $m \ge 2$. The eigenvalues of A are the $m-1$ roots of the polynomial $\phi_m(\lambda) := \prod_{i=1}^{m-1}(\lambda + i) - m!$ plus the multiset*

$$\{-a_i : i = m, m+1, \dots, q\}. \tag{6.3}$$

Proof: We prove by induction on k that the theorem holds for A_k (with q replaced by k in (6.3)), for any k with $m-1 \le k \le q$. The theorem is the case $k = q$.

First, for the initial case $k = m-1$, T_i is a single node with $i-1$ keys, $i = 1, \dots, k$; thus $X_{n,i}^{m-1}$ is the number of nodes with $i-1$ keys, i.e., the number of nodes with i gaps. (In particular, $X_{n,1}^{m-1}$ is the number of external nodes.) This Pólya urn with $m-1$ types has earlier been analyzed, see e.g., [13, Example 7.8] and [17, Section 8.1.3]. The $(m-1) \times (m-1)$ matrix A_{m-1} has elements $a_{i,i} = -i$ for $i \in \{1, \dots, m-1\}$, $a_{i,i-1} = i-1$ for $i \in \{2, \dots, m\}$, $a_{1,m-1} = m \cdot (m-1)$ and all other elements $a_{i,j} = 0$, i.e.,

$$A_{m-1} = \begin{pmatrix} -1 & 0 & 0 & \dots & 0 & m(m-1) \\ 1 & -2 & 0 & \dots & 0 & 0 \\ 0 & 2 & -3 & \dots & 0 & 0 \\ 0 & 0 & 3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & m-2 & -(m-1) \end{pmatrix}. \tag{6.4}$$

As is well-known, the matrix A_{m-1} has characteristic polynomial $\phi_m(\lambda)$; this shows the theorem for $k = m-1$, since the set (6.3) is empty in this case.

We proceed to the induction step. Let $m-1 \le k < q$. By using arguments similar to those that were used in the proof of [10, Lemma 5.1] we will show that A_{k+1} inherits (with multiplicities) the eigenvalues of A_k. We write $a^k = (a_1, \dots, a_k)$ for the activity vector for the Pólya urn with types in $\mathcal{S}_k$.

We have $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{T_{k+1}\}$. Since the vector $\mathcal{X}_n^{k+1}$ obviously determines also the number of subtrees of each type in the decomposition of $\mathcal{T}_n$ into the types in $\mathcal{S}_k$, there is an obvious linear map $T : \mathbb{R}^{k+1} \to \mathbb{R}^k$ that is onto such that $\mathcal{X}_n^k = T\mathcal{X}_n^{k+1}$. Furthermore, starting the urns with an arbitrary (deterministic) non-zero vector $\mathcal{X}_0^{k+1} \in \mathbb{Z}_{\ge 0}^{k+1}$ and $\mathcal{X}_0^k = T\mathcal{X}_0^{k+1}$, the urn dynamics yield

$$\mathbb{E}(\mathcal{X}_1^{k+1} - \mathcal{X}_0^{k+1}) = \frac{A_{k+1}\mathcal{X}_0^{k+1}}{a^{k+1} \cdot \mathcal{X}_0^{k+1}}, \tag{6.5}$$

$$\mathbb{E}(\mathcal{X}_1^{k} - \mathcal{X}_0^{k}) = \frac{A_{k}\mathcal{X}_0^{k}}{a^{k} \cdot \mathcal{X}_0^{k}}. \tag{6.6}$$

Consequently, since also $a^{k+1} \cdot \mathcal{X}_0^{k+1} = a^k \cdot \mathcal{X}_0^k$ (this is the total activity, i.e., the total number of gaps),

$$\begin{aligned} TA_{k+1}\mathcal{X}_0^{k+1} &= (a^{k+1} \cdot \mathcal{X}_0^{k+1}) T\,\mathbb{E}(\mathcal{X}_1^{k+1} - \mathcal{X}_0^{k+1}) = (a^k \cdot \mathcal{X}_0^k)\,\mathbb{E}(\mathcal{X}_1^k - \mathcal{X}_0^k) = A_k\mathcal{X}_0^k \\ &= A_k T\mathcal{X}_0^{k+1}, \end{aligned}$$

and thus, since $\mathcal{X}_0^{k+1}$ is arbitrary,

$$TA_{k+1} = A_k T. \tag{6.7}$$

Let u' be a left generalised eigenvector of rank p corresponding to the eigenvalue λ of the matrix A_k, i.e.,

$$u'(A_k - \lambda I_k)^p = 0.$$

Then, by (6.7),

$$u'T(A_{k+1} - \lambda I_{k+1})^p = u'(A_k - \lambda I_k)^p T = 0,$$

and thus $u'T = (T'u)'$ is a left generalised eigenvector of A_{k+1} for the eigenvalue λ. Since T is onto, T' is injective and thus T' is an injective map of the generalised eigenspace (for λ) of A_k into the generalised eigenspace of A_{k+1}. This shows that λ is an eigenvalue of A_{k+1} with algebraic multiplicity at least as large as for A_k. Consequently, if A_k has eigenvalues $\lambda_1, \dots, \lambda_k$ (including repetitions, if any), then A_{k+1} has eigenvalues $\lambda_1, \dots, \lambda_k, \lambda_{k+1}$ for some complex number λ_{k+1}.

Then the result follows by the following observation. The trace of a matrix is equal to the sum of the eigenvalues; hence,

$$\operatorname{tr} A_{k+1} = \lambda_1 + \cdots + \lambda_{k+1} = \operatorname{tr} A_k + \lambda_{k+1} \tag{6.8}$$

and thus by (6.2) (when $m > 2$) or the corresponding result in [12] (when $m = 2$),

$$\lambda_{k+1} = \operatorname{tr}(A_{k+1}) - \operatorname{tr}(A_k) = -a_{k+1}. \tag{6.9}$$

Thus, by induction, Theorem 6.2 holds for every A_k, with $m-1 \le k \le q$, and in particular for $A = A_q$. □

Theorem 6.2 shows that the eigenvalues of A are the roots of ϕ_m plus some negative numbers $-a_i$; hence the condition $\operatorname{Re}\lambda < \lambda_1/2$ in Theorem 4.1 is satisfied for all eigenvalues of A except λ_1 if the condition is satisfied for the roots of ϕ_m (except λ_1); it is well-known that this holds if $m \le 26$, but not for larger m, see [18] and [7].

In the remainder of this section we assume $m \le 26$. Thus $\operatorname{Re}\lambda < \lambda_1/2$ for every eigenvalue $\lambda \ne \lambda_1$, and Theorem 4.1 applies to the urn defined above.

Proof of Theorem 3.2: By Theorem 4.1(i), (4.6) holds, with $\mu = \lambda_1 v_1 = v_1$.

By (5.2), $\mathbf{Y}_n = \bigl(X_n^{T^1}, X_n^{T^2}, \dots, X_n^{T^d}\bigr) = R\mathcal{X}_n$ for some (explicit) linear operator R. Hence, (4.6) implies

$$n^{-1/2}\bigl(\mathbf{Y}_n - nR\mu\bigr) = R\bigl(n^{-1/2}(\mathcal{X}_n - \mu)\bigr) \xrightarrow{\mathrm{d}} \mathcal{N}\bigl(0, R\Sigma R'\bigr). \tag{6.10}$$

Furthermore, by [14],

$$\mathbb{E}\,\mathcal{X}_n = n\mu + o\bigl(n^{1/2}\bigr), \tag{6.11}$$

and thus

$$\boldsymbol{\mu}_n = \mathbb{E}\,\mathbf{Y}_n = R\bigl(\mathbb{E}\,\mathcal{X}_n\bigr) = nR\mu + o\bigl(n^{1/2}\bigr). \tag{6.12}$$

Hence, (6.10) implies (3.1) (with the covariance matrix $R\Sigma R'$, where Σ is as in (4.6)).

Moreover, as said in Remark 3.3, it follows from [11], to be precise by combining [11, (5.30), Theorem 7.10 and Theorem 7.11], that

$$\mathbb{E}\,\mathbf{Y}_n = n\hat{\mu} + o(n). \tag{6.13}$$

By combining (6.12) and (6.13) we see that $R\mu = \hat{\mu}$ (since neither depends on n), and thus (6.12) yields (3.3).

To see that the covariance matrix $R\Sigma R'$ is non-singular when each T^i has an internal node so $k_i > 0$, suppose that, on the contrary, $u'R\Sigma R'u = 0$ for some vector $u \neq 0$. Then, by [14, Theorem 3.6], $u'\mathbf{Y}_n = u'R\mathcal{X}_n$ is deterministic for every n. We argue as for the case $k = 2$ in the proof of [9, Lemma 3.6]. We may assume that every $u_i \neq 0$, since we may just ignore each T^i with $u_i = 0$; we may also assume that $1 \le k_1 \le k_2 \le \dots$. Let N be a large integer, with $N > k_d$, and let T_1 be a tree consisting of a single path with $N + k_1$ internal nodes, each of them (except the root) the right-most child of the preceding one. Let T_2 consist of a similar right-most path with N internal nodes, together with $m - 1$ copies of T_1, which have their roots as the $m - 1$ first children of T_2. Note that both T_1 and T_2 have $(N + k_1)(m - 1)$ keys, so they are possible realizations of $\mathcal{T}_{(N+k_1)(m-1)}$. Moreover, for any tree T^i, $i \ge 2$, T_1 and T_2 have the same number of fringe trees isomorphic to T^i, while T_1 contains $m - 1$ more copies of T^1 than T_2 does. Hence the linear combination $u'\mathbf{Y}_n = \sum_i u_i X_n^{T^i}$ may take at least two different values when $n = (N+k_1)(m-1)$, which is a contradiction. Consequently, the covariance matrix cannot be singular when all $k_i > 0$. □

Proof of Theorem 3.4: This follows from Theorem 3.2; we refer to [12] for details. □

7 Example of Theorem 3.4 when $m = 3$ and $k = 4$

We consider the case when we want to evaluate σ_4^2 in Theorem 3.4 in the case of a random ternary search tree ($m = 3$).

We use the construction of the Pólya urn in Example 5.1, which gives an urn with the following 6 different (living) types:

1. An empty node.
2. A node with one key.
3. A node with two keys and three external children.
4. A tree with a root holding two keys and one child holding one key, plus two external children.
5. A tree with a root holding two keys and two children holding one key each, plus one external child.
6. A tree with a root holding two keys and one child holding two keys, plus two external children of the root and three external childen of the leaf.

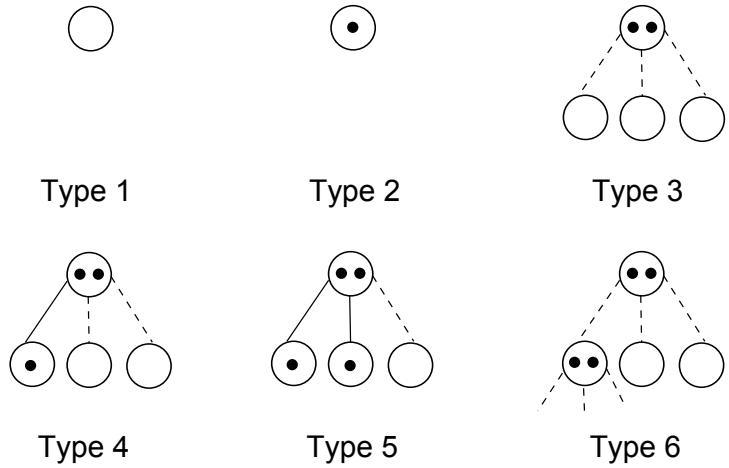


Fig. 1: The different types for counting the number of the fringe subtrees with four keys in a ternary search tree.

See Figure 1 for an illustration of these types.

The activities of the types are $1, 2, 3, 4, 5, 5$. We can easily describe the intensity matrix, first noting that if we draw a type k for $k \le 3$ it is replaced by one of type $k+1$. If we draw a type 4 it is replaced by one of type 5 with probability 1/2 and one of type 6 with probability 1/2. If we draw a type 5 it is replaced by three of type 2 with probability 1/5, and one each of the types 1, 2 and 3 with probability 4/5; see Figure 2 for an illustration. Finally if we draw a type 6 it is replaced by one each of the types 1, 2 and 3 with probability 2/5, and two of type 1 and one of type 4 with probability 3/5.

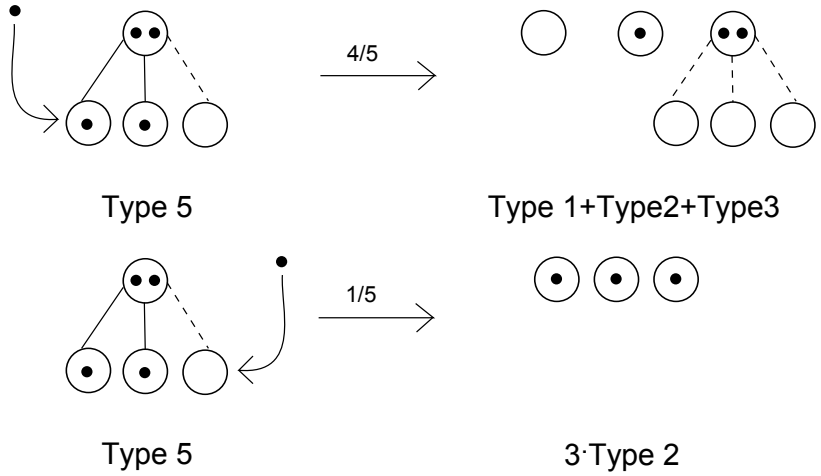


Fig. 2: The two possibilities for adding a key to a node in a tree of type 5 of a ternary search tree.

Thus, we get the intensity matrix A in (4.1) as

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 4 & 8 \\ 1 & -2 & 0 & 0 & 7 & 2 \\ 0 & 2 & -3 & 0 & 4 & 2 \\ 0 & 0 & 3 & -4 & 0 & 3 \\ 0 & 0 & 0 & 2 & -5 & 0 \\ 0 & 0 & 0 & 2 & 0 & -5 \end{pmatrix}. \tag{7.1}$$

The eigenvalues are, by direct calculation or by Theorem 6.2,

$$1, -3, -4, -4, -5, -5. \tag{7.2}$$

(We know already that $\lambda_1 = 1$, as was noted in Section 6 as a consequence of Remark 4.2.)

Furthermore, by Remark 4.2, the left eigenvector u_1 is $u_1 = a = (1, 2, 3, 4, 5, 5)$. The right eigenvector v_1, with the normalization (4.2), is $v_1 = (3/25, 1/10, 2/25, 3/50, 1/50, 1/50)$. Note that $\mu = v_1$ in Theorem 4.1, since $\lambda_1 = 1$. Hence, the asymptotic mean in (3.6) for $X_{n,4}$ is $(\mu_5 + \mu_6)n = \frac{n}{25}$. (However, to get the asymptotic expectation in (3.6) for arbitrary k and m we could instead use branching processes, see [11].)

To calculate the variance σ_4^2, we calculate the covariance matrix Σ in Theorem 4.1 by Theorem 4.1(ii); thus we first calculate B_i, B and Σ_I in in (4.3)–(4.5). Since A is diagonalisable, there is also an alternative way to calculate Σ, see Remark 4.2; see also [12, Theorem 4.1(iii)] and [13, Lemma 5.3].

To calculate B in (4.4) we need to calculate $B_i = \mathbb{E}(\xi_i \xi_i')$ in (4.3). We only describe how to get the matrix B_5 since the other cases are analogous. We get $B_5 = \frac{1}{5} \cdot b_1 b_1' + \frac{4}{5} \cdot b_2 b_2'$, where $b_1 = (0, 3, 0, 0, -1, 0)'$ and $b_2 = (1, 1, 1, 0, -1, 0)'$; see Figure 2. Now we can use Mathematica to evaluate the integral in (4.5), which yields Σ_I. Finally, $\Sigma = \Sigma_I$ by Theorem 4.1 with $c = 1$. The result is given in (7.3).

$$\Sigma = \begin{pmatrix} \frac{29017}{259875} & -\frac{117371}{10395000} & -\frac{44311}{5197500} & -\frac{2143}{945000} & -\frac{28289}{5197500} & -\frac{28289}{5197500} \\ -\frac{117371}{10395000} & \frac{7379}{83160} & -\frac{34927}{5197500} & -\frac{3907}{236250} & -\frac{166037}{20790000} & -\frac{166037}{20790000} \\ -\frac{44311}{5197500} & -\frac{34927}{5197500} & \frac{159241}{2598750} & -\frac{4747}{236250} & -\frac{84709}{10395000} & -\frac{84709}{10395000} \\ -\frac{2143}{945000} & -\frac{3907}{236250} & -\frac{4747}{236250} & \frac{39227}{945000} & -\frac{13309}{1890000} & -\frac{13309}{1890000} \\ -\frac{28289}{5197500} & -\frac{166037}{20790000} & -\frac{84709}{10395000} & -\frac{13309}{1890000} & \frac{22613}{1299375} & -\frac{6749}{2598750} \\ -\frac{28289}{5197500} & -\frac{166037}{20790000} & -\frac{84709}{10395000} & -\frac{13309}{1890000} & -\frac{6749}{2598750} & \frac{22613}{1299375} \end{pmatrix}. \tag{7.3}$$

However to calculate σ_4^2, we only need the submatrix

$$\Delta = \begin{pmatrix} \sigma_{5,5} & \sigma_{5,6} \\ \sigma_{6,5} & \sigma_{6,6} \end{pmatrix} = \begin{pmatrix} \frac{22613}{1299375} & -\frac{6749}{2598750} \\ -\frac{6749}{2598750} & \frac{22613}{1299375} \end{pmatrix}. \tag{7.4}$$

Summing the $\sigma_{i,j}$ in (7.4), which is equivalent to calculating $(1, 1)\Delta(1, 1)'$, we find

$$\sigma_4^2 = \frac{38477}{1299375}.$$

Note that we can use this urn to calculate the asymptotic variance for the total number of leaves in the random ternary search tree, which was evaluated in [10, Theorem 4.1]. We get

$$(0, 1, 1, 1, 2, 1)\Sigma(0, 1, 1, 1, 2, 1)' = \frac{89}{2100}.$$

We could also use this urn to evaluate

$$\sigma_3^2 = (0, 0, 0, 1, 0, 0)\Sigma(0, 0, 0, 1, 0, 0)' = \frac{39227}{945000}, \tag{7.5}$$

$$\sigma_2^2 = (0, 0, 1, 0, 0, 1)\Sigma(0, 0, 1, 0, 0, 1)' = \frac{131}{2100}, \tag{7.6}$$

$$\sigma_1^2 = (0, 1, 0, 1, 2, 0)\Sigma(0, 1, 0, 1, 2, 0)' = \frac{8}{75}. \tag{7.7}$$

References

[1] Aldous D., Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.* **1** (1991), no. 2, 228–266.

[2] Chern, H.-H. and Hwang, H.-K., Phase changes in random m-ary search trees and generalized quicksort. *Random Structures Algorithms* **19**, (2001), no. 3-4, 316–358.

[3] Dennert F. and Grübel R., On the subtree size profile of binary search trees. *Combin. Probab. Comput.* **19** (2010), no. 4, 561–578.

[4] Devroye L., Limit laws for local counters in random binary search trees. *Random Structures Algorithms* **2** (1991), no. 3, 303–315.

[5] Devroye L., Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.* **32** (2002/03), no. 1, 152–171.

[6] Drmota M., *Random Trees*. Springer, Vienna, 2009.

[7] Fill J.A. and Kapur N., Transfer theorems and asymptotic distributional results for m-ary search trees. *Random Structures Algorithms* **26** (2005), no. 4, 359–391.

[8] Fuchs M., Subtree sizes in recursive trees and binary search trees: Berry–Esseen bounds and Poisson approximations. *Combin. Probab. Comput.* **17** (2008), no. 5, 661–680.

[9] Holmgren C. and Janson S., Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electron. J. Probab.* **20** (2015), no. 4, 1–51.

[10] Holmgren C. and Janson S., Asymptotic distribution of two-protected nodes in ternary search trees. *Electron. J. Probab.* **20** (2015), no. 9, 1–20.

[11] Holmgren, C. and Janson S., Fringe trees, Crump–Mode–Jagers branching processes and m-ary search trees. Preprint, 2016. `arXiv:1601.03691`.

[12] Holmgren, C., Janson S and M. Šileikis., Multivariate normal limit laws for the numbers of fringe subtrees in m-ary search trees and preferential attachment trees. Preprint, 2016. `arXiv:1603.08125`

[13] Janson S., Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stoch. Process. Appl.* **110** (2004), 177–245.

[14] Janson S., Mean and variance of balanced Pólya urns. Preprint, 2016. `arXiv:1602.06203`

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

[16] Mahmoud H.M., The size of random bucket trees via urn models. *Acta Inform.* **38** (2002), no. 11-12, 813–838.

[17] Mahmoud H.M., *Pólya Urn Models*. CRC Press, Boca Raton, FL, 2009.

[18] Mahmoud, H.M. and Pittel, B., Analysis of the space of search trees under the random insertion algorithm. *J. Algorithms* **10**, (1989), no. 1, 52–75.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Average Size of a Suffix Tree for Markov Sources

Philippe Jacquet[1], Wojciech Szpankowski[2†]

[1]*Bell Labs, Alcatel-Lucent, France.*
[2]*Department of Computer Science, Purdue University, USA*

We study a suffix tree built from a sequence generated by a Markovian source. Such sources are more realistic probabilistic models for text generation, data compression, molecular applications, and so forth. We prove that the average size of such a suffix tree is asymptotically equivalent to the average size of a trie built over n independent sequences from the same Markovian source. This equivalence is only known for memoryless sources. We then derive a formula for the size of a trie under Markovian model to complete the analysis for suffix trees. We accomplish our goal by applying some novel techniques of analytic combinatorics on words also known as analytic pattern matching.

Keywords: Suffix tree, Markov sources, digital trees, size, pattern matching, number of occurrences.

1 Introduction

Suffix trees are the most popular data structures on words. They find myriad of applications in computer science and telecommunications, most notably in algorithms on strings, data compressions (Lempel-Ziv'77 scheme), and codes. Despite this, little is still known about their typical behaviors for general probabilistic models (see [5, 1, 3]).

A suffix tree is a *trie* (a digital tree; see [9]) built from the suffixes of a single string. In Figure 1 we show the suffix tree constructed for the first four suffixes of the string $X = 0101101110$. More precisely, we actually build a suffix tree on the first n *infinite* suffixes of a string X as shown in Figure 1. We shall call it simply a suffix tree which we study in this paper. Such a tree consists of internal (branching) nodes and external node storing the suffixes. Our goal is to analyze the number of internal nodes called also the *size* of a suffix tree built from a sequence X generated by a Markov source. We accomplish it by employing powerful techniques of analytic combinatorics on words known also as *analytic pattern matching* [9].

In recent years there has been a resurgence of interest in algorithmic and combinatorial problems on words due to a number of novel applications in computer science, telecommunications, and most notably in molecular biology. A few possible applications are listed below. The reader is referred to our recent

[†]W. Szpankowski is also with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland. His work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01, and the NCN grant, grant UMO-2013/09/B/ST6/02258.

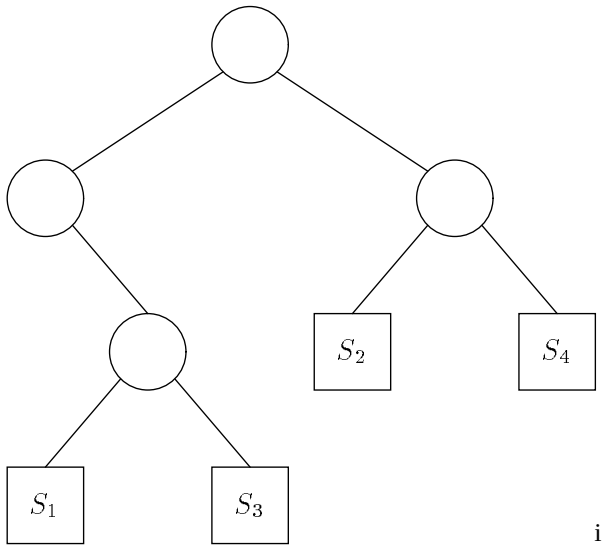


Fig. 1: Suffix tree built from the first five suffixes of $X = 0101101110$, i.e. 0101101110, 101101110, 01101110, 1101110.

book [9] for more details. In computer science and molecular biology many algorithms depend on a solution to the following problem: given a word X and a set of arbitrary $b+1$ suffixes S_1, ... , S_{b+1} of X, what is the longest common prefix of these suffixes. In coding theory (e.g., prefix codes) one asks for the shortest prefix of a suffix S_i which is not a prefix of any other suffixes S_j, $1 \leq j \leq n$ of a given sequence X (cf. [14]). In data compression schemes, the following problem is of prime interest: for a given "data base" sequence of length n, find the longest prefix of the $(n+1)$st suffix S_{n+1} which is not a prefix of any other suffixes S_i ($1 \leq i \leq n$) of the data base sequence. And last but not least, in molecular sequences comparison (e.g., finding homology between DNA sequences), one may search for the longest run of a given motif, a unique sequence, the longest alignment, and the number of common subwords [9]. These, and several other problems on words, can be efficiently solved and analyzed by a clever manipulation of a data structure known as a *suffix tree*. In literature other names have been also coined for this structure, and among these we mention here position trees, subword trees, directed acyclic graphs, *etc*.

The extension of suffix tree analysis to Markov sources is quite significant, especially when the suffix tree is used for natural languages. Indeed, Markov sources of finite memory approximate very well realistic texts. For example, the following quote is generated by a memoryless source with the letter statistic of the *Declaration of Independence*:

> esdehTe,a; psseCed vcenseusirh vra f uetaiapgnuev n cosb mgffgfL itbahhr nijue n S ueef,ru
> s,k smodpztrnno.eeteespfg mtet tr i aur oiyr

which should be compared to the following quote generated by a Markov source of order 3 trained on the same text:

> We hat Government of Governments long that their right of abuses are these rights, it, and or themselves and are disposed according Men, der.

In this paper we analyze the average number of internal nodes (size) of a suffix tree built from n (infinite) suffixes of a string generated by a Markov source with positive transition probabilities. We first prove in Theorem 1 that the average size of a suffix tree under Markovian model is asymptotically equivalent to the size of a *trie* that is built from n *independently* generated strings, each string emitted by the corresponding Markovian source. To accomplish this, we study another quantity, namely the number of occurrences of a given pattern w in a string of length n generated by a Markovian source. We use its properties to establish our asymptotic equivalence between suffix trees and tries. Finally, we compare the average size of suffix trees to trie size under Markovian model (see Theorem 2), which – to the best of our knowledge – is only partially known [2].

In fact, there is extensive literature on tries [9] and very scarce one on suffix trees. An analysis of the depth in a Markovian trie has been presented earlier in [12]. A rigorous analysis of the depth of suffix tree was first presented in [5] for memoryless sources, and then extended in [3] to Markov sources. We should point out that depth grows like $O(\log n)$ which makes the analysis manageable. In fact, height and fillup level for suffix tree – which are also of logarithmic growth – were analyzed in [15] (see also [1, 14]). But the average size grows like $O(n)$ and is harder to study. For memoryless sources it was analyzed in [11] for tries and in [5] for suffix trees. We also know that some parameters of suffix trees (e.g., profile) cannot be inferred from tries, see [4]. Markov sources add additional level of complications in the analysis of suffix trees as well documented in [1]. In fact, the average size of tries under general dynamic sources was analyzed in [2], however, specifications to Markov sources requires extra care, especially for the so called rational Markov sources.

The proof of the convergence of the average size of the suffix tree to the average size of the trie borrows many fundamental elements of the depth analysis in [3], for example the term $q_n(w)$ (see next section), but the extension of the depth analysis to the size analysis require the introduction of a new term $d_n(w)$ which has non trivial properties. The analysis of average size of the trie in a Markovian model has been made by several author before but surprisingly we could not find a clear statement about the periodic case. This is the reason why we have to present a sketched proof here.

2 Main Results

We consider a stationary source generating a sequence of symbols drawn from a finite alphabet $\mathcal{A}$.

We first derive a formula for the average size of a suffix tree in terms of the number of pattern occurrences. Let w be a word over $\mathcal{A}$. We denote by $O_n(w)$ the number of occurrences of word w in a sequence of length n generated by a Markov source with the transition matrix $\mathbf{P}$. We observe [5] that the average size s_n of a suffix tree built over a sequence of length n is

$$s_n = \sum_{w \in \mathcal{A}^*} P(O_n(w) \geq 2). \tag{1}$$

In fact, (1) holds for any probabilistic source. We compare it to the average size t_n of trie built over n independent Markov sequences. If $N_n(w)$ is the number of words which begin with w in a trie build with n words, we have

$$t_n = \sum_{w \in \mathcal{A}^*} P(N_n(w) \geq 2). \tag{2}$$

Let $P(w)$ be the probability of observing w in a Markov sequence, $N_n(w)$ is a Bernoulli $(n, P(w))$ and

random variable t_n can be written as

$$t_n = \sum_{w \in \mathcal{A}^*} 1 - (1 - P(w))^n - nP(w)(1 - P(w))^{n-1} \tag{3}$$

We specifically consider a Markovian source. We assume that the source is stationary and ergodic. We will consider a Markovian process of order 1 with a positive transition matrix $\mathbf{P} = [P(a|b)]_{a,b \in \mathcal{A}}$. Extensions to higher order Markov is possible since a Markovian source of order r is simply a Markovian source of order 1 over the alphabet $\mathcal{A}^r$. Notice that contrary to previous analysis we don't assume that $P(a|b) > 0$ for all $(a, b) \in \mathcal{A}^2$, since we allow that some transition may be forbiden and some other mandatory (while keeping the source ergodic).

Our main result of the paper is formulated next,

Theorem 1 *Consider a suffix tree built over n suffixes of a sequence of length n generated by a Markov source with a positive state transition matrix $\mathbf{P}$. There exists $\varepsilon > 0$ such that*

$$s_n - t_n = O(n^{1-\varepsilon}) \tag{4}$$

for large n.

In order to apply Theorem 1 one needs to estimate the average size of a trie under Markovian model. This seems to be unknown except for some general dynamic sources [2]. In fact, analysis of tries under Markovian sources is quite challenging (see [6]). But we can offer the following result for the average size of a trie under Markovian assumptions. A sketch of the proof is presented in Section 4.

Theorem 2 *Consider a trie built over n independent sequences generated by a Markov source with positive transition probabilities. For $(a, b, c) \in \mathcal{A}^3$ define*

$$\alpha_{abc} = \log \left[\frac{P(a|b)P(c|a)}{P(c|b)} \right]. \tag{5}$$

Then:
(aperiodic case) *If not all $\{\alpha_{abc}\}$ are commensurable, then*

$$t_n = \frac{n}{h} + o(n)$$

where $h = -\sum_{a,b} \pi_a P(b|a) \log P(b|a)$ is the entropy rate of the underlying Markov source with π_a, $a \in \mathcal{A}$, denoting the stationary probability.
(periodic case) *If all $\{\alpha_{abc}\}$ are commensurable, then*

$$t_n = \frac{n}{h}(1 + Q(n)) + O(n^{1-\varepsilon})$$

where $Q(n)$ is a periodic function and some $\varepsilon > 0$.

Remark We recall that a set of real numbers are commensurable (also known as "rationally related") when their ratios are rational numbers. We observe that if for all $(a,b) \in \mathcal{A}^2$, the α_{abc} are commensurable for one $c \in \mathcal{A}$, then α_{abc} are commensurable for all values of c.Furthermore in the aperiodic case the $o(n)$ term can have a growth rate arbitrary close to order n, depending on source settings as shown in [7] in the memoryless case.

In the rest of this section, we present a road map of the proof of (4). For this we will make use of ordinary generating functions. Let $w \in \mathcal{A}^k$ be a word of length k. We also define $N_0(z,w) = \sum_{n>0} P(O_n(w)=0)z^n$ and $N_1(z,w) = \sum_{n>0} P(O_n(w)=1)z^n$ for $z \in \mathbb{C}$. We know from [9] that

$$\begin{aligned} N_0(z,w) &= \frac{S_w(z)}{D_w(z)} \\ N_1(z,w) &= \frac{z^k P(w)}{D_w^2(z)} \end{aligned}$$

where $S_w(z)$ is the autocorrelation polynomial of word w and $D_w(z)$ is defined as follows

$$D_w(z) = S_w(z)(1-z) + z^k P(w)\left(1 + F_w(z)(1-z)\right), \tag{6}$$

The memoryless case considers $F_w(z) = 0$. The addition of a non zero $F_w(z)$ is a significant change from the analysis in the memoryless case. In fact it captures the correlations between characters in the sequence and leads to non trivial developments. Here $F_w(z)$ for $w \in \mathcal{A}^* - \{\varepsilon\}$ is a function that depends on the Markov parameters of the source. It also depends only on the first and last character of w, say respectively a and b for $(a,b) \in \mathcal{A}^2$ as described below.

Let $\mathbf{P}$ be the transition matrix of the Markov source and $\boldsymbol{\pi}$ be its stationary vector with π_a its coefficient at symbol $a \in \mathcal{A}$. The vector $\mathbf{1}$ is the vector with all coefficients equal to 1 and $\mathbf{I}$ is the identity matrix. Assuming that $a \in \mathcal{A}$ (resp. b) is the first (resp. last) symbol of w, we have [13, 9]

$$F_w(z) = \frac{1}{\pi_a}\left[\left(\mathbf{P} - \boldsymbol{\pi}\otimes\mathbf{1}\right)\left(\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi}\otimes\mathbf{1})\right)^{-1}\right]_{b,a} \tag{7}$$

where $[\mathbf{A}]_{a,b}$ indicates the (a,b) coefficient of the matrix $\mathbf{A}$, and $\otimes$ represents the tensor product. An alternative way to express $F_w(z)$ is

$$F_w(z) = \frac{1}{\pi_a}\langle \mathbf{e}_a\left(\mathbf{P} - \boldsymbol{\pi}\otimes\mathbf{1}\right)\left(\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi}\otimes\mathbf{1})\right)^{-1}\mathbf{e}_b\rangle \tag{8}$$

where $\mathbf{e}_c$ for $c \in \mathcal{A}$ is the vector with a 1 at the position corresponding to symbol c and all other coefficients are 0. Here $\langle \mathbf{x}, \mathbf{y}\rangle$ represents the scalar product of $\mathbf{x}$ and $\mathbf{y}$.

Let us define two important quantities:

$$\begin{aligned} d_n(w) &= P(O_n(w)=0) - (1-P(w))^n, \\ q_n(w) &= P(O_n(w)=1) - nP(w)(1-P(w))^{n-1}, \end{aligned}$$

and their corresponding generating functions

$$\begin{aligned} \Delta_w(z) &= \sum_{n>0} d_n(w) z^n \\ Q_w(z) &= \sum_{n>0} q_n(w) z^n. \end{aligned}$$

Observe that $t_n - s_n = \sum_{w\in\mathcal{A}^*} d_n(w) + q_n(w)$. Thus we need to estimate $d_n(w)$ and $q_n(w)$ for all $w \in \mathcal{A}^*$.

We denote $\mathcal{B}_k$ the set of words of length k that do not overlap with themselves over more than $k/2$ symbols (see [9, 5, 3] for more precise definition). To be precise $w \in \mathcal{A}^k - \mathcal{B}_k$ if there exist $j > k/2$ and $v \in \mathcal{A}^j$ and $(u_1, u_2) \in \mathcal{A}^{k-j}$ such that $w = u_1 v = v u_2$. This set plays a fundamental role in the analysis and it is already proven in [3] that

$$\sum_{w\in\mathcal{A}^k-\mathcal{B}_k} P(w) = O(\delta_1^k)$$

where δ_1 is the largest coefficient in the Markovian transition matrix $\mathbf{P}$. Since the authors of [3] only consider strictly positive matrix $\mathbf{P}$ we have $\delta_1 < 1$. Anyhow in the present paper we allow some coefficients to be equal to 1 or 0, as long the source is ergodic. Therefore δ_1 may be equal to 1. To cope with this minor problem we define

$$\begin{aligned} p &= \exp\left(\lim \sup_{k,w\in\mathcal{A}^k} \frac{\log P(w)}{k}\right) \\ q &= \exp\left(\lim \inf_{k,w\in\mathcal{A}^k, P(w)\neq 0} \frac{\log P(w)}{k}\right). \end{aligned}$$

These quantities exist and are smaller than 1 since $\mathcal{A}$ is a finite alphabet. From now we set $\delta = \sqrt{p}$ which replaces the parameter δ_1 in the previous statements.

Now we are in the position to present two crucial lemmas, proved in the next section, from which Theorem 1 follows.

Lemma 1 *There exist $\varepsilon < 1$ such that $\sum_{w\in\mathcal{A}^*} q_n(w) = O(n^\varepsilon)$.*

Lemma 2 *There exists a sequence $R_n(w)$, for $w \in \mathcal{A}^*$ such for all $1 > \varepsilon > 0$ we have*

- *(i) for $w \in \mathcal{B}_k$: $d_n(w) = O((nP(w))^\varepsilon k\delta^k) + R_n(w)$;*
- *(ii) for $w \in \mathcal{A}^k - \mathcal{B}_k$: $d_n(w) = O((nP(w))^\varepsilon) + R_n(w)$,*

where $R_n(w)$ is such that $\sum_{w\in\mathcal{A}^*} R_n(w) = O(1)$.

Remark: The sequence $d_n(w)$ is the main new element which makes the difference between the suffix tree depth analysis done in [3] and the suffix tree size analysis. The later was done in [9] for the memoryless case. The sequence $R_n(w)$ reflects the impact of the Markovian source on the analysis in particular is a consequence of the introduction of a non zero function $F_w(z)$.

Proof of Theorem 1: We already know via Lemma 1 that there exists $\varepsilon < 1$ such that $\sum_{w\in\mathcal{A}^*} q_n(w) = O(n^\varepsilon)$. Let now $d_n^{(1)} = \sum_k \sum_{w\in\mathcal{B}_k} (d_n(w) - R_n(w))$ and since for all $\varepsilon > 0$ observe that

$$d_n^{(1)} = \sum_k \sum_{w\in\mathcal{B}_k} O(n^\varepsilon P^\varepsilon(w) k\delta^k) = \sum_k O(n^\varepsilon k(p^\varepsilon\delta)^k),$$

hence it converges for all $\varepsilon > 0$. Also let $d_n^{(2)} = \sum_k \sum_{w \in \mathcal{A}^k - \mathcal{B}_k} (d_n(w) - R_n(w))$. Observe that

$$\begin{aligned} d_n^{(2)} &= \sum_k \sum_{w \in \mathcal{A}^k - \mathcal{B}_k} O(n^\varepsilon P^{\varepsilon-1}(w) P(w)) \\ &= \sum_k \sum_{w \in \mathcal{A}^k - \mathcal{B}_k} O(n^\varepsilon q^{(\varepsilon-1)k} P(w)) \\ &= \sum_k O(n^\varepsilon (\delta q^{\varepsilon-1})^k), \end{aligned}$$

which converges for all ε such that $\delta q^{\varepsilon-1} < 1$ (take $\varepsilon < 1$ close enough to 1) and is $O(n^\varepsilon)$. Finally $d_n^{(1)} + d_n^{(2)} + \sum_{w \in \mathcal{A}^*} R_n(w)$ is also $O(n^\varepsilon)$ for $\varepsilon > 0$ since $\sum_{w \in \mathcal{A}^*} R_n(w)$ is finitely bounded. This completes the proof of Theorem 1. □

3 Proof of Lemmas

In this section we prove Lemma 1 and Lemma 2. In the proof of Lemma 1 we shall use some facts from [3], however, our proof follows the pattern matching approach developed in [9].

3.1 Proof of Lemma 1

The result is in fact already proven in [3]. Define

$$Q_w(z) = P(w) \left(\frac{z^k}{D_w^2(z)} - \frac{z}{(1-(1-P(w))z)^2} \right). \tag{9}$$

In [3] one defines $Q_n(1) = \frac{1}{n} \sum_{w \in \mathcal{A}^*} q_n(w)$ and it is proven there that $Q_n(1) = O(n^{-\varepsilon})$ for some $\varepsilon > 0$.

3.2 Proof of Lemma 2

First we have the following simple lemma. The largest eigenvalue of $\mathbf{P}$ is 1, let $\lambda_1, \lambda_2, \ldots$ be a sequence of other eigenvalues in the decreasing order of their modulus.

Lemma 3 *Uniformly for all $w \in \mathcal{A}^*$ we find $F_w(z) = O(\frac{1}{1-|\lambda_1 z|})$.*

Proof: By the spectral representation of $\mathbf{P}$ we know that $\mathbf{P} = \boldsymbol{\pi} \otimes \mathbf{1} + \sum_{i>0} \lambda_i \mathbf{u}_i \otimes \boldsymbol{\zeta}_i$ where $\mathbf{u}_i$ (resp. $\boldsymbol{\zeta}_i$) are the corresponding right (resp. left) eigenvectors. In fact we can introduce the matrices $\mathbf{D} = \boldsymbol{\pi} \otimes \mathbf{1}$ and $\mathbf{R} = \sum_{i>0} \lambda_i \mathbf{u}_i \otimes \boldsymbol{\zeta}_i$ whose spectral radius is $|\lambda_1|$ and satisfies the orthogonal property: $\mathbf{RD} = \mathbf{DR} = 0$.

We have Let $\mathbf{M}(z) = \mathbf{P} - \boldsymbol{\pi} \otimes \mathbf{1}) \, (\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi} \otimes \mathbf{1}))^{-1}$ we have $\mathbf{M}(z) = \mathbf{R}(\mathbf{1} - z\mathbf{R})^{-1}$. Since $\mathbf{R}^k = O(|\lambda_1|z) \, \mathbf{R}(\mathbf{I} - z\mathbf{R})^{-1}$ is defined for all z such that $|z| < \frac{1}{|\lambda_1|}$ and is $O(\frac{1}{1-|\lambda_1 z|})$, and so is $F_w(z) = [\mathbf{M}(z)]_{a,b}$.

□

The next lemma is important.

Lemma 4 *For z such that $|\lambda_1 z| < 1$ we have for all integers k*

$$\sum_{w\in\mathcal{A}^{k+1}} P(w)F_w(z) = O(\lambda_1^k). \tag{10}$$

Proof: The function $F_w(z)$ depends only on the first and last symbol of w. Considering a pair of symbols $(a,b) \in \mathcal{A}^2$ the sum of the probabilities of the words of length $k+1$ starting with a and ending with b, $\sum_{awb\in\mathcal{A}^{k+1}} P(w)$, equals $\pi_a\langle \mathbf{e}_b\mathbf{P}^k\mathbf{e}_a\rangle$. Easy algebra leads to

$$\begin{aligned}\sum_{w\in\mathcal{A}^{k+1}} P(w)F_w(z) &= \sum_{(a,b)\in\mathcal{A}^2} \langle \mathbf{e}_a\mathbf{M}(z)\mathbf{e}_b\rangle\langle \mathbf{e}_b\mathbf{P}^k\mathbf{e}_a\rangle &(11)\\ &= \text{trace}\left(\mathbf{M}(z)\mathbf{P}^k\right). &(12)\end{aligned}$$

But since $\mathbf{P}^k = \mathbf{D} + \mathbf{R}^k$ and $\mathbf{M}(z)\mathbf{D} = 0$ and $\mathbf{R}^k = O(|\lambda_1|^k)$, we conclude the proof

□

We now follow a parallel approach to the approach developped in [3] and in [5, 9].
The generating function $\Delta_w(z) = \sum_{n\geq 0} d_n(w)z^n$ becomes

$$\Delta_w(z) = \frac{P(w)z}{1-z}\left(\frac{1+(1-z)F_w(z)}{D_w(z)} - \frac{1}{1-z+P(w)z}\right). \tag{13}$$

We have

$$d_n(w) = \frac{1}{2i\pi}\oint \Delta_w(z)\frac{dz}{z^{n+1}},$$

integrated on any loop encircling the origin in the definition domain of $d_w(z)$. Extending the result in [5], the authors of [3] show that there exists $\rho > 1$ such that the function $D_w(z)$ has a single root in the disk of radius ρ. Let A_w be such a root. We have via the residue formula

$$d_n(w) = \text{Res}(\Delta_w(z), A_w)A_w^{-n} - (1-P(w))^n + d_n(w,\rho), \tag{14}$$

where $\text{Res}(f(z), A)$ denotes the residue of function $f(z)$ on complex number A and

$$d_n(w,\rho) = \frac{1}{2i\pi}\oint_{|z|=\rho} \Delta_w(z)\frac{dz}{z^{n+1}}. \tag{15}$$

We have

$$\text{Res}(\Delta_w(z), A_w) = \frac{P(w)\left(1+(1-A_w)F_w(A_w)\right)}{(1-A_w)C_w} \tag{16}$$

where $C_w = D'_w(A_w)$. But since $D_w(A_w) = 0$ we can write

$$\text{Res}(\Delta_w(z), A_w) = -\frac{A_w^{-k}S_w(A_w)}{C_w} \tag{17}$$

We now consider asymptotic expansion of A_w and C_w as it is described in [9], in Lemma 8.1.8 and Theorem 8.2.2. Although the expansions were presented for memoryless case, but for Markov source we simply replace $S_w(1)$ by $S_w(1) + P(w)F_w(1)$. We find

$$\begin{array}{rcl} A_w & = & 1 + \frac{P(w)}{S_w(1)} \\ & & +P(w)^2 \left(\frac{k - F_w(1)}{S_w^2(1)} - \frac{S'_w(1)}{S_w^3(1)} \right) + O(P(w)^3) \\ C_w & = & -S_w(1) + P(w) \left(k - F_w(1) - 2\frac{S'_w(1)}{S_w(1)} \right) \\ & & +O(P(w)^2) \end{array} \tag{18}$$

Notice that these expansions in the Markov model first appeared in [3].

From now follow the proof of Theorem 8.2.2 in [9]. We define the function

$$\delta_w(x) = \frac{A_w^{-k} S_w(A_w)}{C_w} A_w^{-x} - (1 - P(w))^x. \tag{19}$$

More precisely we define the function

$$\bar{\delta}_w(x) = \delta_w(x) - \delta_w(0)e^{-x}$$

which has a Mellin transform $\delta_w^*(s)\Gamma(s) = \int_0^\infty \bar{\delta}_w(x) x^{s-1} dx$ defined for all $\Re(s) \in (-1, 0)$ with

$$\delta_w^*(s) = \frac{A_w^{-k} S_w(A_w}{C_w} \left[(\log A_w)^{-s} - 1 \right] + 1 - \left[-\log(1 - P(w)) \right]^{-s}. \tag{20}$$

When $w \in \mathcal{B}_k$ with the expansion of A_w and since $S_w(1) = 1 + O(\delta^k)$ and $S'_w(1) = O(k\delta^k)$, we find that similarly as shown in [9]

$$\delta_w^*(s) = O(|s| k \delta^k) P(w)^{1-s}. \tag{21}$$

Therefore, by the reverse Mellin transform, for all $1 > \varepsilon > 0$:

$$\begin{array}{rcl} \bar{\delta}(n, w) & = & \frac{1}{2i\pi} \int_{-\varepsilon - i\infty}^{-\varepsilon + i\infty} \delta_w^*(s) \Gamma(s) n^{-s} ds \\ & = & O(n^{1-\varepsilon} P(w)^{1-\varepsilon} k \delta^k) \end{array} \tag{22}$$

When $w \in \mathcal{A}^k - \mathcal{B}_k$ we don't have the $S_w(1) = 1 + O(\delta^k)$. But it is shown in [3] that there exists $\alpha > 0$ such that for all $w \in \mathcal{A}^*$: $S_w(z) > \alpha$ for all z such that $|z| \leq \rho$. Therefore we get

$$\bar{\delta}(n, w) = O(n^{1-\varepsilon} P(w)^{1-\varepsilon}).$$

We set

$$R_n(w) = d_w(0)e^{-n} + d_n(w, \rho). \tag{23}$$

We first investigate the quantity $d_w(0)$. We need to prove that $\sum_{w \in \mathcal{A}^*} d_w(0)$ converges. For this, noticing that

$$S_w(A_w) = S_w(1) + \frac{P(w)}{S_w(1)} S'_w(1) + O(P(w)^2)$$

we obtain

$$-\frac{A_w^{-k} S_w(A_w)}{C_w} = 1 - \frac{P(w)}{S_w(1)} \left(F_w(1) + \frac{S'_w(1)}{S_w(1)} \right) + O(P(w)^2). \tag{24}$$

Thus

$$d_w(0) = -\frac{P(w)}{S_w(1)} \left(F_w(1) + \frac{S'_w(1)}{S_w(1)} \right) + O(P(w))^2). \tag{25}$$

Without the term $F_w(1)$ we would have the same expression as in [9] whose sum over $w \in \mathcal{A}^*$ converges. Therefore we need to prove that the sum $\sum_{w\in\mathcal{A}^*} \frac{P(w)}{S_w(1)} F_w(1)$ converges. It is clear that the sum

$$\sum_k \sum_{w\in\mathcal{A}^k - \mathcal{B}_k} \frac{P(w)}{S_w(1)} F_w(1)$$

converges since

$$\sum_{w\in\mathcal{A}^k - \mathcal{B}_k} P(w) = O(\delta^k)$$

and $F_w(1)$ is uniformly bounded. Now we consider the other part

$$\sum_k \sum_{w\in\mathcal{B}_k} \frac{P(w)}{S_w(1)} F_w(1).$$

We know that $S_w(1) = 1 + O(\delta^k)$, therefore

$$\sum_{w\in\mathcal{B}_k} \frac{P(w)}{S_w(1)} F_w(1) = \sum_{w\in\mathcal{B}^k} P(w) F_w(1) + O(\delta^k). \tag{26}$$

But

$$\sum_{w\in\mathcal{B}^k} P(w) F_w(1) = \sum_{w\in\mathcal{A}^k} P(w) F_w(1) + O(\delta^k),$$

and we know by Lemma 4 that $\sum_{w\in\mathcal{A}^k} P(w) F_w(1) = O(\lambda_1^k)$. Thus the sum $\sum_k \sum_{w\in\mathcal{A}^k} \frac{P(w)}{S_w(1)} F_w(1)$ converges.

The second and last effort concentrates on the term $d_n(w,\rho)$. We proceed as in the proof of Theorem 8.2.2 in [9]. We first have $d_n(w,\rho) = O(P(w)\rho^{-n})$ which is $O(n^\varepsilon P(w)^\varepsilon)$ without any condition on w. The issue is now to work on $w \in \mathcal{B}_k$. In this case we have $S_w(z) = 1 + O(\delta^k)$ and therefore

$$\begin{aligned} d_n(w,\rho) &= \frac{1}{2i\pi} \oint \frac{P(w)}{1-z} \left(\frac{1}{D_w(z)} - \frac{1}{1-z+zP(w)} \right) \frac{dz}{z^{n+1}} \\ &\quad + \frac{1}{2i\pi} \oint P(w) \frac{F_w(z)}{D_w(z)} \frac{dz}{z^{n+1}}. \end{aligned} \tag{27}$$

We notice that the function

$$\frac{P(w)}{1-z} \left(\frac{1}{D_w(z)} - \frac{1}{1-z+zP(w)} \right)$$

is $O(P(w)\delta^k) + O(P(w)^2)$, therefore the first integral is $O(P(w)\delta^k \rho^{-n})$. The second function $P(w)\frac{F_w(z)}{D_w(z)}$ is equal to $P(w)F_w(z) + O(P(w)\delta^k)$. We already know that $\sum_{w\in\mathcal{B}_k} P(w)F_w(z) = O(\lambda_1^k)$, thus the series converges and the lemma is proven.

4 Sketch of the Proof of Theorem 2

Let $a \in \mathcal{A}$. We denote by $t_{a,n}$ the average size of a trie over n independent Markovian sequences, all starting with the same symbol a. Then for $n \geq 2$

$$t_n = 1 + \sum_{a\in\mathcal{A}} \sum_{k=0}^{n} \binom{n}{k} \pi_a^k (1-\pi_a)^{n-k} t_{a,k}, \tag{28}$$

and similarly for $b \in \mathcal{A}$

$$t_{n,b} = 1 + \sum_{a\in\mathcal{A}} \sum_{k=0}^{n} \binom{n}{k} P(a|b)^k (1-P(a|b))^{n-k} t_{a,k}, \tag{29}$$

where we recall $P(a|b)$ is the element of matrix $\mathbf{P}$. Let $T(z) = \sum_n t_n \frac{z^n}{n!} e^{-z}$ and $T_a(z) = \sum_n t_{a,n} \frac{z^n}{n!} e^{-z}$ be the familiar Poisson transforms. Using (28) and (29) we find

$$T(z) = 1 - (1+z)e^{-z} + \sum_{a\in\mathcal{A}} T_a(\pi_a z), \tag{30}$$

$$T_b(z) = 1 - (1+z)e^{-z} + \sum_{a\in\mathcal{A}} T_a(P(a|b) z). \tag{31}$$

Using dePoissonization arguments (see [8]) we shall obtain $t_n = T(n) + O(\frac{1}{n} T(n))$. Thus we need to study $T(z)$ for large z in a cone around the real axis. For this we apply the Mellin transform that we describe next. In fact the convergence between the quantities t_n and T_n could also be derived by the application of the Rice method on the Mellin transform, since the later as an explicit form.

Let now $\mathbf{T}(z)$ be the vector consisting of $T_a(z)$ for every $a \in \mathcal{A}$. It is not hard to see that its Mellin transform

$$\mathbf{T}^*(s) = \int_0^\infty \mathbf{T}(z) z^{s-1} dz$$

is defined for $-1 > \Re(s) > -2$ (since $\mathbf{T}(z) = O(z^2)$ when $z \to 0$), and

$$\mathbf{T}^*(s) = -(1+s)\Gamma(s)\mathbf{1} + \mathbf{P}(s)\mathbf{T}^*(s) \tag{32}$$

where $\mathbf{P}(s)$ is the matrix consisting of $P(a|b)^{-s}$ if $P(a|b) > 0$ and 0 otherwise. This identity leads to

$$\mathbf{T}^*(s) = -(1+s)\Gamma(s)(\mathbf{I} - \mathbf{P}(s))^{-1}\mathbf{1}$$

where $\mathbf{I}$ is the identity matrix. Similarly the Mellin transform $T^*(s)$ of $T(z)$ satisfies

$$T^*(s) = -(1+s)\Gamma(s) + \langle \boldsymbol{\pi}(s), \mathbf{T}^*(s) \rangle. \tag{33}$$

where $\boldsymbol{\pi}(s)$ is the vector composed of π_a^{-s}.

The inverse Mellin transform of $T^*(s)$ is defined as

$$T(n) = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} T^*(s) n^{-s} ds, \quad -1 > c > -2. \tag{34}$$

In order to find asymptotic behavior of $T(z)$ as $z \to \infty$ we need to study the poles of $T^*(s)$ for $-2 < \Re(s)$. As discussed in [6, 9] this is equivalent to analyzing the poles of $\mathbf{T}^*(s)$. Since $(1+s)\Gamma(s)$ has no pole on $-2 < \Re(s) < 0$ we must consider poles of $(\mathbf{I} - \mathbf{P}(s))^{-1}$. In other words (see [6, 9]) we need to find s for which the eigenvalue of largest modulus $\lambda(s)$ of $\mathbf{P}(s)$ is equal to 1. It is easy to see that $\lambda(-1) = 1$ since $\mathbf{P}(-1) = \mathbf{P}$. The residue at $s = -1$ of $n^{-s}(\mathbf{I} - \mathbf{P}(s))^{-1}\mathbf{1}$ is equal to $\frac{n}{h}\mathbf{1}$ where h is the entropy rate of the Markovian source.

As explained in [6] in the periodic case there are multiple values of s such that $\lambda(s) = 1$ and $\Re(s) = -1$. Since these poles are regularly spaced on the axis $\Re(s) = 0$, they contribute to the oscillating terms (function Q in Theorem 2) in the asymptotic expansion of t_n. Furthermore, the location of zeros of $\lambda(s) = 1$ in the periodic case tells us that there exists ε such that $(\mathbf{I} - \mathbf{P}(s))$ has no pole for $-1 < \Re(s) < -1+\varepsilon$ leading to the error term $O(n^{1-\varepsilon})$.

In the aperiodic case there is only one pole on the line $\Re(s) = -1$, thus the oscillating term disappears. However, zeros of $\lambda(s) = 1$ can lie arbitrarily close to the line $\Re(s) = 1$, therefore the error term is just $o(n)$.

References

[1] P. Cénac, B. Chauvin, F. Paccaut, and N. Pouyanne, Uncommon suffix tries, *Random Structures & Algorithms*, vol. 46, 117-141, 2015

[2] J. Clement, P. Flajolet, and B. Vallée, Dynamic Sources in Information Theory: A General Analysis of Trie Structures, *Algorithmica*, 29, 307-369, 2001.

[3] Fayolle, J., Ward, M. D. Analysis of the average depth in a suffix tree under a Markov model. In *International Conference on Analysis of Algorithms DMTCS*, proc. AD (Vol. 95, p. 104), 2005.

[4] J. Geithner and M. Ward, Variance of the Profile in Suffix Trees, submitted to AofA 2016.

[5] Jacquet, P., Szpankowski, W. Autocorrelation on words and its applications: analysis of suffix trees by string-ruler approach. *J. Combinatorial Theory*, Series A, 66(2), 237-269, 1994.

[6] P. Jacquet, W. Szpankowski, and J. Tang, Average Profile of the Lempel-Ziv Parsing Scheme for a Markovian Source, *Algorithmica*, 2002.

[7] Flajolet, P., Roux, M., & Valle, B., Digital trees and memoryless sources: from arithmetics to analysis. DMTCS Proceedings, (01), 2010.

[8] Jacquet, P., Szpankowski, W. (1998). Analytical depoissonization and its applications. Theoretical Computer Science, 201(1), 1-62.

[9] Jacquet, P., Szpankowski, W. *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, 2015.

[10] N.Merhav and W. Szpankowski, Average Redundancy of the Shannon Code for Markov Sources, *IEEE Trans. Information Theory*, 59, 7186-7193, 2013.

[11] M. Regnier, and P. Jacquet, New Results on the Size of Tries, *IEEE Trans. Information Theory*, 35, 203–205, 1989.

[12] Jacquet, P., Szpankowski, W. (1991). Analysis of digital tries with Markovian dependency. IEEE Transactions on Information Theory, 37(5), 1470-1475.

[13] M. Régnier, and W. Szpankowski, On Pattern Frequency Occurrences in a Markovian Sequence, *Algorithmica*, 22, 631–649, 1998.

[14] P. Shields, *The Ergodic Theory of Discrete Sample Paths*, American Mathematical Society, Providence, 1996.

[15] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176–1198, 1993.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Bootstrap percolation on G(n,p) revisited

Mihyun Kang[†], Tamás Makai [‡]

Graz University of Technology, Institute of Discrete Mathematics, Steyrergasse 30, 8010 Graz, Austria
{*kang, makai*}*@math.tugraz.at*

Bootstrap percolation on a graph with infection threshold $r \in \mathbb{N}$ is an infection process, which starts from a set of initially infected vertices and in each step every vertex with at least r infected neighbours becomes infected. We consider bootstrap percolation on the binomial random graph $G(n,p)$, which was investigated among others by Janson, Łuczak, Turova and Valier (2012). We improve their results by strengthening the probability bounds for the number of infected vertices at the end of the process.

Keywords: Random graph, Bootstrap percolation, Martingale

1 Introduction

Bootstrap percolation on a graph with infection threshold $r \in \mathbb{N}$ is a deterministic infection process which evolves in rounds. In each round every vertex has exactly one of two possible states: it is either infected or uninfected. We denote the set of initially infected vertices by $A(0)$. In each round of the process every uninfected vertex v becomes infected if it has at least r infected neighbours, otherwise it remains uninfected. Once a vertex has become infected, it remains infected forever. The final infected set is denoted by A_f.

Bootstrap percolation was introduced by Chalupa, Leath, and Reich [CLR79] in the context of magnetic disordered systems. Since then bootstrap percolation processes (and extensions) have been used to describe several complex phenomena: from neuronal activity [Ami10, ELP$^+$] to the dynamics of the Ising model at zero temperature [FSS02].

In the context of social networks, bootstrap percolation provides a prototype model for the spread of ideas. In this setting infected vertices represent individuals who have already adopted a new belief and a person adopts a new belief if at least r of his acquaintances have already adopted it.

On the d-dimensional grid $[n]^d$ bootstrap percolation has been studied by Balogh, Bollobás, Duminil-Copin, and Morris [BBDCM12], when the initial infected set contains every vertex independently with probability p. For the size of the final infection set they showed the existence of a sharp threshold. More precisely, they established the threshold probability p_c, such that if $p \leq (1-\varepsilon)p_c$, then the probability that every vertex in $[n]^d$ becomes infected tends to 0, as $n \to \infty$, while if $p \geq (1+\varepsilon)p_c$, then the probability that every vertex in $[n]^d$ becomes infected tends to one, as $n \to \infty$.

[†]Supported by Austrian Science Fund (FWF): P26826, W1230.
[‡]Supported by Austrian Science Fund (FWF): P26826.

Bootstrap percolation has also been studied for several random graph models. For instance Amini and Fountoulakis [AF14] considered the Chung-Lu model [CL02] where the vertex weights follow a power law degree distribution and the presence of an edge $\{u, v\}$ is proportional to the product of the weights of u and v. Taking into account that in this model a linear fraction of the vertices have degree less than r and thus at most a linear fraction of the vertices can become infected, the authors proved the size of the final infected set A_f exhibits a phase transition.

Janson, Łuczak, Turova, and Vallier [JŁTV12] analysed bootstrap percolation on the binomial random graph $G(n, p)$, a graph with vertex set $[n] := \{1, 2, \ldots, n\}$ where every edge appears independently with probability $p = p(n)$, and the set of initially infected vertices $A(0)$ is chosen uniformly at random from the vertex sets of size a. For $r \geq 2$ and p satisfying both $p = \omega(n^{-1})$ and $p = o(n^{-1/r})$, they showed, among other results, that with probability tending to one as $n \to \infty$ either only a few additional vertices are infected or almost every vertex becomes infected. In addition they determined, depending on the number of initially infected vertices, the probability of both of these events up to an additive term tending to zero as $n \to \infty$.

The main contributions of this paper are threefold. First we strengthen this result by showing exponential tail bounds. Second we introduce a martingale in order to determine the number of infected vertices during the early stages of the process. Finally in the supercritical regime we show that the subgraph spanned by the vertices with $r - 1$ infected neighbours grows large enough to contain a giant component. The infection of just one vertex in this giant component leads to every vertex in the component becoming infected and we show that this in fact happens.

Main Results. Throughout the paper we assume that $r \geq 2$ and that both $p = \omega(n^{-1})$ and $p = o(n^{-1/r})$ hold. Set

$$t_0 := \left(\frac{r!}{np^r}\right)^{1/(r-1)}.$$

Let $\hat{\pi}(t) = \mathbb{P}[\mathrm{Bin(t, p)} \geq r]$ and define

$$a_c := -\min_{t \leq t_0} \frac{n\hat{\pi}(t) - t}{1 - \hat{\pi}(t)}.$$

In addition denote by t_c the smallest value t where this minimum is reached. Similarly to [JŁTV12] it can be shown that

$$t_c = (1 + o(1))((r-1)!/(np^r))^{1/(r-1)} \quad \text{and} \quad a_c = (1 + o(1))(1 - 1/r)t_c.$$

Theorem 1 *Let ω_0 be any function satisfying the conditions $\omega_0 = \omega(\sqrt{a_c})$ and $\omega_0 \leq a_c - r$. If $|A(0)| = a_c - \omega_0$, then with probability at least*

$$1 - \exp\left(-\frac{\omega_0^2}{10t_0}\right)$$

we have $|A_f| < t_c$.

Theorem 2 *Let ω_0 be any function satisfying the conditions $\omega_0 = \omega(\sqrt{a_c})$ and $\omega_0 \leq t_0 - a_c$. If $|A(0)| = a_c + \omega_0$, then with probability at least*

$$1 - \exp\left(-\frac{\omega_0^2}{10t_0}\right) - \exp\left(-\frac{a_c + \omega_0}{4}\right)$$

we have $|A_f| = (1 + o(1))n$.

Proof Technique. When the number of infected vertices is small (at most t_0), we introduce a martingale to show that the number of infected vertices is concentrated around its expectation with *exponentially* high probability. The martingale resembles the one introduced in [JŁTV12], however the maximal one step difference in our martingale is significantly smaller and thus provides a tighter concentration bound (Lemma 7).

In the subcritical regime, the expected number of infected vertices is less than $t_c < t_0$ and therefore the martingale argument alone implies the result (Section 4).

In the supercritical regime, this is not enough as the number of infected vertices will reach t_0 with exponentially high probability. In fact, at least $t_0 + a_c$ vertices become infected (Lemma 8). Now take a subset of the infected vertices with size t_0 and consider the vertices with at least $r - 1$ neighbours in this set. The size of this set is roughly rp^{-1} (Lemma 9) and the subgraph spanned by these vertices is also a binomial random graph, $G(rp^{-1}, p)$. Since the seminal work of Erdős and Rényi [ER60], it is known that this graph has with probability $1 + o(1)$ a linear sized giant component. More recently, Bollobás and Riordan [BR] showed that this happens with exponentially high probability (Theorem 5). Should any vertex in the giant component have an additional infected neighbour, then every vertex in the giant will become infected eventually. We show that this happens with exponentially high probability.

Thus we have $\Omega(p^{-1})$ infected vertices. After this, the process ends in two steps and this can be shown by two simple applications of the Chernoff bound (Lemmas 10 and 11).

2 Preliminaries

We will use the following form of the Chernoff bound.

Theorem 3 *[CL06] Let $X \sim \mathrm{Bin}(n, p)$, i.e. a binomial random variable with parameters n and p. Then for any $\lambda > 0$*

$$\mathbb{P}[X - \mathbb{E}(X) \leq -\lambda] \leq \exp\left(-\frac{\lambda^2}{2\mathbb{E}(X)}\right) \quad \textit{and} \quad \mathbb{P}[X - \mathbb{E}(X) \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2(\mathbb{E}(X) + \lambda/3)}\right).$$

Let $M_0, M_1, \ldots, M_i$ be a sequence of random variables and denote by $\mathcal{F}(i)$ the filter generated by $M_0, \ldots, M_i$. We say $M_0, \ldots, M_k$ forms a martingale if for every $0 \leq i \leq k$ we have $\mathbb{E}(|M_i|) < \infty$ and for every $1 \leq i \leq k$

$$\mathbb{E}[M_i | \mathcal{F}(i-1)] = M_{i-1}.$$

The following concentration bound on martingales due to Chung and Lu [CL06] will prove to be vital.

Theorem 4 *[CL06] For $m_0 \in \mathbb{R}$ let $M_0 = m_0, M_1, \ldots, M_k$ be a martingale whose conditional variance and differences satisfy the following: for each $1 \leq i \leq k$,*

- $\mathrm{Var}[M_i | M_{i-1}, \ldots, M_0] \leq \sigma_i^2$*;*
- $|M_i - M_{i-1}| \leq m$ *for some positive m.*

Then for any $\lambda > 0$, we have

$$\mathbb{P}[M_k - M_0 \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2\left(\sum_{i=1}^k \sigma_i^2 + m\lambda/3\right)}\right).$$

We will also need the following Theorem on the appearance of a giant component in $G(n,p)$ by Bollobás and Riordan [BR].

Theorem 5 *[BR] Let $c > 1$ be a constant independent of n and let $\varepsilon > 0$ independent of n. Then with probability $1 - \exp(-\Omega(n))$ the binomial random graph $G(n, c/n)$ has a component of size at least $(1-\varepsilon)\rho n$, where $\rho \in (0,1)$ is the unique positive solution of $1 - \rho = \exp(-c\rho)$.*

3 Setup: Martingale

In order to analyse the bootstrap percolation on $G(n,p)$ we will use the following reformulation due to Scalia-Tomba [ST85] as in [JŁTV12]. Roughly speaking they examine the infected vertices one by one and determine the vertices which have at least r neighbours in the set of previously examined vertices. The set of examined vertices until step t is denoted by $Z(t)$ and the set of infected vertices by $A(t)$. Formally let $A(0)$ be the set of initially infected vertices of size a and without the loss of generality we may assume that $A(0) = \{1, ..., a\}$. Set $Z(0) = \emptyset$. For each step $t \in \mathbb{N}$, if $A(t-1)\backslash Z(t-1) \neq \emptyset$, then let $U_t = \{u_t\}$, where u_t is a vertex in $A(t-1)\backslash Z(t-1)$ selected according to an arbitrary rule, otherwise set $U_t = \emptyset$. Set $Z(t) := Z(t-1) \cup U_t$. Now for $t \geq 0$ and each $i \in [n-a] := \{1, \dots, n-a\}$ let $X(t,i)$ be the indicator random variable for the event that the vertex $a+i$ has at least r neighbours in $Z(t)$ and set

$$A(t) := A(0) \cup \{a + i : X(t,i) = 1, i \in [n-a]\}.$$

The process stops when $t = n$.

Clearly $Z(t) \subset A(t)$. Let T denote the smallest value of t such that $A(t) = Z(t)$. Note that $t \leq T$ implies that $|Z(t)| = t$ and thus T is also the smallest t such that $|A(t)| = t$. Since $|A(t)| \leq n$ for every natural number $0 \leq t \leq n$ we have that $T \leq n$. Note further that $A(T) = A_f$.

In order to have a better control on the maximal number of vertices which can become infected in a single step, we refine the process by dividing every step into rounds, in such a way that in each round *exactly one vertex* $v \in [n]\backslash A(0)$ is examined (regardless whether it was examined in earlier rounds or not). Thus each step $1 \leq t \leq n$ consists of $n-a$ rounds and round i of step t is denoted by (t,i). We denote the step following (t,i) by $(t,i)+1$ and the preceding step by $(t,i)-1$. Also the ordering of the rounds is given by the lexicographical order i.e. $(\tau, \iota) < (t,i)$ if either $\tau < t$ or $\tau = t$ and $\iota < i$.

In round i of step t we examine if vertex $a+i$ has at least r neighbours in $Z(t)$ and if it has we add it to the set of infected vertices. Formally for $(t,i) \geq (1,1)$

$$A((t,i)+1) := A(0) \cup \{a + j : j \leq i, X(t,j) = 1\} \cup \{a + j : j > i, X(t-1,j) = 1\}.$$

Clearly we have $A(t) = A(t, n-a)$. For consistency define $A(0, n-a) := A(0)$.

Define a function $\pi : \mathbb{N} \to [0,1]$ by

$$\pi(t) := \begin{cases} \mathbb{P}[\mathrm{Bin(t,p)} \geq r], & \text{for } t \leq T \\ \mathbb{P}[\mathrm{Bin(T,p)} \geq r], & \text{for } t > T \end{cases}$$

and note that $\pi(t)$ is a random variable.

For $(t,i) \geq (0, n-a)$, define the random variable

$$M(t,i) := \sum_{j=1}^{i} \frac{X(t,j) - \pi(t)}{1 - \pi(t)} + \sum_{j=i+1}^{n-a} \frac{X(t-1,j) - \pi(t-1)}{1 - \pi(t-1)}. \tag{1}$$

We will denote by $\mathcal{F}(t,i)$ the filter generated by $M(0,n-a),\ldots,M(t,i)$.

Lemma 6 *The sequence of random variables $M(0,n-a),\ldots,M(n,n-a)$ forms a martingale.*

Proof: Fix $1 \le t \le n$ and $1 \le i \le n-a$. For every $\tau < t$, we can express from (1) the number of infected vertices in step τ:

$$|A(\tau)| = a + \sum_{\iota=1}^{n-a} X(\tau,\iota) \stackrel{(1)}{=} a + M(\tau,n-a)(1-\pi(\tau)) + (n-a)\pi(\tau). \tag{2}$$

Recall that $\hat{\pi}(t) = \mathbb{P}[\mathrm{Bin}(t,p) \ge r]$. We will denote by T' the smallest value of t which satisfies $a + M(t,n-a)(1-\hat{\pi}(t)) + (n-a)\hat{\pi}(t) = t$. Because $\pi(t) = \hat{\pi}(t)$ when $t \le T$, we have $T = T'$. Given the filter $\mathcal{F}((t,i)-1)$ one can establish if $a + M(\tau,n-a)(1-\hat{\pi}(\tau)) + (n-a)\hat{\pi}(\tau) = \tau$ for some $\tau < t$. Therefore, it can be determined whether the event $T' < t$ or $t \ge T'$ holds. In particular, if $T' < t$, then the exact value of T' can be determined.

For each $\tau \le t$, since $\pi(\tau)$ depends only on the value of $T = T'$, we can also determine the value of $\pi(\tau)$, i.e. $\mathbb{E}[\pi(\tau)|\mathcal{F}((t,i)-1)] = \pi(\tau)$ for $\tau \le t$.

Note that $X(0,i) = 0$ for every $1 \le i \le n-a$ and that for every $(1,1) \le (\tau,\iota) < (t,i)$ we can easily compute from (1)

$$M(\tau,\iota) - M((\tau,\iota)-1) = \frac{X(\tau,\iota)-\pi(\tau)}{1-\pi(\tau)} - \frac{X(\tau-1,\iota)-\pi(\tau-1)}{1-\pi(\tau-1)}. \tag{3}$$

Therefore, based on the filter $\mathcal{F}((t,i)-1)$, the value of $X(\tau,\iota)$ can be determined for every $(\tau,\iota) < (t,i)$.

Next we shall show that

$$\mathbb{E}\left[\left.\frac{X(t,i)-\pi(t)}{1-\pi(t)}\right| \mathcal{F}((t,i)-1)\right] = \frac{\mathbb{E}[X(t,i)|\mathcal{F}((t,i)-1)]-\pi(t)}{1-\pi(t)} = \frac{X(t-1,i)-\pi(t-1)}{1-\pi(t-1)}. \tag{4}$$

To this end, observe that if $X(t-1,i) = 1$, then we have $X(t,i) = 1$ with probability 1 and in this case both sides of equation (4) equal 1.

Now assume that $X(t-1,i) = 0$. When $t > T$, we have $X(t,i) = X(t,i-1) = 0$ with probability 1 and by the definition of $\pi(t)$ we have $\pi(t) = \pi(t-1) = \hat{\pi}(T)$. Evaluating both sides of equation (4) gives us $-\hat{\pi}(T)/(1-\hat{\pi}(T))$. When $t \le T$ and $X((t,i)-1) = 0$, we have $\pi(t) = \hat{\pi}(t)$ and thus

$$\mathbb{P}\left[X(t,i) = 0 | \mathcal{F}((t,i)-1)\right] = \frac{1-\hat{\pi}(t)}{1-\hat{\pi}(t-1)} = 1 - \frac{\hat{\pi}(t)-\hat{\pi}(t-1)}{1-\hat{\pi}(t-1)}.$$

Therefore in this case

$$\mathbb{E}\left[\left.\frac{X(t,i)-\pi(t)}{1-\pi(t)}\right| \mathcal{F}((t,i)-1)\right] = -\frac{\hat{\pi}(t)}{1-\hat{\pi}(t)}\frac{1-\hat{\pi}(t)}{1-\hat{\pi}(t-1)} + 1\cdot\frac{\hat{\pi}(t)-\hat{\pi}(t-1)}{1-\hat{\pi}(t-1)} = -\frac{\hat{\pi}(t-1)}{1-\hat{\pi}(t-1)}$$

and thus (4) holds. According to (3) we have that

$$\begin{aligned}&\mathbb{E}[M(t,i) - M((t,i)-1)|\mathcal{F}((t,i)-1)]\\ &\quad= \mathbb{E}\left[\left.\frac{X(t,i)-\pi(t)}{1-\pi(t)}\right| \mathcal{F}((t,i)-1)\right] - \frac{X(t-1)-\pi(t-1)}{1-\pi(t-1)} \stackrel{(4)}{=} 0.\end{aligned}$$

□

Lemma 7 *Let $t \in \{0, \dots, n\}$ and $\lambda \in \mathbb{R}^+$ be given.*

$$\mathbb{P}\left[\bigwedge_{(0,n-a)\le(\tau,i)\le(t,n-a)} M(\tau,i) > -\lambda\right] \ge 1 - \exp\left(-\frac{\lambda^2(1-\hat{\pi}(t))^3}{2(n\hat{\pi}(t)+\lambda/3)}\right)$$

and

$$\mathbb{P}\left[\bigwedge_{(0,n-a)\le(\tau,i)\le(t,n-a)} M(\tau,i) < \lambda\right] \ge 1 - \exp\left(-\frac{\lambda^2(1-\hat{\pi}(t))^3}{2(n\hat{\pi}(t)+\lambda/3)}\right).$$

Proof: We will only show the bound on the probability that $M(\tau,i) < \lambda$ for each $(\tau,i) \le (t, n-a)$. The other case follows simply from the fact that if the random variables $M(0,n-a), \dots, M(t,n-a)$ form a martingale, then $-M(0,n-a), \dots, -M(t,n-a)$ is also a martingale and they both have the same conditional variance and maximal difference. In order to show that the bounds hold for each round, we introduce the following martingale:

$$\hat{M}(\tau,i) = \begin{cases} M(\tau,i) & \text{if } \hat{M}((\tau,i)-1) < \lambda \\ \hat{M}((\tau,i)-1) & \text{otherwise.} \end{cases}$$

Similarly to $M(t,i)$ we denote the filter generated by $\hat{M}(0,n-a), \dots, \hat{M}(t,i)$ with $\hat{\mathcal{F}}(t,i)$. Note that if there exists a round (τ,i) such that $M(\tau,i) \ge \lambda$, then we have $\hat{M}(\tau',i') \ge \lambda$ for every $(\tau',i') \ge (\tau,i)$. Therefore $\hat{M}(t,n-a) < \lambda$ implies that for every $(\tau,i) \le (t,n-a)$ we have $M(\tau,i) < \lambda$.

By (3) and since $\pi(\tau) \le \hat{\pi}(\tau)$ with probability 1, we have

$$|\hat{M}(\tau,\iota) - \hat{M}((\tau,\iota)-1)| \le \max\left\{\frac{\hat{\pi}(\tau)}{1-\hat{\pi}(\tau)} - \frac{\hat{\pi}(\tau-1)}{1-\hat{\pi}(\tau-1)}, 1 + \frac{\hat{\pi}(\tau-1)}{1-\hat{\pi}(\tau-1)}\right\} = \frac{1}{1-\hat{\pi}(\tau-1)}.$$

Note that $M(0,n-a) = 0$. Since $\mathrm{Var}[\hat{M}(\tau,i)|\hat{\mathcal{F}}((\tau,i)-1)] = 0$ if $\hat{M}((\tau,\iota)-1) \ge \lambda$ and

$$\mathrm{Var}[\hat{M}(\tau,i)|\hat{\mathcal{F}}((\tau,i)-1)] = \mathrm{Var}[M(\tau,i)|\mathcal{F}((\tau,i)-1)]$$

otherwise, Theorem 4 implies that

$$\mathbb{P}[\hat{M}(t,n-a) \ge \lambda] \le \exp\left(-\frac{\lambda^2}{2(S+\lambda/(3(1-\hat{\pi}(\tau-1))))}\right)$$

where

$$S \le \sum_{\tau=1}^{t}\sum_{i=1}^{n-a} \mathrm{Var}[M(\tau,i)|\mathcal{F}((\tau,i)-1)]. \tag{5}$$

Note that

$$\begin{aligned}\mathrm{Var}[M(\tau,i)|\mathcal{F}((\tau,i)-1)] &= \mathrm{Var}\left[\frac{X(\tau,i)}{1-\pi(t)}\middle|\mathcal{F}((\tau,i)-1)\right] \\ &\le \frac{1}{(1-\hat{\pi}(t))^2}\mathrm{Var}[X(\tau,i)|\mathcal{F}((\tau,i)-1)]. \end{aligned}\tag{6}$$

Recall that $X(\tau-1,i)=1$ implies $X(\tau,i)=1$ and that $\tau>T$ implies $X(\tau,i)=X(\tau-1,i)$. In both of these cases we have

$$\mathrm{Var}[X(\tau,i)|\mathcal{F}((\tau,i)-1)]=0 \overset{\hat{\pi}(t)\geq\hat{\pi}(t-1)}{\leq} \frac{\hat{\pi}(t)-\hat{\pi}(t-1)}{1-\hat{\pi}(t)}. \tag{7}$$

Now assume $\tau\leq T$ and $X(\tau-1,i)=0$. Since $X(t,i)$ is an indicator random variable, we have

$$\mathrm{Var}[X(\tau,i)|\mathcal{F}((\tau,i)-1)]\leq\mathbb{E}[X(\tau,i)|\mathcal{F}((\tau,i)-1)]=\frac{\hat{\pi}(\tau)-\hat{\pi}(\tau-1)}{1-\hat{\pi}(\tau-1)}\leq\frac{\hat{\pi}(\tau)-\hat{\pi}(\tau-1)}{1-\hat{\pi}(\tau)}. \tag{8}$$

Putting (5)-(8) together, we obtain

$$S\leq\sum_{\tau=1}^{t}\sum_{i=1}^{n-a}\frac{\hat{\pi}(\tau)-\hat{\pi}(\tau-1)}{(1-\hat{\pi}(\tau))^3}\leq\sum_{\tau=1}^{t}\frac{n(\hat{\pi}(\tau)-\hat{\pi}(\tau-1))}{(1-\hat{\pi}(t))^3}\leq\frac{n\hat{\pi}(t)}{(1-\hat{\pi}(t))^3}.$$

□

The previous lemma allows us to analyse the process in the first t_0 steps. This will be used in the proofs of Theorems 1 and 2.

4 Proof of Theorem 1

We want to investigate the number of infected vertices at time t_c. By the definition of a_c and t_c, we have

$$a_c=-\min_{t\leq t_0}\frac{n\hat{\pi}(t)-t}{1-\hat{\pi}(t)}=\frac{t_c-n\hat{\pi}(t_c)}{1-\hat{\pi}(t_c)}. \tag{9}$$

By the definition of $M(t,i)$, we have

$$|A(t_c)|\overset{(2)}{=}a+(1-\pi(t_c))M(t_c,n-a)+(n-a)\pi(t_c).$$

Since $\pi(t)\leq\hat{\pi}(t)$ and $a=a_c-\omega_0$, we obtain

$$\begin{aligned}|A(t_c)|&\leq a+M(t_c,n-a)+(n-a)\hat{\pi}(t_c)\\&=(a_c-\omega_0)(1-\hat{\pi}(t_c))+n\hat{\pi}(t_c)+M(t_c,n-a)\\&\overset{(9)}{=}t_c-n\hat{\pi}(t_c)+n\hat{\pi}(t_c)-\omega_0(1-\hat{\pi}(t_c))+M(t_c,n-a)\\&\overset{\hat{\pi}(t_c)\leq\hat{\pi}(t_0)}{\leq}t_c-\omega_0(1-\hat{\pi}(t_0))+M(t_c,n-a).\end{aligned} \tag{10}$$

Using $np=\omega(1)$ and $t_0=(r!/(np^r))^{1/(r-1)}$, we have

$$t_0p=O\left(\left(\frac{1}{np^r}\right)^{1/(r-1)}p\right)=O\left(\left(\frac{1}{np}\right)^{1/(r-1)}\right)=o(1). \tag{11}$$

Furthermore,

$$\hat{\pi}(t_0) = \mathbb{P}[\mathrm{Bin}(t_0,p) \geq r] = \sum_{j=r}^{t_0} \binom{t_0}{j} p^j (1-p)^{t_0-j} \overset{(11)}{=} (1+o(1))\frac{t_0^r p^r}{r!}$$
$$= (1+o(1))\frac{t_0^{r-1}p^r}{r!} t_0 = (1+o(1))\frac{t_0}{n} \overset{np=\omega(1)}{=} o\left(t_0 p\right) \overset{(11)}{=} o(1). \tag{12}$$

Applying Lemma 7 with $\lambda = \omega_0/2$, we have that with probability at least

$$1 - \exp\left(-\frac{\omega_0^2}{8((1+o(1))n\hat{\pi}(t_c) + \omega_0/6)}\right) \geq 1 - \exp\left(-\frac{\omega_0^2}{10t_0}\right)$$

$M(t_c, n-a) < \omega_0/2$. This together with (10) implies

$$|A(t_c)| \leq t_c - (1+o(1))\omega_0 + \omega_0/2 < t_c.$$

Therefore $T < t_c$ and thus $|A_f| = T < t_c$.

5 Proof of Theorem 2

Before proving Theorem 2 we begin with an observation on $A(t_0)$, the set of infected vertices after the first t_0 steps.

Lemma 8 *Let ω_0 be any function satisfying the conditions $\omega_0 = \omega(\sqrt{a_c})$ and $\omega_0 \leq t_0 - a_c$. If $|A(0)| = a_c + \omega_0$, then with probability at least*

$$1 - \exp\left(-\frac{\omega_0^2}{9.5t_0}\right)$$

we have $T > t_0$ and $|A(t_0)| \geq t_0 + (1+o(1))a_c + \omega_0/2$.

Proof: By the definition of a_c, for every $t \leq t_0$ we have

$$a_c \geq \frac{t - n\hat{\pi}(t)}{1-\hat{\pi}(t)}. \tag{13}$$

Assume that $M(t,i) > -\omega_0/2$ for every $t \leq t_0$ and $1 \leq i \leq n-a$. First we will show by induction that if $M(t,i)$ satisfies this lower bound, then $T > t_0$. Clearly $T > 0$. Now assume that for some $t \leq t_0 - 1$ we have that $T > t-1$. Therefore $\pi(t) = \hat{\pi}(t)$. For $t \leq t_0$ we have $\hat{\pi}(t) \leq \hat{\pi}(t_0) \overset{(12)}{=} o(1)$ and thus

$$\begin{aligned} |A(t)| &\overset{(2)}{=} a + (1-\pi(t))M(t,n-a) + (n-a)\pi(t) \\ &\overset{M(t,i)>-\omega_0/2}{\geq} (1-\hat{\pi}(t))(a_c+\omega_0) + n\hat{\pi}(t) - (1+o(1))\omega_0/2 \\ &\overset{(13)}{\geq} t + (1+o(1))\omega_0/2. \end{aligned}$$

Therefore $|A(t)| > t$ which together with $T > t-1$ implies $T > t$.

Also note that

$$|A(t_0)| \stackrel{(2)}{=} a + M(t_0, n-a)(1-\pi(t_0)) + (n-a)\hat{\pi}(t_0)$$
$$\stackrel{(12)}{\geq} (1+o(1))a_c + (1+o(1))t_0 + (1+o(1))\omega_0/2. \tag{14}$$

Let $t_1 := ((r-1)!/np^r)^{1/(r-1)}$. Then $t_1 \leq t_0$ and so

$$a_c \geq \frac{t_1 - n\hat{\pi}(t_1)}{1-\hat{\pi}(t_1)}. \tag{15}$$

Also

$$\hat{\pi}(t_1) = \mathbb{P}[\mathrm{Bin}[t_1,p) \geq r] = \sum_{j=r}^{t_0} \binom{t_0}{j} p^j (1-p)^{t_1-j} \stackrel{(11)}{=} (1+o(1))\frac{t_1^r p^r}{r!}$$
$$= (1+o(1))\frac{t_1^{r-1}p^r}{r!}t_1 = (1+o(1))\frac{t_1}{rn}. \tag{16}$$

From this and (15)

$$a_c \geq \frac{t_1 - (1+o(1))t_1/r}{1-\hat{\pi}(t_1)} = (1+o(1))\left(1-\frac{1}{r}\right)t_1 = \Omega(t_0), \tag{17}$$

which together with (14) and $\omega_0 \leq t_0$ implies

$$|A(t_0)| = t_0 + (1+o(1))a_c + \omega_0/2.$$

Lemma 7 with $\lambda = \omega_0/2$ implies the result.

□

We will need to establish the size of the giant component in the set of vertices which have at least $r-1$ neighbours in $Z(t_0)$. For this we first need to determine the number of vertices which have at least $r-1$ neighbours in $Z(t_0)$.

Lemma 9 *Let $A \subset [n]$ with $|A| = o(n)$. Conditional on $T \geq t_0$, $A(t_0)$ and $Z(t_0)$, with probability $1-\exp(-\Omega(p^{-1}))$ we have that the number of vertices in $[n]\backslash(Z(t_0)\cup A)$ with at least $r-1$ neighbours in $Z(t_0)$ is at least $3rp^{-1}/4$.*

Proof: Let X_v be the indicator random variable that a vertex $v \in [n]\backslash(Z(t_0)\cup A)$ has at least $r-1$ neighbours in $Z(t_0)$ and set $X := \sum_{v\in[n]\backslash(Z(t_0)\cup A)} X_v$. Clearly

$$\mathbb{P}[X_v = 1 | v \in A(t_0)] = 1.$$

Note that if $X_v = 1$ and $v \notin A(t_0)$, then v has exactly $r-1$ neighbours in $Z(t_0)$ and thus

$$\mathbb{P}[X_v = 1 | v \notin A(t_0)] \geq \mathbb{P}[X_v = 1, v \notin A(t_0)] = \binom{t_0}{r-1} p^{r-1}(1-p)^{t_0-r+1}$$
$$\stackrel{(11)}{=} (1+o(1))\frac{t_0^{r-1}}{(r-1)!}p^{r-1} = (1+o(1))\frac{r}{np}.$$

Since the set of random variables $\{X_v | v \in [n]\backslash(Z(t_0)\cup A)\}$ are mutually independent, X stochastically dominates the binomial random variable $\hat{X} = \mathrm{Bin}(n - t_0 - |A|, (1 + o(1))r/np)$. Because $t_0 = o(n)$ and $|A| = o(n)$, we have

$$\mathbb{E}[\hat{X}] = (1 + o(1))rp^{-1}$$

and Theorem 3 implies

$$\mathbb{P}[X - \mathbb{E}(\hat{X}) \le -(1 + o(1))rp^{-1}/4] \le \exp\left(-(1 + o(1))\frac{r^2p^{-2}}{32rp^{-1}}\right) \le \exp\left(-\frac{p^{-1}}{32}\right).$$

□

In the following two lemmas we look at the number of vertices with at least r neighbours in a set of order p^{-1} and set of order n. Estimating the probability that a vertex has at least r neighbours in such sets differs significantly and are discussed separately.

Lemma 10 *Let $U, W \subset [n]$ in $G(n,p)$ satisfy $|U| = p^{-1}/2$ and $|W| = o(n)$. With probability at least $1 - \exp(-\Omega(n))$ the number of vertices in $[n]\backslash(U \cup W)$ with at least r neighbours in U is at least $n/(2^r r!\sqrt{\mathrm{e}})$.*

Proof: Let Y_v be the indicator random variable that a vertex $v \in [n]\backslash(U \cup W)$ has at least r neighbours in U and set $Y = \sum_{v\in[n]\backslash(U\cup W)} Y_v$. We have that

$$\begin{aligned}
\mathbb{P}[Y_v = 1] &= 1 - \sum_{j=0}^{r-1} \binom{p^{-1}/2}{j} p^j (1-p)^{p^{-1}/2 - j} \\
&= 1 - (1-p)^{p^{-1}/2} \sum_{j=0}^{r-1} \binom{p^{-1}/2}{j} \left(\frac{p}{1-p}\right)^j \\
&= 1 - (1 + o(1))\mathrm{e}^{-1/2} \sum_{j=0}^{r-1} \frac{(2p)^{-j}}{j!} p^j \\
&= 1 - (1 + o(1))\mathrm{e}^{-1/2} \sum_{j=0}^{r-1} \frac{1}{j!2^j}.
\end{aligned}$$

Clearly $1 \ge \mathbb{P}[Y_v = 1] > 1/(2^r r!\sqrt{\mathrm{e}})$. Set $\eta := \mathbb{P}[Y_v = 1] - 1/(2^r r!\sqrt{\mathrm{e}})$ and note that $\eta = \Omega(1)$. Furthermore the set of random variables $\{Y_v | v \in [n]\backslash(U \cup W)\}$ are mutually independent and $|[n]\backslash(U \cup W)| = (1 + o(1))n$. Therefore, by Theorem 3 we have

$$\mathbb{P}[Y < n/(2^r r!\sqrt{\mathrm{e}})] \le \exp\left(-(1 + o(1))\frac{\eta^2 n^2}{2n}\right) = \exp(-\Omega(n)).$$

□

Lemma 11 *Let $U, W \subset [n]$ in $G(n,p)$ satisfy $|U| = n/(2^r r!\sqrt{\mathrm{e}})$. Then with probability $\exp(-\Omega(p^{-1}))$ all but at most p^{-1} vertices in $[n]\backslash(U \cup W)$ have at least r neighbours in U.*

Proof: Let B_v be the indicator random variable that a vertex $v \in [n]\backslash(U \cup W)$ has less than r neighbours in U and set $B := \sum_{v\in[n]\backslash(U\cup W)} B_v$. Since $np = \omega(1)$, we have

$$\mathbb{P}[B_v = 1] = \sum_{j=0}^{r-1} \binom{|U|}{j} p^j (1-p)^{|U|-j} \leq \exp(-|U|p) \sum_{j=0}^{r-1} \frac{(np)^j}{j!} \leq \exp(-|U|p)(np)^{r-1}.$$

Since $|[n]\backslash(U \cup W)| \leq n$ and $np = \omega(1)$, we have

$$\mathbb{E}[B] \leq \exp(-|U|p)n(np)^{r-1} = \exp(-\Omega(np))n^r p^{r-1} = o(p^{-1}).$$

Note that the set of random variables $\{B_v | v \in [n]\backslash(U \cup W)\}$ are mutually independent and therefore, by Theorem 3 we have

$$\mathbb{P}[B > p^{-1}] \leq \exp\left(-\Omega(p^{-1})\right).$$

□

Proof of Theorem 2: According to Lemma 8 with probability at least $1 - \exp(-\omega_0^2/(9.5t_0))$ we have that $|A(t)| > t$ for every $t \leq t_0$ and $A(t_0) \geq t_0 + (1 + o(1))a_c + \omega_0/2$. Therefore the process runs for at least t_0 steps and there exists a set $A \subseteq |A(t_0)\backslash Z(t_0)|$ of size $a_c/2 + \omega_0/2$.

Lemma 9 implies that conditional on $A(t_0)$ and $Z(t_0)$ with probability at least $1-\exp(-\Omega(p^{-1}))$ there is a set of vertices in $[n]\backslash(Z(t_0) \cup A)$ with size at least $3rp^{-1}/4$ where every vertex in the set has at least $r-1$ neighbours in $Z(t_0)$ and select a subset W of these vertices of size exactly $3rp^{-1}/4$. Note that until this point every event depends only on edges with one end in $Z(t_0)$.

According to Theorem 5, with probability $1 - \exp(-\Omega(p^{-1}))$ there is a set $U \subset W$ such that the vertices in U form a connected component and $|U| \geq (1-\varepsilon)\rho p^{-1}$ for arbitrary $\varepsilon > 0$ independent of n, where ρ is the unique solution of $1 - \rho = \exp(-3\rho r/4)$. Since

$$1 - \rho > \sum_{k=0}^{4} \frac{(-3\rho r/4)^k}{k!} > \exp(-3\rho r/4)$$

when $0 < \rho \leq 1/2$, we have $\rho > 1/2$ and thus we have that $|U| \geq p^{-1}/2$. Also this event depends only on the edges with both endpoints in U and thus it is independent of the previous events.

Note that if a vertex in A is connected to a vertex in U, then every vertex in U will become infected. The probability that no vertex in A is connected to any vertex in U is

$$(1-p)^{|U|(a_c+\omega)/2} \leq \exp(-(a_c+\omega)/4).$$

This event depends on edges between A and U and thus it is independent of the previous events.

Now take $U' \subset U$ such that $|U'| = p^{-1}/2$ and denote the set of vertices in $[n]\backslash(W \cup A \cup Z(t_0))$ which have at least r neighbours in U' with B. Since $|W \cup A \cup Z(t_0)| = o(n)$ by Lemma 10 with probability $1-\exp(-\Omega(n))$ we have that $|B| \geq n/(2^r r!\sqrt{\mathrm{e}})$. Note that $B \subset |A_f|$. This event depends only on edges between U' and $[n]\backslash(U \cup W \cup A \cup Z(t_0))$ and thus it is independent of the previous events.

Finally let $B' \subset B$ with $|B'| = n/(2^r r!\sqrt{\mathrm{e}})$ and consider the set of vertices in $[n]\backslash(B \cup U \cup Z(t_0) \cup A)$ which contain at least r neighbours in B'. Note that all of these vertices will be infected. By Lemma 11 we have with probability $1 - \exp(-\Omega(p^{-1}))$ that all but at most p^{-1} vertices in $[n]\backslash(B \cup U \cup Z(t_0) \cup A)$

will become infected. Similarly as before this event depends only on edges which we haven't considered previously and thus it is independent of the previous events. Recall that $B \cup U \cup Z(t_0) \cup A \subset A_f$ and thus $|A_f| = (1 + o(1))n$.

Since $p^{-1} = \omega(t_0)$ and $n = \omega(p^{-1})$, we have that the probability that almost every vertex becomes infected is at least

$$1 - \exp\left(-\frac{\omega_0^2}{10t_0}\right) - \exp\left(-\frac{a_c + \omega_0}{4}\right).$$

□

References

[AF14] H. Amini and N. Fountoulakis. Bootstrap percolation in power-law random graphs. *Journal of Statistical Physics*, 155(1):72–92, 2014.

[Ami10] H. Amini. Bootstrap percolation in living neural networks. *J. Stat. Phys.*, 141(3):459–475, 2010.

[BBDCM12] J. Balogh, B. Bollobás, H. Duminil-Copin, and R. Morris. The sharp threshold for bootstrap percolation in all dimensions. *Trans. Amer. Math. Soc.*, 364(5):2667–2701, 2012.

[BR] B. Bollobás and O. Riordan. Exploring hypergraphs with martingales. arXiv:1403.6558.

[CL02] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Ann. Comb.*, 6(2):125–145, 2002.

[CL06] F. Chung and L. Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Math.*, 3(1):79–127, 2006.

[CLR79] J. Chalupa, P. Leath, and G. Reich. Bootstrap percolation on a Bethe lattice. *Journal of Physics C: Solid State Physics*, 12:L31–L35, 1979.

[ELP$^+$] H. Einarsson, J. Lengler, K. Panagiotou, F. Mousset, and A. Steger. Bootstrap percolation with inhibition. arXiv:1410.3291.

[ER60] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[FSS02] L. Fontes, R. Schonmann, and V. Sidoravicius. Stretched exponential fixation in stochastic Ising models at zero temperature. *Comm. Math. Phys*, 228:495–518, 2002.

[JŁTV12] S. Janson, T. Łuczak, T. Turova, and T. Vallier. Bootstrap percolation on the random graph $G_{n,p}$. *Ann. Appl. Probab.*, 22(5):1989–2047, 2012.

[ST85] G.-P. Scalia-Tomba. Asymptotic final-size distribution for some chain-binomial processes. *Adv. in Appl. Probab.*, 17(3):477–495, 1985.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

q-Quasiadditive Functions

Sara Kropf[1,2†§], Stephan Wagner[3‡§]

[1] *Institut für Mathematik, Alpen-Adria-Universität Klagenfurt, Austria, sara.kropf@aau.at*
[2] *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, sarakropf@stat.sinica.edu.tw*
[3] *Department of Mathematical Sciences, Stellenbosch University, South Africa, swagner@sun.ac.za*

In this paper, we introduce the notion of q-quasiadditivity of arithmetic functions, as well as the related concept of q-quasimultiplicativity, which generalises strong q-additivity and -multiplicativity, respectively. We show that there are many natural examples for these concepts, which are characterised by functional equations of the form $f(q^{k+r}a+b) = f(a) + f(b)$ or $f(q^{k+r}a+b) = f(a)f(b)$ for all $b < q^k$ and a fixed parameter r. In addition to some elementary properties of q-quasiadditive and q-quasimultiplicative functions, we prove characterisations of q-quasiadditivity and q-quasimultiplicativity for the special class of q-regular functions. The final main result provides a general central limit theorem that includes both classical and new examples as corollaries.

Keywords: q-additive function, q-quasiadditive function, q-regular function, central limit theorem

1 Introduction

Arithmetic functions based on the digital expansion in some base q have a long history (see, e.g., [3–8,11]) The notion of a q-*additive* function is due to [11]: an arithmetic function (defined on nonnegative integers) is called q-additive if

$$f(q^k a + b) = f(q^k a) + f(b)$$

whenever $0 \leq b < q^k$. A stronger version of this concept is *strong* (or *complete*) q-additivity: a function f is said to be strongly q-additive if we even have

$$f(q^k a + b) = f(a) + f(b)$$

whenever $0 \leq b < q^k$. The class of (strongly) q-*multiplicative* functions is defined in an analogous fashion. Loosely speaking, (strong) q-additivity of a function means that it can be evaluated by breaking up the base-q expansion. Typical examples of strongly q-additive functions are the q-ary sum of digits and the number of occurrences of a specified nonzero digit.

†The first author is supported by the Austrian Science Fund (FWF): P 24644-N26.
‡The second author is supported by the National Research Foundation of South Africa under grant number 96236.
§The authors were also supported by the Karl Popper Kolleg "Modeling–Simulation–Optimization" funded by the Alpen-Adria-Universität Klagenfurt and by the Carinthian Economic Promotion Fund (KWF). Part of this paper was written while the second author was a Karl Popper Fellow at the Mathematics Institute in Klagenfurt. He would like to thank the institute for the hospitality received.

There are, however, many simple and natural functions based on the q-ary expansion that are not q-additive. A very basic example of this kind are *block counts*: the number of occurrences of a certain block of digits in the q-ary expansion. This and other examples provide the motivation for the present paper, in which we define and study a larger class of functions with comparable properties.

Definition. An arithmetic function (a function defined on the set of nonnegative integers) is called q-*quasiadditive* if there exists some nonnegative integer r such that

$$f(q^{k+r}a+b) = f(a) + f(b) \tag{1}$$

whenever $0 \leq b < q^k$. Likewise, f is said to be q-*quasimultiplicative* if it satisfies the identity

$$f(q^{k+r}a+b) = f(a)f(b) \tag{2}$$

for some fixed nonnegative integer r whenever $0 \leq b < q^k$.

We remark that the special case $r = 0$ is exactly strong q-additivity, so strictly speaking the term "strongly q-quasiadditive function" might be more appropriate. However, since we are not considering a weaker version (for which natural examples seem to be much harder to find), we do not make a distinction. As a further caveat, we remark that the term "quasiadditivity" has also been used in [1] for a related, but slightly weaker condition.

In the following section, we present a variety of examples of q-quasiadditive and q-quasimultiplicative functions. In Section 3, we give some general properties of such functions. Since most of our examples also belong to the related class of q-regular functions, we discuss the connection in Section 4. Finally, we prove a general central limit theorem for q-quasiadditive and -multiplicative functions that contains both old and new examples as special cases.

2 Examples of q-quasiadditive and q-quasimultiplicative functions

Let us now back up the abstract concept of q-quasiadditivity by some concrete examples.

Block counts

As mentioned in the introduction, the number of occurrences of a fixed nonzero digit is a typical example of a q-additive function. However, the number of occurrences of a given block $B = \epsilon_1\epsilon_2\cdots\epsilon_\ell$ of digits in the expansion of a nonnegative integer n, which we denote by $c_B(n)$, does not represent a q-additive function. The reason is simple: the q-ary expansion of $q^k a + b$ is obtained by joining the expansions of a and b, so occurrences of B in a and occurrences of B in b are counted by $c_B(a) + c_B(b)$, but occurrences that involve digits of both a and b are not.

However, if B is a block different from $00\cdots0$, then c_B is q-quasiadditive: note that the representation of $q^{k+\ell}a + b$ is of the form

$$\underbrace{a_1a_2\cdots a_\mu}_{\text{expansion of } a}\underbrace{00\cdots0}_{\ell \text{ zeros}}\underbrace{b_1b_2\cdots b_\nu}_{\text{expansion of } b}$$

whenever $0 \leq b < q^k$, so occurrences of the block B have to belong to either a or b only. This implies that $c_B(q^{k+\ell}a + b) = c_B(a) + c_B(b)$, with one small caveat: if the block starts and/or ends with a sequence of zeros, then the count needs to be adjusted by assuming the digital expansion of a nonnegative integer to be padded with zeros on the left and on the right.

For example, let B be the block 0101 in base 2. The binary representations of 469 and 22 are 111010101 and 10110, respectively, so we have $c_B(469) = 2$ and $c_B(22) = 1$ (note the occurrence of 0101 at the beginning of 10110 if we assume the expansion to be padded with zeros), as well as

$$c_B(240150) = c_B(2^9 \cdot 469 + 22) = c_B(469) + c_B(22) = 3.$$

Indeed, the block B occurs three times in the expansion of 240150, which is 111010101000010110.

The number of runs and the Gray code

The number of ones in the Gray code of a nonnegative integer n, which we denote by $h_{\mathsf{GRAY}}(n)$, is also equal to the number of runs (maximal sequences of consecutive identical digits) in the binary representations of n (counting the number of runs in the representation of 0 as 0); the sequence defined by $h_{\mathsf{GRAY}}(n)$ is A005811 in Sloane's On-Line Encyclopedia of Integer Sequences [17]. An analysis of its expected value is performed in [10]. The function h_{GRAY} is 2-quasiadditive up to some minor modification: set $f(n) = h_{\mathsf{GRAY}}(n)$ if n is even and $f(n) = h_{\mathsf{GRAY}}(n) + 1$ if n is odd. The new function f can be interpreted as the total number of occurrences of the two blocks 01 and 10 in the binary expansion (considering binary expansions to be padded with zeros at both ends), so the argument of the previous example applies again and shows that f is 2-quasiadditive.

The nonadjacent form and its Hamming weight

The nonadjacent form (NAF) of a nonnegative integer is the unique base-2 representation with digits $0, 1, -1$ (-1 is usually represented as $\overline{1}$ in this context) and the additional requirement that there may not be two adjacent nonzero digits, see [18]. For example, the NAF of 27 is $100\overline{1}0\overline{1}$. It is well known that the NAF always has minimum Hamming weight (i.e., the number of nonzero digits) among all possible binary representations with this particular digit set, although it may not be unique with this property (compare, e.g., [18] with [15]).

The Hamming weight h_{NAF} of the nonadjacent form has been analysed in some detail [13, 20], and it is also an example of a 2-quasiadditive function. It is not difficult to see that h_{NAF} is characterised by the recursions $h_{\mathsf{NAF}}(2n) = h_{\mathsf{NAF}}(n)$, $h_{\mathsf{NAF}}(4n + 1) = h_{\mathsf{NAF}}(n) + 1$, $h_{\mathsf{NAF}}(4n - 1) = h_{\mathsf{NAF}}(n) + 1$ together with the initial value $h_{\mathsf{NAF}}(0) = 0$. The identity

$$h_{\mathsf{NAF}}(2^{k+2}a + b) = h_{\mathsf{NAF}}(a) + h_{\mathsf{NAF}}(b)$$

can be proved by induction. In Section 4, this example will be generalised and put into a larger context.

The number of optimal $\{0, 1, -1\}$-representations

As mentioned above, the NAF may not be the only representation with minimum Hamming weight among all possible binary representations with digits $0, 1, -1$. The number of optimal representations of a given nonnegative integer n is therefore a quantity of interest in its own right. Its average over intervals of the form $[0, N)$ was studied by Grabner and Heuberger [12], who also proved that the number $r_{\mathsf{OPT}}(n)$ of optimal representations of n can be obtained in the following way:

Lemma 1 (Grabner–Heuberger [12])**.** *Let sequences u_i ($i = 1, 2, \ldots, 5$) be given recursively by*

$$u_1(0) = u_2(0) = \cdots = u_5(0) = 1, \qquad u_1(1) = u_2(1) = 1,\ u_3(1) = u_4(1) = u_5(1) = 0,$$

and

$$\begin{aligned}
u_1(2n) &= u_1(n), & u_1(2n+1) &= u_2(n) + u_4(n+1),\\
u_2(2n) &= u_1(n), & u_2(2n+1) &= u_3(n),\\
u_3(2n) &= u_2(n), & u_3(2n+1) &= 0,\\
u_4(2n) &= u_1(n), & u_4(2n+1) &= u_5(n+1),\\
u_5(2n) &= u_4(n), & u_5(2n+1) &= 0.
\end{aligned}$$

The number $r_{\mathsf{OPT}}(n)$ of optimal representations of n is equal to $u_1(n)$.

A straightforward calculation shows that

$$\begin{aligned}
&u_1(8n) = u_2(8n) = \cdots = u_5(8n) = u_1(8n+1) = u_2(8n+1) = u_1(n),\\
&u_3(8n+1) = u_4(8n+1) = u_5(8n+1) = 0.
\end{aligned} \tag{3}$$

This gives us the following result (see the full version of this extended abstract for a detailed proof):

Lemma 2. *The number of optimal $\{0, 1, -1\}$-representations of a nonnegative integer is a 2-quasimultiplicative function. Specifically, for any three nonnegative integers a, b, k with $b < 2^k$, we have*

$$r_{\mathsf{OPT}}(2^{k+3}a + b) = r_{\mathsf{OPT}}(a)r_{\mathsf{OPT}}(b).$$

In Section 4, we will show that this is also an instance of a more general phenomenon.

The run length transform and cellular automata

The *run length transform* of a sequence is defined in a recent paper of Sloane [19]: it is based on the binary representation, but could in principle also be generalised to other bases. Given a sequence $s_1, s_2, \ldots$, its run length transform is obtained by the rule

$$t(n) = \prod_{i \in \mathcal{L}(n)} s_i,$$

where $\mathcal{L}(n)$ is the multiset of run lengths of n (lengths of blocks of consecutive ones in the binary representation). For example, the binary expansion of 1910 is 11101110110, so the multiset $\mathcal{L}(n)$ of run lengths would be $\{3, 3, 2\}$, giving $t(1910) = s_2 s_3^2$.

A typical example is obtained for the sequence of Jacobsthal numbers given by the formula $s_n = \frac{1}{3}(2^{n+2} - (-1)^n)$. The associated run length transform t_n (sequence A071053 in the OEIS [17]) counts the number of odd coefficients in the expansion of $(1 + x + x^2)^n$, and it can also be interpreted as the number of active cells at the n-th generation of a certain cellular automaton. Further examples stemming from cellular automata can be found in Sloane's paper [19].

The argument that proved q-quasiadditivity of block counts also applies here, and indeed it is easy to see that the identity

$$t(2^{k+1}a + b) = t(a)t(b),$$

where $0 \leq b < 2^k$, holds for the run length transform of any sequence, meaning that any such transform is 2-quasimultiplicative. In fact, it is not difficult to show that every 2-quasimultiplicative function with parameter $r = 1$ is the run length transform of some sequence.

3 Elementary properties

Now that we have gathered some motivating examples for the concepts of q-quasiadditivity and q-quasimultiplicativity, let us present some simple results about functions with these properties. First of all, let us state an obvious relation between q-quasiadditive and q-quasimultiplicative functions:

Proposition 3. *If a function f is q-quasiadditive, then the function defined by $g(n) = c^{f(n)}$ for some positive constant c is q-quasimultiplicative. Conversely, if f is a q-quasimultiplicative function that only takes positive values, then the function defined by $g(n) = \log_c f(n)$ for some positive constant $c \neq 1$ is q-quasiadditive.*

The next proposition deals with the parameter r in the definition of a q-quasiadditive function:

Proposition 4. *If the arithmetic function f satisfies $f(q^{k+r}a+b) = f(a)+f(b)$ for some fixed nonnegative integer r whenever $0 \leq b < q^k$, then it also satisfies $f(q^{k+s}a+b) = f(a)+f(b)$ for all nonnegative integers $s \geq r$ whenever $0 \leq b < q^k$.*

Proof. If a, b are nonnegative integers with $0 \leq b < q^k$, then clearly also $0 \leq b < q^{k+s-r}$ if $s \geq r$, and thus

$$f(q^{k+s}a+b) = f(q^{(k+s-r)+r}a+b) = f(a)+f(b).$$

□

Corollary 5. *If two arithmetic functions f and g are q-quasiadditive functions, then so is any linear combination $\alpha f + \beta g$ of the two.*

Proof. In view of the previous proposition, we may assume the parameter r in (1) to be the same for both functions. The statement follows immediately. □

Finally, we observe that q-quasiadditive and q-quasimultiplicative functions can be computed by breaking the q-ary expansion into pieces. A detailed proof can be found in the full version:

Lemma 6. *If f is a q-quasiadditive (q-quasimultiplicative) function, then*

- $f(0) = 0$ *($f(0) = 1$, respectively, unless f is identically 0),*
- $f(qa) = f(a)$ *for all nonnegative integers a.*

Proposition 7. *Suppose that the function f is q-quasiadditive with parameter r, i.e., $f(q^{k+r}a+b) = f(a)+f(b)$ whenever $0 \leq b < q^k$. Going from left to right, split the q-ary expansion of n into blocks by inserting breaks after each run of r or more zeros. If these blocks are the q-ary representations of $n_1, n_2, \ldots, n_\ell$, then we have*

$$f(n) = f(n_1) + f(n_2) + \cdots + f(n_\ell).$$

Moreover, if m_i is the greatest divisor of n_i which are not divisible by q for $i = 1, \ldots, \ell$, then

$$f(n) = f(m_1) + f(m_2) + \cdots + f(m_\ell).$$

Analogous statements hold for q-quasimultiplicative functions, with sums replaced by products.

Proof. This is obtained by a straightforward induction on ℓ together with the fact that $f(q^h a) = f(a)$, which follows from the previous lemma. □

Example 1. Recall that the Hamming weight of the NAF (which is the minimum Hamming weight of a $\{0, 1, -1\}$-representation) is 2-quasiadditive with parameter $r = 2$. To determine $h_{\mathsf{NAF}}(314\,159\,265)$, we split the binary representation, which is $100101011100110110000010100001$, into blocks by inserting breaks after each run of at least two zeros:

$$100|101011100|110110000|1010000|1.$$

The numbers $n_1, n_2, \ldots, n_\ell$ in the statement of the proposition are now $4, 348, 432, 80, 1$ respectively, and the numbers $m_1, m_2, \ldots, m_\ell$ are therefore $1, 87, 27, 5, 1$. Now we use the values $h_{\mathsf{NAF}}(1) = 1$, $h_{\mathsf{NAF}}(5) = 2$, $h_{\mathsf{NAF}}(27) = 3$ and $h_{\mathsf{NAF}}(87) = 4$ to obtain

$$h_{\mathsf{NAF}}(314\,159\,265) = 2h_{\mathsf{NAF}}(1) + h_{\mathsf{NAF}}(5) + h_{\mathsf{NAF}}(27) + h_{\mathsf{NAF}}(87) = 11.$$

Example 2. In the same way, we consider the number of optimal representations r_{OPT}, which is 2-quasimultiplicative with parameter $r = 3$. Consider for instance the binary representation of $204\,280\,974$, namely $1100001011010001010010001110$. We split into blocks:

$$110000|101101000|101001000|1110.$$

The four blocks correspond to the numbers $48 = 16 \cdot 3$, $360 = 8 \cdot 45$, $328 = 8 \cdot 41$ and $14 = 2 \cdot 7$. Since $r_{\mathsf{OPT}}(3) = 2$, $r_{\mathsf{OPT}}(45) = 5$, $r_{\mathsf{OPT}}(41) = 1$ and $r_{\mathsf{OPT}}(7) = 1$, we obtain $r_{\mathsf{OPT}}(204\,280\,974) = 10$.

4 q-Regular functions

In this section, we introduce q-regular functions and examine the connection to our concepts. See [2] for more background on q-regular sequences.

A function f is *q-regular* if it can be expressed as $f = \boldsymbol{u}^t \boldsymbol{f}$ for a vector $\boldsymbol{u}$ and a vector-valued function $\boldsymbol{f}$, and there are matrices M_i, $0 \le i < q$, satisfying

$$\boldsymbol{f}(qn + i) = M_i \boldsymbol{f}(n) \tag{4}$$

for $0 \le i < q$, $qn + i > 0$. We set $\boldsymbol{v} = \boldsymbol{f}(0)$.

Equivalently, a function f is q-regular if and only if f can be written as

$$f(n) = \boldsymbol{u}^t \prod_{i=0}^{L} M_{n_i} \boldsymbol{v} \tag{5}$$

where $n_L \cdots n_0$ is the q-ary expansion of n.

The notion of q-regular functions is a generalisation of q-additive and q-multiplicative functions. However, we emphasise that q-quasiadditive and q-quasimultiplicative functions are not necessarily q-regular: a q-regular sequence can always be bounded by $O(n^c)$ for a constant c, see [2, Thm. 16.3.1]. In our setting however, the values of $f(n)$ can be chosen arbitrarily for those n whose q-ary expansion does not contain 0^r. Therefore a q-quasiadditive or -multiplicative function can grow arbitrarily fast.

We call $(\boldsymbol{u}, (M_i)_{0\le i<q}, \boldsymbol{v})$ a *linear representation* of the q-regular function f. Such a linear representation is called *zero-insensitive* if $M_0\boldsymbol{v} = \boldsymbol{v}$, meaning that in (5), leading zeros in the q-ary expansion of n do not change anything. We call a linear representation *minimal* if the dimension of the matrices M_i is minimal among all linear representations of f.

Following [9], every q-regular function has a zero-insensitive minimal linear representation.

4.1 When is a q-regular function q-quasimultiplicative?

We now give a characterisation of q-regular functions that are q-quasimultiplicative. Proofs of the results in this and the following subsection can be found in the full version.

Theorem 8. *Let f be a q-regular sequence with zero-insensitive minimal linear representation* (5)*. Then the following two assertions are equivalent:*

- *The sequence f is q-quasimultiplicative with parameter r.*
- $M_0^r = \boldsymbol{v}\boldsymbol{u}^t$.

Example 3 (The number of optimal $\{0, 1, -1\}$-representations). The number of optimal $\{0, 1, -1\}$-representations as described in Section 2 is a 2-regular sequence by Lemma 1. A minimal zero-insensitive linear representation for the vector $(u_1(n), u_2(n), u_3(n), u_1(n+1), u_4(n+1), u_5(n+1))^t$ is given by

$$M_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0\\ 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$\boldsymbol{u}^t = (1, 0, 0, 0, 0, 0)$ and $\boldsymbol{v} = (1, 1, 1, 1, 0, 0)^t$.

As $M_0^3 = vu^t$, this sequence is 2-quasimultiplicative with parameter 3, which is the same result as in Lemma 2.

Remark. The condition on the minimality of the linear representation in Theorem 8 is necessary as illustrated by the following example:

Consider the sequence $f(n) = 2^{s_2(n)}$ where $s_2(n)$ is the binary sum of digits function. This sequence is 2-regular and 2-(quasi-)multiplicative with parameter $r = 0$. A minimal linear representation is given by $M_0 = 1$, $M_1 = 2$, $v = 1$ and $u = 1$. As stated in Theorem 8, we have $M_0^0 = vu^t = 1$.

If we use the zero-insensitive non-minimal linear representation defined by $M_0 = \left(\begin{smallmatrix} 1 & 13 \\ 0 & 2 \end{smallmatrix}\right)$, $M_1 = \left(\begin{smallmatrix} 2 & 27 \\ 0 & 5 \end{smallmatrix}\right)$, $v = (1, 0)^t$ and $u^t = (1, 0)$ instead, we have $\operatorname{rank} M_0^r = 2$ for all $r \ge 0$. Thus $M_0^r \ne vu^t$.

4.2 When is a q-regular function q-quasiadditive?

The characterisation of q-regular functions that are also q-quasiadditive is somewhat more complicated. Again, we consider a zero-insensitive (but not necessarily minimal) linear representation. We let U be the smallest vector space such that all vectors of the form $\boldsymbol{u}^t \prod_{i\in I} M_{n_i}$ lie in the affine subspace $\boldsymbol{u}^t + U^t$ (U^t is used as a shorthand for $\{\boldsymbol{x}^t : \boldsymbol{x} \in U\}$). Such a vector space must exist, since $\boldsymbol{u}^t$ is a vector of this form (corresponding to the empty product, where $I = \emptyset$). Likewise, let V be the smallest vector space such that all vectors of the form $\prod_{j\in J} M_{n_j}\boldsymbol{v}$ lie in the affine subspace $\boldsymbol{v} + V$.

Theorem 9. *Let f be a q-regular sequence with zero-insensitive linear representation* (5)*. The sequence f is q-quasiadditive with parameter r if and only if all of the following statements hold:*

- $\boldsymbol{u}^t\boldsymbol{v} = 0$*,*
- U^t *is orthogonal to* $(M_0^r - I)\boldsymbol{v}$*, i.e.,* $\boldsymbol{x}^t(M_0^r - I)\boldsymbol{v} = \boldsymbol{x}^t M_0^r\boldsymbol{v} - \boldsymbol{x}^t\boldsymbol{v} = 0$ *for all* $\boldsymbol{x} \in U$*,*
- V *is orthogonal to* $\boldsymbol{u}^t(M_0^r - I)$*, i.e.,* $\boldsymbol{u}^t(M_0^r - I)\boldsymbol{y} = \boldsymbol{u}^t M_0^r\boldsymbol{y} - \boldsymbol{u}^t\boldsymbol{y} = 0$ *for all* $\boldsymbol{y} \in V$*,*
- $U^t M_0^r V = 0$*, i.e.,* $\boldsymbol{x}^t M_0^r\boldsymbol{y} = 0$ *for all* $\boldsymbol{x} \in U$ *and* $\boldsymbol{y} \in V$*.*

Example 4. For the Hamming weight of the nonadjacent form, a zero-insensitive (and also minimal) linear representation for the vector $(h_{\mathsf{NAF}}(n), h_{\mathsf{NAF}}(n+1), h_{\mathsf{NAF}}(2n+1), 1)^t$ is

$$M_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad M_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$\boldsymbol{u}^t = (1, 0, 0, 0)$ and $\boldsymbol{v} = (0, 1, 1, 1)^t$.

The three vectors $\mathbf{w}_1 = \boldsymbol{u}^t M_1 - \boldsymbol{u}^t$, $\mathbf{w}_2 = \boldsymbol{u}^t M_1^2 - \boldsymbol{u}^t$ and $\mathbf{w}_3 = \boldsymbol{u}^t M_1 M_0 M_1 - \boldsymbol{u}^t$ are linearly independent. If we let W be the vector space spanned by those three, it is easily verified that M_0 and M_1 map the affine subspace $\boldsymbol{u}^t + W^t$ to itself, so $U = W$ is spanned by these vectors.

Similarly, the three vectors $M_1\boldsymbol{v} - \boldsymbol{v}$, $M_1^2\boldsymbol{v} - \boldsymbol{v}$ and $M_1 M_0 M_1\boldsymbol{v} - \boldsymbol{v}$ span V.

The first condition of Theorem 9 is obviously true. We only have to verify the other three conditions with $r = 2$ for the basis vectors of U and V, which is done easily. Thus h_{NAF} is a 2-regular sequence that is also 2-quasiadditive, as was also proved in Section 2.

Finding the vector spaces U and V is not trivial. But in a certain special case of q-regular functions, we can give a sufficient condition for q-additivity, which is easier to check. These q-regular functions are output sums of transducers as defined in [14]: a transducer transforms the q-ary expansion of an integer n (read from the least significant to the most significant digit) deterministically into an output sequence and leads to a state s. The output sum is then the sum of this output sequence together with the final output of the state s. This defines the value of the q-regular function evaluated at n. The function h_{NAF} discussed in the example above, as well as many other examples, can be represented in this way.

Proposition 10. *The output sum of a connected transducer is q-additive with parameter r if the following conditions are satisfied:*

- *The transducer has the reset sequence 0^r going to the initial state, i.e., reading r zeros always leads to the initial state of the transducer.*
- *For every state, the output sum along the path of the reset sequence 0^r equals the final output of this state.*
- *Additional zeros at the end of the input sequence do not change the output sum.*

5 A central limit theorem for q-quasiadditive and -multiplicative functions

In this section, we prove a central limit theorem for q-quasimultiplicative functions taking only positive values. By Proposition 3, this also implies a central limit theorem for q-quasiadditive functions.

To this end, we define a generating function: let f be a q-quasimultiplicative function with positive values, let $\mathcal{M}_k$ be the set of all nonnegative integers less than q^k (i.e., those positive integers whose q-ary expansion needs at most k digits), and set

$$F(x,t) = \sum_{k\geq 0} x^k \sum_{n\in\mathcal{M}_k} f(n)^t.$$

The decomposition of Proposition 7 now translates directly to an alternative representation for $F(x,t)$: let $\mathcal{B}$ be the set of all positive integers not divisible by q whose q-ary representation does not contain the block 0^r, let $\ell(n)$ denote the length of the q-ary representation of n, and define the function $B(x,t)$ by

$$B(x,t) = \sum_{n\in\mathcal{B}} x^{\ell(n)} f(n)^t.$$

We remark that in the special case where $q = 2$ and $r = 1$, this simplifies greatly to

$$B(x,t) = \sum_{k\geq 1} x^k f(2^k-1)^t. \tag{6}$$

Proposition 11. *The generating function $F(x,t)$ can be expressed as*

$$F(x,t) = \frac{1}{1-x}\cdot\frac{1}{1-\frac{x^r}{1-x}B(x,t)}\Big(1+(1+x+\cdots+x^{r-1})B(x,t)\Big) = \frac{1+(1+x+\cdots+x^{r-1})B(x,t)}{1-x-x^rB(x,t)}.$$

Proof. The first factor stands for the initial sequence of leading zeros, the second factor for a (possibly empty) sequence of blocks consisting of an element of $\mathcal{B}$ and r or more zeros, and the last factor for the final part, which may be empty or an element of $\mathcal{B}$ with up to $r-1$ zeros (possibly none) added at the end. □

Under suitable assumptions on the growth of a q-quasiadditive or q-quasimultiplicative function, we can exploit the expression of Proposition 11 to prove a central limit theorem in the following steps (full proofs can again be found in the full version).

Definition. We say that a function f has *at most polynomial growth* if $f(n) = O(n^c)$ and $f(n) = \Omega(n^{-c})$ for a fixed $c \geq 0$. We say that f has *at most logarithmic growth* if $f(n) = O(\log n)$.

Note that our definition of at most polynomial growth is slightly different than usual: the extra condition $f(n) = \Omega(n^{-c})$ ensures that the absolute value of $\log f(n)$ does not grow too fast.

Lemma 12. *Assume that the positive, q-quasimultiplicative function f has at most polynomial growth. There exist positive constants δ and ϵ such that*

- *$B(x,t)$ has radius of convergence $\rho(t) > \frac{1}{q}$ whenever $|t| \leq \delta$.*

- *For $|t| \leq \delta$, the equation $x + x^r B(x,t) = 1$ has a complex solution $\alpha(t)$ with $|\alpha(t)| < \rho(t)$ and no other solutions with modulus $\leq (1+\epsilon)|\alpha(t)|$.*
- *Thus the generating function $F(x,t)$ has a simple pole at $\alpha(t)$ and no further singularities of modulus $\leq (1+\epsilon)|\alpha(t)|$.*
- *Finally, α is an analytic function of t for $|t| \leq \delta$.*

Lemma 13. *Assume that the positive, q-quasimultiplicative function f has at most polynomial growth. With δ and ϵ as in the previous lemma, we have, uniformly in t,*

$$[x^k]F(x,t) = \kappa(t) \cdot \alpha(t)^{-k}\big(1 + O((1+\epsilon)^{-k})\big)$$

for some function κ. Both α and κ are analytic functions of t for $|t| \leq \delta$, and $\kappa(t) \neq 0$ in this region.

Theorem 14. *Assume that the positive, q-quasimultiplicative function f has at most polynomial growth. Let N_k be a randomly chosen integer in $\{0, 1, \dots, q^k - 1\}$. The random variable $L_k = \log f(N_k)$ has mean $\mu k + O(1)$ and variance $\sigma^2 k + O(1)$, where the two constants are given by*

$$\mu = \frac{B_t(1/q, 0)}{q^{2r}}$$

and

$$\begin{aligned}\sigma^2 = {} & -B_t(1/q,0)^2 q^{-4r+1}(q-1)^{-1} + 2B_t(1/q,0)^2 q^{-3r+1}(q-1)^{-1} - B_t(1/q,0)^2 q^{-4r}(q-1)^{-1} \\ & - 4rB_t(1/q,0)^2 q^{-4r} + B_{tt}(1/q,0)q^{-2r} - 2B_t(1/q,0)B_{tx}(1/q,0)q^{-4r-1}. \quad (7)\end{aligned}$$

If f is not the constant function $f \equiv 1$, then $\sigma^2 \neq 0$ and the normalised random variable $(L_k - \mu k)/(\sigma\sqrt{k})$ converges weakly to a standard Gaussian distribution.

Corollary 15. *Assume that the q-quasiadditive function f has at most logarithmic growth. Let N_k be a randomly chosen integer in $\{0, 1, \dots, q^k - 1\}$. The random variable $L_k = f(N_k)$ has mean $\hat{\mu} k + O(1)$ and variance $\hat{\sigma}^2 k + O(1)$, where the two constants μ and $\sigma^2$are given by the same formulas as in Theorem 14, with $B(x,t)$ replaced by*

$$\hat{B}(x,t) = \sum_{n \in \mathcal{B}} x^{\ell(n)} e^{f(n)t}.$$

If f is not the constant function $f \equiv 0$, then the normalised random variable $(L_k - \hat{\mu} k)/(\hat{\sigma}\sqrt{k})$ converges weakly to a standard Gaussian distribution.

Remark. By means of the Cramér-Wold device (and Corollary 5), we also obtain joint normal distribution of tuples of q-quasiadditive functions.

We now revisit the examples discussed in Section 2 and state the corresponding central limit theorems. Some of them are well known while others are new. We also provide numerical values for the constants in mean and variance.

Example 5 (see also [8, 16]). The number of blocks 0101 occurring in the binary expansion of n is a 2-quasiadditive function of at most logarithmic growth. Thus by Corollary 15, the standardised random variable is asymptotically normally distributed, the constants being $\hat{\mu} = \frac{1}{16}$ and $\hat{\sigma}^2 = \frac{17}{256}$.

Example 6 (see also [13, 20]). The Hamming weight of the nonadjacent form is 2-quasiadditive with at most logarithmic growth (as the length of the NAF of n is logarithmic). Thus by Corollary 15, the standardised random variable is asymptotically normally distributed. The associated constants are $\hat{\mu} = \frac{1}{3}$ and $\hat{\sigma}^2 = \frac{2}{27}$.

Example 7 (see Section 2). The number of optimal $\{0, 1, -1\}$-representations is 2-quasimultiplicative. As it is always greater or equal to 1 and 2-regular, it has at most polynomial growth. Thus Theorem 14 implies that the standardised logarithm of this random variable is asymptotically normally distributed with numerical constants given by $\mu \approx 0.060829$, $\sigma^2 \approx 0.038212$.

Example 8 (see Section 2). Suppose that the sequence $s_1, s_2, \ldots$ satisfies $s_n \geq 1$ and $s_n = O(c^n)$ for a constant $c \geq 1$. The run length transform $t(n)$ of s_n is 2-quasimultiplicative. As $s_n \geq 1$ for all n, we have $t(n) \geq 1$ for all n as well. Furthermore, there exists a constant A such that $s_n \leq Ac^n$ for all n, and the sum of all run lengths is bounded by the length of the binary expansion, thus

$$t(n) = \prod_{i \in \mathcal{L}(n)} s_i \leq \prod_{i \in \mathcal{L}(n)} (Ac^i) \leq (Ac)^{1+\log_2 n}.$$

Consequently, $t(n)$ is positive and has at most polynomial growth. By Theorem 14, we obtain an asymptotic normal distribution for the standardised random variable $\log t(N_k)$. The constants μ and σ^2 in mean and variance are given by

$$\mu = \sum_{i \geq 1} (\log s_i) 2^{-i-2}$$

and

$$\sigma^2 = \sum_{i \geq 1} (\log s_i)^2 \left(2^{-i-2} - (2i-1)2^{-2i-4}\right) - \sum_{j > i \geq 1} (\log s_i)(\log s_j)(i+j-1)2^{-i-j-3}.$$

These formulas can be derived from those given in Theorem 14 by means of the representation (6), and the terms can also be interpreted easily: write $\log t(n) = \sum_{i \geq 1} X_i(n) \log s_i$, where $X_i(n)$ is the number of runs of length i in the binary representation of n. The coefficients in the two formulas stem from mean, variance and covariances of the $X_i(n)$.

In the special case that s_n is the Jacobsthal sequence ($s_n = \frac{1}{3}(2^{n+2} - (-1)^n)$, see Section 2), we have the numerical values $\mu \approx 0.429947$, $\sigma^2 \approx 0.121137$.

References

[1] Jean-Paul Allouche and Olivier Salon, *Sous-suites polynomiales de certaines suites automatiques*, J. Théor. Nombres Bordeaux **5** (1993), no. 1, 111–121.

[2] Jean-Paul Allouche and Jeffrey Shallit, *Automatic sequences: Theory, applications, generalizations*, Cambridge University Press, Cambridge, 2003.

[3] Nader L. Bassily and Imre Kátai, *Distribution of the values of q-additive functions on polynomial sequences*, Acta Math. Hungar. **68** (1995), no. 4, 353–361.

[4] Richard Bellman and Harold N. Shapiro, *On a problem in additive number theory*, Ann. of Math. (2) **49** (1948), no. 2, 333–340.

[5] Emmanuel Cateland, *Suites digitales et suites k-régulières*, 1992, Doctoral thesis, Université de Bordeaux.

[6] Hubert Delange, *Sur les fonctions q-additives ou q-multiplicatives*, Acta Arith. **21** (1972), 285–298.

[7] ———, *Sur la fonction sommatoire de la fonction "somme des chiffres"*, Enseignement Math. (2) **21** (1975), 31–47.

[8] Michael Drmota, *The distribution of patterns in digital expansions*, Algebraic Number Theory and Diophantine Analysis (F. Halter-Koch and R. F. Tichy, eds.), de Gruyter (Berlin), 2000, pp. 103–121.

[9] Philippe Dumas, *Asymptotic expansions for linear homogeneous divide-and-conquer recurrences: Algebraic and analytic approaches collated*, Theoret. Comput. Sci. **548** (2014), 25–53.

[10] Philippe Flajolet and Lyle Ramshaw, *A note on Gray code and odd-even merge*, SIAM J. Comput. **9** (1980), 142–158.

[11] A. O. Gel′fond, *Sur les nombres qui ont des propriétés additives et multiplicatives données*, Acta Arith. **13** (1967/1968), 259–265.

[12] Peter J. Grabner and Clemens Heuberger, *On the number of optimal base 2 representations of integers*, Des. Codes Cryptogr. **40** (2006), no. 1, 25–39.

[13] Clemens Heuberger and Sara Kropf, *Analysis of the binary asymmetric joint sparse form*, Combin. Probab. Comput. **23** (2014), 1087–1113.

[14] Clemens Heuberger, Sara Kropf, and Helmut Prodinger, *Output sum of transducers: Limiting distribution and periodic fluctuation*, Electron. J. Combin. **22** (2015), no. 2, 1–53.

[15] Marc Joye and Sung-Ming Yen, *Optimal left-to-right binary signed digit recoding*, IEEE Trans. Comput. **49** (2000), no. 7, 740–748.

[16] Peter Kirschenhofer, *Subblock occurrences in the q-ary representation of n*, SIAM J. Algebraic Discrete Methods **4** (1983), no. 2, 231–236.

[17] *The On-Line Encyclopedia of Integer Sequences*, `http://oeis.org`, 2016.

[18] George W. Reitwiesner, *Binary arithmetic*, Advances in Computers, Vol. 1, Academic Press, New York, 1960, pp. 231–308.

[19] Neil J. A. Sloane, *The number of ON cells in cellular automata*, arXiv:1503.01168 [math.CO].

[20] Jörg M. Thuswaldner, *Summatory functions of digital sums occurring in cryptography*, Period. Math. Hungar. **38** (1999), no. 1-2, 111–130.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Combinatorial analysis of growth models for series-parallel networks

Markus Kuba[1] and Alois Panholzer[2†]

[1]*Institute of Applied Mathematics and Natural Sciences, University of Applied Sciences - Technikum Wien, Höchstädtplatz 5, 1200 Wien, Austria,* kuba@technikum-wien.at
[2]*Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien, Wiedner Hauptstr. 8-10/104, 1040 Wien, Austria,* Alois.Panholzer@tuwien.ac.at

We give combinatorial descriptions of two stochastic growth models for series-parallel networks introduced by Hosam Mahmoud by encoding the growth process via recursive tree structures. Using decompositions of the tree structures and applying analytic combinatorics methods allows a study of quantities in the corresponding series-parallel networks. For both models we obtain limiting distribution results for the degree of the poles and the length of a random source-to-sink path, and furthermore we get asymptotic results for the expected number of source-to-sink paths.

Keywords: series-parallel networks, growth models, distributional analysis, source-to-sink paths, node degrees

1 Introduction

Series-parallel networks are two-terminal graphs, i.e., they have two distinguished vertices called the source and the sink, that can be constructed recursively by applying two simple composition operations, namely the parallel composition (where the sources and the sinks of two series-parallel networks are merged) and the series composition (where the sink of one series-parallel network is merged with the source of another series-parallel network). Here we will always consider series-parallel networks as digraphs with edges oriented in direction from the north-pole, the source, towards the south-pole, the sink. Such graphs can be used to model the flow in a bipolar network, e.g., of current in an electric circuit or goods from the producer to a market. Furthermore series-parallel networks and series-parallel graphs (i.e., graphs which are series-parallel networks when some two of its vertices are regarded as source and sink; see, e.g., [2] for exact definitions and alternative characterizations) are of interest in computational complexity theory, since some in general NP-complete graph problems are solvable in linear time on series-parallel graphs (e.g., finding a maximum independent set).

Recently there occurred several studies concerning the typical behaviour of structural quantities (as, e.g., node-degrees, see [6]) in series-parallel graphs and networks under a uniform model of randomness, i.e., where all series-parallel graphs of a certain size (counted by the number of edges) are equally likely. In contrast to these uniform models, Mahmoud [11, 12] introduced two interesting growth models for

†Supported by the Austrian Science Fund FWF, grant P25337-N23

series-parallel networks, which are generated by starting with a single directed arc from the source to the sink and iteratively carrying out serial and parallel edge-duplications according to a stochastic growth rule; we call them uniform Bernoulli edge-duplication rule ("Bernoulli model" for short) and uniform binary saturation edge-duplication rule ("binary model" for short). A formal description of these models is given in Section 2. Using the defining stochastic growth rules and a description via Pólya-Eggenberger urn models (see, e.g., [10]), several quantities for series-parallel networks (as the number of nodes of small degree and the degree of the source for the Bernoulli model, and the length of a random source-to-sink path for the binary model) are treated in [11, 12].

The aim of this work is to give an alternative description of these growth models for series-parallel networks by encoding the growth of them via recursive tree structures, to be precise, via edge-coloured recursive trees and so-called bucket-recursive trees (see [9] and references therein). The advantage of such a modelling is that these objects allow not only a stochastic description (the tree evolution process which reflects the growth rule of the series-parallel network), but also a combinatorial one (as certain increasingly labelled trees or bucket trees), which gives rise to a top-down decomposition of the structure. An important observation is that indeed various interesting quantities for series-parallel networks can be studied by considering certain parameters in the corresponding recursive tree model and making use of the combinatorial decomposition. We focus here on the quantities degree D_n of the source and/or sink, length L_n of a random source-to-sink path and the number P_n of source-to-sink paths in a random series-parallel network of size n, but mention that also other quantities (as, e.g., the number of ancestors, node-degrees, or the number of paths through a random or the j-th edge) could be treated in a similar way. We obtain limiting distribution results for D_n and L_n (thus answering questions left open in [11, 12]), whereas for the r.v. P_n (whose distributional treatment seems to be considerably more involved) we are able to give asymptotic results for the expectation.

Mathematically, an analytic combinatorics treatment of the quantities of interest leads to studies of first and second order non-linear differential equations. In this context we want to mention that another model of series-parallel networks called increasing diamonds has been introduced recently in [1]. A treatment of quantities in such networks inherently also yields a study of second order non-linear differential equations; however, the definition as well as the structure of increasing diamonds is quite different from the models treated here as can be seen also by comparing the behaviour of typical graph parameters (e.g., the number of source-to-sink paths P_n in increasing diamonds is trivially bounded by n, whereas in the models studied here the expected number of paths grows exponentially). We mention that the analysis of the structures considered here has further relations to other objects; e.g., it holds that the Mittag-Leffler limiting distributions occurring in Theorem 3.1 & 3.2 also appear in other combinatorial contexts as in certain triangular balanced urn models (see [8]) or implicitly in the recent study of an extra clustering model for animal grouping [5] (after scaling, as continuous part of the characterization given in [5, Theorem 2], since it is possible to simplify some of the representations given there). Also the characterizations of the limiting distribution for D_n and L_n of binary series-parallel networks via the sequence of r-th integer moments satisfies a recurrence relation of "convolution type" similar to ones occurring in [3], for which asymptotic studies have been carried out. Furthermore, the described top-down decomposition of the combinatorial objects makes these structures amenable to other methods, in particular, it seems that the contraction method, see, e.g., [13, 14], allows an alternative characterization of limiting distributions occurring in the analysis of binary series-parallel networks. Moreover, the combinatorial approach presented is flexible enough to allow also a study of series-parallel networks generated by modifications of the edge-duplication rules, in particular, one could treat also a Bernoulli model with a "preferential

edge-duplication rule", or a b-ary saturation model by encoding the growth process via other recursive tree structures (edge-coloured plane increasing trees and bucket-recursive trees with bucket size $b \geq 2$, respectively); the authors plan to comment on that in a journal version of this work.

2 Series-parallel networks and description via recursive tree structures

2.1 Bernoulli model

In the Bernoulli model in step 1 one starts with a single edge labelled 1 connecting the source and the sink, and in step n, with $n > 1$, one of the $n-1$ edges of the already generated series-parallel network is chosen uniformly at random, let us assume it is edge $j = (x, y)$; then either with probability p, $0 < p < 1$, this edge is doubled in a parallel way[(i)], i.e., an additional edge (x, y) labelled n is inserted into the graph (let us say, right to edge e), or otherwise, thus with probability $q = 1 - p$, this edge is doubled in a serial way, i.e., edge (x, y) is replaced by the series of edges (x, z) and (z, y), with z a new node, where (x, z) gets the label j and (z, y) will be labelled by n.

The growth of series-parallel networks corresponds to the growth of random recursive trees, where one starts in step 1 with a node labelled 1, and in step n one of the $n-1$ nodes is chosen uniformly at random and node n is attached to it as a new child. Thus, a doubling of edge j in step n when generating the series-parallel network corresponds in the recursive tree to an attachment of node n to node j. Additionally, in order to keep the information about the kind of duplication of the chosen edge, the edge incident to n is coloured either blue encoding a parallel doubling, or coloured red encoding a serial doubling. Such combinatorial objects of edge-coloured recursive trees can be described via the formal equation

$$\mathcal{T} = \mathcal{Z}^{\square} * \mathrm{SET}(B \cdot \mathcal{T} + R \cdot \mathcal{T}),$$

with B and R markers (see [7]). Of course, one has to keep track of the number of blue and red edges to get the correct probability model according to

$$\mathbb{P}\{T \in \mathcal{T}_n \text{ is chosen}\} = \frac{p^{\#\text{blue edges of } T} \cdot q^{\#\text{red edges of } T}}{T_n},$$

where $\mathcal{T}_n = \{T \in \mathcal{T} : T \text{ has order } n\}$ and $T_n := |\mathcal{T}_n| = (n-1)!$. Throughout this work the term order of a tree T shall denote the number of labels contained in T, which, of course, for recursive trees coincides with the number of nodes of T. Then, each edge-coloured recursive tree of order n and the corresponding series-parallel network of size n occur with the same probability. An example for a series-parallel network grown via the Bernoulli model and the corresponding edge-coloured recursive tree is given in Figure 1.

2.2 Binary model

In the binary model again in step 1 one starts with a single edge labelled 1 connecting the source and the sink, and in step n, with $n > 1$ one of the $n-1$ edges of the already generated series-parallel network is chosen uniformly at random; let us assume it is edge $j = (x, y)$; but now whether edge j is doubled in a parallel or serial way is already determined by the out-degree of node x: if node x has out-degree 1 then

[(i)] In the original work [11] the rôles of p and q are switched, but we find it catchier to use p for the probability of a parallel doubling.

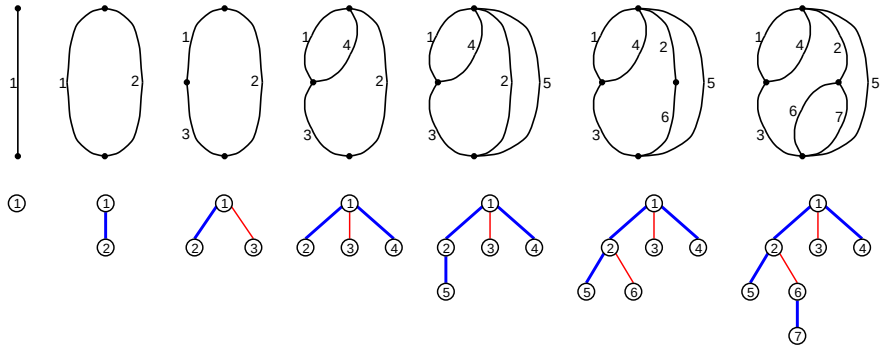


Fig. 1: Growth of a series-parallel network under the Bernoulli model and of the corresponding edge-coloured recursive tree. In the resulting graph the degree of the source is 4, the length of the leftmost source-to-sink path is 2 and there are 5 different source-to-sink paths.

we carry out a parallel doubling by inserting an additional edge (x, y) labelled n into the graph right to edge j, but otherwise, i.e., if node x has out-degree 2 and is thus already saturated, then we carry out a serial doubling by replacing edge (x, y) by the edges (x, z) and (z, y), with z a new node, where (x, z) gets the label j and (z, y) will be labelled by n.

It turns out that the growth model for binary series-parallel networks corresponds with the growth model for bucket-recursive trees of maximal bucket size 2, i.e., where nodes in the tree can hold up to two labels: in step 1 one starts with the root node containing label 1, and in step n one of the $n-1$ labels in the tree is chosen uniformly at random, let us assume it is label j, and attracts the new label n. If the node x containing label j is saturated, i.e., it contains already two labels, then a new node containing label n will be attached to x as a new child, otherwise, label n will be inserted into node x, then containing the labels j and n. As has been pointed out in [9] such random bucket-recursive trees can also be described in a combinatorial way by extending the notion of increasing trees: namely a bucket-recursive tree is either a node labelled 1 or it consists of the root node labelled $(1, 2)$, where two (possibly empty) forests of (suitably relabelled) bucket-recursive trees are attached to the root as a left forest and a right forest. A formal description of the family $\mathcal{B}$ of bucket-recursive trees (of bucket size at most 2) is in modern notation given as follows:

$$\mathcal{B} = \mathcal{Z}^{\square} + \mathcal{Z}^{\square} * (\mathcal{Z}^{\square} * (\mathrm{SET}(\mathcal{B}) * \mathrm{SET}(\mathcal{B}))) .$$

It follows from this formal description that there are $T_n = (n-1)!$ different bucket-recursive trees with n labels, i.e., of order n, and furthermore it has been shown in [9] that this combinatorial description (assuming the uniform model, where each of these trees occurs with the same probability) indeed corresponds to the stochastic description of random bucket-recursive trees of order n given before. An example for a binary series-parallel network and the corresponding bucket-recursive tree is given in Figure 2.

In our analysis of binary series-parallel networks the following link between the decomposition of a bucket-recursive tree T into its root $(1, 2)$ and the left forest (consisting of the trees $T_1^{[L]}, \ldots, T_\ell^{[L]}$) and the right forest (consisting of the trees $T_1^{[R]}, \ldots, T_r^{[R]}$), and the subblock-structure of the corresponding binary network G is important: G consists of a left half $G^{[L]}$ and a right half $G^{[R]}$ (which share the source and the sink), where $G^{[L]}$ is formed by a series of blocks (i.e., maximal 2-connected components)

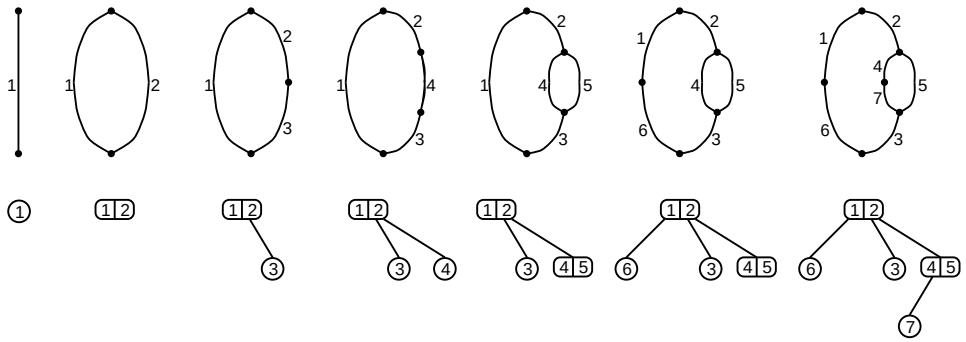


Fig. 2: Growth of a binary series-parallel network and of the corresponding bucket-recursive tree. In the resulting graph the degree of the sink is 2, the length of the leftmost source-to-sink path is 2 and there are 3 different source-to-sink paths.

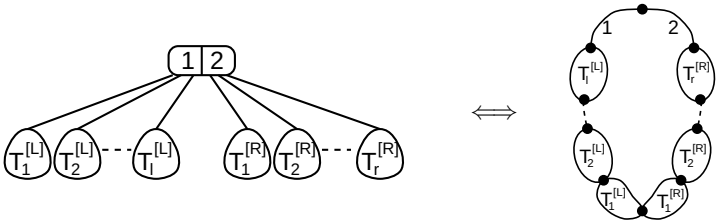


Fig. 3: Decomposition of a bucket recursive tree T into its root and the left and right forest, respectively, and the subblock-structure of the corresponding binary network.

consisting of the edge labelled 1 followed by binary networks corresponding to $T_\ell^{[L]}$, $T_{\ell-1}^{[L]}$, …, $T_1^{[L]}$, and $G^{[R]}$ is formed by a series of blocks consisting of the edge labelled 2 followed by binary networks corresponding to $T_r^{[R]}$, $T_{r-1}^{[R]}$, …, $T_1^{[R]}$; see Figure 3 for an example.

3 Uniform Bernoulli edge-duplication growth model

3.1 Degree of the source

Let $D_n = D_n(p)$ denote the r.v. measuring the degree of the source in a random series-parallel network of size n for the Bernoulli model, with $0 < p < 1$. A first analysis of this quantity has been given in [11], where the exact distribution of D_n as well as exact and asymptotic results for the expectation $\mathbb{E}(D_n)$ could be obtained. However, questions concerning the limiting behaviour of D_n and the asymptotic behaviour of higher moments of D_n have not been touched; in this context we remark that the explicit results for the probabilities $\mathbb{P}\{D_n = m\}$ as obtained in [11] and restated in (6) are not easily amenable for asymptotic studies, because of large cancellations of the alternating summands in the corresponding formula. We will reconsider this problem by applying the combinatorial approach introduced in Section 2, and in order to get limiting distribution results we apply methods from analytic combinatorics. As has been already remarked in [11] the degree of the sink is equally distributed as D_n due to symmetry reasons, although a

simple justification of this fact via direct "symmetry arguments" does not seem to be completely trivial (the insertion process itself is a priori not symmetric w.r.t. the poles, since edges are always inserted towards the sink); however, it is not difficult to show this equality by establishing and treating a recurrence for the distribution of the sink, which is here omitted.

When considering the description of the growth process of these series-parallel networks via edge-coloured recursive trees it is apparent that the degree of the source in such a graph corresponds to the order of the maximal subtree containing the root node and only blue edges, i.e., we have to count the number of nodes in the recursive tree that can be reached from the root node by taking only blue edges; for simplicity we denote this maximal subtree by "blue subtree". Thus, in the recursive tree model, D_n measures the order of the blue subtree in a random edge-coloured recursive tree of order n. To treat D_n we introduce the r.v. $D_{n,k}$, whose distribution is given as the conditional distribution $D_n\big|\{\text{the tree has exactly } k \text{ blue edges}\}$, and the trivariate generating function

$$F(z,u,v) := \sum_n \sum_k \sum_m T_n \binom{n-1}{k} \mathbb{P}\{D_{n,k}=m\} \frac{z^n}{n!} u^k v^m, \tag{1}$$

with $T_n = (n-1)!$ the number of recursive trees of order n. Thus $T_n\binom{n-1}{k}\mathbb{P}\{D_{n,k}=m\}$ counts the number of edge-coloured recursive trees of order n with exactly k blue edges, where the blue subtree has order m. Additionally we introduce the auxiliary function $N(z,u) := \sum_n \sum_k T_n\binom{n-1}{k}\frac{z^n}{n!}u^k = \frac{1}{1+u}\log\left(\frac{1}{1-z(1+u)}\right)$, i.e., the exponential generating function of the number of edge-coloured recursive trees of order n with exactly k blue edges.

The decomposition of a recursive tree into its root node and the set of branches attached to it immediately can be translated into a differential equation for $F(z,u,v)$, where we only have to take into account that the order of the blue subtree in the whole tree is one (due to the root node) plus the orders of the blue subtrees of the branches which are connected to the root node by a blue edge (i.e., only branches which are connected to the root node by a blue edge will contribute). Namely, with $F := F(z,u,v)$ and $N := N(z,u)$, we get the first order separable differential equation

$$F' = v \cdot e^{uF+N}, \tag{2}$$

with initial condition $F(0,u,v) = 0$. Throughout this work, the notation f' for (multivariate) functions $f(z,\dots)$ shall always denote the derivative w.r.t. the variable z. The exact solution of (2) can be obtained by standard means and is given as follows:

$$F(z,u,v) = \frac{1}{u}\log\left(\frac{1}{1-v+v(1-z(1+u))^{\frac{u}{1+u}}}\right). \tag{3}$$

Since we are only interested in the distribution of D_n we will actually consider the generating function

$$F(z,v) := \sum_n \sum_m T_n \mathbb{P}\{D_n=m\}\frac{z^n}{n!}v^m = \sum_n \sum_m \mathbb{P}\{D_n=m\}\frac{z^n}{n}v^m. \tag{4}$$

According to the definition of the conditional r.v. $D_{n,k}$ it holds that $\mathbb{P}\{D_n=m\} = \sum_{k=0}^{n-1}\mathbb{P}\{D_{n,k}=m\}\binom{n-1}{k}p^k q^{n-1-k}$, which, after simple computations, gives the relation $F(z,v) = \frac{1}{q}F(qz,\frac{p}{q},v)$. Thus

we obtain the following explicit formula for $F'(z,v)$, which has been obtained already in [11] by using a description of D_n via urn models:

$$F'(z,v) = \frac{v}{v(1-z)+(1-v)(1-z)^{1-p}}. \tag{5}$$

Extracting coefficients from (5) immediately yields the explicit result for the probability distribution of D_n, with $1 \leq m \leq n$, stated in [11]:

$$\mathbb{P}\{D_n = m\} = [z^{n-1}v^m]F'(z,v) = \sum_{j=0}^{m-1} \binom{m-1}{j} (-1)^{n+j-1} \binom{p(j+1)-1}{n-1}. \tag{6}$$

In order to describe the limiting distribution behaviour of D_n we first study the integer moments. To do this we introduce $\tilde{F}(z,w) := F(z,1+w)$, since we get for its derivative the relation $\tilde{F}'(z,w) = \sum_n \sum_r \mathbb{E}(D_n^r) z^{n-1} \frac{w^r}{r!}$, with $\mathbb{E}(D_n^r) = \mathbb{E}(D_n \cdot (D_n-1) \cdots (D_n-r+1))$ the r-th factorial moment of D_n. Plugging $v = 1+w$ into (5), extracting coefficients and applying Stirling's formula for the factorials easily gives the following explicit and asymptotic result for the r-th factorial moments of D_n, with $r \geq 1$:

$$\mathbb{E}(D_n^r) = r! \sum_{j=0}^{r-1} \binom{r-1}{j} (-1)^{r-1-j} \binom{n+p(j+1)-1}{n-1} \sim \frac{r! \cdot n^{rp}}{\Gamma(rp+1)},$$

from which we further deduce

$$\mathbb{E}\left(\left(\frac{D_n}{n^p}\right)^r \right) \sim \frac{r!}{\Gamma(rp+1)}. \tag{7}$$

Thus, the r-th integer moments of the suitably scaled r.v. D_n converge to the integer moments of a so-called Mittag-Leffler distribution $D = D(p)$ with parameter p (see, e.g., [8]), which, by an application of the theorem of Fréchet and Shohat, indeed characterizes the limiting distribution of D_n.

From the explicit formula (5) it is also possible to characterize the density function $f(x)$ of D (We remark that alternatively we can obtain $f(x)$ from the moment generating function $M(z) = \mathbb{E}(e^{Dz}) = \sum_{r \geq 0} \mathbb{E}(D^r) \frac{z^r}{r!}$ and applying the inverse Laplace transform.). Namely, it holds

$$\mathbb{P}\{D_n = m\} = [z^{n-1}v^m]F'(z,v) = \frac{1}{2\pi i} \oint \frac{(1-(1-z)^p)^{m-1}}{z^n(1-z)^{1-p}} dz, \tag{8}$$

where we have to choose as contour a positively oriented simple closed curve around the origin, which lies in the domain of analyticity of the integrand. To evaluate the integral asymptotically (and uniformly) for $m = O(n^{p+\delta})$, $\delta > 0$ and $n \to \infty$ one can adapt the considerations done in [15] for the particular instance $p = \frac{1}{2}$. After straightforward computations one obtains the following asymptotic equivalent of these probabilities, which determines the density function $f(x)$ of the limiting distribution D:

$$\mathbb{P}\{D_n = m\} \sim \frac{1}{n^p} \cdot \frac{1}{2\pi i} \int_{\mathcal{H}} \frac{e^{-t-\frac{m}{n^p}(-t)^p}}{(-t)^{1-p}} dt,$$

with $\mathcal{H}$ a Hankel contour starting from $e^{2\pi i}\infty$, passing around 0 and terminating at $+\infty$. The results are collected in the following theorem.

Theorem 3.1. *The degree D_n of the source or the sink in a randomly chosen series-parallel network of size n generated by the Bernoulli model converges after scaling, for $n \to \infty$, in distribution to a Mittag-Leffler distribution $D = D(p)$ with parameter p: $\frac{D_n}{n^p} \xrightarrow{(d)} D$, where D is characterized by the sequence of its r-th integer moments:*

$$\mathbb{E}(D^r) = \frac{r!}{\Gamma(rp+1)}, \quad \text{for } r \geq 0,$$

as well as by its density function $f(x)$ (with $\mathcal{H}$ a Hankel contour):

$$f(x) = \frac{1}{2\pi i} \int_{\mathcal{H}} \frac{e^{-t-x(-t)^p}}{(-t)^{1-p}} dt, \quad \text{for } x > 0.$$

We remark that after simple manipulations we can also write $f(x)$ as the following real integral:

$$f(x) = \frac{1}{\pi p} \int_0^\infty e^{-w^{\frac{1}{p}} - xw\cos(\pi p)} \sin(\pi p - xw\sin(\pi p)) dw, \quad \text{for } x > 0.$$

We further remark that for the particular instance $p = \frac{1}{2}$ one can evaluate the Hankel integral above and obtains that the limiting distribution D is characterized by the density function $f(x) = \frac{1}{\sqrt{\pi}} \cdot e^{-\frac{x^2}{4}}$, $x > 0$. Thus, $f(x)$ is the density function of a so-called half-normal distribution with parameter $\sigma = \sqrt{2}$.

3.2 Length of a random path from source to sink

We consider the length $L_n = L_n(p)$ (measured by the number of edges) of a random path from the source to the sink in a randomly chosen series-parallel network of size n for the Bernoulli model. In this context, the following definition of a random source-to-sink path seems natural: we start at the source and walk along outgoing edges, such that whenever we reach a node of out-degree d, $d \geq 1$, we choose one of these outgoing edges uniformly at random, until we arrive at the sink.

The following two observations are very helpful in the analysis of this parameter. First it holds that the length L_n of a random path is distributed as the length $L_n^{[L]}$ of the leftmost source-to-sink path in a random series-parallel network of size n; the meaning of the leftmost path is, that whenever we reach a node of out-degree d, we choose the first (i.e., leftmost) outgoing edge. Unfortunately, so far we do not see a simple symmetry argument to show this fact (such an argument easily shows that the rightmost path has the same distribution as the leftmost path, but it does not seem to explain the general situation). However, we are able to show this in a somehow indirect manner: namely, it is possible to establish a more involved distributional recurrence for the length of a random path L_n and show that the explicit solution for the probability distribution of the length of the leftmost path $L_n^{[L]}$ is indeed the solution of the recurrence for L_n. These computations will be given in the journal version of this work, here we have to omit them.

Second we use that the length of the leftmost source-to-sink path in a series-parallel network has the following simple description in the corresponding edge-coloured recursive tree: namely, an edge is lying on the leftmost source-to-sink path if and only if the corresponding node in the recursive tree can be reached from the root by using only red edges (i.e., edges that correspond to serial edges). This means that the length ℓ of the leftmost source-to-sink path corresponds in the edge-coloured recursive tree model to the order of the maximal subtree containing the root node and only red edges. If we switch the colours red and blue in the tree we obtain an edge-coloured recursive tree where the maximal blue subtree has the

same order, i.e., where the source-degree of the corresponding series-parallel network is ℓ. But switching colours in the tree model corresponds to switching the probabilities p and $q = 1 - p$ for generating a parallel and a serial edge, respectively, in the series-parallel network. Thus it simply holds $L_n^{[L]}(p) \stackrel{(d)}{=} D_n(1-p)$, where D_n denotes the source-degree in a random series-parallel network of size n.

Combining these considerations and the results for D_n obtained in Section 3.1 we obtain the following theorem.

Theorem 3.2. *The length L_n of a random path from the source to the sink in a randomly chosen series-parallel network of size n generated by the Bernoulli model has the following probability distribution:*

$$\mathbb{P}\{L_n = m\} = \sum_{j=0}^{m-1} \binom{m-1}{j} (-1)^{n+j-1} \binom{j - p(j+1)}{n-1}, \quad \textit{for } 1 \le m \le n.$$

Moreover, L_n has, for $n \to \infty$, the following limiting distribution behaviour: $\frac{L_n}{n^{1-p}} \xrightarrow{(d)} L$, where the limiting distribution L is a Mittag-Leffler distribution with parameter $1-p$, i.e., L is characterized by the sequence of its r-th integer moments:

$$\mathbb{E}(L^r) = \frac{r!}{\Gamma(r(1-p)+1)}, \quad \textit{for } r \ge 0,$$

as well as by its density function $g(x)$ (with $\mathcal{H}$ a Hankel contour):

$$g(x) = \frac{1}{2\pi i} \int_{\mathcal{H}} \frac{e^{-t - x(-t)^{1-p}}}{(-t)^p} dt, \quad \textit{for } x > 0,$$

3.3 Number of paths from source to sink

Let $P_n = P_n(p)$ denote the r.v. measuring the number of different paths from the source to the sink in a randomly chosen series-parallel network of size n for the Bernoulli model. Again we use the description of the growth of the graphs via edge-labelled recursive trees, but in contrast to the previous studies of parameters, here it seems advantageous to use an alternative decomposition of recursive trees with respect to the edge connecting nodes 1 and 2, which allows to establish a stochastic recurrence for the r.v. P_n. Namely, it is not difficult to show (see, e.g., [4]) that when starting with a random recursive tree T of order $n \ge 2$ and removing the edge $1-2$, both resulting trees T' and T'' are (after an order-preserving relabelling) again random recursive trees of smaller orders; moreover, if U_n denotes the order of the resulting tree T' rooted at the former label 2 (and thus $n - U_n$ gives the order of the tree T'' rooted at the original root of the tree T), it holds that U_n follows a discrete uniform distribution on the integers $\{1, \ldots, n-1\}$, i.e., $\mathbb{P}\{U_n = k\} = \frac{1}{n-1}$, for $1 \le k \le n-1$. Depending on the colour of the edge $1-2$ in the edge-labelled recursive tree, it corresponds to a parallel edge (colour blue, which occurs with probability p) or a serial edge (colour red, which occurs with probability $q = 1-p$) in the series-parallel network: if it is a parallel edge then the number of source-to-sink paths in the corresponding substructures have to be added, whereas for a serial edge they have to be multiplied in order to obtain the total number of source-to-sink paths in the whole graph. Thus P_n satisfies the following stochastic recurrence:

$$P_n \stackrel{(d)}{=} \mathbf{1}_{\{B_n=1\}} \cdot \left(P'_{U_n} + P''_{n-U_n}\right) + \mathbf{1}_{\{B_n=0\}} \cdot \left(P'_{U_n} \cdot P''_{n-U_n}\right), \quad \text{for } n \ge 2, \quad P_1 = 1, \tag{9}$$

where B_n and U_n are independent of each other and independent of (P_n), (P'_n) and (P''_n), and where (P'_n) and (P''_n) are independent copies of (P_n). Here B_n is the indicator variable of the event that $1-2$ is a blue edge in the recursive tree, thus B_n is a Bernoulli distributed r.v. with success probability p, i.e., $\mathbb{P}\{B_n = 1\} = p$. Furthermore, the r.v. U_n measuring the order of the subtree rooted at 2, is uniformly distributed on $\{1, 2, \ldots, n-1\}$, i.e., $\mathbb{P}\{U_n = k\} = \frac{1}{n-1}$, for $1 \leq k \leq n-1$.

Starting with (9) and taking the expectations yields after simple manipulations the following recurrence:

$$\mathbb{E}(P_n) = \frac{2p}{n-1}\sum_{k=1}^{n-1}\mathbb{E}(P_k) + \frac{1-p}{n-1}\sum_{k=1}^{n-1}\mathbb{E}(P_k)\mathbb{E}(P_{n-k}), \quad n \geq 2, \qquad \mathbb{E}(P_1) = 1. \tag{10}$$

To treat this recurrence we introduce the generating function $E(z) := \sum_{n\geq 1}\mathbb{E}(P_n)z^{n-1}$, which gives the following non-linear first order differential equation of Bernoulli type:

$$E'(z) = \frac{2p}{1-z}E(z) + (1-p)\big(E(z)\big)^2, \quad E(0) = 1. \tag{11}$$

Equation (11) can be treated by a standard technique for Bernoulli type differential equations and leads to the following solution, where we have to distinguish whether $p = \frac{1}{2}$ or not:

$$E(z) = \begin{cases} \frac{1-2p}{(1-p)(1-z)-p(1-z)^{2p}}, & \text{for } p \neq \frac{1}{2}, \\ \frac{2}{2(1-z)-(1-z)\log\left(\frac{1}{1-z}\right)}, & \text{for } p = \frac{1}{2}. \end{cases} \tag{12}$$

From the formula (12) for the generating function $E(z)$ one can easily deduce explicit results for the expected value $\mathbb{E}(P_n) = [z^{n-1}]E(z)$, which, however, due to alternating signs of the summands are not easily amenable for asymptotic considerations. Instead, in order to obtain the asymptotic behaviour of $\mathbb{E}(P_n)$ we consider the formulæ for the generating function $E(z)$ stated in (12) and describe the structure of the singularities: for $0 < p < 1$ the dominant singularity at $z = \rho < 1$ is annihilating the denominator; there $E(z)$ has a simple pole, which due to singularity analysis [7] yields the main term of $\mathbb{E}(P_n)$, i.e., the asymptotically exponential growth behaviour; the (algebraic or logarithmic) singularity at $z = 1$ determines the second and higher order terms in the asymptotic behaviour of $\mathbb{E}(P_n)$, which differ according to the ranges $0 < p < \frac{1}{2}$, $p = \frac{1}{2}$, and $\frac{1}{2} < p < 1$. This yields the following theorem.

Theorem 3.3. *The expectation $\mathbb{E}(P_n)$ of the number of paths P_n from source to sink in a random series-parallel network of size n generated by the Bernoulli model is given by the following explicit formula:*

$$\mathbb{E}(P_n) = \begin{cases} \sum_{j=0}^{n-1}(-1)^{n+j-1}\binom{(2p-1)j-1}{n-1}\sum_{k=0}^{n-1}\binom{k}{j}\left(\frac{p}{2p-1}\right)^k, & \text{for } p \neq \frac{1}{2}, \\ \sum_{k=0}^{n-1}\frac{(-1)^k}{2^k}\cdot B_k(-H_{n-1}^{(1)}, -H_{n-1}^{(2)}, -2H_{n-1}^{(3)}, \ldots, -(k-1)!H_{n-1}^{(k)}), & \text{for } p = \frac{1}{2}, \end{cases}$$

where $B_k(x_1, x_2, \ldots, x_k)$ denotes the k-th complete Bell polynomial and where $H_n^{(m)} := \sum_{j=1}^{n}\frac{1}{j^m}$ denote the m-th order harmonic numbers (see [16]).

The asymptotic behaviour of $\mathbb{E}(P_n)$ *is, for* $n \to \infty$*, given as follows:*

$$\mathbb{E}(P_n) = \frac{1}{1-p} \cdot \alpha_p^n + R_p(n),$$

where $\alpha_p = \dfrac{1}{1-\left(\frac{p}{1-p}\right)^{\frac{1}{1-2p}}}$*, for* $p \neq \dfrac{1}{2}$*, and* $\alpha_p = \dfrac{1}{1-e^{-2}} = \lim\limits_{p\to\frac{1}{2}} \dfrac{1}{1-\left(\frac{p}{1-p}\right)^{\frac{1}{1-2p}}}$*, for* $p = \dfrac{1}{2}$*,*

and $R_p(n) = -\dfrac{1-2p}{p\Gamma(2p)} n^{2p-1} + \mathcal{O}(n^{2(2p-1)})$*, for* $0 < p < \dfrac{1}{2}$*,* $R_p(n) = -\dfrac{2}{\log n} + \mathcal{O}(\dfrac{1}{\log^2 n})$*, for* $p = \dfrac{1}{2}$*,*

$R_p(n) = -\dfrac{2p-1}{1-p} + \mathcal{O}(n^{1-2p})$*, for* $\dfrac{1}{2} < p < 1$.

4 Uniform binary saturation edge-duplication growth model

4.1 Length of a random path from source to sink

We are interested in the length of a typical source-to-sink path in a series-parallel network of size n. Again, it is natural to start at the source of the graph and move along outgoing edges, in a way that whenever we have the choice of two outgoing edges we use one of them uniformly at random to enter a new node, until we finally end at the sink. Let us denote by L_n the length of such a random source-to-sink path in a random series-parallel network of size n for the binary model. Due to symmetry reasons it holds that $L_n \stackrel{(d)}{=} L_n^{[L]}$, where $L_n^{[L]}$ denotes the length of the leftmost source-to-sink path in a random series-parallel network of size n, i.e., the source-to-sink path, where in each node we choose the left outgoing edge to enter the next node.

In order to analyse $L_n^{[L]}$ we use the description of the growth of series-parallel networks via bucket-recursive trees: the length of the left path is equal to 1 (coming from the root node of the tree, i.e., stemming from the edge 1 in the graph) plus the sum of the lengths of the left paths in the subtrees contained in the left forest (which correspond to the blocks of the left part of the graph). When we introduce the generating function

$$F(z,v) := \sum_{n\geq 1}\sum_{m\geq 0} T_n \mathbb{P}\{L_n = m\}\frac{z^n}{n!}v^m = \sum_{n\geq 1}\sum_{m\geq 0} \mathbb{P}\{L_n = m\}\frac{z^n}{n}v^m, \tag{13}$$

then the above description yields the following differential equation:

$$F''(z,v) = ve^{F(z,v)}e^{N(z)} = \frac{v}{1-z}e^{F(z,v)}, \quad F(0,v) = 0, \quad F'(0,v) = v, \tag{14}$$

where $N(z) = \log\frac{1}{1-z}$ is the exponential generating function of the number $T_n = (n-1)!$ of bucket-recursive trees of order n. In order to compute the expectation we consider $E(z) := \frac{\partial}{\partial v}F(z,v)\big|_{v=1} = \sum_{n\geq 1}\mathbb{E}(L_n)\frac{z^n}{n}$, which satisfies the following linear second order differential equation of Eulerian type:

$$E''(z) = \frac{1}{(1-z)^2}E(z) + \frac{1}{(1-z)^2}, \quad E(0) = 0, \quad E'(0) = 1.$$

The explicit solution can be obtained by a standard technique and is given as follows:

$$E(z) = \frac{3+\sqrt{5}}{2\sqrt{5}}\frac{1}{(1-z)^{\frac{\sqrt{5}-1}{2}}} - \frac{3-\sqrt{5}}{2\sqrt{5}}(1-z)^{\frac{1+\sqrt{5}}{2}} - 1.$$

Extracting coefficients and applying Stirling's formula immediately yields the following explicit and asymptotic result for the expectation:

$$\mathbb{E}(L_n) = n\left(\frac{3+\sqrt{5}}{2\sqrt{5}}\binom{n+\frac{\sqrt{5}}{2}-\frac{3}{2}}{n} - \frac{3-\sqrt{5}}{2\sqrt{5}}\binom{n-\frac{\sqrt{5}}{2}-\frac{3}{2}}{n}\right) \sim \frac{1+\sqrt{5}}{2\sqrt{5}}\frac{n^{\frac{\sqrt{5}-1}{2}}}{\Gamma(\frac{\sqrt{5}-1}{2})}. \tag{15}$$

In order to characterize the limiting distribution of L_n we will compute iteratively the asymptotic behaviour of all its integer moments. To this aim it is advantageous to consider $G(z,v) := F'(z,v)$. Differentiating (14) shows that $G(z,v)$ satisfies the following differential equation:

$$G''(z,v) = G'(z,v)G(z,v) + \frac{1}{1-z}G'(z,v), \quad G(0,v) = v, \quad G'(0,v) = v. \tag{16}$$

Introducing $M_r(z) := \frac{\partial^r}{\partial v^r}G(z,v)\big|_{v=1} = \sum_{n\geq 1}\mathbb{E}(L_n^r)z^{n-1}$, differentiating (16) r times w.r.t. v and evaluating at $v=1$ yields

$$M_r''(z) = \frac{2}{1-z}M_r'(z) + \frac{1}{(1-z)^2}M_r(z) + R_r(z),$$

with $R_r(z) = \sum_{k=1}^{r-1}\binom{r}{k}M_k'(z)M_{r-k}(z)$. Thus $M_r(z)$ satisfies for each r an Eulerian differential equation, where the inhomogeneous part $R_r(z)$ depends on the functions $M_k(z)$, with $k<r$. The asymptotic behaviour of $M_r(z)$ around the dominant singularity $z=1$ can be established inductively, namely it holds:

$$M_r(z) \sim \frac{c_r}{(1-z)^{r\tilde{\phi}+1}},$$

with $\tilde{\phi} = \frac{\sqrt{5}-1}{2}$, and where the constants c_r satisfy a certain recurrence of "convolution type". Singularity analysis and an application of the theorem of Fréchet and Shohat shows then the following limiting distribution result.

Theorem 4.1. *The length L_n of a random path from the source to the sink in a random series-parallel network of size n generated by the binary model satisfies, for $n\to\infty$, the following limiting distribution behaviour, with $\tilde{\phi} = \frac{\sqrt{5}-1}{2}$:*

$$\frac{L_n}{n^{\tilde{\phi}}} \xrightarrow{(d)} L,$$

where the limiting distribution L is characterized by its sequence of r-th integer moments via

$$\mathbb{E}(L^r) = \frac{r!\cdot\tilde{c}_r}{\Gamma(r\tilde{\phi}+1)}, \quad r\geq 0,$$

and where the sequence $\tilde{c}_r$ satisfies the recurrence $\tilde{c}_r = \frac{1}{\tilde{\phi}(r-1)((r+1)\tilde{\phi}+1)}\sum_{k=1}^{r-1}(k\tilde{\phi}+1)\tilde{c}_k\tilde{c}_{r-k}$, for $r\geq 2$, with $\tilde{c}_0 = 1$ and $\tilde{c}_1 = \frac{3+\tilde{\phi}}{5}$.

4.2 Degree of the sink

Whereas the (out-)degree of the source of a binary series-parallel network is two (if the graph has at least two edges), typically the (in-)degree of the sink is quite large, as will follow from our treatments. Let us denote by D_n the degree of the sink in a random series-parallel network of size n for the binary model. For a binary series-parallel network, the value of this parameter can be determined recursively by adding the degrees of the sinks in the last block of each half of the graph; in the case that a half only consists of one edge then the contribution of this half is of course 1. When considering the corresponding bucket-recursive tree this means that the degree of the sink can be computed recursively by adding the contributions of the left and the right forest attached to the root, where the contribution of a forest is either given by 1 in case that the forest is empty (then the corresponding root node contributes to the degree of the sink) or it is the contribution of the first tree in the forest (which corresponds to the last block), see Figure 3. Introducing the generating functions

$$F(z,v) := \sum_{n\geq 1}\sum_{m\geq 0} T_n \mathbb{P}\{D_n = m\}\frac{z^n}{n!}v^m, \quad A(z,v) := \sum_{n\geq 0}\sum_{m\geq 0} \tilde{T}_n \mathbb{P}\{\tilde{D}_n = m\}\frac{z^n}{n!}v^m, \tag{17}$$

with $\tilde{D}_n$ denoting the corresponding quantity for the left or right forest and $\tilde{T}_n = n!$ counting the number of forests of order n, the combinatorial decomposition of bucket-recursive trees yields the following system of differential equations:

$$F''(z,v) = \big(A(z,v)\big)^2, \quad A'(z,v) = \frac{1}{1-z}\cdot F'(z,v). \tag{18}$$

From system (18) the following non-linear differential equation for $F(z,v)$ can be obtained:

$$F'''(z,v) = \frac{2}{1-z}\sqrt{F''(z,v)}F'(z,v), \quad F(0,v) = 0, F'(0,v)v, F''(0,v) = v^2,$$

which, by considering $E(z) := \frac{\partial}{\partial v}F(z,v)\big|_{v=1}$ and solving an Eulerian differential equation, allows to compute an exact and asymptotic expression for the expectation; namely it holds

$$\mathbb{E}(D_n) = \frac{1+\sqrt{2}}{2}\binom{n+\sqrt{2}-2}{n-1} - \frac{\sqrt{2}-1}{2}\binom{n-\sqrt{2}-2}{n-1} \sim \frac{1+\sqrt{2}}{2}\frac{n^{\sqrt{2}-1}}{\Gamma(\sqrt{2})}. \tag{19}$$

However, for asymptotic studies of higher moments it seems to be advantageous to consider the following second order non-linear differential equation for $A(z,v)$, which follows immediately from (18):

$$A''(z,v) = \frac{1}{1-z}A'(z,v) + \frac{1}{1-z}\big(A(z,v)\big)^2, \quad A(0,v) = v, \quad A'(0,v) = v. \tag{20}$$

Introducing the functions $\tilde{M}_r(z) := \frac{\partial^r}{\partial v^r}A(z,v)\big|_{v=1}$ and differentiating (20) r times, one obtains that $\tilde{M}_r(z)$ satisfies for $r \geq 1$ the following second order Eulerian differential equation:

$$\tilde{M}_r''(z) = \frac{1}{1-z}\tilde{M}_r'(z) + \frac{2}{(2-z)^2}\tilde{M}_r(z) + R_r(z), \tag{21}$$

with $R_r(z) := \frac{1}{1-z}\sum_{k=1}^{r-1}\binom{r}{k}\tilde{M}_k(z)\tilde{M}_{r-k}(z)$. From (21) one can inductively show that the local behaviour of the functions $\tilde{M}_r(z)$ around the (unique) dominant singularity $z=1$ is given as follows:

$$\tilde{M}_r(z) \sim \frac{c_r}{(1-z)^{r\sqrt{2}-(r-1)}}, \quad r \geq 1,$$

where the constants c_r are determined recursively. Actually we are interested in the functions $M_r(z) := \frac{\partial^r}{\partial v^r}F(z,v)\big|_{v=1} = \sum_{n\geq 1}\mathbb{E}(D_n^r)\frac{z^n}{n}$, which are, due to (18), related to $\tilde{M}_r(z)$ via $M_r''(z) = (1-z)\tilde{M}_r''(z) - \tilde{M}_r'(z)$. Singularity analysis as well as the theorem of Fréchet and Shohat show then the following limiting distribution result.

Theorem 4.2. *The degree D_n of the sink in a random series-parallel network of size n generated by the binary model satisfies, for $n\to\infty$, the following limiting distribution behaviour:*

$$\frac{D_n}{n^{\sqrt{2}-1}} \xrightarrow{(d)} D,$$

where the limiting distribution D is characterized by its sequence of r-th integer moments via

$$\mathbb{E}(D^r) = \frac{r!(r(\sqrt{2}-1)+1)\tilde{c}_r}{\Gamma(r(\sqrt{2}-1)+1)}, \quad r \geq 0,$$

where the sequence $\tilde{c}_r$ satisfies the recurrence $\tilde{c}_r = \frac{1}{(r(\sqrt{2}-1)+1)^2-2}\sum_{k=1}^{r-1}\tilde{c}_k\tilde{c}_{r-k}$, for $r\geq 2$, with $\tilde{c}_0 = 1$ and $\tilde{c}_1 = \frac{1+\sqrt{2}}{2\sqrt{2}}$.

4.3 Number of paths from source to sink

As for the Bernoulli model we are interested in results concerning the number of different paths from the source to the sink in a series-parallel network and denote by P_n the number of source-to-sink paths in a random series-parallel network of size n for the binary model. In order to study P_n it seems advantageous to start with a stochastic recurrence for this random variable obtained by decomposing the bucket-recursive tree into the root node and the left and right forest (of bucket-recursive trees) attached to the root node. As auxiliary r.v. we introduce Q_n, which denotes the number of source-to-sink paths in the series-parallel network corresponding to a forest (i.e., a set) of bucket-recursive trees, where each tree in the forest corresponds to a subblock in the left or right half of the graph. By decomposing the forest into its leftmost tree and the remaining set of trees and taking into account that the number of source-to-sink paths in the forest is the product of the number of source-to-sink paths in the leftmost tree and the corresponding paths in the remaining forest, we obtain the following system of stochastic recurrences:

$$P_n \overset{(d)}{=} Q'_{U_n} + Q''_{n-2-U_n}, \quad \text{for } n\geq 2, \qquad Q_n \overset{(d)}{=} P'_{V_n}\cdot Q'''_{n-V_n}, \quad \text{for } n\geq 1, \tag{22}$$

with $P_0 = 0, P_1 = 1, Q_0 = 1$, and where the r.v. U_n and V_n are independent of each other and independent of (P_n), $(P_n)'$, (Q_n), $(Q_n)'$, $(Q_n)''$ and $(Q_n)'''$. Furthermore, they are distributed as follows:

$$\mathbb{P}\{U_n = k\} = \frac{1}{n-1}, \quad 0\leq k\leq n-2, \qquad \mathbb{P}\{V_n = k\} = \frac{1}{n}, \quad 1\leq k\leq n.$$

Introducing $E_n := \mathbb{E}(P_n)$ and $\tilde{E}_n := \mathbb{E}(Q_n)$, the stochastic recurrence above yields the following system of equations for E_n and $\tilde{E}_n$ (with $E_0 = 0$, $E_1 = 1$ and $\tilde{E}_0 = 1$):

$$E_n = \frac{2}{n-1}\sum_{k=0}^{n-2} \tilde{E}_k, \quad n \geq 2, \qquad \tilde{E}_n = \frac{1}{n}\sum_{k=1}^{n} E_k \tilde{E}_{n-k}, \quad n \geq 1.$$

Introducing $E(z) := \sum_{n\geq 1} E_n z^{n-1}$ and $\tilde{E}(z) := \sum_{n\geq 0} \tilde{E}_n z^n$ one obtains that $E(z)$ satisfies the following non-linear second order differential equation:

$$E''(z) = \frac{1}{1-z}E'(z) + E(z)E'(z), \quad E(0) = 1, \quad E'(0) = 2. \tag{23}$$

Differential equation (23) is not explicitly solvable; furthermore, the so-called Frobenius method to determine a singular expansion fails for $E(z)$. However, it is possible to apply the so-called Psi-series method in the setting introduced in [3], i.e., assuming a logarithmic Psi-series expansion of $E(z)$ when z lies near the (unique) dominant singularity ρ on the positive real axis. This yields the following result.

Theorem 4.3. *The expectation $\mathbb{E}(P_n)$ of the number P_n of paths from source to sink in a random series-parallel network of size n generated by the binary model has, for $n \to \infty$, the following asymptotic behaviour, with $\rho \approx 0.89\ldots$:*

$$\mathbb{E}(P_n) = \frac{2}{\rho^n} \cdot \left(1 - \frac{\rho^2}{(\rho-1)^2(n-1)(n-2)} + \mathcal{O}\Big(\frac{\log n}{n^4}\Big)\right).$$

References

[1] O. Bodini, M. Dien, X. Fontaine, A. Genitrini, and H.-K. Hwang, Increasing diamonds, accepted for *LATIN 2016*, 2015.

[2] A. Brandstädt, V. B. Le, and J. Spinrad, *Graph classes: a survey*, SIAM Monographs on Discrete Mathematics and Applications 3, SIAM, Philadelphia, 1999.

[3] H.-H. Chern, M.-I. Fernández-Camacho, H.-K. Hwang, and C. Martínez, Psi-series method for equality of random trees and quadratic convolution recurrences, *Random Structures & Algorithms* 44, 67–108, 2014.

[4] R. Dobrow and J. Fill, Total path length for random recursive trees, *Combinatorics, Probability and Computing* 8, 317—333, 1999.

[5] M. Drmota, M. Fuchs, and Y.-W. Lee, Stochastic analysis of the extra clustering model for animal grouping, *Journal of Mathematical Biology*, to appear.

[6] M. Drmota, O. Giménez, and M. Noy, Vertices of given degree in series-parallel graphs, *Random Structures & Algorithms* 36, 273–314, 2010.

[7] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.

[8] S. Janson, Moments of Gamma type and the Brownian supremum process area, *Probability Surveys* 7, 1—52, 2010.

[9] M. Kuba and A. Panholzer, A combinatorial approach to the analysis of bucket recursive trees, *Theoretical Computer Science* 411, 3255–3273, 2010.

[10] H. Mahmoud, *Pólya urn models*, Texts in Statistical Science Series, CRC Press, Boca Raton, 2009.

[11] H. Mahmoud, Some node degree properties of series-parallel graphs evolving under a stochastic growth model, *Probability in the Engineering and Informational Sciences* 27, 297–307, 2013.

[12] H. Mahmoud, Some properties of binary series-parallel graphs, *Probability in the Engineering and Informational Sciences* 28, 565–572, 2014.

[13] R. Neininger and L. Rüschendorf, A general limit theorem for recursive algorithms and combinatorial structures, *Annals of Applied Probability* 14, 378–418, 2004.

[14] R. Neininger and H. Sulzbach, On a functional contraction method, *Annals of Probability* 43, 1777–1822, 2015.

[15] A. Panholzer, The distribution of the size of the ancestor-tree and of the induced spanning subtree for random trees, *Random Structures & Algorithms* 25, 179–207, 2004.

[16] D. Zave, A series expansion involving the harmonic numbers, *Information Processing Letters* 5, 75–77, 1976.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Corners in tree–like tableaux

Paweł Hitczenko[1†] and Amanda Lohss[1]

[1] *Department of Mathematics, Drexel University, Philadelphia, PA 19104, USA*

Abstract. In this paper, we study tree–like tableaux, combinatorial objects which exhibit a natural tree structure and are connected to the partially asymmetric simple exclusion process (PASEP). There was a conjecture made on the total number of corners in tree–like tableaux and the total number of corners in symmetric tree–like tableaux. In this paper, we prove the first conjecture leaving the proof of the second conjecture to the full version of this paper. Our proofs are based on the bijection with permutation tableaux or type–B permutation tableaux and consequently, we also prove results for these tableaux.

Keywords: Tree–like tableaux, permutation tableaux, type–B permutation tableaux

1 Introduction

Tree–like tableaux are relatively new objects which were introduced in Aval et al. (2013). They are in bijection with permutation tableaux and alternative tableaux but are interesting in their own right as they exhibit a natural tree structure (see Aval et al. (2013)). They also provide another avenue in which to study the partially asymmetric simple exclusion process (PASEP), an important model from statistical mechanics. See Aval et al. (2013) and Laborde Zubieta (2015a) for more details on the connection between tree–like tableaux and the PASEP. See also Burstein (2007), Corteel and Nadeau (2009), Corteel and Williams (2007b), Corteel and Williams (2007a), Nadeau (2011), Steingrímsson and Williams (2007) and Viennot (2008) for more details on permutation and alternative tableaux.

In Laborde Zubieta (2015a), the expected number of occupied corners in tree–like tableaux and the number of occupied corners in symmetric tree–like tableaux were computed (see Section 2 for definitions). In addition, it was conjectured (see Conjectures 4.1 and 4.2 in Laborde Zubieta (2015a)) that the total number of corners in tree–like tableaux of size n is $n! \times \frac{n+4}{6}$ and the total number of corners in symmetric tree–like tableaux of size $2n+1$ is $2^n \times n! \times \frac{4n+13}{12}$.

We have proven both conjectures and in this paper, we will present the proof of the first conjecture (note that Laborde Zubieta (2015b) was able to prove the first conjecture independently using a different method). The proof of the second conjecture will be given in the full version of this paper Hitczenko and Lohss (2015). Our proofs are based on the bijection with permutation tableaux or type–B permutation tableaux and consequently, we also have results for these tableaux (see Theorems 4 and 11 below for precise statements).

[†] Partially supported by a grant from Simons Foundation (grant #208766)

The rest of the paper is organized as follows. In the next section we introduce the necessary definitions and notation. Section 3 contains the proof of the conjecture for tree–like tableaux. Section 4 develops the tools necessary to prove the second conjecture for symmetric tree–like tableaux. The proof then follows similarly to the proof of the first conjecture and will be left to the full version of this paper Hitczenko and Lohss (2015).

2 Preliminaries

A *Ferrers diagram*, F, is a left–aligned sequence of cells with weakly decreasing rows. The *half–perimeter* of F is the number of rows plus the number of columns. The *border edges* of a Ferrers diagram are the edges of the southeast border, and the number of border edges is equal to the half–perimeter. We will occasionally refer to a border edge as a step (south or west). A *shifted Ferrers diagram* is a diagram obtained from a Ferrers diagram with k columns by adding k rows above it of lengths $k, (k-1), \ldots, 1$, respectively. The half–perimeter of the shifted Ferrers diagram is the same as the original Ferrers diagram (and similarly, the border edges are the same). The right–most cells of added rows are called *diagonal cells*.

Let us recall the following two definitions introduced in Aval et al. (2013) and Steingrímsson and Williams (2007), respectively.

Definition 1 *A tree–like tableau of size n is a Ferrers diagram of half-perimeter $n+1$ with some cells (called pointed cells) filled with a point according to the following rules:*

1. *The cell in the first column and first row is always pointed (this point is known as the root point).*
2. *Every row and every column contains at least one pointed cell.*
3. *For every pointed cell, all the cells above are empty or all the cells to the left are empty.*

Definition 2 *A permutation tableau of size n is a Ferrers diagram of half–perimeter n filled with 0's and 1's according to the following rules:*

1. *There is at least one 1 in every column.*
2. *There is no 0 with a 1 above it and a 1 to the left of it simultaneously.*

We will also need a notion of type–B tableaux originally introduced in Lam and Williams (2008). Our definition follows a more explicit description given in (Corteel and Kim, 2011, Section 4).

Definition 3 *A type–B permutation tableau of size n is a shifted Ferrers diagram of half–perimeter n filled with 0's and 1's according to the following rules:*

1. *There is at least one 1 in every column.*
2. *There is no 0 with a 1 above it and a 1 to the right of it simultaneously.*
3. *If one of the diagonal cells contains a 0 (called a diagonal 0), then all the cells in that row are 0.*

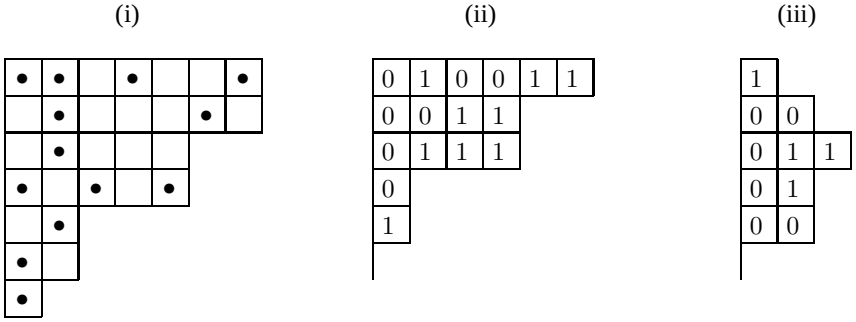


Fig. 1: (i) A tree–like tableau of size 13. (ii) A permutation tableau of size 12. (iii) A type-B permutation tableau of size 6.

Let $\mathcal{T}_n$ be the set of all tree–like tableaux of size n, $\mathcal{P}_n$ denote the set of all permutation tableaux of size n, and $\mathcal{B}_n$ denote the set of all type–B permutation tableaux of size n. In addition to these tableaux, we are also interested in *symmetric tree–like tableaux*, a subset of tree–like tableaux which are symmetric about their main diagonal (see (Aval et al., 2013, Section 2.2) for more details). As noticed in Aval et al. (2013), the size of a symmetric tree–like tableaux must be odd, and thus, we let $\mathcal{T}^{sym}_{2n+1}$ denote the set of all symmetric tree–like tableaux of size $2n+1$. It is a well–known fact that $|\mathcal{P}_n| = n!$ and $|\mathcal{B}_n| = 2^n n!$. Consequently, $|\mathcal{T}_n| = n!$ and $|\mathcal{T}^{sym}_{2n+1}| = 2^n n!$ since by Aval et al. (2013), there are bijections between these objects. We let $\mathcal{X}_n \in \{\mathcal{T}_n, \mathcal{T}^{sym}_{2n+1}, \mathcal{P}_n, \mathcal{B}_n\}$ be any of the four sets of tableaux defined above.

In permutation tableaux and type–B permutation tableaux, a *restricted* 0 is a 0 which has a 1 above it in the same column. An *unrestricted row* is a row which does not contain any restricted 0's (and for type–B permutation tableaux, also does not contain a diagonal 0). We let $U_n(T)$ denote the number of unrestricted rows in a tableau T of size n. It is also convenient to denote a topmost 1 in a column by 1_T and a right-most restricted 0 by 0_R.

Corners of a Ferrers diagram (or the associated tableau) are the cells in which both the right and bottom edges are border edges (i.e. a south step followed by a west step). In tree–like tableaux (symmetric or not) *occupied corners* are corners that contain a point.

Our proofs will rely on techniques developed in Corteel and Hitczenko (2007) (see also Hitczenko and Janson (2010)). These two papers used probabilistic language and we adopt it here, too. Thus, instead of talking about the number of corners in tableaux we let $\mathbb{P}_n$ be a probability distribution on $\mathcal{X}_n$ defined by

$$\mathbb{P}_n(T) = \frac{1}{|\mathcal{X}_n|}, \quad T \in \mathcal{X}_n, \tag{1}$$

and we consider a random variable C_n on the probability space $(\mathcal{X}_n, \mathbb{P}_n)$ defined by

$$C_n(T) = k \quad \text{if and only if } T \text{ has } k \text{ corners}, \quad T \in \mathcal{X}_n, \quad k \geq 0.$$

For convenience, let S_k indicate that the k^{th} step (border edge) is south and W_k indicate that the k^{th} step is west. Thus,

$$C_n = \sum_{k=1}^{n-1} I_{S_k, W_{k+1}}, \tag{2}$$

where I_A is the indicator random variable of the event A.

A tableau chosen from $\mathcal{X}_n$ according to the probability measure $\mathbb{P}_n$ is usually referred to as a random tableau of size n and C_n is referred to as the number of corners in a random tableau of size n. We let $\mathbb{E}_n$ denote the expected value with respect to the measure $\mathbb{P}_n$. If $c(\mathcal{X}_n)$ denotes the total number of corners in tableaux in $\mathcal{X}_n$ then, in view of (1), we have the following simple relation:

$$\mathbb{E}_n C_n = \frac{c(\mathcal{X}_n)}{|\mathcal{X}_n|} \quad \text{or, equivalently,} \quad c(\mathcal{X}_n) = |\mathcal{X}_n|\, \mathbb{E}_n C_n. \tag{3}$$

3 Corners in Tree-Like Tableaux

The main result of this section is the proof of the first conjecture of Laborde Zubieta.

Theorem 1 *(see (Laborde Zubieta, 2015a, Conjecture 4.1)) For $n \geq 2$ we have*

$$c(\mathcal{T}_n) = n! \times \frac{n+4}{6}.$$

To prove this, we will use the bijection between tree–like tableaux and permutation tableaux. According to Proposition 1.3 of Aval et al. (2013), there exists a bijection between permutation tableaux and tree–like tableaux which transforms a tree–like tableau of shape F to a permutation tableau of shape F' which is obtained from F by removing the SW–most edge from F and the cells of the left–most column (see Figure 2).

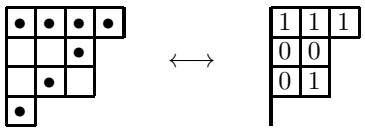


Fig. 2: An example of the bijection between permutation tableaux and tree–like tableaux of size 7.

The number of corners in F is the same as the number of corners in F' if the last edge of F' is horizontal and it is one more than the number of corners in F' if the last edge of F' is vertical. Furthermore, as is clear from a recursive construction described in (Corteel and Hitczenko, 2007, Section 2), any permutation tableau of size n whose last edge is vertical is obtained as the unique extension of a permutation tableau of size $n-1$. Therefore, there are $(n-1)!$ such tableaux and we have a simple relation

$$c(\mathcal{T}_n) = c(\mathcal{P}_n) + |\{P \in \mathcal{P}_n : \ S_n\}| = c(\mathcal{P}_n) + (n-1)!. \tag{4}$$

Thus, it suffices to determine the number of corners in permutation tableaux of size n. Since $|\mathcal{P}_n| = n!$, Equation (3) becomes

$$c(\mathcal{P}_n) = n!\, \mathbb{E}_n C_n. \tag{5}$$

In order to determine the number of corners in permutation tableaux, we first have the following result.

Theorem 2 *For permutation tableaux of size n, the probability of having a corner with border edges k and $k+1$ is given by*

$$\mathbb{P}_n\left(I_{S_k,W_{k+1}}\right) = \frac{n-k+1}{n} - \frac{(n-k)^2}{n(n-1)}.$$

Proof: The theorem can be proven by using techniques developed in Corteel and Hitczenko (2007). Specifically, if $k+1 \leq n-1$ then $I_{S_k,W_{k+1}}$ is a random variable on $\mathcal{P}_{n-1}$ (denoted by $\mathcal{T}_{n-1}$ in Corteel and Hitczenko (2007) and Hitczenko and Janson (2010)). A relationship between the measures on $\mathcal{P}_n$ and $\mathcal{P}_{n-1}$ was derived in Corteel and Hitczenko (2007) and is given by (see (Corteel and Hitczenko, 2007, Equation (7)) and (Hitczenko and Janson, 2010, Section 2, Equation (2.1))),

$$\mathbb{E}_n X_{n-1} = \frac{1}{n}\mathbb{E}_{n-1}(2^{U_{n-1}} X_{n-1}) \tag{6}$$

where X_{n-1} is any random variable defined on $\mathcal{P}_{n-1}$.

Therefore,

$$\begin{aligned}\mathbb{P}_n\left(I_{S_k,W_{k+1}}\right) &= \mathbb{E}_n\left(I_{S_k,W_{k+1}}\right) = \frac{1}{n}\mathbb{E}_{n-1}\left(2^{U_{n-1}} I_{S_k,W_{k+1}}\right)\\ &= \frac{1}{n}\mathbb{E}_{n-1}\mathbb{E}\left(2^{U_{n-1}} I_{S_k,W_{k+1}}|\mathcal{F}_{n-2}\right),\end{aligned}$$

where $\mathcal{F}_{n-2}$ is a σ–subalgebra on $\mathcal{P}_{n-1}$ obtained by grouping into one set all tableaux in $\mathcal{P}_{n-1}$ that are obtained by extending the same tableau in $\mathcal{P}_{n-2}$ (we refer to (Hitczenko and Janson, 2010, Section 2) for a detailed explanation). Now, if $k+1 \leq n-2$ then $I_{S_k,W_{k+1}}$ is measurable with respect to the σ-algebra $\mathcal{F}_{n-2}$. Thus by the properties of conditional expectation the above is:

$$\mathbb{E}_n\left(I_{S_k,W_{k+1}}\right) = \frac{1}{n}\mathbb{E}_{n-1} I_{S_k,W_{k+1}}\mathbb{E}\left(2^{U_{n-1}}|\mathcal{F}_{n-2}\right).$$

By (Corteel and Hitczenko, 2007, Equation (4)), the conditional distribution of U_n given U_{n-1} is given by

$$\mathcal{L}(U_n|\mathcal{F}_{n-1}) = 1 + \mathrm{Bin}(U_{n-1}),$$

where $\mathrm{Bin}(m)$ denotes a binomial random variable with parameters m and $1/2$. By this result and the fact that $\mathbb{E}a^{\mathrm{Bin}(m)} = \left(\frac{a+1}{2}\right)^m$,

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{n-1} I_{S_k,W_{k+1}}\mathbb{E}\left(2^{U_{n-1}}|\mathcal{F}_{n-2}\right) &= \frac{1}{n}\mathbb{E}_{n-1} I_{S_k,W_{k+1}}\mathbb{E}\left(2^{1+\mathrm{Bin}(U_{n-2})}|\mathcal{F}_{n-2}\right)\\ &= \frac{2}{n}\mathbb{E}_{n-1} I_{S_k,W_{k+1}}\left(\frac{3}{2}\right)^{U_{n-2}}\\ &= \frac{2}{n(n-1)}\mathbb{E}_{n-2} I_{S_k,W_{k+1}} 3^{U_n-2}\end{aligned} \tag{7}$$

where the last step follows from (6). Iterating $(n-1)-(k+1)$ times, we obtain

$$\frac{2\cdot 3\cdot\cdots\cdot(n-k-1)}{n(n-1)\cdot\cdots\cdot(k+2)}\mathbb{E}_{k+1} I_{S_k,W_{k+1}}(n-k)^{U_{k+1}}. \tag{8}$$

Thus, we need to compute

$$\mathbb{E}_{k+1} I_{S_k, W_{k+1}} (n-k)^{U_{k+1}} \tag{9}$$

for $1 \leq k \leq n-1$ (note that $k+1 = n$ gives $\mathbb{E}_n I_{S_{n-1}, W_n}$ which is exactly the summand omitted earlier by the restriction $k+1 \leq n-1$). This can be computed as follows. First, by the tower property of the conditional expectation and the fact that S_k is $\mathcal{F}_k$–measurable, we obtain

$$\mathbb{E}_{k+1} I_{S_k, W_{k+1}} (n-k)^{U_{k+1}} = \mathbb{E}_{k+1} I_{S_k} \mathbb{E}(I_{W_{k+1}} (n-k)^{U_{k+1}} | \mathcal{F}_k).$$

And now

$$\mathbb{E}(I_{W_{k+1}} (n-k)^{U_{k+1}} | \mathcal{F}_k) = \mathbb{E}((n-k)^{U_{k+1}} | \mathcal{F}_k) - \mathbb{E}(I_{S_{k+1}} (n-k)^{U_{k+1}} | \mathcal{F}_k)$$

because the two indicators are complementary. The first conditional expectation on the right–hand side, by a computation similar to (7) (see also (Hitczenko and Janson, 2010, Equation (2.2))) is

$$(n-k)\mathbb{E}\left((n-k)^{U_{k+1}} | \mathcal{F}_k\right) = (n-k)\left(\frac{n-k+1}{2}\right)^{U_k}. \tag{10}$$

To compute the second conditional expectation, note that on the set S_{k+1}, $U_{k+1} = 1 + U_k$ so that

$$\begin{aligned} \mathbb{E}(I_{S_{k+1}} (n-k)^{U_{k+1}} | \mathcal{F}_k) &= (n-k)^{1+U_k} \mathbb{E}(I_{S_{k+1}} | \mathcal{F}_k) \\ &= (n-k)^{1+U_k} \mathbb{P}(I_{S_{k+1}} | \mathcal{F}_k) \\ &= (n-k)^{1+U_k} \frac{1}{2^{U_k}} \end{aligned}$$

where the last equation follows from the fact that for every tableau $P \in \mathcal{P}_k$ only one of its $2^{U_k(P)}$ extensions to a tableau in $\mathcal{P}_{k+1}$ has S_{k+1} (see Corteel and Hitczenko (2007); Hitczenko and Janson (2010) for more details and further explanation). Combining with (10) yields

$$\mathbb{E}(I_{W_{k+1}} (n-k)^{U_{k+1}} | \mathcal{F}_k) = (n-k)\left(\left(\frac{n-k+1}{2}\right)^{U_k} - \left(\frac{n-k}{2}\right)^{U_k}\right)$$

and thus (9) equals

$$(n-k)\mathbb{E}_{k+1}\left(I_{S_k}\left(\left(\frac{n-k+1}{2}\right)^{U_k} - \left(\frac{n-k}{2}\right)^{U_k}\right)\right).$$

The expression inside the expectation is a random variable on $\mathcal{P}_k$ so that we can use the same argument as above (based on (Corteel and Hitczenko, 2007, Equation 5) or (Hitczenko and Janson, 2010, Equation (2.1))) to reduce the size by one and obtain that the expression above is

$$\frac{n-k}{k+1}\mathbb{E}_k I_{S_k}\left((n-k+1)^{U_k} - (n-k)^{U_k}\right).$$

Furthermore, on the set S_k, $U_k = U_{k-1} + 1$ so that the above is

$$\frac{n-k}{k+1}\mathbb{E}_k\left(\left((n-k+1)^{1+U_{k-1}} - (n-k)^{1+U_{k-1}}\right)\mathbb{E}(I_{S_k} | \mathcal{F}_{k-1})\right),$$

which, by the same argument as above, equals

$$\frac{n-k}{k+1}\mathbb{E}_k\left(\left((n-k+1)^{1+U_{k-1}}-(n-k)^{1+U_{k-1}}\right)\frac{1}{2^{U_{k-1}}}\right).$$

After reducing the size one more time we obtain

$$\frac{n-k}{(k+1)k}\left(\mathbb{E}_{k-1}\,(n-k+1)^{1+U_{k-1}}-\mathbb{E}_{k-1}\,(n-k)^{1+U_{k-1}}\right). \tag{11}$$

As computed in (Hitczenko and Janson, 2010, Equation (2.4)) for a positive integer m the generating function of U_m is given by

$$\mathbb{E}_m z^{U_m}=\frac{\Gamma(z+m)}{\Gamma(z)m!}.$$

(There is an obvious omission in (2.4) there; the $z+n$ in the third expression should be $z+n-1$.) Using this with $m=k-1$ and $z=n-k+1$ and then with $z=n-k$ we obtain

$$\mathbb{E}_{k-1}\left((n-k+1)^{1+U_{k-1}}\right)=(n-k+1)\frac{(n-1)!}{(n-k)!(k-1)!} \tag{12}$$

and

$$\mathbb{E}_{k-1}\left((n-k)^{1+U_{k-1}}\right)=(n-k)\frac{(n-2)!}{(n-k-1)!(k-1)!}. \tag{13}$$

Combining Equations (8), (11), (12), and (13),

$$\begin{aligned}&\mathbb{E}_n\left(I_{S_k,W_{k+1}}\right)=\\&\quad\frac{(n-k-1)!(k+1)!}{n!}\cdot\frac{n-k}{k(k+1)}\left(\frac{(n-k+1)(n-1)!}{(k-1)!(n-k)!}-\frac{(n-k)(n-2)!}{(k-1)!(n-k-1)!}\right)\\&\quad=\frac{n-k+1}{n}-\frac{(n-k)^2}{n(n-1)},\end{aligned}$$

and the conclusion follows. □

The relationship between permutation tableaux and tree–like tableaux given by (4) allows us to deduce the following corollary to Theorem 6.

Corollary 3 *For tree–like tableaux of size n, $n\geq 2$, the probability of having a corner with border edges k and $k+1$ is given by*

$$\mathbb{P}_n\left(I_{S_k,W_{k+1}}\right)=\begin{cases}\frac{n-k+1}{n}-\frac{(n-k)^2}{n(n-1)} & k=1,\ldots,n-1;\\ \frac{1}{n} & k=n.\end{cases}$$

Finally, we establish the following result which, when combined with (4) and (5), completes the proof of Theorem 1.

Theorem 4 *For permutation tableaux of size n we have*

$$\mathbb{E}_n C_n=\frac{n+4}{6}-\frac{1}{n}.$$

Proof: In view of (2) we are interested in

$$\mathbb{E}_n\left(\sum_{k=1}^{n-1} I_{S_k,W_{k+1}}\right) = \sum_{k=1}^{n-1} \mathbb{E}_n\left(I_{S_k,W_{k+1}}\right).$$

Therefore, the result is obtained by summing the expression from Theorem 2 from $k = 1$ to $n - 1$. □

To conclude this section, note that Theorem 1 could also be obtained by summing the expression from Corollary 3 from $k = 1$ to n.

4 Corners in Symmetric Tree-Like Tableaux

The main result of this section concerns the second conjecture of Laborde Zubieta.

Theorem 5 *(see (Laborde Zubieta, 2015a, Conjecture 4.2)) For $n \geq 2$ we have*

$$c(\mathcal{T}_{2n+1}^{sym}) = 2^n \times n! \times \frac{4n+13}{12}.$$

As in Section 3, we will use a bijection between symmetric tree–like tableaux and type–B permutation tableaux to relate the corners of $\mathcal{T}_{2n+1}^{sym}$ to the corners of $\mathcal{B}_n$. In Section 2.2 of Aval et al. (2013), it was mentioned that there exists such a bijection; however, no details were given. Thus, we give a description of one such bijection which will be useful to us (see Figure 3).

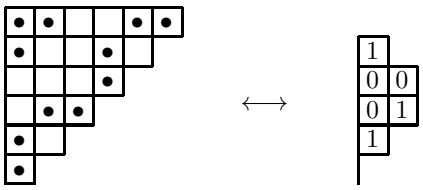


Fig. 3: An example of the bijection Fas defined in Lemma 6 between type–B permutation tableaux of size 5 and symmetric tree–like tableaux of size 11.

Lemma 6 *Consider $F : \mathcal{T}_{2n+1}^{sym} \to \mathcal{B}_n$ defined by the following rules,*

1. *Replace the topmost point in each column with 1_T's.*
2. *Replace the leftmost points in each row with 0_R's*
3. *Fill in the remaining cells according to the rules of type–B permutation tableaux.*
4. *Remove the cells above the diagonal.*
5. *Remove the first column.*

and $F^{-1} : \mathcal{B}_n \to \mathcal{T}_{2n+1}^{sym}$ defined by:

1. *Add a column and point all cells except those in a restricted row.*
2. *Replace all 0_R's with points unless that 0_R is in the same row as a diagonal 0.*
3. *Replace all non-diagonal 1_T's with points.*
4. *Delete the remaining numbers, add a pointed box in the upper–left–hand corner (the root point), and then add the boxes necessary to make the tableau symmetric.*

Then F is a bijection between $\mathcal{T}^{sym}_{2n+1}$ and $\mathcal{B}_n$.

Proof: The details of this proof are straightforward and will be given in the full version of this paper Hitczenko and Lohss (2015). □

As mentioned earlier, Lemma 6 will allow us to relate the corners of symmetric tree–like tableaux to the corners of type–B permutation tableaux. To carry out the calculations for type–B permutation tableaux we will develop techniques similar to those developed in Corteel and Hitczenko (2007) for permutation tableaux. First, we briefly describe an extension procedure for B–type tableaux that mimics a construction given in (Corteel and Hitczenko, 2007, Section 2). Fix any $B \in \mathcal{B}_{n-1}$ and let $U_{n-1} = U_{n-1}(B)$ be the number of unrestricted rows in B. We can extend the size of B to n by inserting a new row or a new column. The details of this insertion will be left for the full version of this paper. However, if U_n is the number of unrestricted rows in the extended tableaux, $U_n = 1, \ldots, U_{n-1}+1$, the (conditional) probability that $U_n = U_{n-1} + 1$ is given by inserting a row,

$$\mathbb{P}(U_n = U_{n-1} + 1 | \mathcal{F}_{n-1}) = \mathbb{P}(S_n | \mathcal{F}_{n-1}) = \frac{1}{2^{U_{n-1}+1}}. \tag{14}$$

(Here, analogously to permutation tableaux (see the proof of Theorem 4 above or (Hitczenko and Janson, 2010, Section 2)) $\mathcal{F}_{n-1}$ is a σ–subalgebra on $\mathcal{B}_n$ obtained by grouping together all tableaux in $\mathcal{B}_n$ that are obtained as the extension of the same tableau from $\mathcal{B}_{n-1}$.) The (conditional) probability of the remaining cases is given by inserting a column,

$$\mathbb{P}(U_n = k | \mathcal{F}_{n-1}) = \frac{1}{2^{U_{n-1}+1}} \left(\binom{U_{n-1}}{k-1} + \binom{U_{n-1}}{k-1} \right) = \frac{1}{2^{U_{n-1}}} \binom{U_{n-1}}{k-1},$$

for $k = 1, \ldots, U_{n-1}$. This agrees with (14) when $k = U_{n+1}$. Thus,

$$\mathcal{L}(U_n | \mathcal{F}_{n-1}) = 1 + \mathrm{Bin}(U_{n-1}),$$

where the left–hand side means the conditional distribution of U_n given U_{n-1} and $\mathrm{Bin}(m)$ denotes a binomial random variable with parameters m and $1/2$. Note that this is the same relationship as for permutation tableaux (see (Hitczenko and Janson, 2010, Equation (2.2)) or (Corteel and Hitczenko, 2007, Equation 4)).

As in the case of permutation tableaux, the uniform measure $\mathbb{P}_n$ on $\mathcal{B}_n$ induces a measure (still denoted by $\mathbb{P}_n$) on $\mathcal{B}_{n-1}$ via a mapping $\mathcal{B}_n \to \mathcal{B}_{n-1}$ that assigns to any $B' \in \mathcal{B}_n$ the unique tableau of size $n-1$ whose extension is B'. These two measures on $\mathcal{B}_{n-1}$ are not identical, but the relationship between them

can be easily calculated (see (Corteel and Hitczenko, 2007, Section 2) or (Hitczenko and Janson, 2010, Section 2) for more details and calculations for permutation tableaux). Namely,

$$\mathbb{P}_n(B) \quad = \quad 2^{U_{n-1}(B)+1}\frac{|\mathcal{B}_{n-1}|}{|\mathcal{B}_n|}\mathbb{P}_{n-1}(B), \quad B \in \mathcal{B}_{n-1}.$$

This relationship implies that for any random variable X on $\mathcal{B}_{n-1}$,

$$\mathbb{E}_n X \quad = \quad \frac{2|\mathcal{B}_{n-1}|}{|\mathcal{B}_n|}\mathbb{E}_{n-1}(2^{U_{n-1}(B_{n-1})}X). \tag{15}$$

This allows us to provide a direct proof of the following well known fact,

Proposition 7 *For all $n \geq 0$, $|\mathcal{B}_n| = 2^n n!$.*

Proof: By considering all the extensions of a type–B permutation tableau of size $n-1$, we have the following relationship,

$$|\mathcal{B}_n| = \sum_{B\in\mathcal{B}_{n-1}} 2^{U_{n-1}(B)+1}.$$

Thus,

$$\begin{aligned}
|\mathcal{B}_n| &= |\mathcal{B}_{n-1}|\mathbb{E}_{n-1}\left(2^{U_{n-1}+1}\right)\\
&= 2|\mathcal{B}_{n-1}|\mathbb{E}_{n-1}\mathbb{E}\left(2^{1+\mathrm{Bin}(U_{n-2})}|U_{n-2}\right)\\
&= 2\cdot 2|\mathcal{B}_{n-1}|\mathbb{E}_{n-1}\left(\frac{3}{2}\right)^{U_{n-2}}\\
&= 2\cdot 2|\mathcal{B}_{n-1}|\frac{2|\mathcal{B}_{n-2}|}{|\mathcal{B}_{n-1}|}\mathbb{E}_{n-2}\left(2^{U_{n-2}}\left(\frac{3}{2}\right)^{U_{n-2}}\right)\\
&= 2^2\cdot 2!\,|\mathcal{B}_{n-2}|\mathbb{E}_{n-2}3^{U_{n-2}}.
\end{aligned}$$

Iterating n times,

$$\begin{aligned}
|\mathcal{B}_n| &= 2^3\cdot 3!\,|\mathcal{B}_{n-3}|\mathbb{E}_{n-3}4^{U_{n-3}} = 2^{n-1}(n-1)!|\mathcal{B}_1|\mathbb{E}_1 n^{U_1}\\
&= 2^n n!,
\end{aligned}$$

where the final equality holds because $|\mathcal{B}_1| = 2$ and $U_1 \equiv 1$. □

Given Proposition 7, (15) reads

$$\mathbb{E}_n X = \frac{1}{n}\mathbb{E}_{n-1}(2^{U_{n-1}(B_{n-1})}X). \tag{16}$$

This is exactly the same expression as (Corteel and Hitczenko, 2007, Equation (7)) which means that the relationship between $\mathbb{E}_n$ and $\mathbb{E}_{n-1}$ is the same regardless of whether we are considering $\mathcal{P}_n$ or $\mathcal{B}_n$. Thus, any computation for B–type tableaux based on (16) will lead to the same expression as the analogous computation for permutation tableaux based on (Corteel and Hitczenko, 2007, Equation (7)).

Now we have the tools necessary to obtain a relationship between corners in symmetric tree–like tableaux and type–B permutation tableaux which is analogous to (4).

Lemma 8 *The number of corners in symmetric tree–like tableaux is given by,*

$$c(\mathcal{T}_{2n+1}^{sym}) \quad = \quad 2c(\mathcal{B}_n) + 2^n(n-1)! + 2^{n-1}n!. \tag{17}$$

Proof: The bijection described in Lemma 6 leads to the following relationship,

$$c(\mathcal{T}_{2n+1}^{sym}) = 2c(\mathcal{B}_n) + 2|\{B \in \mathcal{B}_n : \ S_n\}| + |\{B \in \mathcal{B}_n : \ W_1\}|. \tag{18}$$

The result is then obtained by the extension process described above. The details will be given in the full version of this paper Hitczenko and Lohss (2015). □

It follows from Lemma 8 that to prove Theorem 5, it suffices to determine the number of corners in type–B permutation tableaux of size n. Since $|\mathcal{B}_n| = 2^n n!$, Equation (3) becomes

$$c(\mathcal{B}_n) = 2^n n! \, \mathbb{E}_n C_n. \tag{19}$$

In order to determine the number of corners in type–B permutation tableaux, we first have the following result.

Theorem 9 *For type–B permutation tableaux of size n, the probability of having a corner with border edges k and $k+1$ is given by*

$$\mathbb{P}_n\left(I_{M_k=S, M_{k+1}=W}\right) \quad = \quad \frac{n-k+1}{2n} - \frac{(n-k)^2}{4n(n-1)}.$$

Proof: The proof is similar to the proof of Theorem 2, using the techniques developed in this section for type–B permutation tableaux. The details will be given in the full version of this paper Hitczenko and Lohss (2015). □

The relationship between permutation tableaux and tree–like tableaux given by (17) allows us to deduce the following corollary to Theorem 9.

Corollary 10 *For symmetric tree–like tableaux of size $2n+1$, $n \geq 2$, the probability of having a corner with border edges k and $k+1$ is given by*

$$\mathbb{P}_n\left(I_{S_k, W_{k+1}}\right) = \begin{cases} \frac{1}{2n} & k = 1 \\ \frac{k}{2n} - \frac{(k-1)^2}{4n(n-1)} & k = 2, \ldots n, \\ \frac{1}{2} & k = n+1 \\ \frac{2n-k+2}{2n} - \frac{(2n-k+1)^2}{4n(n-1)} & k = n+2, \ldots 2n \\ \frac{1}{2n} & k = 2n+1. \end{cases}$$

Finally, we establish the following result which, when combined with (17) and (19), completes the proof of Theorem 5.

Theorem 11 *For type–B permutation tableaux of size n we have*

$$\mathbb{E}_n C_n = \frac{4n+7}{24} - \frac{1}{2n}.$$

Proof: The result is obtained by summing the expression from Theorem 9 from $k = 1$ to $n-1$. □

To conclude this section, note that Theorem 5 could also be obtained by summing the expression from Corollary 10 from $k = 1$ to $2n+1$.

References

J.-C. Aval, A. Boussicault, and P. Nadeau. Tree-like tableaux. *Electron. J. Combin.*, 20(4):Paper 34, 24, 2013.

A. Burstein. On some properties of permutation tableaux. *Ann. Comb.*, 11(3-4):355–368, 2007.

S. Corteel and P. Hitczenko. Expected values of statistics on permutation tableaux. In *2007 Conference on Analysis of Algorithms, AofA 07*, Discrete Math. Theor. Comput. Sci. Proc., AH, pages 325–339. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2007.

S. Corteel and J. S. Kim. Combinatorics on permutation tableaux of type A and type B. *European J. Combin.*, 32(4):563–579, 2011.

S. Corteel and P. Nadeau. Bijections for permutation tableaux. *European J. Combin.*, 30(1):295–310, 2009.

S. Corteel and L. K. Williams. Tableaux combinatorics for the asymmetric exclusion process. *Adv. in Appl. Math.*, 39(3):293–310, 2007a.

S. Corteel and L. K. Williams. A Markov chain on permutations which projects to the PASEP. *Int. Math. Res. Not. IMRN*, (17):Art. ID rnm055, 27, 2007b.

P. Hitczenko and S. Janson. Asymptotic normality of statistics on permutation tableaux. In *Algorithmic probability and combinatorics*, volume 520 of *Contemp. Math.*, pages 83–104. Amer. Math. Soc., Providence, RI, 2010.

P. Hitczenko and A. Lohss. Corners in tree–like tableaux, 2015. `arXiv:1511.04989v1`.

P. Laborde Zubieta. Occupied corners in tree–like tableaux. *Séminaire Lotharingien de Combinatoire*, 74, 2015a. Article B74b.

P. Laborde Zubieta. Personal communication, 2015b.

T. Lam and L. Williams. Total positivity for cominuscule Grassmannians. *New York J. Math.*, 14:53–99, 2008.

P. Nadeau. The structure of alternative tableaux. *J. Combin. Theory Ser. A*, 118(5):1638–1660, 2011.

E. Steingrímsson and L. K. Williams. Permutation tableaux and permutation patterns. *J. Combin. Theory Ser. A*, 114(2):211–234, 2007.

X. Viennot. Alternative tableaux, permutations, and partially asymmetric exclusion process, 2008. Slides of a talk at the Isaac Newton Institute in Cambridge.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Variance of additive functions defined on random assemblies

Eugenijus Manstavičius and Vytautas Stepas

Department of Mathematics and Informatics, Vilnius University, Naugarduko str. 24, LT-03225 Vilnius, Lithuania

Abstract. An inequality for the variance of an additive function defined on random decomposable structures, called assemblies, is established. The result generalizes estimates obtained earlier in the cases of permutations and mappings of a finite set into itself. It is analogous to the Turán-Kubilius inequality for additive number-theoretic functions.

Keywords: Labeled decomposable structure, additive function, moments, Turán-Kubilius inequality.

We deal with additive functions defined on combinatorial structures such as permutations, mappings of a finite set into itself, 2-regular graphs etc. If a structure is taken at random, such functions are sums of dependent random variables; sometimes, they are called *separable statistics*. Their value distribution is a complex problem. One of the useful tools in analysing it are estimates of the variance. This is our main objective. On the other hand, our interest has been highly stimulated by the Turán-Kubilius inequality in probabilistic number theory or by analogous inequalities in the theory of additive arithmetical semigroups.

An assembly is a construction defined on a set by its partition and some structure introduced in all subsets, afterwards called components of the assembly. Assume that given a subset of size j we can introduce $g_j < \infty$ structures, then the number of assemblies spanned over an n set (assemblies of the order n) equals

$$G(n) = n! \sum_{\ell(\bar{s})=n} \prod_{j=1}^{n} \Big(\frac{g_j}{j!}\Big)^{s_j} \frac{1}{s_j!} =: n!Q(n).$$

Here $n \in \mathbf{N}$, $\ell(\bar{s}) := 1s_1 + \cdots + ns_n$ if $\bar{s} = (s_1, \ldots, s_n) \in \mathbf{N}_0^n$ and the summation is over such vectors satisfying $\ell(\bar{s}) = n$. We will denote the class of assemblies by $\mathcal{G}$ and the set of assemblies of the order n by $\mathcal{G}_n \subset \mathcal{G}$.

Let $\lambda_j := g_j/j!$. In the past decades much attention was paid to the *logarithmic* class defined by the asymptotic condition $\rho^j j\lambda_j \sim \theta$ for some positive constants θ, Θ and ρ as $j \to \infty$ (see [1]). Extensions were initiated in the first author's paper [2], where a condition

$$0 < \theta \le \rho^j j\lambda_j \le \Theta, \quad j \ge 1, \tag{1}$$

was used. The lower bound excluded, for example, the class of 2-regular graphs, however. Basing upon the experience, in the present paper we confine ourselves to a class of assemblies characterized by some positive constants ρ, Θ, θ, θ', and $n_0 \ge 1$.

Definition. *We say that a class of assemblies is weakly logarithmic if the following conditions are satisfied:*

$$\rho^j j\lambda_j \le \Theta, \quad j \ge 1; \tag{2}$$

$$\sum_{j\le n} \rho^j j\lambda_j \ge \theta n, \quad n \ge n_0; \tag{3}$$

$$nQ(n)\rho^n \ge \theta' \exp\Big\{\sum_{j\le n} \lambda_j \rho^j\Big\}, \quad n \ge 1. \tag{4}$$

Let $k_j(\sigma) \ge 0$ be the number of components of size j in $\sigma \in \mathcal{G}_n$ and $1 \le j \le n$. An *additive function* $h : \mathcal{G}_n \to \mathbf{R}$ is defined by a real two-dimensional array $\{h_j(k)\}$, where $j, k \in \mathbf{N}$, $jk \le n$, and $h_j(0) = 0$ for all $j \le n$, by setting

$$h(\sigma) = \sum_{j=1}^{n} h_j\big(k_j(\sigma)\big). \tag{5}$$

Apart from the most popular example of the number-of-components function $w(\sigma) = k_1(\sigma) + \cdots + k_n(\sigma)$, they appear in many algebraic and combinatorial problems. Particular additive functions appear in physical models as a part of Hamiltonians in the Bose gas theory.

Let $\mathbf{E}_n h$ and $\mathbf{V}_n h$ denote the expectation and the variance of $h = h(\sigma)$ with respect to uniform probability measure. The problem is to estimate

$$\mathbf{V}_n h = \frac{1}{G(n)} \sum_{\sigma\in\mathcal{G}_n} \big(h(\sigma) - \mathbf{E}_n h\big)^2 = \mathbf{E}_n h^2 - (\mathbf{E}_n h)^2$$

in terms of the values $h_j(k)$ where $jk \le n$ and parameters characterizing the class of assemblies.

In the sequel, let $Q^{\{j\}}(n)$ be defined by

$$Q^{\{j\}}(n) = \sum_{\substack{\ell(\bar{s})=n \\ s_j=0}} \prod_{i\le n} \frac{\lambda_i^{s_i}}{s_i!}, \quad j \le n$$

and $\ll$ be an analog of the symbol $O(\cdot)$.

Now, the results.

Theorem 1. *Assume that $\mathcal{G}$ is weakly logarithmic and $h : \mathcal{G}_n \to \mathbf{R}$ is an arbitrary additive function. Then*

$$\mathbf{V}_n h \ll \sum_{jk\le n} \frac{\lambda_j^k h_j(k)^2}{k!} \frac{Q^{\{j\}}(n-jk)}{Q(n)} \tag{6}$$

for $n \ge 1$.

Inequality (6) sharpens a bit Theorem 3 in [3] proved for an arbitrary additive function defined on weighted permutations under condition (1).

A *completely additive function* h is defined by the array $h_j(k) = a_j k$, where $a_j \in \mathbf{R}$ and $jk \le n$. For such functions, inequality (6) takes a simpler form.

Theorem 2. *Assume that $\mathcal{G}$ is weakly logarithmic and $h : \mathcal{G}_n \to \mathbf{R}$ is a completely additive function defined via $h_j(s) = a_j s$ where $js \le n$. Then*

$$\mathbf{V}_n h \ll \sum_{j \le n} \lambda_j a_j^2 \frac{Q(n-j)}{Q(n)}. \tag{7}$$

Inequality (7) for weighted permutations satisfying condition (1) has a longer history (see [3])).

References

[1] R. Arratia, A.D. Barbour and S. Tavaré, *Logarithmic Combinatorial Structures: A Probabilistic Approach*. EMS Monographs in Mathematics, EMS Publishing House, Zürich, 2003.

[2] E. Manstavičius, Mappings on decomposable combinatorial structures: analytic approach. *Comb. Probab. Comput.*, **11**, 61–78 (2002).

[3] E. Manstavičius and V. Stepanauskas, On variance of an additive function with respect to a generalized Ewens probability, *Discrete Math. Th. Comp. Sci.*, AofA14, 301–312 (2014).

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Recent results on permutations without short cycles

Robertas Petuchovas

Vilnius University
Naugarduko str. 24, LT-03225 Vilnius, Lithuania
e-mail: robertas.petuchovas@mif.vu.lt

Abstract. The density, denoted by $\kappa(n,r)$, of permutations having no cycles of length less than $r+1$ in a symmetric group S_n is explored. New asymptotic formulas for $\kappa(n,r)$ are obtained using the saddle-point method when $5 \le r < n$ and $n \to \infty$.

Keywords: symmetric group, long cycles, Buchstab's function, Dickman's function, saddle-point method

The probability $\kappa(n,r)$ that a permutation sampled from the symmetric group S_n uniformly at random has no cycles of length less than $r+1$, where $1 \le r < n$ and $n \to \infty$, is explored. New asymptotic formulas valid in specified regions are obtained using the saddle-point method. One of the results is applied to show that estimate of the total variation distance for permutations can be expressed only through the function $\nu(n,r)$ which is a probability that a permutation sampled from the S_n uniformly at random has no cycles of length greater than r.

To address the problem, we need recollect the following functions. Buchstab's function $\omega(v)$ is defined as a solution to difference-differential equation

$$(v\omega(v))' = w(v-1)$$

for $v > 2$ with the initial condition $\omega(v) = 1/v$ if $1 \le v \le 2$. Dickman's function $\varrho(v)$ is the unique continuous solution to the equation

$$v\varrho'(v) + \varrho(v-1) = 0$$

for $v > 1$ with initial condition $\varrho(v) = 1$ if $0 \le v \le 1$.

The interest to the problem begins with the classical example of derangements

$$\kappa(n,1) \;=\; \sum_{j=0}^{n} \frac{(-1)^j}{j!} = \mathrm{e}^{-1} + O\left(\frac{1}{n!}\right)$$

and the trivial case $\kappa(n,r) = 1/n$ if $n/2 \le r < n$. There was a series of works concerning general asymptotic formulas of the probability $\kappa(n,r)$ the strongest of which are presented here as Proposition 1 and Proposition 2.

Proposition 1 *For $1 \leq r < n$, we have*

$$\kappa(n,r) = \mathrm{e}^{-H_r+\gamma}\omega(n/r) + O\left(\frac{1}{r^2}\right).$$

See [6, Theorem 3].

Proposition 2 *Let $u = n/r$. For $1 \leq r \leq n/\log n$,*

$$\kappa(n,r) = \mathrm{e}^{-H_r} + O\left(\frac{(u/\mathrm{e})^{-u}}{r^2}\right).$$

If $r \geq 3$, we can replace e by 1 in the error term.

See [12, Proposition 2]. Together these propositions provide stronger estimates of $\kappa(n,r)$ than those in [2], [3], [4]. New results are the following theorems:

Theorem 1 *For $\sqrt{n\log n} \leq r < n$, we have*

$$\kappa(n,r) = \mathrm{e}^{-H_r+\gamma}\omega(n/r) + O\left(\frac{\varrho(n/r)}{r^2}\right).$$

Proof. The result is a corollary of Theorem 1 in [7]. It is obtained from the probability generating function using saddle-point method, the technique is elaborated in [11].

Theorem 2 *For $(\log n)^4 \leq r < n$, we have*

$$\kappa(n,r) = \mathrm{e}^{-H_r} + O\left(\frac{\varrho(n/r)}{r}\right).$$

Proof. The saddle-point method is applied to the Cauchy's integral representation of $\kappa(n,r)$, as in the proof of Theorem 1. However, there are some other technical difficulties one must to overcome.

Theorem 3 *For $5 \leq r < n$, we have*

$$\kappa(n,r) = \mathrm{e}^{-H_r} + O\left(\frac{\nu(n,r)}{r}\right).$$

Proof. Quite the same technique to that used in the proof of Theorem 2 is employed, just a different approximation of the saddle point is taken and Corollary 5 of [8] is applied.

Theorem 1 and Theorem 2 (see also Corollary 2.3 in [5]) improve on Proposition 1 and Proposition 2. Theorem 3 is of separate interest; as we see, it can be useful in formulas where both probabilities $\kappa(n,r)$ and $\nu(n,r)$ are involved. Here is an example.

Let $k_j(\sigma)$ equal the number of cycles of length j in a permutation $\sigma \in \mathrm{S}_n$, $\overline{k}(\sigma) = (k_1(\sigma), k_2(\sigma), \ldots, k_n(\sigma))$, and $\overline{Z} = (Z_1, Z_2, \ldots, Z_n)$, where Z_j are Poisson random variables such that $\mathrm{E}Z_j = 1/j$, $j \in \mathrm{N}$. Thus, if $5 \le r < n$, we have (see Lemma 3.1 on p. 69 of [1])

$$\begin{aligned} d_{TV}(n,r) &= \sup_{V \subseteq \mathrm{Z}_+^r} \left| \frac{\#\{\sigma : \overline{k}(\sigma) \in V\}}{n!} - \Pr(\overline{Z} \in V) \right| \\ &= \frac{1}{2} \sum_{m=0}^{\infty} \nu(m,r) \left| \kappa(n-m,r) - \mathrm{e}^{-H_r} \right| \\ &= \frac{\mathrm{e}^{-H_r}}{2} \sum_{m=n-r}^{\infty} \nu(m,r) + \frac{1}{2}\nu(n,r) + O\left(\frac{1}{r} \sum_{m=0}^{n-r-1} \nu(m,r)\nu(n-m,r) \right). \end{aligned}$$

Consequently, only results on the probability $\nu(n,r)$ are needed attempting to improve on the order of notable estimate for $d_{TV}(n,r)$ in [2].

References

[1] R. Arratia, A.D. Barbour, and S. Tavaré, *Logarithmic Combinatorial Structures: A Probabilistic Approach*, EMS Publishing House, Zürich, 2003.

[2] R. Arratia and S. Tavaré, The cycle structure of random permutations, *Ann. Probab.*, 1992, **20**, 3, 1567-1591.

[3] E.A. Bender, A. Mashatan, D. Panario, and L.B. Richmond, Asymptotics of combinatorial structures with large smallest component, *J. Comb. Th.*, Ser. A, 2004, **107**, 117–125.

[4] A. Granville, Cycle lengths in a permutation are typically Poisson, *Electronic J. Comb.*, 2006, **13**, #R107.

[5] A. Hildebrand and G. Tenenbaum, Integers without large prime factors, *J. Théorie des Nombres de Bordeaux*, 1993, **5**, 411–484.

[6] E. Manstavičius, On permutations missing short cycles, *Lietuvos matem. rink.*, spec. issue, 2002, **42**, 1–6.

[7] E. Manstavičius and R. Petuchovas, Local probabilities and total variation distance for random permutations, *The Ramanujan J.*, 2016, DOI 10.1007/s11139-016-9786-0.

[8] E. Manstavičius and R. Petuchovas, *Local probabilities for random permutations without long cycles*, Electron. J. Combin., 2016, **23(1)**, #P1.58.

[9] E. Manstavičius and R. Petuchovas, Permutations without long or short cycles, *Electronic Notes in Discrete Mathematics*, 2015, **49**, 153-158.

[10] G. Tenenbaum, *Introduction to Analytic and Probabilistic Number Theory*, Cambridge Univ. Press, 1995.

[11] G. Tenenbaum, Crible d'Ératosthène et modèle de Kubilius. In: *Number Theory in Progress, Proc. Conf. in Honor of Andrzej Schinzel*, Zakopane, Poland, 1997. K. Gyory, H. Iwaniec, J. Urbanowicz (Eds.), Walter de Gruyter, Berlin, New York, 1999, 1099–1129.

[12] A. Weingartner, On the degrees of polynomial divisors over finite fields, arXiv:1507.01920.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Robin Hood Hashing really *has constant average search cost and variance in full tables*

Patricio V. Poblete[1]†, Alfredo Viola[2] ‡

[1]*Dept. of Computer Science, University of Chile, Chile*
[2]*Universidad de la República, Uruguay*

Thirty years ago, the Robin Hood collision resolution strategy was introduced for open addressing hash tables, and a recurrence equation was found for the distribution of its search cost. Although this recurrence could not be solved analytically, it allowed for numerical computations that, remarkably, suggested that the variance of the search cost approached a value of 1.883 when the table was full. Furthermore, by using a non-standard mean-centered search algorithm, this would imply that searches could be performed in expected constant time even in a full table.

In spite of the time elapsed since these observations were made, no progress has been made in proving them. In this paper we introduce a technique to work around the intractability of the recurrence equation by solving instead an associated differential equation. While this does not provide an exact solution, it is sufficiently powerful to prove a bound for the variance, and thus obtain a proof that the variance of Robin Hood is bounded by a small constant for load factors arbitrarily close to 1. As a corollary, this proves that the mean-centered search algorithm runs in expected constant time.

We also use this technique to study the performance of Robin Hood hash tables under a long sequence of insertions and deletions, where deletions are implemented by marking elements as *deleted*. We prove that, in this case, the variance is bounded by $1/(1-\alpha)+O(1)$, where α is the load factor.

To model the behavior of these hash tables, we use a unified approach that can be applied also to study the First-Come-First-Served and Last-Come-First-Served collision resolution disciplines, both with and without deletions.

Keywords: Robin Hood Hashing, full tables, constant variance, constant expected search time

1 Introduction

In 1986, Celis *et al* [3, 4] introduced the Robin Hood collision resolution strategy for open addressing hash tables. Under this discipline, collisions are decided in favor of the element that is farthest from its home location. While this does not change the expected search cost, it turns out to have a dramatic effect

†Supported in part by NIC Chile
‡This work has been partially supported by Project CSIC I+D "Combinatoria Analítica y aplicaciones en criptografía, comunicaciones y recuperación de la información", fondos 2015-2016.

on its *variance*. In effect, unlike other disciplines where the variance tends to infinity as the table becomes full, the variance of Robin Hood seems to remain constant, and very small. This fact, conjectured from numerical computations, has not been proved in the years since it was observed, and is the main focus of our work. This problem has been hard to solve because the distribution of the search cost obeys a nonlinear recurrence equation for which no successful line of attack has been found.

To show the kind of recurrence involved, we quote now Theorem 3.1 from [3] (our notation will be slightly different):

Theorem 3.1 *In the asymptotic model for an infinite Robin Hood hash table with load factor α ($\alpha < 1$), the probability $p_i(\alpha)$ that a record is placed in the i-th or further position in its probe sequence is equal to*

$$p_1(\alpha) = 1, \quad p_{i+1}(\alpha) = 1 - \left(\frac{1-\alpha}{\alpha}\right)\left(e^{\alpha(p_1(\alpha)+\cdots+p_i(\alpha))}\right). \tag{1}$$

They then go on to define another function $r_i(\alpha) = \alpha(p_i(\alpha) + \cdots + p_\infty(\alpha))$, in terms of which the variance can be expressed as

$$V(\alpha) = \frac{2}{\alpha}\sum_{i=1}^{\infty} r_i(\alpha) + \frac{\ln(1-\alpha)}{\alpha} - \frac{\ln^2(1-\alpha)}{\alpha^2}. \tag{2}$$

They show that $r_i(\alpha)$satisfies the following recurrence equation:

$$r_i(\alpha) - r_{i+1}(\alpha) = 1 - e^{-r_i(\alpha)} \tag{3}$$

with $r_1(\alpha) = -\ln(1-\alpha)$. By leaving the "$(\alpha)$" implicit and using the Δ operator (defined as $\Delta r_i = r_{i+1} - r_i$), this can be rewritten as $\Delta r_i = f(r_i)$ where f is the function $f(x) = -1 + e^{-x}$.

This seemingly simpler equation has, nonetheless, so far remained unsolved.

In this paper, we will introduce a technique applicable to equations of this form, and we will use it first to prove a bound on the variance of Robin Hood hashing. Then we will use it to study another recurrence equation of the same type arising from the problem of hashing with deletions.

2 Modeling hashing algorithms

In this paper we will study the search cost of a random element in a hash table, using the *random probing model*. This is an open addressing hashing scheme in which collisions are resolved by additional probes into the table. The sequence of these probes are considered to be random and depends only on the value of the key. The difference with uniform probing is that positions may be repeated in this sequence. We use the *asymptotic model* for a hash table with load factor α [9, 8, 4, 12], where we assume that the number of keys n and the table size m both tend to infinity, maintaining constant their ratio $\alpha = n/m$.

Each element has associated with it an infinite probe sequence consisting of i.i.d. integers uniformly distributed over $\{0, \ldots, m-1\}$, representing the consecutive places of probes for that element. The probe sequence for element x is denoted by $h_1(x), h_2(x), h_3(x), \ldots$. Elements are inserted sequentially into the table. If element x is placed in position $h_j(x)$, then we say that element x has age j, as it requires j probes to reach the element in case of a search. When an element x of age j and an element y of age k compete for the same slot ($h_j(x) = h_k(y)$), a collision resolution strategy is needed.

In the standard method, a collision is resolved in favor of the incumbent key, so the incoming key continues probing to its next location. We call this a First-Come-First-Served (FCFS) collision resolution

discipline. Several authors [2, 1, 7] observed that a collision could be resolved in favor of *any* of the keys involved, and used this additional degree of freedom to decrease the expected search time in the table.

Celis *et al* [3, 4] were the first to observe that collisions could be resolved having instead *variance reduction* as a goal. They defined the Robin Hood (RH) heuristic, in which each collision occurring during an insertion is resolved in favor of the key that is farthest away from its home location (i.e., oldest in terms of *age*). Later, Poblete and Munro [14] defined the Last-Come-First-Served heuristic, where collisions are resolved in favor of the *incoming* key.

In both cases, the variance is reduced, and this can be used to speed up searches by replacing the standard search algorithm by a *mean-centered* one that first searches in the vicinity of where we would expect the element to have *drifted* to, rather than in its initial probe location. This *mean-centered* approach was introduced in [3] (and called "organ-pipe search") to speed up successful searches in the Robin Hood heuristic, with expected cost bounded by the standard deviation of this random variable. Numerical computations in [3] suggest that for full tables the variance of the search cost for RH is constant, but no formal proof is given.

In this paper we formally settle this conjecture, by proving that this is in fact the case, and give an explicit upper bound (although not as tight as the numerical results seem to suggest). As a consequence we prove that the mean-centered searching algorithm in [3] has constant expected cost for full tables.

In section 4 we extend this approach to perform the analysis of hashing with deletions. Deletions in open addressing hash tables are often handled by marking the cells as *deleted* instead of *empty*, because otherwise the search algorithm might fail to find some of the keys. The space used by deleted cells may be reused by subsequent insertions. Intuitively, search times should deteriorate as tables become contaminated with deleted cells and, as Knuth[11] points out, in the long run the average successful search time should approach the average *unsucessful* search time.

In this paper we analize the effect of a long sequence of insertions and deletions in the asymptotic regime (α-full tables with $0 \leq \alpha < 1$) and prove a bound for the variance of RH with deletions that is close to numerical results.

There is an alternative algorithm designed to keep variance low in the presence of deletions. This method marks cells as deleted, but keeps the key values (these cells are called *tombstones*). In this paper we do not study the algorithm with tombstones. We note that [12] derives equations for this algorithm, but only obtains numerical solutions.

3 Analysis without deletions

To analyze the cost of searching for a random element, we begin by presenting a general framework, based on the one used in [5]. This framework applies also to FCFS and LCFS, but in this paper we use it to analyze RH, which has been a long standing open problem. As stated before, we use the asymptotic model for a hash table with load factor α and random probing.

Under this model, if collisions are resolved without "looking ahead" in the table, the cost of inserting a random element is 1 plus a random variable that follows a geometric distribution with parameter $1 - \alpha$, and therefore its expected cost is $1/(1 - \alpha)$, independently of the collision resolution discipline used.

Let $p_i(\alpha)$ be the probability that a randomly chosen key has age i when the table has load factor α.

Suppose we insert a new element. Depending on the insertion discipline used, a number of keys will change locations and therefore increase their ages as a consequence of the arrival of the new element. Let us call $t_i(\alpha)$ the expected number of probes made by keys of age i during the course of the insertion. It is

easy to see that

$$t_1(\alpha) = 1, \quad \sum_{i\geq 1} t_i(\alpha) = \frac{1}{1-\alpha}. \tag{4}$$

Before the insertion, the expected number of keys of age i is $\alpha m p_i(\alpha)$. After the insertion, it is

$$(\alpha m + 1) p_i(\alpha + \frac{1}{m}) = \alpha m p_i(\alpha) + t_i(\alpha) - t_{i+1}(\alpha) \tag{5}$$

If we write $\Delta\alpha = 1/m$ and $q_i(\alpha) = \alpha p_i(\alpha)$, this equation becomes

$$\frac{q_i(\alpha + \Delta\alpha) - q_i(\alpha)}{\Delta\alpha} = t_i(\alpha) - t_{i+1}(\alpha) \tag{6}$$

and, as $\Delta\alpha \to 0$ (i.e. $m \to \infty$),

$$\partial_\alpha q_i(\alpha) = t_i(\alpha) - t_{i+1}(\alpha), \tag{7}$$

where ∂_α denotes a derivative with respect to α, and with the initial condition $q_i(0) = 0$.

We introduce a notation that we will use throughout the paper. For any sequence a_i we define its *tail* $\overline{a}_i$ as

$$\overline{a}_i = \sum_{j\geq i} a_j. \tag{8}$$

Using this, equation (7) can be rewitten as

$$\partial_\alpha \overline{q}_i(\alpha) = t_i(\alpha). \tag{9}$$

We note that this equation is valid for all three collision resolution strategies, and it generalizes formula (10) in [12], where it is proved only for RH.

The mean of the search cost can be obtained using the tail notation, as

$$\mu_\alpha = \overline{\overline{p}}_1(\alpha) = \frac{1}{\alpha}\overline{\overline{q}}_1(\alpha) \tag{10}$$

and the variance as

$$\sigma_\alpha^2 = 2\overline{\overline{\overline{p}}}_1(\alpha) - \mu_\alpha - \mu_\alpha^2 = \frac{2}{\alpha}\overline{\overline{\overline{q}}}_1(\alpha) - \mu_\alpha - \mu_\alpha^2 \tag{11}$$

We note that we can already compute the expected search cost, without needing to know the exact form of the function $t_i(\alpha)$. Taking tails in both sides of (9), we have $\partial_\alpha \overline{\overline{q}}_i(\alpha) = \overline{t}_i(\alpha)$.

Now setting $i = 1$ and using (10), we obtain $\partial_\alpha(\alpha\mu_\alpha) = \frac{1}{1-\alpha}$, and from this we obtain

$$\mu_\alpha = \frac{1}{\alpha} \ln \frac{1}{1-\alpha} \tag{12}$$

independently of the collision resolution discipline used.

The fact that the mean search cost is independent of the collision resolution discipline used does not necessarily carry over to higher moments or to the distribution of the search cost. To compute them, we need to know the $t_i(\alpha)$ for the specific discipline.

For RH, a key will be forced to try its $(i+1)$st probe location or higher each time there is a collision between an incoming key of age i or higher and another key in the table that is also of age i or higher. Therefore, and leaving the "(α)" implicit, to simplify notation, we have:

$$\overline{t}_{i+1} = \overline{t}_i \overline{q}_i \tag{13}$$

Together with equation (7) this implies $\partial_\alpha \overline{q}_i = (1-\overline{q}_i)\partial_\alpha \overline{\overline{q}}_i$. Then, after integrating both sides of the equation we have $\ln \frac{1}{1-\overline{q}_i} = \overline{\overline{q}}_i$ from where we obtain $\overline{q}_i = 1 - e^{-\overline{\overline{q}}_i}$. Moreover, by expressing $\overline{q}$ as the difference of two $\overline{\overline{q}}$, we arrive at

Theorem 1 *Under the asymptotic model for an infinite hash table with random probing, and Robin Hood collision resolution discipline, the double tail of the probability distribution of the search cost of a random element satisfies the recurrence*

$$\overline{\overline{q}}_{i+1} = \overline{\overline{q}}_i - 1 + e^{-\overline{\overline{q}}_i} \tag{14}$$

with the initial condition $\overline{\overline{q}}_1 = \ln \frac{1}{1-\alpha}$. □

This is exactly equation (3) that we quoted from [3], but we obtained it through a completely different derivation. As we mentioned before, numerical computations performed in [4] indicate that as $\alpha \to 1$, the variance converges to a small constant, approximately equal to 1.883.

3.1 Bounding the variance of RH

Since we are interested in the behavior of the method as $\alpha \to 1$, we will introduce a variable β defined as $\beta = \frac{1}{1-\alpha}$, so that $\alpha = 1 - \frac{1}{\beta} \to 1$ as $\beta \to \infty$. Now we rewrite equation (14) as

$$\Delta \overline{\overline{q}}_i = -1 + e^{-\overline{\overline{q}}_i}, \tag{15}$$

with $\overline{\overline{q}}_1 = \ln \beta$. This equation is of the form

$$\Delta \overline{\overline{q}}_i = f(\overline{\overline{q}}_i), \tag{16}$$

where f is the function $f(x) = -1 + e^{-x}$. This recurrence equation seems very hard to solve exactly, but we will be able to obtain useful information about its solution by studying instead the differential equation

$$Q'(x) = f(Q(x)) \tag{17}$$

with the same initial condition $Q(1) = \ln \beta$. The solution to this equation is

$$Q(x) = \ln(\beta - 1 + e^{x-1}) - x + 1. \tag{18}$$

Figure 1 compares the solution $\overline{\overline{q}}_i$ (polygonal line) of recurrence equation (16) to the solution $Q(x)$ (smooth line) of differential equation (17). This plot suggests that $Q(i)$ is an upper bound for $\overline{\overline{q}}_i$. This is true, and will follow from the following lemma.

Lemma 1 *Let* a_i *satisfy the recurrence equation* $\Delta a_i = f(a_i)$*, and* $A(x)$ *satisfy the differential equation* $A'(x) = f(A(x))$*, where* $f : [0, +\infty) \to (-\infty, 0]$ *is a decreasing function. Then*

$$A(i) \geq a_i \implies A(i+1) \geq a_{i+1} \tag{19}$$

for all $i \geq 1$.

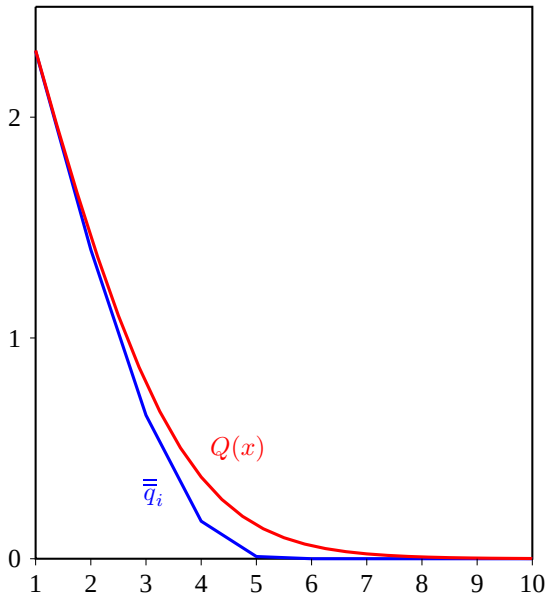


Fig. 1: Comparison of $\overline{\overline{q}}_i$ and $Q(x)$ for $\beta = 10$

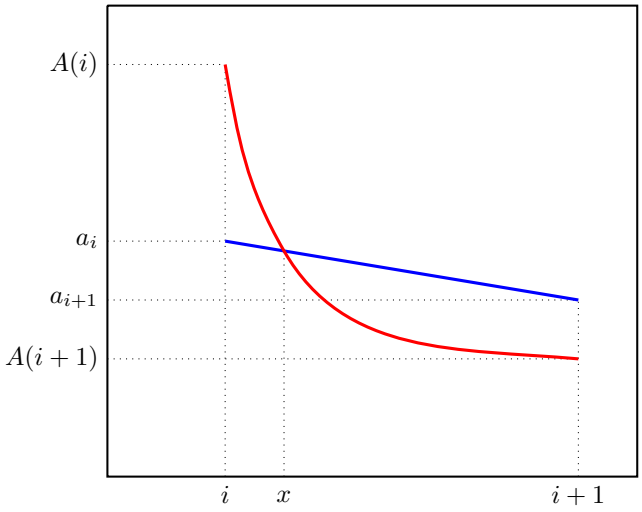


Fig. 2: Proof of Lemma 1

Proof: We begin by noting that both a and A are decreasing functions, because f is negative. Reasoning by contradiction, suppose that $A(i) \geq a_i$ but $A(i+1) < a_{i+1}$. Therefore, there exists an $x \in (i, i+1)$ such that $A(x)$ intersects the straight line joining points (i, a_i) and $(i+1, a_{i+1})$, as illustrated in Figure 2. The slope of this line at x is $f(a_i)$ and the slope of A at point x is $f(A(x))$. At the intersection we must have $f(a_i) > f(A(x))$. But $a_i > A(x)$ implies $f(a_i) < f(A(x))$, a contradiction. □

Corollary 1

$$\overline{\overline{q}}_i \leq Q(i) \quad \forall i \geq 1. \tag{20}$$

Using this, we can rewrite equation (11) to obtain the following upper bound for the variance:

$$\sigma_\alpha^2 \leq \frac{2}{\alpha} \sum_{i \geq 1} Q(i) - \mu_\alpha - \mu_\alpha^2 \tag{21}$$

To approximate the summation, we use Euler's summation formula [10],

$$\sum_{i \geq 1} Q(i) = \int_1^\infty Q(x)dx + \sum_{k=1}^{m} \frac{B_k}{k!}(Q^{(k-1)}(\infty) - Q^{(k-1)}(1)) + R_m, \tag{22}$$

where the B_k are the Bernoulli numbers ($B_0 = 1, B_1 = -\frac{1}{2}, B_2 = \frac{1}{6}, B_3 = 0, B_4 = -\frac{1}{30}, \ldots$). From [10] Exercise 1.2.11.2-3, we know that for even m, if $Q^{(m)}(x) \geq 0$ for $x \geq 1$ then

$$\mid R_m \mid \;\leq\; \mid \frac{B_m}{m!}(Q^{(m-1)}(\infty) - Q^{(m-1)}(1)) \mid . \tag{23}$$

We note that, as $x \to \infty$, all derivatives of $Q(x)$ tend to zero, because they all contain the factor $f(Q(x))$, by repeated differentiation of equation (17), and since $Q(\infty) = 0$, we have $f(Q(\infty)) = f(0) = 0$.

In our case, we will apply this formula with $m = 2$. We note that $Q(1) = \overline{\overline{q}}_1 = \alpha\mu_\alpha$ and $Q'(1) = f(Q(1)) = f(\overline{\overline{q}}_1) = \Delta\overline{\overline{q}}_1 = -\overline{q}_1 = -\alpha$. Furthermore, $Q^{(2)}(x) \geq 0$ for $x \geq 1$ because $Q'(x) = f(Q(x))$ is an increasing function. Therefore, we have

$$\sum_{i \geq 1} Q(i) = \int_1^\infty Q(x)dx + \frac{1}{2}Q(1) - \frac{1}{12}Q'(1) + R_2 \leq \int_1^\infty Q(x)dx + \frac{1}{2}\alpha\mu_\alpha + \frac{1}{6}\alpha \tag{24}$$

and therefore the bound for the variance can be written as

$$\sigma_\alpha^2 \leq \frac{2}{\alpha}\int_1^\infty Q(x)dx + \frac{1}{3} - \mu_\alpha^2 \tag{25}$$

Note that, until now, we have not made use of the specific form of the function $Q(x)$. Using now formulas (18) and (12), we obtain the following upper bound for the variance:

Theorem 2 *Under the asymptotic model for an infinite α-full hash table with random probing and RH collision resolution discipline, the variance of the search cost of a random element satisfies (with $\beta = 1/(1-\alpha)$)*

$$\sigma_\alpha^2 \leq \frac{\pi^2}{3} + \frac{1}{3} + O\left(\frac{\ln\beta}{\beta}\right). \tag{26}$$

□

This gives us an upper bound of $3.6232\ldots$ for the variance of Robin Hood Hashing. Although a numerically computed value of approximately 1.883 has been known for a long time, this is the first proof that this variance is bounded by a small constant as $\alpha \to 1$. As Celis *et al.* observed, the fact that the variance is very small can be used to carry out a more efficient *mean-centered search*. If we call X the random variable "search cost of a random key" the expected cost of this modified search is $\Theta(\mathbb{E}|X-\mu_\alpha|)$. But Jensen's inequality implies that

$$\mathbb{E}|X-\mu_\alpha| = \mathbb{E}\sqrt{(X-\mu_\alpha)^2} \leq \sqrt{\mathbb{E}(X-\mu_\alpha)^2} = \sigma_\alpha \tag{27}$$

so, the *mean value* of the search cost of a mean-centered search is proportional to the *standard deviation* of the cost of a standard seach. Theorem 2 then implies that this search algorithm runs in expected constant time in a full table.

3.2 Bounding the tail of RH

We focus now on the tail of the distribution of the search cost, i.e. we study

$$\Pr\{X \geq i\} = \overline{p}_i = \frac{1}{\alpha}\overline{q}_i = \frac{\beta}{\beta-1}\overline{q}_i. \tag{28}$$

We proved earlier that $\overline{\overline{q}}_i \leq Q(i)$. By applying f to both sides and recalling that f is a decreasing function, we have $f(\overline{\overline{q}}_i) \geq f(Q(i))$. Using equations (16) and (17), we have $\Delta\overline{\overline{q}}_i = -\overline{q}_i \geq Q'(i)$, and therefore

$$\Pr\{X \geq i\} \leq -\frac{\beta}{\beta-1}Q'(i) = \frac{\beta}{\beta-1+e^{i-1}}. \tag{29}$$

If we take the upper bound as the tail $\frac{\beta}{\beta-1+e^{x-1}}$ of a continuous probability function, its density function would be

$$p(x) = \frac{\beta e^{x-1}}{(\beta-1+e^{x-1})^2}, \tag{30}$$

which is symmetric around its mean (and mode) located at the point x such that $e^{x-1} = \beta - 1$, i.e., $x = 1 + \ln(\beta-1)$.

As a consequence, by equation (29), the probability that the search cost will exceed this amount by a given number of steps k:

$$\Pr\{X \geq 1 + \ln(\beta-1) + k\} \leq \frac{\beta}{\beta-1}\frac{1}{e^k+1} \to \frac{1}{e^k+1} \tag{31}$$

as $\beta \to \infty$.

Therefore, as the table becomes full, the mean moves to the right without bound, but the distribution remains tightly packed to the right of the mean, and the probability that the search cost exceeds the mean by a given amount decreases exponentially with the distance.

Finally, it is interesting to note that if we shift to the left the density function (30) so it is centered around zero, we obtain

$$p(1 + \ln(\beta-1) + x) = \frac{\beta}{\beta-1}\frac{e^x}{(1+e^x)^2} \tag{32}$$

which, as $\beta \to \infty$, converges to $\frac{e^x}{(1+e^x)^2}$, or, equivalently, $\frac{e^{-x}}{(1+e^{-x})^2}$, the density function of a Logistic(0,1) distribution.

4 Analysis with deletions

We assume a process where we first insert keys until the table reaches load factor α, and then we enter an infinite cycle where we alternate one random insertion followed by one random deletion.

If the distribution of the retrieval cost is given by $p_i(\alpha)$ and a random element is inserted, the effect is described by equation (5). If we then perform a random deletion, the following classical lemma[6] shows that the distribution remains unchanged:

Lemma 2 *Suppose a set contains n balls of colors $1, 2, \ldots, k$, such that the probability that a ball chosen at random is of color i is p_i. Then, if one ball is chosen at random and discarded, the* a posteriori *probability that a random ball is of color i is still p_i.*

Proof: Call p'_i the probability that a random ball is of color i after the deletion. The expected number of balls of color i afterwards is $(n-1)p'_i$, but that number can also be obtained as the expected number before, np_i, minus the expected number of balls of color i lost, i.e.,

$$(n-1)p'_i = np_i - 1 \cdot p_i. \tag{33}$$

The result follows. □

Therefore, equation (5) describes also the probability distribution after one insert-delete step. Now, assume the process reaches a steady state. In that case, the distribution after the insert-delete must be equal to the distribution before, i.e. $p_i(\alpha + \frac{1}{m}) = p_i(\alpha)$, and replacing this in (5) we have

$$p_i(\alpha) = t_i(\alpha) - t_{i+1}(\alpha). \tag{34}$$

and equivalently,

$$\overline{p}_i(\alpha) = t_i(\alpha). \tag{35}$$

These equations play the role that equation (7) did for the case without deletions. Taking tails in both sides of this equation and setting $i = 1$, we can obtain the expected search cost μ_α as

$$\mu_\alpha = \overline{\overline{p}}_1 = \overline{t}_1 = \frac{1}{1-\alpha}, \tag{36}$$

confirming the prediction that the expected successful search cost should approach the expected *unsuccessful* search cost when deletions are allowed.

For RH, from (35) we get $\overline{\overline{p}}_i = \overline{t}_i$, and combining this with (13) we obtain

$$\overline{\overline{p}}_1 = \frac{1}{1-\alpha}, \quad \overline{\overline{p}}_{i+1} = \frac{\alpha \overline{\overline{p}}_i^2}{1 + \alpha \overline{\overline{p}}_i} \tag{37}$$

We can use this recurrence to compute numerically the distribution for RH.

Figure 3 shows the value of the variance of RH as a function of $\beta = 1/(1-\alpha)$, and from the plot we may see that the variance is very close to β. Moreover, Figure 4 shows the distribution of the search cost for the three methods, for $\alpha = 0.99$. As proven in [13] it can be seen that FCFS and LCFS are now identical and have very large dispersion ($\sigma^2_\alpha = \frac{\alpha}{(1-\alpha)^2}$), while RH retains a much more concentrated shape. We prove that this is indeed the case.

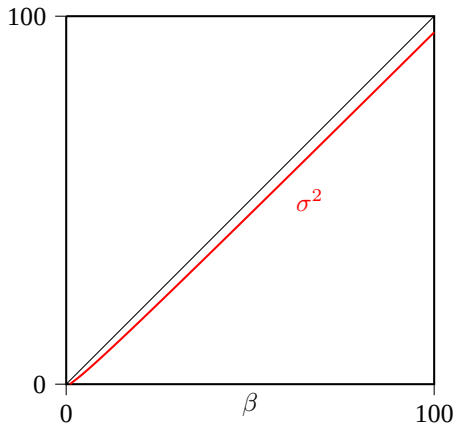


Fig. 3: The variance of RH with deletions as a function of β

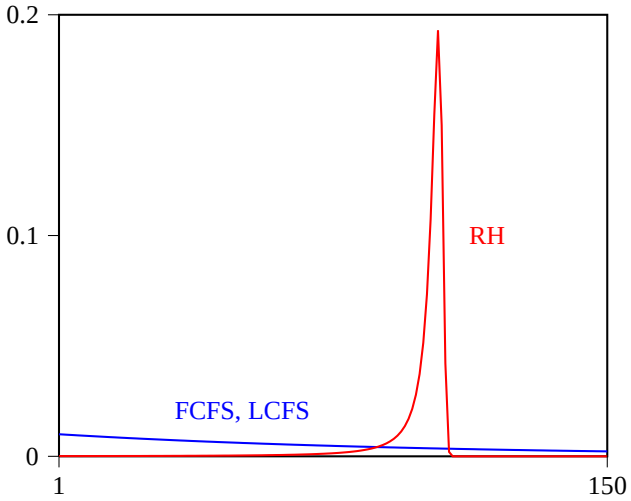


Fig. 4: Distribution of search costs for FCFS, LCFS and RH for $\alpha = 0.99$

4.1 Bounding the variance of RH with deletions

We begin by rewriting the recurrence equation (37) as

$$\overline{\overline{q}}_1 = \beta - 1, \quad \Delta\overline{\overline{q}}_i = -\frac{\overline{\overline{q}}_i}{1+\overline{\overline{q}}_i} \tag{38}$$

This equation is of the form $\Delta\overline{\overline{q}}_i = f(\overline{\overline{q}}_i)$ for $f(x) = -\frac{x}{1+x}$, and all the conditions required in section 3.1 are satisfied, so we can apply the exact same technique used there. Solving the associated differential equation

$$Q'(x) = f(Q(x)), \quad Q(1) = \beta - 1 \tag{39}$$

we find the solution

$$Q(x) = W((\beta-1)e^{\beta-x}), \tag{40}$$

where W is Lambert's function satisfying $x = W(x)e^{W(x)}$. As a consequence, proceeding as in the proof of Theorem 2, we obtain the following result:

Theorem 3 *Under the asymptotic model for an infinite α-full hash table with random probing and RH collision resolution discipline, in the steady state of a sequence of insert-delete operations, the variance of the search cost of a random element satisfies (with $\beta = 1/(1-\alpha)$)*

$$\sigma_\alpha^2 \leq \beta + \frac{1}{3} = \frac{1}{1-\alpha} + \frac{1}{3}. \tag{41}$$

□

This proves our earlier conjecture that the variance was very close to $\frac{1}{1-\alpha}$.

5 Acknowledgements

We are grateful to the anonymous reviewers, for their valuable comments and suggestions, that helped us improve the paper.

References

[1] O. Amble and D. E. Knuth. Ordered-hash-tables. *Computer Journal*, 17(2):135–142, 1974.

[2] R. P. Brent. Reducing-the-retrieval-time-of-scatter-storage-techniques. *CACM*, 16(2):105–109, 1973.

[3] P. Celis. *Robin Hood Hashing*. PhD thesis, University of Waterloo, 1986. Technical Report CS-86-14.

[4] P. Celis, P.-Å. Larson, and J.I. Munro. Robin Hood Hashing. In *26th IEEE Symposium on the Foundations of Computer Science*, pages 281–288, 1985.

[5] Walter Cunto and Patricio V Poblete. Two hybrid methods for collision resolution in open addressing hashing. In *SWAT 88*, pages 113–119. Springer, 1988.

[6] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968.

[7] G. H. Gonnet and J. I. Munro. Efficient-ordering-of-hash-tables. *SIAM Journal on Computing*, 8(3):463–478, 1979.

[8] Leo J. Guibas. The analysis of hashing techniques that exhibit k-ary clustering. *J. ACM*, 25(4):544–555, October 1978.

[9] L.J. Guibas. *The analysis of hashing algorithms*. STAN-CS. Stanford University, 1976.

[10] D.E. Knuth. *Art of Computer Programming Volume 1: Fundamental Algorithms*. Addison-Wesley Publishing Company, 1972.

[11] D.E. Knuth. *The Art of Computer Programming vol. 3 Sorting and Searching*. Addison-Wesley Publishing Company, 1998.

[12] M. Mitzenmacher. A new approach to analyzing robin hood hashing. Preliminary version in *http://arxiv.org/abs/1401.7616*, 2014.

[13] P. Poblete and A. Viola. The effect of deletions on different insertion disciplines for hash tables. In *Brazilian Symposium on Graphs, Algorithms and Combinatorics (GRACO)*, 2001.

[14] P.V. Poblete and J.I. Munro. Last-Come-First-Served Hashing. *Journal of Algorithms*, 10:228–248, 1989.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Additive functionals of d-ary increasing trees

Dimbinaina Ralaivaosaona[1], Stephan Wagner[1†]

Department of Mathematical Sciences, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa, {naina,swagner}@sun.ac.za

A tree functional is called additive if it satisfies a recursion of the form $F(T) = \sum_{j=1}^{k} F(B_j) + f(T)$, where $B_1, \dots, B_k$ are the branches of the tree T and $f(T)$ is a toll function. We prove a general central limit theorem for additive functionals of d-ary increasing trees under suitable assumptions on the toll function. The same method also applies to generalised plane-oriented increasing trees (GPORTs). One of our main applications is a log-normal law that we prove for the size of the automorphism group of d-ary increasing trees, but many other examples (old and new) are covered as well.

Keywords: additive tree functional, increasing trees, random trees, central limit theorem, automorphisms

1 Introduction

In this paper, we are interested in functionals of rooted trees that satisfy an *additive* relation, i.e. a recursion of the form

$$F(T) = \sum_{j=1}^{k} F(B_j) + f(T), \tag{1}$$

where $B_1, \dots, B_k$ are the branches of the tree T and $f(T)$ is a so-called toll function, which often only depends on specific features of the tree such as the size or the root degree, but can in principle be arbitrary. The trees in our context will be labelled; it is assumed that the toll function only depends on the relative order of the labels, not the labels themselves, so that it is also well-defined if the labels are not necessarily $1, 2, \dots, n$. It is consistent with (1) to assume that we have the identity $F(\odot) = f(\odot)$ for the tree $T = \odot$ consisting only of a single labelled vertex. Important examples include

- the number of leaves, which corresponds to the toll function

$$f(T) = \begin{cases} 1 & |T| = 1, \\ 0 & \text{otherwise.} \end{cases}$$

†This material is based upon work supported financially by the National Research Foundation of South Africa under grant number 96236.

- the number of vertices of outdegree k, in which case one can simply take

$$f(T) = \begin{cases} 1 & \text{if the root of } T \text{ has outdegree } k, \\ 0 & \text{otherwise.} \end{cases}$$

- the internal path length, i.e., the sum of the distances from the root to all vertices, which can be obtained from the toll function $f(T) = |T| - 1$.

- the log-product of the subtree sizes [11], also called the "shape functional" [6], corresponding to $f(T) = \log |T|$,

- the logarithm of the size of the automorphism group: here, it is not difficult to see that the toll function is $f(T) = \log(R(T))$, where $R(T)$ is the size of the symmetry group of the collection of root branches.

Such functionals also arise frequently in the study of divide-and-conquer algorithms, e.g. quicksort [9]. An alternative viewpoint is based on the notion of *fringe subtrees*: a fringe subtree of a tree is a subtree induced by a vertex and all its descendants. If we let $\mathcal{F}(T)$ denote the collection of all fringe subtrees of a tree T, then it is easy to verify that

$$F(T) = \sum_{S \in \mathcal{F}(T)} f(S).$$

In particular, the number of occurrences of a specific tree as a fringe subtree is an additive functional (corresponding to the case that the toll function f is an indicator function), and every additive functional can be obtained as a linear combination of such special functionals.

There are several recent papers providing central limit theorems for rather general additive tree functionals [3, 5, 6, 8, 10, 14]. Specifically, Holmgren and Janson [8] proved such a central limit theorem for binary increasing trees (which are also equivalent to binary search trees) and recursive trees. Both are instances of so-called *increasing trees*: labelled trees with the additional property that the labels increase along any path starting at the root.

Varieties of increasing trees were studied systematically in [1] (see also [4, Section 1.3.3]). The exponential generating function $Y(x)$ associated with a variety of increasing trees satisfies a differential equation of the characteristic shape

$$Y'(x) = \Phi(Y(x)), \qquad Y(0) = 0 \tag{2}$$

for some function $\Phi(t)$. Varieties of increasing trees for which a uniformly random tree with a given number of vertices can also be generated by a growth process have been of particular interest. There are three such types [12]:

- The variety of recursive trees is perhaps the most basic instance: these are simply labelled rooted unordered trees ("unordered" meaning that the order of branches does not matter) with the aforementioned property that the labels increase along paths starting at the root. Uniformly random recursive trees can be obtained by the following growth process: starting from a single vertex (the

root, carrying label 1), the vertex labelled n is attached in the n-th step to one of the previous vertices, chosen uniformly at random. As mentioned earlier, the order of children attached to a vertex does not matter. To obtain a canonical representation, one can e.g. always make the newly added vertex the rightmost child. In this example, the function Φ is the exponential function. It is easy to see that the generating function is $Y(x) = -\log(1-x)$, and for every positive integer n, there are $(n-1)!$ recursive trees.

- Plane-oriented recursive trees (PORTs) differ from recursive trees in only one aspect: trees are regarded as embedded in the plane, the order of branches is taken into account. The growth process to generate uniformly random PORTs follows a "preferential attachment" rule: it is essentially the same as for recursive trees, but the probability that the vertex labelled n is attached to a specific vertex v is proportional to 1 plus the current outdegree of v. Here, we have $\Phi(t) = (1-t)^{-1}$, so the generating function is $Y(x) = 1 - \sqrt{1-2x}$, and the number of plane oriented recursive trees with n vertices is $(2n-3)!!$.

  Generalised plane oriented recursive trees (GPORTs) are obtained by introducing an additional parameter: for some positive real number α, we let the probability that the vertex labelled n is attached to a specific vertex v be proportional to α plus the current outdegree of v. An equivalent description uses weighted PORTs: to each PORT T, we associate a weight based on its outdegrees. If $N_j(T)$ is the number of vertices whose outdegree is j, we set

$$w(T) = \prod_{j\geq 1} \binom{\alpha+j-1}{j}^{N_j(T)}.$$

  In choosing a random GPORT, the probability of a tree to be chosen is proportional to its weight. In the exponential generating function $Y(x)$, each tree is also weighted with $w(T)$. The function Φ in (2) is now given by $\Phi(t) = (1-t)^{-\alpha}$. It follows that $Y(x) = 1 - (1-(\alpha+1)x)^{1/(1+\alpha)}$, the total weight of all trees with n vertices is $\prod_{j=1}^{n-1}((\alpha+1)j-1)$.

- Finally, we have the variety of d-ary increasing trees, which will be the focus of this paper: here, every vertex has d possible places to which a child can be attached (for example, in the binary case, there are left and right children). In the construction of uniform d-ary increasing trees by a growth process, we simply attach the vertex labelled n to one of the $(d-1)(n-1)+1$ places available in total, once again selected uniformly at random. Therefore, the probability that the new vertex is attached to an existing vertex v is proportional to d minus the current outdegree of v (in particular, if v already has d children, no further vertices can be attached to it). Here, $\Phi(t) = (1+t)^d$ and $Y(x) = (1-(d-1)x)^{-1/(d-1)} - 1$. The total number of d-ary increasing trees with n vertices is $\prod_{j=1}^{n-1}((d-1)j+1)$.

Remark 1 *We remark that recursive trees and d-ary increasing trees can also be seen as weighted PORTs, with weights*

$$w(T) = \prod_{j\geq 1} c_j^{N_j(T)},$$

where $c_j = \frac{1}{j!}$ for recursive trees (to factor out the different ways of ordering the branches) and $c_j = \binom{d}{j}$ (to take the d possible points of attachment into account) respectively.

In the following, we state and prove a central limit theorem for additive tree functionals of uniformly random d-ary increasing trees under certain technical conditions on the toll function. As mentioned earlier, binary increasing trees (as well as recursive trees) have already been covered in [8]. Since the approach in [8] is based on representations of binary increasing trees and recursive trees that are not available for other classes of increasing trees, and the generating function method of [14] requires the resulting differential equations to be explicitly solvable, which is also not the case, we use a different approach based on moments, as in a paper of Fuchs [7] on the number of fringe subtrees of given size (which is also an additive functional). Although we only discuss the case of d-ary increasing trees in (some) detail, our method also applies to GPORTs, for which we only state the corresponding result in the following section.

This extended abstract only summarises the proof our main theorem and lists some interesting examples to which our result can be applied. Technical details and proofs of all intermediate lemmas will be provided in the full version of this paper.

2 The general central limit theorem

Let us now formulate our main result. In the following, d is fixed, and T_n always denotes a random d-ary increasing tree of order n (except for Theorem 2). We assume that the toll function $f(T)$ satisfies the following conditions:

(C1) $f(T)$ is bounded,

(C2) $\displaystyle\sum_{k\geq 1} \frac{\mathbb{E}|f(T_k)|}{k} < \infty$ and $\mathbb{E}|f(T_n)| \to 0$ as $n \to \infty$.

Under these assumptions, our central limit theorem for additive functionals reads as follows:

Theorem 1 *Let T_n be a uniformly random d-ary increasing tree with n vertices. If the toll function $f(T)$ satisfies (C1) and (C2), then there exist constants μ and σ such that the mean and variance of $F(T_n)$ are asymptotically*

$$\mathbb{E}(F(T_n)) = \mu n + \frac{\mu}{d-1} + o(1), \quad \mathrm{Var}(F(T_n)) = \sigma^2 n + o(n).$$

The constants μ and σ can be represented as

$$\mu = (d-1)\sum_T f(T) \prod_{j=1}^{|T|} \frac{1}{(d-1)j+d} \tag{3}$$

and

$$\sigma^2 = -\frac{\mu^2}{d-1} - (d-1)\sum_T \frac{f(T)^2 - 2f(T)(F(T)-\mu|T|)}{\prod_{j=1}^{|T|}((d-1)j+d)} + d\sum_{T_1}\sum_{T_2} \frac{(d-1)^{1-|T_1|-|T_2|} f(T_1) f(T_2)}{(|T_1|-1)!(|T_2|-1)!} \int_0^1 \phi_{|T_1|}(x)\phi_{|T_2|}(x)dx,$$

where

$$\phi_k(x) = (1-x)^{-1} \int_x^1 (1-w)^{d/(d-1)} w^{k-1} dw.$$

The sums are taken over all d-ary increasing trees. If $\sigma \neq 0$, then the renormalised random variable $(F(T_n) - \mu n)/\sqrt{\sigma^2 n}$ converges weakly to a standard normal distribution.

Remark 2 *We remark that the result remains true if conditions (C1) and (C2) hold for a shifted version $f(T) + c$ (c any constant) of the toll function rather than the toll function itself, since this changes $F(T)$ only by the deterministic quantity $c|T|$.*

Remark 3 *By means of the Cramér-Wold device, we also obtain joint normal distribution of tuples of additive functionals.*

As mentioned earlier, the method used in proving Theorem 1 also applies to GPORTs. Without going into detail, let us just state the corresponding theorem:

Theorem 2 *Let T_n be a random GPORT (with fixed parameter α) with n vertices. If the toll function $f(T)$ satisfies (C1) and (C2), then there exist constants μ and σ such that the mean and variance of $F(T_n)$ are asymptotically*

$$\mathbb{E}(F(T_n)) = \mu n - \frac{\mu}{\alpha+1} + o(1), \quad \mathrm{Var}(F(T_n)) = \sigma^2 n + o(n).$$

The constants μ and σ can be represented as

$$\mu = (\alpha+1) \sum_T w(T) f(T) \prod_{j=1}^{|T|} \frac{1}{(\alpha+1)j + \alpha}$$

and

$$\begin{aligned}\sigma^2 =& \frac{\mu^2}{\alpha+1} - (\alpha+1) \sum_T w(T) \frac{f(T)^2 - 2f(T)(F(T) - \mu|T|)}{\prod_{j=1}^{|T|}((\alpha+1)j+\alpha)} + \\ & \alpha \sum_{T_1} \sum_{T_2} w(T_1) w(T_2) \frac{(\alpha+1)^{1-|T_1|-|T_2|} f(T_1) f(T_2)}{(|T_1|-1)!(|T_2|-1)!} \int_0^1 \varphi_{|T_1|}(x) \varphi_{|T_2|}(x) dx,\end{aligned}$$

where

$$\varphi_k(x) = \int_x^1 (1-w)^{\alpha/(\alpha+1)} w^{k-1} dw.$$

The sums are taken over all PORTs, weighted by $w(T)$. If $\sigma \neq 0$, then the renormalised random variable $(F(T_n) - \mu n)/\sqrt{\sigma^2 n}$ converges weakly to a standard normal distribution.

3 Preliminaries

Recall that the exponential generating function $Y(x)$ of d-ary increasing trees satisfies the differential equation

$$Y'(x) = \Phi(Y(x)), \qquad Y(0) = 0, \tag{4}$$

where $\Phi(t) = (1+t)^d$. The explicit solution is given by $Y(x) = (1-(d-1)x)^{-1/(d-1)} - 1$, and the total number of d-ary increasing trees with n vertices is

$$Y_n = n! \cdot [x^n]Y(x) = \prod_{j=1}^{n-1}((d-1)j+1).$$

Let us first define a multivariate generating function that also incorporates the tree functional F and its toll function f. Specifically, we set

$$Y(x,a,b) = \sum_T \frac{x^{|T|}}{|T|!} e^{aF(T)-bf(T)}.$$

In view of the recursion satisfied by F, (4) becomes

$$\frac{\partial}{\partial x}Y(x,a,a) = \sum_T \frac{x^{|T|-1}}{(|T|-1)!} e^{a(F(T)-f(T))} = \Phi(Y(x,a,0)), \qquad Y(0,a,b) = 0.$$

We set

$$Z(x,a,b) = 1 + Y(xe^{-a\mu},a,b) = 1 + \sum_T \frac{x^{|T|}}{|T|!} e^{aF(T)-a\mu|T|-bf(T)},$$

where μ will be determined later, so that

$$\frac{\partial}{\partial x}Z(x,a,a) = e^{-a\mu}\Phi(Y(xe^{-a\mu},a,0)) = e^{-a\mu}\Phi(Z(x,a,0)-1) = e^{-a\mu}Z(x,a,0)^d.$$

Note that

$$M_n(a) = \frac{[x^n]Z(x,a,0)}{[x^n]Z(x,0,0)} = \frac{n![x^n]Z(x,a,0)}{Y_n}$$

is the moment generating function for the random variable $F(T_n) - \mu|T_n| = F(T_n) - \mu n$ when a random d-ary increasing tree T_n with n vertices is generated. Its derivatives with respect to a, evaluated at 0, yield the moments.

Let the r-th derivative of Z with respect to a be denoted by $Z^{(r)}(x,a,b)$. Our first goal is to determine a differential equation for the function $Z^{(r)}(x,0,0)$. This is done by means of Faà di Bruno's formula. First, we need some further notation regarding integer partitions: we represent partitions of a positive integer r as sequences $\ell = (\ell_1, \ell_2, \ldots)$, where ℓ_j denotes the multiplicity of j. Thus ℓ is a partition of r if $\sum_j j\ell_j = r$. The set of all partitions of r is denoted by $\mathcal{P}(r)$, and we write $|\ell| = \ell_1 + \ell_2 + \cdots$ for the total number of parts in the partition ℓ.

Lemma 3 *The function $Z^{(r)}(x,0,0)$ satisfies the differential equation*

$$\begin{aligned}&\frac{\partial}{\partial x}\Big(Z(x,0,0)^{-d}Z^{(r)}(x,0,0)\Big)\\ &\quad = -Z(x,0,0)^{-d}H_r(x) + \sum_{s=0}^{r}\binom{r}{s}(-\mu)^{r-s}s!\sum_{\substack{\ell\in\mathcal{P}(s)\\ \ell_r\neq 1}}\frac{d!}{(d-|\ell|)!}\prod_{j\geq 1}\frac{1}{\ell_j!j!^{\ell_j}}\Big(\frac{Z^{(j)}(x,0,0)}{Z(x,0,0)}\Big)^{\ell_j}, \quad (5)\end{aligned}$$

where

$$H_r(x) = \sum_{s=1}^{r} \binom{r}{s} \sum_{T} \frac{x^{|T|-1}}{(|T|-1)!} (F(T) - \mu|T|)^{r-s} (-f(T))^s.$$

Note that at this stage, $H_r(x)$ is only considered as a formal power series, convergence is not taken into account. We first analyse this differential equation in the special cases $r = 1$ and $r = 2$ corresponding to mean and variance before we move on to the central limit theorem.

4 Mean and variance

Let us now determine mean and variance of $F(T_n)$. Since the values of the toll function $f(T)$ for $|T| > n$ will not affect the distribution of $F(T_n)$, we can assume in this section that $f(T) = 0$ for $|T| > n$. This means in particular that the functions $H_r(x)$ also depend on n, so we write $H_r^{(n)}(x)$ to emphasize this dependence. For $r = 1$, Equation (5) becomes

$$\frac{\partial}{\partial x}\Big(Z(x,0,0)^{-d} Z^{(1)}(x,0,0) \Big) = -Z(x,0,0)^{-d} H_1^{(n)}(x) - \mu,$$

so

$$Z^{(1)}(x,0,0) = Z(x,0,0)^d \int_0^x \Big(-Z(w,0,0)^{-d} H_1^{(n)}(w) - \mu \Big)\, dw,$$

where

$$H_1^{(n)}(x) = -\sum_{|T| \le n} \frac{x^{|T|-1}}{(|T|-1)!} f(T). \tag{6}$$

If we choose $\mu = \mu^{(n)}$ in such a way that

$$\mu^{(n)} = -(d-1) \int_0^{1/(d-1)} Z(w,0,0)^{-d} H_1^{(n)}(w)\, dw,$$

then we can write

$$Z^{(1)}(x,0,0) = \frac{\mu^{(n)}}{d-1} Z(x,0,0) + R(x), \tag{7}$$

where

$$R(x) = Z(x,0,0)^d \int_x^{1/(d-1)} Z(w,0,0)^{-d} H_1^{(n)}(w)\, dw.$$

The first term on the right side of (7) contributes $\frac{\mu^{(n)}}{d-1}$ to the mean, so it suffices to determine the contribution from $R(x)$. Note that $R(x)$ is a polynomial of degree n whose coefficients can be computed explicitly using the following lemma:

Lemma 4 *If*

$$P(x) = \sum_{k=0}^{n-1} a_k x^k \;\; \textit{and} \;\; Q(x) = (1-x)^{-\beta} \int_x^1 (1-w)^\beta P(w) dw,$$

then $Q(x)$ is a polynomial of degree n with

$$[x^m]Q(x) = -\frac{a_{m-1}}{m+\beta} + \sum_{k=m}^{n-1} \binom{\beta+m-1}{m} \cdot \frac{\Gamma(\beta+1)k!a_k}{\Gamma(\beta+k+2)}$$
$$= \mathcal{O}\left(\frac{|a_{m-1}|}{m} + m^{\beta-1}\sum_{k=m}^{n-1} k^{-\beta-1}|a_k|\right).$$

This lemma gives us in particular an expression for $[x^n]R(x)$, since

$$[x^n]R(x) = [x^n](1-(d-1)x)^{-d/(d-1)} \int_x^{1/(d-1)} (1-(d-1)w)^{d/(d-1)} H_1^{(n)}(w)\, dw$$
$$= (d-1)^{n-1}[x^n](1-x)^{-d/(d-1)} \int_x^1 (1-u)^{d/(d-1)} H_1^{(n)}\Big(\frac{u}{d-1}\Big)\, du.$$

Evaluating the integral in the expression for $\mu^{(n)}$ explicitly gives us

$$\mu^{(n)} = d(d-1)\sum_{m\le n} \frac{1}{((d-1)m+1)((d-1)m+d)Y_m} \sum_{|T|=m} f(T)$$
$$= d(d-1)\sum_{m\le n} \frac{\mathbb{E}(f(T_m))}{((d-1)m+1)((d-1)m+d)}.$$

Putting everything together, we arrive at an explicit formula for the mean:

$$\mathbb{E}(F(T_n)) = \mu^{(n)}n + \frac{\mu^{(n)}}{d-1} + \frac{n![x^n]R(x)}{Y_n}$$
$$= (d(d-1)n+d)\sum_{m\le n} \frac{\mathbb{E}(f(T_m))}{((d-1)m+1)((d-1)m+d)} + \frac{n}{(n+d/(d-1))Y_n} \sum_{|T|=n} f(T)$$
$$= (d(d-1)n+d)\sum_{m<n} \frac{\mathbb{E}(f(T_m))}{((d-1)m+1)((d-1)m+d)} + \mathbb{E}(f(T_n)).$$

If we complete the series and make use of conditions (C1) and (C2), we arrive exactly at the desired asymptotic formula for the mean in Theorem 1. The variance, which is obtained by using Equation (5) for $r = 2$, can be treated in a similar fashion. Without going into detail, under our conditions (C1) and (C2), we can show that

$$\mathrm{Var}(F(T_n)) = c^{(n)}n + o(n),$$

where $c^{(n)}$ is a truncated double series which converges to the constant σ^2 in Theorem 1 as $n \to \infty$.

5 The central limit theorem

We first consider the case that $f(T)$ has finite support. Conditions (C1) and (C2) are then automatically satisfied, hence the results in the previous section for the mean and variance are valid in this case as well. For the central limit theorem, we also need higher moments, for which we have the following statement.

Lemma 5 *If the toll function f has finite support, i.e. there exists a constant K such that $f(T) = 0$ whenever $|T| > K$, then the centred moments of the functional F are asymptotically given by*

$$\mathbb{E}((F(T_n) - \mu n)^r) = \begin{cases} (r-1)!!\sigma^r n^{r/2} + O\left(n^{r/2-1}\right) & r \text{ even}, \\ O\left(n^{(r-1)/2}\right) & r \text{ odd}. \end{cases}$$

Here, μ and σ are as in Theorem 1. Consequently, if $\sigma \neq 0$, then the renormalised random variable

$$\frac{F(T_n) - \mu n}{\sqrt{\sigma^2 n}}$$

converges weakly to a standard normal distribution.

The key to proving this lemma is the fact that the functions H_r are now given by finite sums and therefore trivially represent entire functions. This enables us to apply singularity analysis to the functions $Z^{(r)}(x, 0, 0)$ for arbitrary r, which yields the asymptotics of the centred moments.

To deal with toll functions that are not finitely supported, we employ a trick that was already used in [8, 10]: we approximate them by truncated versions to which we can apply Lemma 5. This approach is based on the following simple yet general lemma.

Lemma 6 *If $(X_n)_{n\geq 1}$ and $(W_{m,n})_{m,n\geq 1}$ are sequences of centred random variables such that*

- *$W_{m,n} \xrightarrow{d}_n W_m$, and $W_m \xrightarrow{d}_m W$, where W has a continuous distribution function,*
- *$\mathrm{Var}(X_n - W_{m,n}) \to_n \gamma_m^2$ and $\gamma_m \to_m 0$,*

then $X_n \xrightarrow{d}_n W$.

We return to additive functionals and assume that the toll function $f(T)$ satisfies conditions (C1) and (C2). For every positive integer m, consider the truncated toll function f_m and the corresponding function F:

$$f_m(T) = \begin{cases} f(T) & |T| \leq m, \\ 0 & \text{otherwise}, \end{cases} \qquad \text{and} \qquad F_m(T) = \sum_{S \in \mathcal{F}(T)} f_m(S) = \sum_{S \in \mathcal{F}(T), |S| \leq m} f(S).$$

From Section 4, we know that the mean and variance of $F_m(T)$ have the asymptotic estimates

$$\mathbb{E}(F_m(T)) = \mu_m n + \frac{\mu_m}{d-1} + o(1) \quad \text{and} \quad \mathrm{Var}(F_m(T)) = \sigma_m^2 n + o(n)$$

as $n \to \infty$. Furthermore, for each m, if $\sigma_m^2 \neq 0$ then $F_m(T)$ satisfies the central limit theorem, and $\mu_m \to \mu$ and $\sigma_m^2 \to \sigma$ as $m \to \infty$. On the other hand, the functional $F(T) - F_m(T)$ is also additive with toll function $f(T) - f_m(T)$. The conditions (C1) and (C2) are both satisfied by the latter toll function, so from the asymptotic formula for the variance we know that

$$\gamma_m^2 = \lim_{n\to\infty} \frac{\mathrm{Var}(F(T_n) - F_m(T_n))}{n} \to_m 0$$

under the conditions on the toll function f. Hence, Lemma 6 applies to the sequences

$$W_{m,n} = \frac{F_m(T_n) - \mathbb{E}(F_m(T_n))}{n} \text{ and } X_n = \frac{F(T_n) - \mathbb{E}(F(T_n))}{n},$$

which proves Theorem 1 for arbitrary toll functions f that satisfy (C1) and (C2).

6 Some applications

6.1 Fringe subtrees of given size and occurrences of specific fringe subtrees

The simplest example of a toll function is perhaps the indicator function of a specific tree S:

$$f(T) = \begin{cases} 1 & T = S, \\ 0 & \text{otherwise.} \end{cases}$$

The associated additive functional is simply the number of occurrences of S on the fringe of a random tree: by an occurrence of S, we mean a fringe subtree that is isomorphic to S (including the relative order of the labels).

In this case, we obtain a central limit theorem with mean and variance only depending on the size of S: if S has k vertices, then

$$\mu = \frac{d-1}{\prod_{j=1}^{k}((d-1)j+d)} \qquad \text{and} \qquad \sigma^2 = -\mu^2\Big(2k + \frac{1}{d-1}\Big) + \mu + \frac{d(d-1)^{1-2k}}{(k-1)!^2}\int_0^1 \phi_k(x)^2\,dx.$$

A closely related functional is the number of fringe subtrees of some given size k (equivalently, the number of vertices with exactly $k-1$ descendants). In particular, the special case $k = 1$ corresponds to the number of leaves. Here, the toll function is given by

$$f(T) = \begin{cases} 1 & |T| = k, \\ 0 & \text{otherwise,} \end{cases}$$

and we obtain a central limit theorem with

$$\mu = \frac{d(d-1)}{((d-1)k+d)((d-1)k+1)} \quad \text{and} \quad \sigma^2 = -\mu^2\Big(2k+\frac{1}{d-1}\Big)+\mu+\frac{d(d-1)^{1-2k}Y_k^2}{(k-1)!^2}\int_0^1 \phi_k(x)^2\,dx.$$

This was already shown by Fuchs [7], who also considered the case that k is not fixed but rather tends to infinity with the size of the tree as well.

6.2 The number of subtrees

The number of subtrees is already a somewhat more complicated example: for Galton-Watson trees, binary increasing trees and recursive trees, it was already studied in [14]. Here, we count all subtrees, i.e. all induced subgraphs that are again trees, not just those on the fringe. It is useful to study an auxiliary quantity first, namely the number of subtrees containing the root: we write $s(T)$ for this number. It is not difficult to see that

$$s(T) = \prod_{j=1}^{k}(1 + s(B_j)),$$

since each subtree induces either the empty set or a subtree containing the root in each of the branches. Taking the logarithm gives us

$$\log(1 + s(T)) = \sum_{j=1}^{k}\log(1 + s(B_j)) + \log(1 + s(T)^{-1}),$$

so $\log(1+s(T))$ is an additive functional with toll function $f(T)=\log(1+s(T)^{-1})$. Simple a priori estimates show that the technical conditions of our general central limit theorem are satisfied: this is because $s(T)\geq|T|$ (since every path from the root to a vertex is also a subtree), which implies that $f(T)=\mathcal{O}(|T|^{-1})$ for all T (even deterministically, not just on average). Thus our main result applies to the functional $s(T)$. As it was shown in [14], the difference between $F(T)=\log(1+s(T))$ and the logarithm of the total number of subtrees (not necessarily containing the root) is $\mathcal{O}(\log|T|)$, so the central limit theorem remains correct for the total number of subtrees.

6.3 The size of the automorphism group

An important motivating example for this paper is the size of the automorphism group. In their article [2], Bóna and Flajolet proved, motivated by questions in phylogenetics, that the logarithm of the size of the automorphism group of uniformly random binary trees is asymptotically normally distributed (they proved this limit law for the number of nodes for which the two branches are isomorphic, which is equivalent). Here, we obtain an analogous statement for d-ary increasing trees. We remark that binary increasing trees are also essentially equivalent to the Yule-Harding model (as opposed to the uniform model) of phylogenetics [13, Section 2.5].

As it was mentioned in the introduction, the relevant toll function is $f(T)=\log(R(T))$, where $R(T)$ is the size of the symmetry group of the collection of root branches. This simplifies considerably in the case of binary trees, where we only have two branches B_1 and B_2. In this case, it follows that

$$f(T)=\begin{cases}\log 2 & \text{if } B_1 \text{ and } B_2 \text{ are isomorphic,}\\ 0 & \text{otherwise.}\end{cases}$$

As one would expect, it is very unlikely for large trees that the two branches are actually isomorphic, which is why the technical condition on the toll function is satisfied. In fact, one can show that $\mathbb{E}(|f(T_n)|)$ decays exponentially for binary increasing trees. We find that the number of automorphisms of a random binary increasing tree asymptotically follows a log-normal law, which parallels the aforementioned result of Bóna and Flajolet.

The same holds more generally for d-ary trees, although the expected value of the toll function does not decay as quickly: in this case, the probability that two branches are isomorphic only decreases at a rate of $\mathcal{O}(|T|^{-2/(d-1)})$, which however is still sufficient.

6.4 The number of orbits

Two vertices of a rooted tree (or generally any graph) are said to belong to the same orbit if there exists an automorphism that maps one of the vertices to the other. The vertex set can thus be partitioned in a natural way into orbits, and the number of orbits can also be regarded as an additive functional. Let us illustrate this for binary increasing trees: if the two branches B_1 and B_2 of T are isomorphic, then $F(T)=F(B_1)+1=F(B_2)+1$, otherwise, $F(T)=F(B_1)+F(B_2)+1$. Hence, the toll function is given by

$$f(T)=\begin{cases}1-F(B_1) & \text{if } B_1 \text{ and } B_2 \text{ are isomorphic,}\\ 1 & \text{otherwise.}\end{cases}$$

Our technical conditions (C1) and (C2) are not completely satisfied in this examples, but it is possible to work around that. First, the expected value of $|f(T_n)|$ does not tend to 0, but the expected value of

$|f(T_n) - 1|$ does, cf. Remark 2. Second, $f(T)$ is not bounded, but this condition can be replaced by the fact that $f(T)$ and $F(T)$ are both $O(|T|)$. Again, everything also remains valid for d-ary increasing trees, although the details are somewhat more intricate.

References

[1] F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. In *CAAP '92 (Rennes, 1992)*, volume 581 of *Lecture Notes in Comput. Sci.*, pages 24–48. Springer, Berlin, 1992.

[2] M. Bóna and P. Flajolet. Isomorphism and symmetries in random phylogenetic trees. *J. Appl. Probab.*, 46(4):1005–1019, 2009.

[3] L. Devroye. Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput.*, 32(1):152–171, 2002/03.

[4] M. Drmota. *Random trees*. SpringerWienNewYork, Vienna, 2009.

[5] J. A. Fill, P. Flajolet, and N. Kapur. Singularity analysis, Hadamard products, and tree recurrences. *J. Comput. Appl. Math.*, 174(2):271–313, 2005.

[6] J. A. Fill and N. Kapur. Limiting distributions for additive functionals on Catalan trees. *Theoret. Comput. Sci.*, 326(1-3):69–102, 2004.

[7] M. Fuchs. Limit theorems for subtree size profiles of increasing trees. *Combin. Probab. Comput.*, 21(3):412–441, 2012.

[8] C. Holmgren and S. Janson. Limit laws for functions of fringe trees for binary search trees and random recursive trees. *Electron. J. Probab.*, 20:no. 4, 51, 2015.

[9] H.-K. Hwang and R. Neininger. Phase change of limit laws in the quicksort recurrence under varying toll functions. *SIAM J. Comput.*, 31(6):1687–1722, 2002.

[10] S. Janson. Asymptotic normality of fringe subtrees and additive functionals in conditioned galton-watson trees. *Random Structures Algorithms*, 48(1):57–101, 2016.

[11] A. Meir and J. W. Moon. On the log-product of the subtree-sizes of random trees. *Random Structures Algorithms*, 12(2):197–212, 1998.

[12] A. Panholzer and H. Prodinger. Level of nodes in increasing trees revisited. *Random Structures Algorithms*, 31(2):203–226, 2007.

[13] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.

[14] S. Wagner. Central limit theorems for additive tree parameters with small toll functions. *Combin. Probab. Comput.*, 24(1):329–353, 2015.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

The Maximum Block Size of Critical Random Graphs[†]

Vonjy Rasendrahasina[1], Andry Rasoanaivo[2], Vlady Ravelomanana[3]

[1]*ENS – Université d'Antananarivo, Madagascar.* rasendrahasina@gmail.com
[2]*LIMA – Université d'Antananarivo, Madagascar.* r.andry.rasoanaivo@gmail.com
[3]*IRIF UMR CNRS 8243 – Université Denis Diderot, France.* vlad@liafa.univ-paris-diderot.fr

Abstract. Let $G(n,\, M)$ be the uniform random graph with n vertices and M edges. Let $\wp_{n,\, M}$ be the maximum block-size of $G(n,\, M)$ or the maximum size of its maximal 2-connected induced subgraphs. We determine the expectation of $\wp_{n,\, M}$ near the critical point $M = n/2$. As $n - 2M \gg n^{2/3}$, we find a constant c_1 such that

$$c_1 = \lim_{n\to\infty} \left(1 - \frac{2M}{n}\right) \mathbb{E}(\wp_{n,\, M})\,.$$

Inside the window of transition of $G(n,\, M)$ with $M = \frac{n}{2}(1 + \lambda n^{-1/3})$, where λ is any real number, we find an exact analytic expression for

$$c_2(\lambda) = \lim_{n\to\infty} \frac{\mathbb{E}\left(\wp_{n,\, \frac{n}{2}(1+\lambda n^{-1/3})}\right)}{n^{1/3}}\,.$$

This study relies on the symbolic method and analytic tools coming from generating function theory which enable us to describe the evolution of $n^{-1/3}\,\mathbb{E}\left(\wp_{n,\, \frac{n}{2}(1+\lambda n^{-1/3})}\right)$ as a function of λ.

Keywords: Random graph, Analytic Combinatorics, Maximum block-size

1 Introduction

Random graph theory Frieze and C. (1997); Bollobás (2001); Janson et al. (2000) is an active area of research that combines algorithmics, combinatorics, probability theory and graph theory. The uniform random graph model $G(n,\, M)$ studied in Erdos and Renyi (1960) consists in n vertices with M edges drawn uniformly at random from the set of $\binom{n}{2}$ possible edges. Erdős and Rényi showed that for many properties of random graphs, graphs with a number of edges slightly less than a given threshold are unlikely to have a certain property, whereas graphs with slightly more edges are almost guaranteed to satisfy the same property, showing paramount changes inside their structures (refer to as *phase transition*). As shown in their seminal paper Erdos and Renyi (1960), when $M = \frac{cn}{2}$ for constant c the largest component of $G(n,\, M)$ has a.a.s. $O(\log n)$, $\Theta(n^{2/3})$ or $\Theta(n)$ vertices according to whether $c < 1$, $c = 1$ or $c > 1$.

[†]Supported by ANR 2010 BLAN 0204 (MAGNUM) and PEPS FASCIDO INSMI-INS2I 2015.

This *double-jump* phenomenon about the structures of $G(n,\ M)$ was one of the most spectacular results in Erdos and Renyi (1960) which later became a cornerstone of the random graph theory. Due to such a dramatic change, researchers worked around the critical value $\frac{n}{2}$ and one can distinguish three different phases: *sub-critical* when $(M - n/2)n^{-2/3} \to -\infty$, *critical* $M = n/2 + O(n^{2/3})$ and *supercritical* as $(M - n/2)n^{-2/3} \to \infty$. We refer to Bollobás Bollobás (2001) and Janson, Łuczak and Ruciński Janson et al. (2000) for books devoted to the random graphs $G(n,\ M)$ and $G(n,\ p)$. If the $G(n,\ p)$ model is the one more commonly used today, partly due to the independence of the edges, the $G(n, M)$ model has more enumerative flavors allowing generating functions based approaches. By setting $p = \frac{1}{n} + \frac{\lambda}{n^{4/3}}$, the stated results of this paper can be extended to the $G(n,\ p)$ model.

Previous works. In graph theory, a block is a maximal 2-connected subgraph (formal definitions are given in Section 2). The problem of estimating the maximum block size has been well studied for some class of graphs. For a graph drawn uniformly from the class of simple labeled planar graphs with n vertices, the expectation of the number of vertices in the largest block is αn asymptotically almost surely (a.a.s) where $\alpha \approx 0.95982$ Panagiotou and Steger (2010); Giménez et al. (2013). They found that the largest block in random planar graphs is related to a distribution of the exponential-cubic type, corresponding to distributions that involve the Airy function Banderier et al. (2001).

For the labeled connected class, these authors proved also independently that a connected random planar graph has a unique block of linear size.

When we restrict to sub-critical graph (graph that the block-decomposition looks tree-like), Drmota and Noy Drmota and Noy (2013) proved that the maximum block size of a random connected graph in an aperiodic[(i)] sub-critical graph class is $O(\log n)$.

For random maps (a map is a planar graph embedded in the plane), Gao and Wormald Gao and Wormald (1999) proved that a random map with n edges has almost surely $n/3$ edges. That is, the probability that the size of the largest block is about $n/3$ tends to 1 as n goes to infinity. This result is improved by Banderier *et al.* Banderier et al. (2001) by finding the density Airy distribution of the map type.

Panagiotou Panagiotou (2009) obtained more general results for any graph class $\mathcal{C}$. He showed that the size of largest block of a random graph from $\mathcal{C}$ with n vertices and m edges belongs to one of the two previous categories ($\Theta(n)$ and $O(\log n)$). In particular, the author pointed out that random planar graphs with cn edges belong to the first category, while random outerplanar and series-parallel graphs with fixed average degree belong to the second category.

For the Erdős-Rényi $G(n, M)$ model, the maximum block-size is implicitly a well-studied graph property when $M = \frac{cn}{2}$ for fixed $c < 1$. For this range, $G(n, M)$ contains only trees and unicyclic components a.a.s. Erdos and Renyi (1960). So, studying maximum block-size and the largest cycle are the same in this case. Denote by $\wp_{n,\,M}$ the maximum block-size of $G(n,\ M)$. It is shown in (Bollobás, 2001, Corollary 5.8) that as $M = \frac{cn}{2}$ for fixed $c < 1$ then $\wp_{n,\,M}$ is a.a.s at most ω for any function $\omega = \omega(n) \to \infty$. Pittel Pittel (1988) then obtained the limiting distribution (amongst other results) for $\wp_{n,\,M}$ for $c < 1$. Note that the results of Pittel are extremely precise and include other parameters of random graphs with c satisfying $c < 1 - \varepsilon$ for fixed $\varepsilon > 0$.

Our results. In this paper, we study the fine nature of the Erdős and Rényi phase transition, with emphasis on what happens as the number of edges is close to $\frac{n}{2}$: within the window of the phase transition and near to it, we quantify the maximum block-size of $G(n,\ M)$.

[(i)] In the periodic case, $n \equiv 1 \mod d$ for some $d > 1$ (see Drmota and Noy (2013) for more details)

For sub-critical random graphs, our finding can be stated precisely as follows :

Theorem 1 *If $n - 2M \gg n^{2/3}$, the maximum block-size $\wp_{n,\,M}$ of $G(n,\,M)$ satisfies*

$$\mathbb{E}(\wp_{n,\,M}) \sim c_1 \left(\frac{n}{n-2M}\right), \tag{1}$$

where $c_1 \approx 0.378\,911$ is the constant given by

$$c_1 = \int_0^\infty \left(1 - e^{-E_1(v)}\right) dv \text{ with } E_1(x) = \frac{1}{2}\int_x^\infty e^{-t}\frac{dt}{t}\,. \tag{2}$$

For critical random graphs, we have the following :

Theorem 2 *Let λ be any real constant and $M = \frac{n}{2}(1 + \lambda n^{-1/3})$. The maximum block-size $\wp_{n,\,M}$ of $G(n,\,M)$ verifies :*

$$\mathbb{E}(\wp_{n,\,M}) \sim c_2(\lambda)\, n^{1/3}, \tag{3}$$

where

$$c_2(\lambda) = \frac{1}{\alpha}\int_0^\infty \left(1 - \sqrt{2\pi}\sum_{r\geq 0}\sum_{d\geq 0} A\left(3r+\frac{1}{2},\lambda\right) e^{-E_1(u)}\, e_{r,d}\left(e^{-u}\right)\right) du \tag{4}$$

$E_1(x)$ is defined in (2), α is the positive solution of

$$\lambda \;=\; \alpha^{-1} \,-\, \alpha\,, \tag{5}$$

the function A is defined by

$$A(y,\lambda) = \frac{e^{-\lambda^3/6}}{3^{(y+1)/3}}\sum_{k\geq 0}\frac{\left(\frac{1}{2}3^{2/3}\lambda\right)^k}{k!\,\Gamma\left((y+1-2k)/3\right)}\,, \tag{6}$$

and the $(e_{r,d}(z))$ are polynomials with rational coefficients defined recursively by (22).

The accuracy of our results is of the same vein as the one on the probability of planarity of the Erdős-Rényi critical random graphs Noy et al. (2015) or on the finite size scaling for the core of large random hypergraphs Dembo and Montanari (2008) which have been also expressed in terms of the Airy function. This function has been encountered in the physics of random graphs Janson et al. (1993) and is shown in Flajolet et al. (1989) related to $A(y,\,\lambda)$ defined by (6) and appearing in our formula (4).

It is important to note that there is *no discontinuity* between Theorems 1 and 2. First, observe that as $M = \frac{n}{2} - \frac{\lambda(n)n^{2/3}}{2}$ with $1 \ll \lambda(n) \ll n^{1/3}$, equation (1) states that $\mathbb{E}(\wp_{n,\,M})$ is about $c_1\frac{n^{1/3}}{\lambda(n)}$. Next, to see that this value matches the one from (3), we argue briefly as follows. In (5), as $\lambda(n) \to -\infty$ we have $\alpha \sim |\lambda(n)|$ and (see (Janson et al., 1993, equation (10.3)))

$$A\left(3r+\frac{1}{2},\lambda\right) \sim \frac{1}{\sqrt{2\pi}|\lambda(n)|^{3r}}\,.$$

Thus, all the terms in the inner double summation 'vanish' except the one corresponding to $r = 0$ and $d = 0$ (this term is the coefficient for graphs without multicyclic components $e_{0,0}^{[k]} = 1$). It is then remarkable that as $\lambda(n) \to -\infty$, $c_2(\lambda(n))$ behaves as $\frac{c_1}{|\lambda(n)|}$.

Outline of the proofs and organization of the paper. In (Flajolet and A., 1990, Section 4), Flajolet and Odlyzko described generating functions based methods to study extremal statistics on random mappings. Random graphs are obviously harder structures but as shown in the masterful work of Janson *et al.* Janson et al. (1993), analytic combinatorics can be used to study in depth the development of the connected components of $G(n,\, M)$. As in Flajolet and A. (1990), we will characterize the expectation of $\wp_{n,\, M}$ by means of truncated generating functions.

Given a family $\mathcal{F}$ of graphs, denote by (F_n) the number of graphs of $\mathcal{F}$ with n vertices. The *exponential generating function* (EGF for short) associated to the sequence (F_n) (or family $\mathcal{F}$) is $F(z) = \sum_{n\geq 0} F_n \frac{z^n}{n!}$. Let $F^{[k]}(z)$ be the EGF of the graphs in $\mathcal{F}$ but *with all blocks of size at most* k. From the formula for the mean value of a discrete random variable X,

$$\mathbb{E}(X) = \sum_{k\geq 0} k\mathbb{P}\left[X = k\right] = \sum_{k\geq 0} \left(1 - \mathbb{P}\left[X \leq k\right]\right),$$

we get a generating function version to obtain

$$\Xi(z) = \sum_{k\geq 0} \left[F(z) - F^{[k]}(z)\right]$$

and the expectation of the maximum block-size of graphs of $\mathcal{F}$ is[ii] $\frac{n![z^n]\Xi(z)}{F_n}$. Turning back to $G(n,\, M)$, realizations of random graphs when M is close to $\frac{n}{2}$ contain a set of trees, some components with one cycle and complex components with 3-regular 3-cores a.a.s. In this paper, our plan is to apply this scheme above by counting realizations of $G(n,\, M)$ with all blocks of size less than a certain value. Once we get the forms of their generating functions, we will use complex analysis techniques to get our results.

This extended abstract is organized as follows. Section 2 starts with the enumeration of trees of given degree specification. We then show how to enumerate 2-connected graphs with 3-regular 3-cores. Combining the trees and the blocks graphs lead to the forms of the generating functions of connected graphs under certain conditions. Section 2 ends with the enumeration of complex connected components with all blocks of size less than a parameter k. Based on the previous results and by means of analytic methods, Section 3 (resp. 4) offers the proof of Theorem 1 (resp. 2).

2 Enumerative tools

Trees of given degree specification. Let $U(z)$ be the exponential generating function of labelled unrooted trees and $T(z)$ be the EGF of rooted labelled trees, it is well-known that[iii]:

$$U(z) = \sum_{n=1}^{\infty} n^{n-2}\frac{z^n}{n!} = T(z) - \frac{T(z)^2}{2} \quad \text{and} \quad T(z) = \sum_{n=1}^{\infty} n^{n-1}\frac{z^n}{n!} = ze^{T(z)}. \tag{7}$$

[ii] For any power series $A(z) = \sum a_n z^n$, $[z^n]A(z)$ denotes the n-th coefficient of $A(z)$, viz. $[z^n]A(z) = a_n$.

[iii] We refer for instance to Goulden and Jackson Goulden and Jackson (1983) for combinatorial operators, to Harary and Palmer Harary and Palmer (1973) for graphical enumeration and to Flajolet and Sedgewick Flajolet and Sedgewick (2009) for the symbolic method of generating functions.

For a tree with exactly m_i vertices of degree i, define its *degree specification* as the $(n-1)$-tuple $(m_1, m_2, \cdots, m_{n-1})$. We have the following.

Lemma 1 *The number of labeled trees with n vertices and degree specification $(m_1, m_2, \cdots m_{n-1})$ with $\sum_{i=1}^{n} m_i = n$ and $\sum_{i=1}^{n} i m_i = 2n-2$ is*

$$a_n(m_1, m_2, \ldots, m_{n-1}) = \frac{(n-2)!}{\prod_{i=1}^{n-1} ((i-1)!)^{m_i}} \binom{n}{m_1,\, m_2, \cdots,\, m_{n-1}}.$$

Proof (sketched). Using Prüfer code, the number of trees with degree sequence $d_1, d_2, \cdots, d_n$ (that is with node numbered i of degree d_i) is $\frac{(n-2)!}{\prod_{i=1}^{n}(d_i-1)!}$. The result is obtained by regrouping nodes of the same degree. ■

Define the associated EGF to $a_n(m_1, m_2, \ldots, m_{n-1})$ with

$$U(\delta_1, \delta_2, \cdots; z) = \sum_{n=2}^{\infty} \sum a_n(m_1, m_2, \ldots, m_{n-1}) \delta_1^{m_1} \delta_2^{m_2} \cdots \delta_{n-1}^{m_{n-1}} \frac{z^n}{n!} \tag{8}$$

where the inner summation is taken other all i such that $\sum i m_i = 2n-2$ and $\sum m_i = n$. Define $U_n(\delta_1, \delta_2, \ldots, \delta_{n-1})$ as

$$U_n(\delta_1, \delta_2, \ldots, \delta_{n-1}) = [z^n] U(\delta_1, \delta_2, \ldots, \delta_{n-1}; z)\,. \tag{9}$$

The following result allows us to compute recursively $U_n(\delta_1, \cdots, \delta_{n-1})$.

Lemma 2 *The generating functions U_n defined in* (9) *satisfy $U_2(\delta_1) = \frac{\delta_1^2}{2}$ and for any $n \geq 3$,*

$$\begin{aligned} U_n(\delta_1, \ldots, \delta_{n-1}) &= \delta_2 U_{n-1}(\delta_1, \ldots, \delta_{n-2}) \\ &+ \sum_{i=2}^{n-2} \delta_{i+1} \int_0^{\delta_1} \frac{\partial}{\partial \delta_i} U_{n-1}(x, \delta_2, \ldots, \delta_{n-2}) dx\,. \end{aligned}$$

Proof. Postponed in the Appendix – 6.1.

Enumerating 2-connected graphs whose kernels are 3-regular. A *bridge* or *cut-edge* of a graph is an edge whose removal increases its number of connected components. Especially, the deletion of such an edge disconnects a connected graph. Similarly an *articulation point* or *cut-vertex* of a connected graph is a vertex whose removal disconnects a graph. A connected graph without an articulation point is called a *block* or a 2-*connected* graph.

Following the terminology of Janson et al. (1993), a connected graph has *excess* r if it has r edges more than vertices. Trees (resp. *unicycles* or *unicyclic components*) are connected components with excess $r = -1$ (resp. $r = 0$). Connected components with excess $r > 0$ are called *complex connected components*. A graph (not necessarily connected) is called *complex* when all its components are complex. The *total excess* of a graph is the number of edges plus the number of acyclic components, minus the number of vertices.

Given a graph, its 2-*core* is obtained by deleting recursively all nodes of degree 1. A *smooth* graph is a graph without vertices of degree one.

The 3-*core* (also called *kernel*) of a complex graph is the graph obtained from its 2-core by repeating the following process on any vertex of degree two : for a vertex of degree two, we can remove it and splice

together the two edges that it formerly touched. A graph is said *cubic* or 3-*regular* if all of its vertices are of degree 3. Denote by $\mathcal{B}_r$ the family of 2-connected smooth graphs of excess r with 3-regular 3-cores and let

$$\mathcal{B} = \bigcup_{r=1}^{\infty} \mathcal{B}_r \,. \tag{10}$$

In this paragraph, we aim to enumerate asymptotically the graphs of $\mathcal{B}_r$. In Chae et al. (2007), the authors established recurrence relations for the numbers of labeled cubic multigraphs with given connectivity, number of double edges and number of loops. For instance, they were able to rederive Wormald's result about the numbers of labeled connected simple cubic graphs with $3n$ simple edges and $2n$ vertices (Chae et al., 2007, equation (24)). They proved that the number of such objects is given by

$$\frac{(2n)!}{3n2^n}\,(t_n - 2t_{n-1}),\ n \geq 2 \tag{11}$$

with

$$t_1 = 0, t_2 = 1 \text{ and } t_n = 3nt_{n-1} + 2t_{n-2} + (3n-1)\sum_{i=2}^{n-3} t_i t_{n-1-i},\ n \geq 2\,. \tag{12}$$

From the sequence (t_n), they found the number of 2-connected multigraphs.

Lemma 3 (Chae, Palmer, Robinson) *Let $g(s,d)$ be the number of cubic block (2-connected labelled) multigraphs with s single edges and d double-edges. Then, the numbers $g(s,d)$ satisfy*

$$g(s,\,d) = 0 \text{ if } s < 2\,, g(s,\,s) = (2s-1)! \text{ and } g(3s,\,0) = \frac{(2s)!}{3s2^s}\,(t_s - 2t_{s-1})$$

with t_s defined as in (12). In all other cases,

$$g(s,\,d) = 2n(2n-1)\left(\frac{(s-1)}{d}g(s-1,d-1) + g(s-3,d)\right)\,.$$

We are now ready to enumerate asymptotically the family $\mathcal{B}_r$. Throughout the rest of this paper if $A(z)$ and $B(z)$ are two EGFs we write

$$A(z) \asymp B(z) \text{ if and only if } [z^n]A(z) \sim [z^n]B(z) \text{ as } n \to +\infty\,.$$

Lemma 4 *For $r \geq 1$, let $B_r(z)$ be the EGF of smooth graphs of excess r whose kernels are 3-regular and 2-connected. $B_r(z)$satisfies $B_r(z) \asymp \frac{b_r}{(1-z)^{3r}}$ where $b_1 = \frac{1}{12}$ and for $r \geq 2$*

$$b_r = \sum_{s+2d=3r} \frac{g(s,\,d)}{2^d(2r)!} \tag{13}$$

with the $g(s,\,d)$ defined as in lemma 3.

Proof. Postponed in the Appendix – 6.2.
We need to count graphs of excess r with at most k vertices so that all the blocks of such structures are of size at most k. We begin our task with the graphs with cubic and 2-connected kernels.

Lemma 5 *Let $\mathcal{B}_r^{[k]}$ be the family of 2-connected graphs of excess r, with at most $k-2r$ vertices of degree two in their 2-cores and whose 3-cores are cubic. For any fixed $r \geq 1$, we have*

$$B_r^{[k]}(z) \asymp b_r \frac{1-z^k}{(1-z)^{3r}}\,.$$

Proof. Postponed in the Appendix – 6.3.
Let $\mathcal{B}_r^{\bullet s}$ be the set of graphs of $\mathcal{B}_r$ such that s vertices of degree two of their 2-cores are distinguished amongst the others. In other words, an element of $\mathcal{B}_r^{\bullet s}$ can be obtained from an element of $\mathcal{B}_r$ by marking (or pointing) s unordered vertices of its 2-core. In terms of generating functions, we simply get (see Harary and Palmer (1973); Goulden and Jackson (1983); Flajolet and Sedgewick (2009)) :

$$B_r^{\bullet s}(z) = \frac{z^s}{s!}\frac{\partial^s}{\partial z^s} B_r(z,\,t)\Big|_{t=z} = \frac{z^s}{s!}\frac{\partial^s}{\partial z^s}\left(b_r \frac{t^{2r}}{(1-z)^{3r}}\right)\Big|_{t=z}, \tag{14}$$

where $B_r(z,t)$ is the bivariate EGF of $\mathcal{B}_r$ with t the variable for the vertices of degree 3. (The substitution $t=z$ is made after the derivations.)
Define

$$b_r^{\bullet s} = \frac{1}{s!} b_r \prod_{i=1}^{s} [3r + (s-i)]$$

so that $B_r^{\bullet s}(z) \asymp \frac{b_r^{\bullet s}}{(1-z)^{3r+s}}$. Now if we switch to the class of graphs with blocks of size at most k then by similar arguments, the asymptotic number of graphs of $\mathcal{B}_r^{\bullet s}$ with s distinguished vertices and at most k vertices on their 2-cores behaves as

$$B_r^{\bullet s,\,[k]}(z) \asymp b_r^{\bullet s} \frac{1-z^k}{(1-z)^{3r+s}}\,.$$

Counting 2-cores with cubic kernels by number of bridges. In this paragraph, we aim to enumerate connected smooth graphs whose 3-cores are 3-regular according to their number of bridges (or cut-edges) and their excess. To that purpose, let $\mathcal{C}_r$ be the family of such graphs with excess $r \geq 0$, and for any $d \geq 0$ let

$$\mathcal{C}_{r,d} \stackrel{\text{def}}{=} \{G \in \mathcal{C}_r : G \text{ is a cycle or its 3-core is 3-regular and has } d \text{ bridges}\}\,.$$

Clearly, we have $\mathcal{C}_{r,0} = \mathcal{B}_r$. If we want to mark the excess of these graphs by the variable w, we simply have

$$C_{r,d}(w,z) = w^r C_{r,d}(z)\,.$$

Lemma 6 *For any $r \geq 1$ and $d \geq 1$,*

$$\begin{aligned} C_{r,d}(z) \;=\; & [w^r] U_{d+1}\Big(B^{\bullet 1}(w,z),\, 2!B^{\bullet 2}(w,z),\, 3!B^{\bullet 3}(w,z) + w^{-1}z, \\ & 4!B^{\bullet 4}(w,z),\,\ldots,\, d!B^{\bullet d}(w,z)\Big)\,\frac{w^d}{(1-z)^d}, \end{aligned}$$

where U_{d+1} are the EGF given by lemma 2, $B_0(w,z) = -\frac{1}{2}\log(1-z) - z/2 - z^2/4$, $B_0^{\bullet s}(w,z) = \frac{1}{s!}\frac{\partial^s}{\partial z^s} B_0(w,z)$ and $B^{\bullet s}(w,z) = \sum_{r\geq 0} w^r B_r^{\bullet s}(z)$.

Proof. Postponed in the Appendix – 6.4

Lemma 7 *For $r \geq 1$ and $d \geq 1$, we have*

$$C_{r,\,d}(z) \asymp \frac{c_{r,\,d}}{(1-z)^{3r}}$$

where the coefficients $c_{r,\,d}$ are defined by

$$c_{r,\,d} = [w^r] U_{d+1}\left(\beta_1(w), \beta_2(w), \beta_3(w) + w^{-1}, \beta_4(w), \ldots, \beta_d(w)\right) w^d,$$

with b_ℓ given by (13) and

$$\beta_s(w) = \frac{(s-1)!}{2} + \sum_{\ell=1}^{r-1} w^\ell b_\ell \prod_{i=1}^{s} [3\ell + (s-i)] \quad \textit{with } s \geq 1.$$

Proof. Postponed in the Appendix – 6.5.

Let us restrict our attention to elements of $\mathcal{C}_{r,\,d}$ with blocks of size at most k. Denote by $\mathcal{C}_{r,\,d}^{[k]}$ this set of graphs. Since they can be obtained from a tree with $d+1$ vertices by replacing each vertex of degree s by a s-marked block (block with a distinguished degree of degree two) of the family $\bigcup_{r=0}^{\infty} \mathcal{B}^{\bullet s,\,[k]}$, we infer the following :

Lemma 8 *For fixed values of r, the EGF of graphs of $\mathcal{C}_{r,d}^{[k]}$ verifies*

$$C_{r,\,d}^{[k]} \asymp c_{r,\,d} \frac{(1-z^k)^{d+1}}{(1-z)^{3r}}\,.$$

From connected components to complex components. Denote by $\mathcal{E}_r^{[k]}$ the family of complex graphs (not necessarily connected) of total excess r with all blocks of sized $\leq k$. Let $E_r^{[k]}$ be the EGF of $\mathcal{E}_r^{[k]}$. Using the symbolic method and sprouting the rooted trees from the smooth graphs counted by $C_{r,\,d}^{[k]}(z)$, we get

$$\sum_{r=0}^{\infty} w^r E_r^{[k]}(z) = \exp\left(\sum_{r=1}^{\infty} w^r \sum_{d\geq 0}^{2r-1} C_{r,\,d}^{[k]}\left(\,T(z)\,\right)\right).$$

We now use a general scheme which relates behavior of connected components and complex components (see for instance (Janson et al., 1993, Section 8)). If $E(w,z) = 1 + \sum_{r\geq 1} w^r E_r(z)$ with $E_r(z) \asymp \frac{e_r}{(1-T(z))^{3r}}$ and $C_r(z) \asymp \frac{c_r}{(1-T(z))^{3r}}$ are EGFs satisfying

$$1 + \sum_{r\geq 1} w^r E_r(z) = \exp\left(\sum_{r=1}^{\infty} w^r C_r(z)\right).$$

then the coefficients (e_r) and (c_r) are related by

$$e_0 = 1 \text{ and } e_r = c_r + \frac{1}{r}\sum_{j=1}^{r-1} j c_j e_{r-j} \text{ as } r \geq 1\,.$$

Similarly, after some algebra we get

Lemma 9 *For fixed $r \geq 1$,*

$$E_r^{[k]}(z) \asymp \sum_{d=0}^{2r-1} \frac{e_{r,d}^{[k]}(T(z))}{(1-T(z))^{3r}}$$

where the functions $(e_{r,d}^{[k]})$ are defined recursively by $e_{0,0}^{[k]}(z) = 1$, $e_{r,d}^{[k]}(z) = 0$ if $d > 2r-1$ and

$$e_{r,d}^{[k]}(z) = c_{r,d}\left(1-z^k\right)^{d+1} + \frac{1}{r}\sum_{j=1}^{r-1} j c_{j,d}\, e_{r-j,d}^{[k]}(z)\,\left(1-z^k\right)^{d+1}\,. \tag{15}$$

Remark. Note that the maximal range $2r-1$ of d appears when the 2-core is a cacti graph (each edge lies on a path or on a unique cycle), each cycle have exactly one vertex of degree three and its 3-core is 3-regular.

3 Proof of Theorem 1

Following the work of Flajolet and Odlyzko Flajolet and A. (1990) on extremal statistics of random mappings, let us introduce the relevant EGF for the expectation of the maximum block-size in $G(n,M)$.

On the one hand, if there are n vertices, M edges and with a total excess r there must be exactly $n-M+r$ acyclic components. Thus, the number of (n,M)-graphs[(iv)] of total excess r without blocks of size larger than k is

$$n![z^n]\frac{U(z)^{n-M+r}}{(n-M+r)!}\left(e^{W_0(z)-\sum_{i=k+1}^{\infty}\frac{T(z)^i}{2i}}\right)E_r^{[k]}(z)\,.$$

where $W_0(z) = -\frac{1}{2}\log(1-T(z)) - \frac{T(z)}{2} - \frac{T(z)^2}{4}$ is the EGF of connected graphs of excess $r=0$ (see (Janson et al., 1993, equation (3.5))).

On the other hand, the EGF of all (n,M)-graphs is

$$G_{n,M}(z) = \sum_{n\geq 0}\binom{\binom{n}{2}}{M}\frac{z^n}{n!}\,.$$

Define

$$\Xi(z) = \sum_{k\geq 0}\left[G_{n,M}(z) - \sum_{n\geq 0}\left(n![z^n]\frac{U(z)^{n-M+r}}{(n-M+r)!}\left(e^{W_0(z)-\sum_{i=k+1}^{\infty}\frac{T(z)^i}{2i}}\right)E_r^{[k]}(z)\right)\frac{z^n}{n!}\right]\,, \tag{16}$$

so that

$$\frac{n![z^n]}{\binom{\binom{n}{2}}{M}}\Xi(z) = \sum_{k\geq 0}\left[1 - \frac{n!}{\binom{\binom{n}{2}}{M}}[z^n]\frac{U(z)^{n-M+r}}{(n-M+r)!}\left(e^{W_0(z)-\sum_{i=k+1}^{\infty}\frac{T(z)^i}{2i}}\right)E_r^{[k]}(z)\right]\,, \tag{17}$$

(iv) Graph with n vertices and M edges

is the expectation of $\wp_{n,\,M}$.
We know from the theory of random graphs that in the sub-critical phase when $n-2M \gg n^{2/3}$ $G(n,M)$ has no complex components with probability $1-O\left(\frac{n^2}{(n-2M)^3}\right)$ (cf (Daudé and Ravelomanana, 2009, Theorem 3.2)). In this abstract, we restrict our attention to the typical random graphs. Otherwise, we will obtain the same result as stated by bounds on the $E_r^{[k]}(z)$ in (16) since

$$1 \le E_r^{[k]}(z) \le E_r(z) \le \frac{e_r T(z)}{(1-T(z))^{3r}}$$

(where inequality between the EGFs means that the coefficients of every power of z obeys the same relation and the last inequality is (Janson et al., 1993, equation (15.2)) with $e_r = \frac{(6r)!}{2^{5r}3^{2r}(3r)!\,(2r)!}$). Assuming that the graphs are typical (i.e. without complex components), $\Xi(z)$ behaves as

$$\Xi(z) \asymp \sum_{k\ge 0}\left[G_{n,M}(z) - \sum_{n\ge 0}\left(n![z^n]\frac{U(z)^{n-M}}{(n-M)!}\frac{e^{-\frac{T(z)}{2}-\frac{T(z)^2}{4}}}{(1-T(z))^{1/2}}\exp\left(-\sum_{j\ge k+1}\frac{T(z)^j}{2j}\right)\right)\frac{z^n}{n!}\right] \quad (18)$$

We need the following Lemma to quantify large coefficients of (18).

Lemma 10 *Let a and b be any fixed rational numbers. For any sequence of integers $M(n)$ such that $\delta n < M$ for some $\delta \in \left[0,\frac{1}{2}\right]$ but $n-2M \gg n^{2/3}$, define*

$$f_{a,b}(n,M) = \frac{n!}{\binom{\binom{n}{2}}{M}}\,[z^n]\,\frac{U(z)^{n-M}}{(n-M)!}\,\frac{U(z)^b\,e^{-T(z)/2-T(z)^2/4}}{(1-T(z))^a}\,.$$

We have

$$f_{a,b}(n,M) \sim 2^b\left(\frac{M}{n}\right)^b\left(1-\frac{M}{n}\right)^b\left(1-\frac{2M}{n}\right)^{1/2-a}.$$

Proof. Postponed in the Appendix – 6.6.
Using Lemma 10 with $a = 1/2$ and $b = 0$, after a bit of algebra (change of variable $u = T(z)$ and approximating the sum by an integral), we first obtain

$$\mathbb{E}(\wp_{n,\,M}) \sim \sum_{k\ge 0}\left(1-\exp\left(-\frac{1}{2}\int_{(k+1)(1-\frac{2M}{n})}^{\infty} e^{-v}\frac{dv}{v}\right)\right).$$

Then by Euler-Maclaurin summation formula and after a change of variable ($(k+1)(1-\frac{2M}{n}) = u$ so $dk = (1-\frac{2M}{n})^{-1}du$), we get the result.

4 Proof of Theorem 2

The following technical result is essentially (Janson et al., 1993, Lemma 3). We give it here in a modified version tailored to our needs (namely involving truncated series). We refer also to the proof of (Flajolet et al., 1989, Theorem 5) and Banderier et al. (2001) for integrals related to the Airy function.

Lemma 11 *Let $M = \frac{n}{2}\left(1+\lambda n^{-1/3}\right)$. Then for any natural integers a, k and r we have*

$$\begin{aligned}&\frac{n!}{\binom{\binom{n}{2}}{M}}\,[z^n]\,\frac{U(z)^{n-M+r}}{(n-M+r)!}\frac{T(z)^a\left(1-T(z)^k\right)}{(1-T(z))^{3r}}\,\exp\left(W_0(z)-\sum_{i=k}^{\infty}\frac{T(z)^i}{2i}\right)\\ =\;&\sqrt{2\,\pi}\,\exp\left(-\sum_{j=k}^{\infty}\frac{e^{-j\alpha n^{-1/3}}}{2j}\right)\left(1-e^{-k\alpha n^{-1/3}}\right)A\left(3r+\frac{1}{2},\,\lambda\right)\left(1+O\left(\frac{\lambda^4}{n^{1/3}}\right)\right),\end{aligned} \tag{19}$$

uniformly for $|\lambda| \le n^{1/12}$ where $A(y,\mu)$ is defined by (6) and α is given by (5).

Proof. Postponed in the Appendix – 6.7.
Using this lemma, equation (17) and next approximating a sum by an integral using Euler-Maclaurin summation, the expectation of $\wp_{n,\,M}$ is about

$$\sum_{k=0}^{n}\left(1-\sum_{r}\sum_{d}\sqrt{2\pi}\exp\left(-\sum_{j=k}^{\infty}\frac{e^{-j\alpha n^{-1/3}}}{2j}\right)e_{r,d}^{[k]}\left(e^{-k\alpha n^{-1/3}}\right)A\left(3r+\frac{1}{2},\lambda\right)\right) \tag{20}$$

$$=\alpha^{-1}n^{1/3}\int_0^{\alpha n^{2/3}}\left(1-\sum_{r}\sum_{d}\sqrt{2\pi}\exp\left(-\int_u^{\infty}\frac{e^{-v}}{2v}dv\right)e_{r,d}\left(e^{-u}\right)A\left(3r+\frac{1}{2},\lambda\right)\right)du \tag{21}$$

where

$$e_{r,d}(z) = c_{r,d}\,(1-z)^{d+1} + \frac{1}{r}\sum_{j=1}^{r-1} j c_{j,d}\,e_{r-j,d}(z)\;(1-z)^{d+1}\;. \tag{22}$$

5 Conclusion

We have shown that the generating function approach is well suited to make precise the expectation of maximum block-size of random graphs. Our analysis is a first step towards a fine description of the various graph parameters inside the window of transition of random graphs.

Acknowledgements: The authors thank the reviewers for their thorough reviews and highly appreciate the comments, remarks and suggestions, which significantly improve the quality of the paper. The authors express their gratitude to the support of the project ANR 2010 BLAN 0204 – MAGNUM and the project PEPS FASCIDO INSMI-INS2I-2015.

References

C. Banderier, P. Flajolet, G. Schaeffer, and M. Soria. Random maps, coalescing saddles, singularity analysis, and airy phenomena. *Random Structures and Algorithms*, 19(3-4):194–246, 2001. ISSN 1098-2418.

B. Bollobás. *Random graphs (2nd edn)*. Cambridge University Press, 2001.

G. B. Chae, E. M. Palmer, and R. R. W. Counting labeled general cubic graphs. *Discrete Math.*, 307: 2979–2992, 2007.

H. Daudé and V. Ravelomanana. Random 2-XORSAT phase transition. *Algorithmica*, Special issue of LATIN 2008:1–18, 2009.

A. Dembo and A. Montanari. Finite size scaling for the core of large random hypergraphs. *Annals of Appl. Probab*, 18:1993 – 2040, 2008.

M. Drmota and M. Noy. Extremal parameters in sub-critical graph classes. In *Proceedings of the 10th Meeting on Analytic Algorithmics and Combinatorics, ANALCO 2013, New Orleans, Louisiana, USA, January 6, 2013*, pages 1 – 7, 2013.

P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

P. Flajolet and O. A. Random mapping statistics. *Advances in Cryptology EUROCRYPT , Lecture Notes in Computer Science*, 434:329–354, 1990.

P. Flajolet and B. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

P. Flajolet, D. E. Knuth, and B. Pittel. The first cycles in an evolving graph. *Discrete Mathematics*, 75 (1-3):167–215, 1989.

A. M. Frieze and M. C. Algorithmic theory of random graphs. *Random Struct. Algorithms*, 10(1-2):5–42, 1997.

Z. Gao and N. C. Wormald. The size of the largest components in random planar maps. *SIAM J. Discrete Math.*, 12(2):217 – 228, 1999.

O. Giménez, M. Noy, and J. J. Rué. Graph classes with given 3-connected components: Asymptotic enumeration and random graphs. *Random Struct. Algorithms*, 42(4):438 – 479, 2013.

I. P. Goulden and D. M. Jackson. *Combinatorial Enumeration*. Wiley, New York, 1983.

F. Harary and E. Palmer. *Graphical Enumeration*. Academic Press, New-York and London, 1973.

S. Janson, D. E. Knuth, T. Łuczak, and B. Pittel. The birth of the giant component. *Random Structures and Algorithms*, 4(3):233–358, 1993.

S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. Wiley-Interscience, 2000.

M. Noy, V. Ravelomanana, and J. J. Rué. On the probability of planarity of a random graph near the critical point. In *Proc. Amer. Math. Soc.*, volume 143, pages 925 – 936, 2015.

K. Panagiotou. Blocks in constrained random graphs with fixed average degree. In *21st International Conference on Formal Power Series and Algebraic Combinatorics (FPSAC '09), Hagenberg*, 2009.

K. Panagiotou and A. Steger. Maximal biconnected subgraphs of random planar graphs. *ACM Transactions on Algorithms Volume 6 Issue 2, Article 31*, 2010.

B. Pittel. A random graph with a sub-critical number of edges. *Trans. Amer. Math. Soc.*, 309:51 – 75, 1988.

E. M. Wright. The number of connected sparsely edged graphs. *Journal of Graph Theory*, 1:317–330, 1977.

E. M. Wright. The number of connected sparsely edged graphs III: Asymptotic results. *Journal of Graph Theory*, 4(4):393–407, 1980.

6 Appendix

6.1 Proof of Lemma 2

The case $n = 2$ is immediate. Let $\mathcal{U}_n$ be the family of trees of size n and $\mathcal{U}_n^\bullet$ be the family of rooted trees of size n whose roots are of degree 1. Deleting the root of the latter trees gives unrooted trees of size $n-1$. Conversely, an element of $\mathcal{U}_n^\bullet$ can be obtained from an element of $\mathcal{U}_{n-1}$, by choosing any vertex and by attaching to this vertex a new vertex which is the root of the newly obtained tree. In terms of EGF, we have :

$$U_n^\bullet(\delta_1, \ldots, \delta_{n-1}) = \sum_{i=1}^{n-2} \delta_1 \delta_{i+1} \frac{\partial}{\partial \delta_i} U_{n-1}(\delta_1, \ldots, \delta_{n-2}).$$

The combinatorial operator that consists to choose a vertex of degree i and add the root is $\delta_1 \delta_i \frac{\partial}{\partial \delta_i}$. The multiplication by the terms $\delta_{i+1}\delta_i^{-1}$ reflects the fact that we have a vertex of degree i that becomes a vertex of degree $i+1$ after the addition of the new vertex of degree 1 (thus the term δ_1). Next, we have to unmark the root which is by construction of degree 1. After a bit of algebra, we obtain the result. ■

6.2 Proof of Lemma 4

The numbers $g(s,\, d)$ count labeled cubic multigraphs. If $s + 2d = 3r$, these multigraphs are exactly the 3-cores of the graphs of the family $\mathcal{B}_r$. Starting from the EGF $g(s,\, d)\frac{w^{3r}z^{2r}}{(2r)!}$ – with the variable w (resp. z) marking the edges (resp. vertices) – if we want to reconstruct from these multigraphs the graphs of the family $\mathcal{B}_r$ each edge w of these multigraphs is substituted by a sequence of vertices of degree 2 introducing the term $\frac{1}{(1-z)}$ (for each of the $3r$ edges of the multigraphs). Next, we have to compensate the symmetry of each double-edge introducing d times the factor $\frac{1}{2!}$. ■

6.3 Proof of Lemma 5

The 3-cores of the graphs of $\mathcal{B}_r$ have as bivariate EGF $b_r w^{3r} t^{2r}$ (with w the variable for the edges and t for the vertices of degree 3). In order to reconstruct the 2-cores of $\mathcal{B}_r^{[k]}$, we insert at most $k - 2r$ vertices on each of the $3r$ paths between the vertices of degree 3. Hence,

$$\begin{aligned} b_r \sum_{i=0}^{k-2r} \binom{3r+i-1}{i} z^i t^{2r} &= b_r \sum_{i=0}^{k-2r} \frac{(3r+i-1)(3r+i-2)\cdots(i+1)}{(3r-1)!}\, z^i t^{2r} \\ &\asymp b_r \frac{1-z^{k+1-2r}}{(1-z)^{3r}} t^{2r} \asymp b_r \frac{1-z^k}{(1-z)^{3r}} t^{2r} \end{aligned}$$

■

6.4 Proof of Lemma 6

Any element of the family $\mathcal{C}_{rd}$ can be obtained from a tree with $d+1$ vertices as follows. Consider a tree $\mathcal{T}$ of size $d+1$. For each vertex v of $\mathcal{T}$ of degree s, we can substitute v by an element of $\mathcal{B}^{\bullet s}$ in $s!$ manners. We distinguish two cases according to the degree of v : vertices of degree 3 can be left unchanged or substituted by elements of $\mathcal{B}^{\bullet 3}$. Thus, the term $3!B^{\bullet s}(w,z) + w^{-1}z$ in (15). Next, each edge of $\mathcal{T}$ can be substituted by a path of length at least 1 with a factor w which parametrizes the excess of the obtained graph. Thus, the factor $\frac{w^d}{(1-z)^d}$. ■

6.5 Proof of Lemma 7

Applying the operator of $\frac{z^s}{s!}\frac{\partial^s}{\partial z^s}$ on unicyclic components gives $b_0^{\bullet s} = \frac{1}{s!}\frac{(s-1)!}{2}$. Define the ordinary generating function of $(b_\ell^{\bullet s})_{\ell\geq 0}$ as

$$b^{\bullet s}(w) = \sum_{\ell=0}^{\infty} b_\ell^{\bullet s} w^\ell = \frac{1}{s!}\left(\frac{(s-1)!}{2} + \sum_{\ell=1}^{\infty} b_\ell \prod_{i=1}^{s}[3\ell + (s-i)]\, w^\ell\right). \tag{23}$$

After a bit of algebra, we get

$$c_{r,\,d} = [w^r]U_{d+1}\left(b^{\bullet 1}(w), 2!b^{\bullet 2}(w), 3!b^{\bullet 3}(w) + w^{-1}, 4!b^{\bullet 4}(w), \ldots, d!b^{\bullet d}(w)\right) w^d. \tag{24}$$

Observe that for any $d \geq 1$, each involved block to obtain an element of $\mathcal{C}_{r,\,d}$ is necessarily of excess at most $r-1$. So, the summation in (23) can be truncated to $r-1$. ■

6.6 Proof of Lemma 10

We split the formula in two parts : $f_{a,b}(m,n) = St(m,n)\cdot Ca(m,n)$ with

$$St(m,n) = \frac{n!}{\binom{\binom{n}{2}}{m}\,(n-m)!} \quad \text{and} \quad Ca(m,n) = [z^n]\,\frac{U(z)^{n-m}}{(n-m)!}\,\frac{U(z)^b\, e^{-t(z)/2 - t(z)^2/4}}{(1-T(z))^a}\,.$$

Using Stirling's formula, we have for the stated range of m

$$\frac{n!m!}{(n-m)!} = \sqrt{2\,\pi}\frac{n^{n+1/2}m^{m+1/2}}{(n-m)^{n-m+1/2}}e^{-2m}\left(1 + O\left(\frac{1}{n}\right)\right).$$

We also have

$$\binom{\binom{n}{2}}{m} = \frac{n^{2m}}{2^m m!}\exp\left(-\frac{m}{n} - \frac{m^2}{n^2} + O\left(\frac{m}{n^2}\right) + O\left(\frac{m^3}{n^4}\right)\right).$$

Next, we get

$$St(m,n) = \left(\frac{2\pi nm}{n-m}\right)^{1/2}\frac{2^m n^n m^m}{n^{2m}(n-m)^{n-m}}\,\exp\left(-2m + \frac{m}{n} + \frac{m^2}{n^2}\right)\left(1 + O\left(\frac{1}{n}\right)\right). \tag{25}$$

For $Ca(m,n)$, in using Cauchy integral's formula and substituting z by ze^{-z}, we obtain :

$$Ca(m,n) = \frac{2^{m-n}}{2\pi i} \oint \left(2T(z) - T(z)^2\right)^{n-m} \frac{U(z)^b\, e^{-T(z)/2 - T(z)^2/4}}{(1-T(z))^a} \frac{dz}{z^{n+1}} \tag{26}$$

$$= \frac{2^{m-n}}{2\pi i} \oint g(z) e^{nh(z)} \frac{dz}{z} \tag{27}$$

where

$$\begin{aligned} g(z) &= \frac{(z - z^2/2)^b\, e^{-z/2 - z^2/4}}{(1-z)^{a-1}}\,, \\ h(z) &= z - \frac{m}{n}\log z + \left(1 - \frac{m}{n}\right)\log(2-z)\,. \end{aligned}$$

$h'(z) = 0$ for $z = 1$ or $z = 2m/n$. $h''(1) = 2m/n - 1 < 0$ and $h''(2m/n) = \frac{n(n-2m)}{4m(n-m)} > 0$. As in Flajolet et al. (1989), we can apply the saddle-point method integrating around a circular path $|z| = 2m/n$. Let $\Phi(\theta)$ be the real part of $h(2m/ne^{i\theta})$. We have

$$\Phi(\theta) = 2\frac{m}{n}\cos\theta + \left(1 - 2\frac{m}{n}\right)\log 2 - \frac{m}{n}\log\left(\frac{m}{n}\right) + \frac{\left(1 - \frac{m}{n}\right)}{2}\log\left(1 + \frac{m^2}{n^2} - 2\frac{m}{n}\cos\theta\right)$$

and

$$\Phi'(\theta) = -2\frac{m}{n}\sin\theta + \frac{(1-m/n)m}{n\left(1 + m^2/n^2 - 2m/n\cos\theta\right)}\sin\theta\,.$$

We note that $\Phi(\theta)$ is a symmetric function of θ. Fix sufficiently small positive constant θ_0. Then, $\Phi(\theta)$ takes its maximum value at $\theta = \theta_0$ as $\theta \in [-\pi, -\theta_0] \cup [\theta_0, \pi]$. In fact,

$$\Phi(\theta) - \Phi(\pi) = 4\frac{m}{n} + \left(1 - \frac{m}{n}\right)\log\left(\frac{n-m}{n+m}\right) + O(\theta^2)\,.$$

Therefore, if $\theta \to 0$ $\Phi(\theta) > \Phi(\pi)$. Also, $\Phi'(\theta) = 0$ for $\theta = 0$ and $\theta = \theta_1$ (for some $\theta_1 > 0$). Standard calculus show that $\Phi(\theta)$ is decreasing from 0 to θ_1 and then increasing from θ_1 to π. We also have

$$h^{(p)}(z) = (p-1)!\left((-1)^p\frac{m}{n\,z^p} - \frac{(n-m)}{n\,(2-z)^p}\right)\,, \quad p \geq 2\,.$$

Hence,

$$h(2me^{i\theta}/n) = h(2m/n) + \sum_{p\geq 2}\xi_p(e^{i\theta} - 1)^p\,,$$

where $\xi_p = \frac{(2m/n)^p}{p!}h^{(p)}(2m/n)$ and $|\xi_p| \leq \frac{m}{np}\left(\frac{2m}{n}\right)^p + \frac{n-m}{np}$. We then have

$$|\sum_{p\geq 4}\xi_p(e^{i\theta} - 1)^p| = O(\theta^4)\,.$$

This allows us to write

$$h(2m/ne^{i\theta}) = h(2m/n) - \frac{m(n-2m)}{2\,n(n-m)}\theta^2 - i\frac{(n^2-5nm+2m^2)m}{6n(n-m)^2}\theta^3 + O(\theta^4)\,.$$

Let $\tau = n(n-m)/\left(m(n-2m)\right)$ and

$$\theta_0 = \left(\frac{(n-m)}{(n-2m)m}\right)^{1/2}\cdot\omega(n) = \sqrt{\frac{\tau}{n}}\cdot\omega(n)$$

where we need a function $\omega(n)$ satisfying $n\theta_0^2 \gg 1$ but $n\theta_0^3 \ll 1$ as $n\to\infty$. We choose

$$\omega(n) = \frac{(n-2m)^{1/4}}{n^{1/6}}\,. \tag{28}$$

We can now use the magnitude of the integrand at θ_0 to bound the error and our choice of θ_0 verifies

$$|g(2m/ne^{i\theta_0})\left(\exp\left(nh(2m/ne^{i\theta_0})\right) - \exp\left(nh(2m/n)\right)\right)| = O\left(e^{-\omega(n)^2/2}\right)\,.$$

Thus, we obtain

$$Ca(m,n) = \frac{2^{m-n}}{2\pi}\int_{-\theta_0}^{\theta_0} g\left(2\frac{m}{n}e^{i\theta}\right)\exp\left(nh(2m/ne^{i\theta})\right)d\theta \times \left(1+O\left(e^{-\omega(n)^2/2}\right)\right)\,.$$

We replace θ by $\frac{\tau^{1/2}}{n^{1/2}}t$. The integral in the above equation leads to

$$\left(\frac{\tau}{n}\right)^{1/2}\int_{-\omega(n)}^{\omega(n)} g\left(\frac{2m}{n}\exp\left(it\sqrt{\tau/n}\right)\right)\exp\left(nh\left(\frac{2m}{n}\exp\left(it\sqrt{\tau/n}\right)\right)\right)dt\,.$$

Expanding $g(2m/ne^{it\sqrt{\tau/n}})$, we obtain

$$\begin{aligned}\left(\frac{\tau}{n}\right)^{1/2}\int_{-\omega(n)}^{\omega(n)} & g\left(2m/n\right)\left(1 - i\,\frac{2m\tau^{1/2}(n^2-2m^2)}{n^{5/2}(n-2m)}\,t + O\left(\frac{n^2}{(n-2m)^3}t^2\right)\right)\\ & \times \qquad \exp\left(nh\left(\frac{2m}{n}\exp\left(it\sqrt{\tau/n}\right)\right)\right)dt\,.\end{aligned}$$

Observe that our choice of $\omega(n)$ in (28) and the hypothesis $n-2m \gg n^{2/3}$ justify such an expansion. Similarly, using the expansion of $h(2m/ne^{it\sqrt{\tau/n}})$ yields

$$\begin{aligned}\left(\frac{\tau}{n}\right)^{1/2}\int_{-\omega(n)}^{\omega(n)} & g\left(2m/n\right)\left(1 - i\,\frac{2m\tau^{1/2}(n^2-2m^2)}{n^{5/2}(n-2m)}\,t + O\left(\frac{n^2}{(n-2m)^3}t^2\right)\right)\\ \times \quad & \exp\left(nh\left(\frac{2m}{n}\right) - \frac{1}{2}t^2\right)\\ \times \quad & \left(1 - i\,\frac{(n^2-5nm+2m^2)}{6(n-m)^{1/2}m^{1/2}(n-2m)^{3/2}}\,t^3 + O\left(\frac{n}{(n-2m)^2}t^4\right)\right)dt\,.\end{aligned}$$

Using the symmetry of the function, we can cancel terms such as it and it^3 (in fact all odd powers of t). Standard calculations show also that for m in the stated ranges, the multiplication of the factors of it and it^3 leads to a term of order of magnitude $O(n^2/(n-2m)^3 t^4)$. Therefore we obtain,

$$\begin{aligned} Ca(m,n) &= \frac{2^{m-n}}{2\pi}\left(\frac{\tau}{n}\right)^{1/2} g\left(2m/n\right) e^{nh(2m/n)} \int_{-\omega(n)}^{\omega(n)} e^{-t^2/2}\left(1-O\left(\frac{n^2}{(n-2m)^3}t^4\right)\right) dt \\ Ca(m,n) &= 2^{m-n}\left(\frac{\tau}{2\pi\, n}\right)^{1/2} g\left(2m/n\right) e^{nh(2m/n)}\left(1-e^{-O(\omega(n)^2)}-O\left(\frac{n^2}{(n-2m)^3}\right)\right). \end{aligned} \tag{29}$$

Multiplying (25) and (29) leads to the result after nice cancellations. (Note that the error terms $e^{-O(\omega(n)^2)}$ and $O(1/n)$ can be regrouped with the $O(n^2(n-2m)^{-3})$.) ■

6.7 Proof of Lemma 11

Proof. Using Stirling's formula, we get

$$\begin{aligned} \mathrm{St}(M,n) &= \frac{n!}{\binom{n}{2}\choose M}\frac{1}{(n-M+r)!} \\ &= \sqrt{2\pi n}\,\frac{2^{n-M+r}}{n^r}\,\exp\left(-\frac{\lambda^3}{6}+\frac{3}{4}-n\right) \\ &\qquad \times\left(1+O\left(\frac{\lambda^4}{n^{1/3}}\right)\right). \end{aligned} \tag{30}$$

Using Cauchy integral's formula and substituting z by ze^{-z}, we obtain :

$$\begin{aligned} \mathrm{Ca}(M,n) &= [z^n]\,U(z)^{n-M+r}\,\frac{T(z)^a\,(1-T(z)^k)}{(1-T(z))^{3r}}\,e^{(V(z)-\sum_{j=k}^{\infty}\frac{T(z)^j}{2j})} \\ &= \frac{1}{2\pi i}\oint\left(T(z)-\frac{T(z)^2}{2}\right)^{n-M+r}\frac{T(z)^a\,e^{-T(z)/2-T(z)^2/4-\sum_{j=k}^{\infty}T(z)^j/2j}}{(1-T(z))^{3r+1/2}}\frac{dz}{z^{n+1}} \\ &= \frac{2^{M-n-r}e^n}{2\pi i}\oint g(u)\,\exp\left(nh(u)\right)\frac{du}{u}, \end{aligned} \tag{31}$$

where the integrand has been splitted into

$$g(u)=\frac{u^a\,(2u-u^2)^r\,e^{-u/2-u^2/4-\sum_{j=k}^{\infty}u^j/2j}\,(1-u^k)}{(1-u)^{3r-1/2}}$$

and

$$h(u)=u-1-\log u-\left(1-\frac{M}{n}\right)\log\,\frac{1}{1-(u-1)^2}\,.$$

The contour in (31) should keep $|u|<1$. Precisely at the critical value $M=\frac{n}{2}$ we also have $h(1)=h'(1)=h''(1)=0$. This triple zero accounts in the procedure Janson, Knuth, Łuczak and Pittel used

when investigating the value of the integral for large n. Let $\nu = n^{-1/3}$, and let α be the positive solution of (5). Following the proof of (Janson et al., 1993, Lemma 3), we will evaluate (31) on the path $z = e^{-(\alpha+it)\nu}$, where t runs from $-\pi n^{1/3}$ to $\pi n^{1/3}$:

$$\oint f(z)\,\frac{dz}{z} = i\nu \int_{-\pi n^{1/3}}^{\pi n^{1/3}} f(e^{-(\alpha+it)\nu})\,dt\,.$$

The main contribution to the value of this integral comes from the vicinity of $t = 0$. The magnitude of $e^{h(z)}$ depends on the real part of $h(z)$, viz. $\Re h(z)$. $\Re h(e^{-(\alpha+it)\nu})$ decreases as $|t|$ increases and $|e^{nh(z)}|$ has its maximum on the circle $z = e^{-(\alpha+it)\nu}$ when $t = 0$.

We have $nh(e^{-s\nu})$

$$n\,h(e^{-s\nu}) = \tfrac{1}{3}\,s^3 + \tfrac{1}{2}\lambda s^2 + O\big((\lambda^2 s^2 + s^4)\nu\big)\,,$$

uniformly in any region such that $|s\nu| < \log 2$. In (Janson et al., 1993, equation (10.7)), the authors define

$$A(y,\mu) = \frac{1}{2\pi i}\int_{\Pi(1)} s^{1-y} e^{K(\mu,s)}\,ds\,,$$

where $K(\mu, s)$ is the polynomial

$$K(\mu,s) = \frac{(s+\mu)^2(2s-\mu)}{6} = \frac{s^3}{3} + \frac{\mu s^2}{2} - \frac{\mu^3}{6}$$

and $\Pi(\alpha)$ is a path in the complex plane that consists of the following three straight line segments:

$$s(t) = \begin{cases} -e^{-\pi i/3}\,t, & \text{for} -\infty < t \le -2\alpha; \\ \alpha + it\sin\pi/3, & \text{for} -2\alpha \le t \le +2\alpha; \\ e^{+\pi i/3}\,t, & \text{for} +2\alpha \le t < +\infty\,. \end{cases}$$

In particular, they proved that $A(y,\mu)$ can be expressed as (6).
For the function $g(u)$, we have

$$\begin{aligned} g(e^{-s\nu}) &= \frac{\left(2e^{-s\nu} - e^{-2s\nu}\right)^r}{\left(1-e^{-s\nu}\right)^{3r-1/2}}\, e^{-as\nu - e^{-s\nu}/2 - e^{-2s\nu}/4 - \sum_{j=k}^{\infty} e^{\frac{-js\nu}{2j}}}\,\left(1-e^{-ks\nu}\right) \\ &= (s\nu)^{1/2-3r} e^{-3/4-\sum_{j=k}^{\infty} e^{-js\nu}/2j}\,\left(1-e^{-ks\nu}\right)\,\left(1+O(s\nu)\right)\,. \end{aligned}$$

For $g(u)e^{nh(u)}$ in the integrand of (31), we have

$$\begin{aligned} e^{-\lambda^3/6} f(e^{-s\nu}) &= e^{-3/4-\sum_{j=k}^{\infty} e^{-js\nu}/2j} \nu^{1/2-3r}\,\left(1-e^{-ks\nu}\right)\,, s^{1-(3r+1/2)} e^{K(\lambda,\,s)} \\ &\times \left(1 + O(s\nu) + O(\lambda^2 s^2 \nu) + O(s^4\nu)\right) \end{aligned}$$

when $s = O(n^{1/12})$. Finally,

$$\frac{e^{-\lambda^3/6}}{2\pi i}\oint g(u)e^{nh(u)}\frac{du}{u} = \exp\left(-3/4 - \sum_{j=k}^{\infty} e^{-j\alpha\nu}/2j\right)\,\left(1-e^{-k\alpha\nu}\right)$$

$$\times \quad \nu^{3/2-3r}\, A(3r+\frac{1}{2},\, \lambda) + O\left(\nu^{5/2-3r} e^{-\lambda^3/6} \lambda^{3r/2+1/4}\right)$$

where the error term has been derived from those already in Janson et al. (1993). The proof of the lemma is completed by multiplying (30) and (31). ■

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

The Cover Time of Deterministic Random Walks for General Transition Probabilities

Takeharu Shiraga[1†]

[1] *Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan*
takeharu.shiraga@inf.kyushu-u.ac.jp

Abstract. The *deterministic random walk* is a deterministic process analogous to a random walk. While there are some results on the cover time of the *rotor-router* model, which is a deterministic random walk corresponding to a simple random walk, nothing is known about the cover time of deterministic random walks emulating general transition probabilities. This paper is concerned with the *SRT-router* model with multiple tokens, which is a deterministic process coping with general transition probabilities possibly containing irrational numbers. For the model, we give an upper bound of the cover time, which is the first result on the cover time of deterministic random walks for general transition probabilities. Our upper bound also improves the existing bounds for the rotor-router model in some cases.

Keywords: rotor router model, stack walk, multiple random walk, mixing time, cover time

1 Introduction

Previous works for the cover time of random walks A *random walk* is a fundamental stochastic process on a graph, in which a token successively transits to neighboring vertices chosen at random. The expected cover time (this paper simply says *cover time*) of a random walk on a finite graph is the expected time until every vertex has been visited by the token. The cover time is a fundamental measure of a random walk, and it has been well investigated.

Aleliunas et al. [3] showed that the cover time of a *simple random walk*, in which a neighboring vertex is chosen uniformly at random, is upper bounded by $2m(n-1)$ for any connected graph, where m denotes the number of edges and n denotes the number of vertices. Feige [16, 17] showed that the cover time is lower bounded by $\big(1-o(1)\big)n\log n$ and upper bounded by $\big(1+o(1)\big)(4/27)n^3$ for any graph.

Motivated by a faster cover time, the cover time by more than one token has also been investigated. Broder et al. [8] gave an upper bound of the cover time of k independent parallel simple random walks (k-simple random walks) when tokens start from stationary distribution. For an arbitrary initial configuration of tokens, Alon et al. [4] showed that the cover time of k-simple random walks is upper bounded by $\big((\mathrm{e}+o(1))/k\big)t_{\mathrm{hit}}\log n$ for any graph if $k \le \log n$, where e is Napier's constant and t_{hit} denotes the (maximum) *hitting time*. Elsasser and Sauerwald [15] gave an better upper bound for large k of $\mathrm{O}\big(t^* + (t_{\mathrm{hit}}\log n)/k\big)$ for any graph if $k \le n$, where t^* is the *mixing time*.

†Supported by JSPS KAKENHI Grant Number 15J03840.

Ikeda et al. [21] took another approach for speeding up, which uses *general* transition probabilities (beyond simple random walks). They invented *β-random walk*, consisting of irrational transition probabilities in general, and showed that the cover time is $\mathrm{O}(n^2 \log n)$. Nonaka et al. [26] showed that the cover time of a *Metropolis-walk*, which is based on the *Metropolis-Hastings algorithm*, is $\mathrm{O}(n^2 \log n)$ for any graph.

Little is known about the cover time by multiple tokens with general transition probabilities. Elsasser and Sauerwald [15] gave a general lower bound of $\Omega\big((n \log n)/k\big)$ for any transition probabilities and for any $n^\varepsilon \leq k \leq n$, where $0 < \varepsilon < 1$ is a constant.

Previous works for the cover time of deterministic random walks From the view point of the *deterministic* graph exploration, the *rotor-router model*, which is a deterministic process analogous to a simple random walk, is well studied recently. In this model, each vertex u sends tokens one by one to neighboring vertices in the round robin fashion, i.e., u serves tokens to a neighboring vertex v with a ratio about $1/\delta(u)$, where $\delta(u)$ is the number of neighbors.

Yanovski et al. [31] studied the asymptotic behavior of the rotor-router model, and proved that any rotor-router model always stabilizes to a traversal of an Eulerian cycle after $2mD$ steps at most, where D denotes the diameter of the graph. Bampas et al. [6] gave examples of which the stabilization time gets $\Omega(mD)$. Their results imply that the cover time of a single token version of a rotor-router model is $\Theta(mD)$ in general. Another approach to examine the cover time of the rotor-router model is connecting qualities of a random walk and the *visit frequency* $X_v^{(T)}$ of the rotor-router model, where $X_v^{(T)}$ denotes the total number of times that tokens visited vertex v by time T. Holroyd and Propp [20] showed that $|\pi_v - X_v^{(T)}/T| \leq K\pi_v/T$, where K is an constant independent of T, and π is the stationary distribution of the corresponding random walk. This theorem says that $X_v^{(T)}/T$ converges to π_v as T increasing. Using this fact, Friedrich and Sauerwald [19] gave upper bounds of the cover time for many classes of graphs.

To speed up the cover time, the rotor-router model with $k > 1$ tokens is studied by Dereniowski et al. [13]. They gave an upper bound $\mathrm{O}(mD/\log k)$ for any graph when $k = \mathrm{O}\big(\mathrm{poly}(n)\big)$ or $2^{\mathrm{O}(D)}$, and also gave an example of $\Omega(mD/k)$ as a lower bound. Kosowski and Pajak [24] gave a modified upper bound of the cover time for many graph classes by connecting $X_v^{(T)}$ and the corresponding simple random walk. They showed that the upper bound is $\mathrm{O}\big(t^* + (\Delta/\delta)(mt^*/k)\big)$ for general graphs, where Δ/δ is the maximum/minimum degree.

Beyond the rotor-router model, which corresponds to a simple random walk, the *deterministic random walk* for general transition probabilities has been invented, that is each vertex u deterministically serves tokens on u to a neighboring vertex v with a ratio about $P_{u,v}$, where $P_{u,v}$ denotes the transition probability from u to v of a corresponding random walk (See Section 2.2 for the details). Holroyd and Propp [20] provides the *stack walk* (Shiraga et al. [28] called it *SRT-router model*), and showed a connection between the visit frequency and hitting probabilities. Shiraga et al. [28] investigated functional-router model, which is a more general framework, and gave an analysis on the *single vertex discrepancy* between the SRT-router model and its corresponding random walk. As far as we know, nothing is known about the cover time of deterministic random walks for general transition probabilities.

Our results This paper is concerned with the cover time of the deterministic random walk according to general transition probabilities with k tokens, while previous results studied the rotor-router model (corresponding to simple transition probabilities). We give an upper bound of the cover time for any SRT-

router model imitating any ergodic and reversible transition matrix possibly containing irrational numbers (Theorem 4.1). Precisely, the upper bound is $\mathrm{O}\big(t^* + m't^*/k\big)$ for any number of tokens $k \geq 1$, where $m' = \max_{u\in V}(\delta(u)/\pi_u)$. This is the first result of an upper bound of the cover time for deterministic random walks imitating general transition probabilities, as far as we know. Theorem 4.1 implies that the upper bound of the cover time of the rotor-router model is $\mathrm{O}\big(t^* + mt^*/k\big)$ for any graph (Corollary 4.2). For $k = 1$, this bound matches to the existing bound $\mathrm{O}(mD)$ by [31] when $t^* = \mathrm{O}(D)$. This bound is better than $\mathrm{O}(mD/\log k)$ by [13] when t^* is small or k is large. Our bound also improves the bound $\mathrm{O}\big(t^* + (\Delta/\delta)(mt^*/k)\big)$ by [24] in Δ/δ factor for inhomogeneous graphs.

In our proof, we investigate the connection between the visit frequency $X_v^{(T)}$ of the SRT-router model and the corresponding multiple random walks with general transition probabilities. This approach is an extension of [20, 19, 24]. In precise, we show that $|\pi_v - (X_v^{(T)}/kT)| < K\pi_v/T$ holds for any reversible and ergodic transition matrices, where π_v is the stationary distribution of the corresponding transition matrix and K is constant independent of T. This upper bound extends the result of [20] to $k > 1$ tokens and general transition probabilities.

Related topics for deterministic random walks As a highly related topic, there are several results on the single vertex discrepancy between a configuration of tokens of a multiple deterministic random walk and an expected configuration of tokens of its corresponding random walk. Rabani et al. [27] gave an upper bound of the single vertex discrepancy of the *diffusive model*, and gave the framework of the analysis. The single vertex discrepancy on several basic structures were widely studied, e.g., constant upper bound for the lattice [12, 11, 14], lower bound for the tree [10], d-dimensional hyper cube [18, 1], etc. Berenbrink et al. [7] gave a sophisticated upper bound on d-regular graphs. To cope with general rational transition probabilities, rotor-router model on multidigraphs is studied in [23, 22]. The SRT-router model is investigated in [28, 29]. They examined the discrepancy between this model and general Markov chains under natural assumptions. Recently, Chalopin et al. [9] gave the upper and lower bound of the stabilization time for the rotor-router model with many tokens.

2 Preliminaries

2.1 Random walk / Markov chain

Let $V = \{1, 2, \ldots, n\}$ be a finite state set, and let $P \in \mathbb{R}_{>0}^{n\times n}$ be a transition matrix on V. P satisfies $\sum_{v\in V} P_{u,v} = 1$ for any $u \in V$, where $P_{u,v}$ denotes the (u, v)-entry of P. It is well known that any ergodic[(i)] P has a unique *stationary distribution* $\pi \in \mathbb{R}_{>0}^n$ (i.e., $\pi P = \pi$), and the limit distribution is π (i.e., $\lim_{t\to\infty} \xi P = \pi$ for any probability distribution ξ on V). To discuss the *convergence* formally, we introduce the *total variation distance* and the *mixing time*. Let ξ and ζ be probability distributions on V, then the total variation distance $\mathcal{D}_{\mathrm{tv}}$ between ξ and ζ is defined by

$$\mathcal{D}_{\mathrm{tv}}(\xi, \zeta) \stackrel{\text{def.}}{=} \max_{A\subseteq V} \left| \sum_{v\in A} (\xi_v - \zeta_v) \right| = \frac{1}{2} \|\xi - \zeta\|_1 = \frac{1}{2} \sum_{v\in V} |\xi_v - \zeta_v|. \tag{1}$$

[(i)] P is ergodic if P is irreducible ($\forall u, v \in V, \exists t > 0, P_{u,v}^t > 0$) and aperiodic ($\forall v \in V, \mathrm{GCD}\{t \in \mathbb{Z}_{>0} \mid P_{v,v}^t > 0\} = 1$).

The *mixing time* of P is defined by[ii]

$$\tau(\varepsilon) \stackrel{\text{def.}}{=} \max_{u \in V} \min \left\{ t \in \mathbb{Z}_{\geq 0} \mid \mathcal{D}_{\mathrm{tv}}(P^t_{u,\cdot}, \pi) \leq \varepsilon \right\} \tag{2}$$

for $\varepsilon > 0$, and

$$t^* \stackrel{\text{def.}}{=} \tau(1/4), \tag{3}$$

which is often used as an important characterization of P (cf.[25]).

In this paper, we assume P is ergodic and *reversible*. We call a P is reversible if $\pi_u P_{u,v} = \pi_v P_{v,u}$ holds for any $u, v \in V$. For example, transition matrices of the β-random walk [21] and the Metropolis walk [26] are both reversible.

Notations of multiple random walks Let $\mu^{(0)} = (\mu^{(0)}_1, \ldots, \mu^{(0)}_n) \in \mathbb{Z}^n_{\geq 0}$ denote an initial configuration of k tokens over V. At each time step $t \in \mathbb{Z}_{\geq 0}$, each token on $v \in V$ moves independently to $u \in V$ with probability $P_{u,v}$. Let $\mu^{(t)} = (\mu^{(t)}_1, \ldots, \mu^{(t)}_n) \in \mathbb{R}^n_{\geq 0}$ denote the *expected* configuration of tokens at time $t \in \mathbb{Z}_{\geq 0}$: then $\mu^{(t)} = \mu^{(0)} P^t$ holds[iii]. Note that the definitions of the mixing times say that $\mathcal{D}_{\mathrm{tv}}(\mu^{(t)}/k, \pi) \leq \varepsilon$ after $t \geq \tau(\varepsilon)$.

2.2 SRT-router model

To imitate random walks with general transition probabilities possibly containing irrational numbers, the deterministic process based on *low-discrepancy sequences* (cf. [5, 30]) were proposed, called *stack walk* in [20] and *SRT-router model* in [28]. In this section, we describe the definition of this model.

Let $\mathcal{N}(v)$ denote the (out-)neighborhood[iv] of v, i.e., $\mathcal{N}(v) = \{u \in V \mid P_{v,u} > 0\}$. In this model, k tokens move according to *SRT-router* $\sigma_v : \mathbb{Z}_{\geq 0} \to \mathcal{N}(v)$ defined on each $v \in V$ for a given P. Given $\sigma_v(0), \ldots, \sigma_v(i-1)$, inductively $\sigma_v(i)$ is defined as follows. First, let

$$T_i(v) = \{u \in \mathcal{N}(v) \mid |\{j \in [0, i) \mid \sigma_v(j) = u\}| - (i+1)P_{v,u} < 0\},$$

where $[z, z') = \{z, z+1, \ldots, z'-1\}$ (and we remark $[z, z) = \emptyset$). Then, let $\sigma_v(i)$ be $u^* \in T_i(v)$ minimizing the value

$$\frac{\left|\{j \in [0, i) \mid \sigma_v(j) = u\}\right| + 1}{P_{v,u}}$$

over choices $u \in T_i(v)$. If there are two or more such $u \in T_i(v)$, then let u^* be the minimum in them in an arbitrary prescribed order. Then, the sequence $\sigma_v(0), \sigma_v(1), \ldots$ satisfies the following *low-discrepancy property* for any v and P (cf. [5, 30]).

Proposition 2.1 *[5, 30] For any P,*

$$\left| \left|\{j \in [0, z) \mid \sigma_v(j) = u\}\right| - z \cdot P_{v,u} \right| < 1$$

holds for any $v, u \in V$ and for any integer $z > 0$.

[ii] $P^t_{u,v}$ denotes the (u, v) entry of P^t, and $P^t_{u,\cdot}$ denotes the u-th row vector of P^t.

[iii] In this paper, $(\mu^{(0)} P^t)_v$ denotes the v-th element of the vector $\mu^{(0)} P^t$, i.e., $(\mu^{(0)} P^t)_v = \sum_{u \in V} \mu^{(0)}_u P^t_{u,v}$.

[iv] If P is reversible, $u \in \mathcal{N}(v)$ if and only if $v \in \mathcal{N}(u)$, and then we abuse $\mathcal{N}(v)$ for in-neighborhood of $v \in V$

Let $\chi^{(0)} = \mu^{(0)}$ and $\chi^{(t)} \in \mathbb{Z}^n_{\geq 0}$ denote the configuration of k tokens at time $t \in \mathbb{Z}_{\geq 0}$ in a SRT-router model ($\sum_{v \in V} \chi_v^{(t)} = k$). Then, SRT-router model works as follows. At first time step ($t = 0$), there are $\chi_v^{(0)}$ tokens on vertex v, and each v serves tokens to neighbors according to $\sigma_v(0), \sigma_v(1), \ldots, \sigma_v(\chi_v^{(0)}-1)$. In other words, $|\{j \in [0, \chi_v^{(0)}) \mid \sigma_v(j) = u\}|$ tokens move from v to u, and $\chi_u^{(1)} = \sum_{v \in V} |\{j \in [0, \chi_v^{(0)}) \mid \sigma_v(j) = u\}|$. Next time step ($t = 1$), there are $\chi_v^{(1)}$ tokens on vertex v, and each v serves tokens to neighbors according to $\sigma_v(\chi_v^{(0)}), \sigma_v(\chi_v^{(0)}+1), \ldots, \sigma_v(\chi_v^{(0)}+\chi_v^{(1)}-1)$, and $\chi^{(2)}$ is defined in a similar way. In general, let $Z_{v,u}^{(t)}$ denote the number of tokens moving from v to u at time t. Then $Z_{v,u}^{(t)}$ is defined as

$$Z_{v,u}^{(t)} = \left| \left\{ j \in [0, \chi_v^{(t)}) \mid \sigma_v(X_v^{(t)} + j) = u \right\} \right|, \tag{4}$$

where $X^{(T)} = \sum_{t=0}^{T-1} \chi^{(t)}$ (and we remark $X_v^{(0)} = 0$ for any $v \in V$), and $\chi^{(t+1)}$ is defined by

$$\chi_u^{(t+1)} = \sum_{v \in V} Z_{v,u}^{(t)} = \sum_{v \in \mathcal{N}(u)} Z_{v,u}^{(t)} \tag{5}$$

for any $u \in V$. Note that

$$\sum_{u \in V} Z_{v,u}^{(t)} = \sum_{u \in \mathcal{N}(v)} Z_{v,u}^{(t)} = \chi_v^{(t)} \tag{6}$$

holds for any $v \in V$. For the SRT-router model, we have the following basic proposition, based on Proposition 2.1.

Proposition 2.2

$$\left| \sum_{t=0}^{T} (Z_{v,u}^{(t)} - \chi_v^{(t)} P_{v,u}) \right| < 1$$

holds for any P and for any $T \geq 0$.

Proof: From the definition of $Z_{v,u}^{(t)}$, it is not difficult to check that

$$\begin{aligned} \sum_{t=0}^{T} Z_{v,u}^{(t)} &= \sum_{t=0}^{T} \left| \left\{ j \in [0, \chi_v^{(t)}) \mid \sigma_v(X_v^{(t)} + j) = u \right\} \right| \\ &= \sum_{t=0}^{T} \left| \left\{ j \in [X_v^{(t)}, X_v^{(t)} + \chi_v^{(t)}) \mid \sigma_v(j) = u \right\} \right| = \left| \left\{ j \in [0, X_v^{(T+1)}) \mid \sigma_v(j) = u \right\} \right| \end{aligned}$$

and $\sum_{t=0}^{T} \chi_v^{(t)} P_{v,u} = X_v^{(T+1)} P_{v,u}$. Then, Proposition 2.2 is obtained by Proposition 2.1 by letting $z = X_v^{(T+1)}$. □

3 Analysis of the Visit Frequency

As a preliminary of the analysis of the cover time of the SRT-router model, we investigate the upper bound of $|X_w^{(T)} - M_w^{(T)}|$, where $M^{(T)} = \sum_{t=0}^{T-1} \mu^{(t)}$ (and we remark that $M_v^{(0)} = 0$ for any $v \in V$). Let $\delta(v) = |\mathcal{N}(v)|$ and $\Delta = \max_{v \in V} \delta(v)$.

Theorem 3.1 *Suppose that P is ergodic and reversible. Then,*

$$\left| X_w^{(T)} - M_w^{(T)} \right| \leq 3\pi_w t^* \max_{u \in V} \frac{\delta(u)}{\pi_u} = \mathrm{O}\left(\frac{\pi_{\max}}{\pi_{\min}} t^* \Delta \right)$$

holds for any $w \in V$ and for any $T > 0$.

From Theorem 3.1, we get the following corollary 3.2, like Theorem 4 of [20].

Corollary 3.2 *Suppose that P is ergodic and reversible. Then,*

$$\left| \pi_w - \frac{X_w^{(T)}}{kT} \right| \leq \frac{3t^*}{2T} + \frac{3\pi_w t^* \max_{u \in V} \frac{\delta(u)}{\pi_u}}{kT} = \frac{K\pi_w}{T}$$

holds for any $w \in V$ and for any $T > 0$, where $K = \mathrm{O}(\frac{t^*}{\pi_w} + \frac{t^* \Delta}{\pi_{\min} k})$ is a constant independent of T.

Note that Corollary 3.2 gives the upper bound for SRT-router models with k tokens, while Theorem 4 of [20] is for rotor-router models with a single token. Corollary 3.2 also means that $\left| \pi_w - \frac{X_w^{(T)}}{kT} \right| \leq \varepsilon$ if $T \geq 3 \left(\frac{1}{2} + \frac{\pi_w \Delta}{\pi_{\min} k} \right) t^* \varepsilon^{-1}$.

To prove the Theorem 3.1, we begin with the following lemma. In the following arguments, we assume that P is ergodic and reversible.

Lemma 3.3

$$X_w^{(T)} - M_w^{(T)} = \sum_{t=0}^{T-2} \sum_{u \in V} \sum_{v \in \mathcal{N}(u)} \sum_{t'=0}^{T-t-2} (Z_{v,u}^{(t')} - \chi_v^{(t')} P_{v,u})(P_{u,w}^t - \pi_w)$$

holds for any $w \in V$ and for any $T > 1$.

Proof: We use the following lemma to prove Lemma 3.3.

Lemma 3.4 *[28] (Lemma 4.1.)*

$$\chi_w^{(T)} - \mu_w^{(T)} = \sum_{t=0}^{T-1} \sum_{u \in V} \sum_{v \in \mathcal{N}(u)} (Z_{v,u}^{(t)} - \chi_v^{(t)} P_{v,u})(P_{u,w}^{T-t-1} - \pi_w)$$

holds for any $w \in V$ and for any $T > 0$. ■

By the definitions of $X^{(T)}, M^{(T)}$ and Lemma 3.4,

$$\begin{aligned} X_w^{(T)} - M_w^{(T)} &= \sum_{t'=0}^{T-1}(\chi^{(t')} - \mu^{(t')}) = \sum_{t'=1}^{T-1}(\chi^{(t')} - \mu^{(t')}) \\ &= \sum_{t'=1}^{T-1}\sum_{t=0}^{t'-1}\sum_{u\in V}\sum_{v\in\mathcal{N}(u)}(Z_{v,u}^{(t)} - \chi_v^{(t)}P_{v,u})(P_{u,w}^{t'-t-1} - \pi_w) \end{aligned} \tag{7}$$

holds. The second equation holds since $\chi^{(0)} = \mu^{(0)}$. Let $\phi_u^{(t)} = \sum_{v\in\mathcal{N}(u)}(Z_{v,u}^{(t)} - \chi_v^{(t)}P_{v,u})$, for convenience. Then,

$$(7) = \sum_{t'=1}^{T-1}\sum_{t=0}^{t'-1}\sum_{u\in V}\phi_u^{(t)}(P_{u,w}^{t'-t-1} - \pi_w) = \sum_{u\in V}\sum_{t'=1}^{T-1}\sum_{t=0}^{t'-1}\phi_u^{(t'-t-1)}(P_{u,w}^{t} - \pi_w) \tag{8}$$

holds. Carefully exchanging the variables of the summation, we obtain

$$\sum_{t'=1}^{T-1}\sum_{t=0}^{t'-1}\phi_u^{(t'-t-1)}(P_{u,w}^{t} - \pi_w) = \sum_{t=0}^{T-2}\sum_{t'=t+1}^{T-1}\phi_u^{(t'-t-1)}(P_{u,w}^{t} - \pi_w) = \sum_{t=0}^{T-2}\sum_{t'=0}^{T-t-2}\phi_u^{(t')}(P_{u,w}^{t} - \pi_w). \tag{9}$$

Combining (8) and (9), we obtain

$$\begin{aligned} (8) &= \sum_{u\in V}\sum_{t=0}^{T-2}\sum_{t'=0}^{T-t-2}\phi_u^{(t')}(P_{u,w}^{t} - \pi_w) \\ &= \sum_{u\in V}\sum_{t=0}^{T-2}\sum_{t'=0}^{T-t-2}\sum_{v\in\mathcal{N}(u)}(Z_{v,u}^{(t')} - \chi_v^{(t')}P_{v,u})(P_{u,w}^{t} - \pi_w). \end{aligned}$$

□

Proof of Theorem 3.1: It is trivial for $T = 1$, hence we assume $T > 1$. By Lemma 3.3 and Proposition 2.2,

$$\begin{aligned} \left|X_w^{(T)} - M_w^{(T)}\right| &= \left|\sum_{t=0}^{T-2}\sum_{u\in V}\sum_{v\in\mathcal{N}(u)}\sum_{t'=0}^{T-t-2}(Z_{v,u}^{(t')} - \chi_v^{(t')}P_{v,u})(P_{u,w}^{t} - \pi_w)\right| \\ &\le \sum_{t=0}^{T-2}\sum_{u\in V}\sum_{v\in\mathcal{N}(u)}\left|\sum_{t'=0}^{T-t-2}(Z_{v,u}^{(t')} - \chi_v^{(t')}P_{v,u})\right|\left|P_{u,w}^{t} - \pi_w\right| \\ &< \sum_{t=0}^{T-2}\sum_{u\in V}\sum_{v\in\mathcal{N}(u)}\left|P_{u,w}^{t} - \pi_w\right| = \sum_{t=0}^{T-2}\sum_{u\in V}\delta(u)\left|P_{u,w}^{t} - \pi_w\right| \end{aligned} \tag{10}$$

holds. By the reversibility of P,

$$(10) = \sum_{t=0}^{T-2}\sum_{u\in V}\delta(u)\left|\frac{\pi_w}{\pi_u}(P_{w,u}^{t} - \pi_u)\right| \le \pi_w \max_{u\in V}\frac{\delta(u)}{\pi_u}\sum_{t=0}^{T-2}\sum_{u\in V}\left|P_{w,u}^{t} - \pi_u\right| \tag{11}$$

holds. By the definition of total variation distance (1),

$$\sum_{t=0}^{T-2}\sum_{u\in V}\left|P^t_{w,u}-\pi_u\right| = 2\sum_{t=0}^{T-2}\mathcal{D}_{\mathrm{tv}}(P^t_{w,\cdot},\pi) \tag{12}$$

holds. Now, we use the following lemma.

Lemma 3.5 *[28] (Lemma 4.2.) For any $v\in V$ and for any $T>0$,*

$$\sum_{t=0}^{T-1}\mathcal{D}_{\mathrm{tv}}\left(P^t_{v,\cdot},\pi\right) \le \frac{1-\gamma}{1-2\gamma}\,\tau(\gamma)$$

holds for any γ $(0<\gamma<1/2)$.

Thus, we have

$$(12) \le 2\cdot\frac{1-(1/4)}{1-2\cdot(1/4)}\tau(1/4) = 3t^* \tag{13}$$

and we obtain the claim. □

Proof of Corollary 3.2: Notice that

$$\begin{aligned}\left|\pi_w-\frac{X^{(T)}_w}{kT}\right| &= \frac{\left|kT\pi_w-X^{(T)}_w\right|}{kT} \le \frac{\left|kT\pi_w-M^{(T)}_w\right|+\left|M^{(T)}_w-X^{(T)}_w\right|}{kT}\\ &\le \frac{\left|M^{(T)}_w-kT\pi_w\right|}{kT}+\frac{3\pi_w t^*\max_{u\in V}\frac{\delta(u)}{\pi_u}}{kT},\end{aligned}$$

where the last inequality follows Theorem 3.1. Thus, it is sufficient to prove that $|M^{(T)}_w-kT\pi_w|\le 3kt^*/2$. Note that $\sum_{t=0}^{T-1}\sum_{u\in V}\mu^{(0)}\pi_w = kT\pi_w$ holds since $\sum_{v\in V}\mu^{(0)}=k$ from the definition, and also note that $M^{(T)}_w=\sum_{t=0}^{T-1}\mu^{(t)}_w=\sum_{t=0}^{T-1}\sum_{u\in V}\mu^{(0)}_u P^t_{u,w}$ holds by the definitions. Then,

$$\begin{aligned}\left|M^{(T)}_w-kT\pi_w\right| &= \left|\sum_{t=0}^{T-1}\sum_{u\in V}\mu^{(0)}_u P^t_{u,w}-\sum_{t=0}^{T-1}\sum_{u\in V}\mu^{(0)}\pi_w\right| = \left|\sum_{t=0}^{T-1}\sum_{u\in V}\mu^{(0)}_u(P^t_{u,w}-\pi_w)\right|\\ &\le \sum_{u\in V}\mu^{(0)}_u\sum_{t=0}^{T-1}|P^t_{u,w}-\pi_w| \end{aligned}\tag{14}$$

holds. By Lemma 3.5 and the definition of total variation distance (1),

$$\sum_{t=0}^{T-1}|P^t_{u,w}-\pi_w| \le \sum_{t=0}^{T-1}\mathcal{D}_{\mathrm{tv}}(P^t_{u,\cdot},\pi) \le \frac{3}{2}t^*. \tag{15}$$

Combining (14) and (15), $|M^{(T)}_w-kT\pi_w|\le 3kt^*/2$ holds, and we obtain the claim. □

4 Bound of the Cover Time

Combining techniques of the analysis of the visit frequency and reversible Markov chains, we obtain the cover time of SRT-router models. Let

$$T_{cover} = \min\left\{T \in \mathbb{Z}_{\geq 0} \mid X_v^{(T)} \geq 1 \; holds \; for \; any \; v \in V\right\}. \tag{16}$$

First, we show the following theorem.

Theorem 4.1 *Suppose P be ergodic and reversible. Then,*

$$T_{cover} \leq 2t^* + 1 + \frac{12\max_{u\in V}\frac{\delta(u)}{\pi_u}\cdot t^*}{k} = \mathrm{O}\left(\max\left\{\frac{t^*\Delta}{\pi_{\min}k}, t^*\right\}\right)$$

holds for any initial configuration of $k \geq 1$ tokens.

Theorem 4.1 is the first result of the cover time for deterministic random walks imitating general transition probabilities possibly containing irrational transition probabilities. Applying Theorem 4.1 to the transition matrix of simple random walk on G, we obtain the following corollary.

Corollary 4.2 *For any G and for any initial configuration of $k \geq 1$ tokens,*

$$T_{cover} \leq 2t^* + 1 + \frac{24mt^*}{k} = \mathrm{O}\left(\max\left\{\frac{mt^*}{k}, t^*\right\}\right)$$

holds for any rotor-router model on G, where t^* is the mixing time of the simple random walk on G.

The upper bound of [24] (Theorem 4.1, proposition 4.2, and Theorem 4.5) is $\mathrm{O}\big(t^* + (\Delta/\delta)(mt^*/k)\big)$, where Δ/δ is the maximum/minimum degree of the graph. Hence Corollary 4.2 improves this bound for inhomogeneous graphs. Compare to the $\mathrm{O}(mD/\log k)$ by [13] (Theorem 3.3 and 3.7), our bound is better when $t^* = \mathrm{O}\big(D(k/\log k)\big)$ (when t^* is small or k is large).

To prove Theorem 4.1, we check the following lemma.

Lemma 4.3 *Suppose P is ergodic and reversible. Then,*

$$P_{u,w}^t \geq \frac{\pi_w}{4}$$

folds for any $u, w \in V$ if $t \geq 2t^*$.

Proof: The *separation distance* [2] is defined by

$$s(t) = \max_{u,v\in V}\left(1 - \frac{P_{u,v}^t}{\pi_v}\right). \tag{17}$$

This distance satisfies $s(t+t') \leq s(t)s(t')$ for any $t, t' \geq 1$ (submultiplicativity property, Lemma 3.7 of [2]). We have the following lemma for the reversible P.

Lemma 4.4 *[25] (Lemma 19.3.) Suppose P is reversible. then,*

$$s(2t) \leq 1 - \big(1 - \bar{d}(t)\big)^2$$

holds for any $t \geq 0$, where $\bar{d}(t) = \max_{u,v\in V}\mathcal{D}_{\mathrm{tv}}(P_{u,\cdot}^t, P_{v,\cdot}^t)$.

It is known that

$$\bar{d}(t^*) \leq \frac{1}{2} \tag{18}$$

holds when P is ergodic (see (4.34) of [25]). Combining these facts, we have

$$1 - \frac{P^t_{u,w}}{\pi_w} \quad \leq \quad s(t) \leq s(2t^*) \leq 1 - (1 - \bar{d}(t^*))^2 \leq 1 - \left(1 - \frac{1}{2}\right)^2 = \frac{3}{4},$$

and we obtain the claim. □

Proof of Theorem 4.1: Lemma 4.3 gives us a lower bound of $P^t_{u,w}$ for any $u, w \in V$, $t \geq 2t^*$ and for any reversible and ergodic P. It provides a lower bound of $M^{(T)}_w$, like [24].

$$M^{(T)}_w \quad = \quad \sum_{t=0}^{T-1} \sum_{u \in V} \mu^{(0)}_u P^t_{u,w} \geq \sum_{t=2t^*}^{T-1} \sum_{u \in V} \mu^{(0)}_u P^t_{u,w} \geq \sum_{t=2t^*}^{T-1} \sum_{u \in V} \mu^{(0)}_u \frac{\pi_w}{4} = \frac{k\pi_w(T - 2t^*)}{4}. \tag{19}$$

By Theorem 3.1 and (19), we obtain that

$$X^{(T)}_w \quad \geq \quad M^{(T)}_w - 3\pi_w t^* \max_{u \in V} \frac{\delta(u)}{\pi_u} \geq \frac{k\pi_w(T - 2t^*)}{4} - 3\pi_w t^* \max_{u \in V} \frac{\delta(u)}{\pi_u}. \tag{20}$$

Notice that (20) implies

$$X^{(T')}_w > 0$$

for any $w \in V$ and for *any* $T' \in \mathbb{Z}_{\geq 0}$ satisfying

$$T' > 2t^* + \frac{12t^* \max_{u \in V} \frac{\delta(u)}{\pi_u}}{k}.$$

The fact (21) implies that $T_{cover} \leq T'$, and we obtain the claim. □

Proof of Corollary 4.2: Note that a SRT-router model corresponding to a simple random walk on G is exactly a rotor-router model on G, and we see that $\max_{u \in V} \frac{\delta(u)}{\pi_u} = 2m$, since $\pi_u = \frac{\delta(u)}{2m}$. Thus,

$$T_{cover} \leq 2t^* + 1 + \frac{24mt^*}{k}$$

holds by Theorem 4.1. □

5 Concluding Remarks

In this paper, we gave techniques to examine the visit frequency $X^{(T)}_v$ of the SRT-router model with $k > 1$ tokens, and gave an upper bound of the cover time for any ergodic and reversible P. Also, our upper bound improve the upper bound of the previous results of the rotor-router model with $k > 1$ tokens in many cases. A better upper bound of the cover time by derandomizing a specific *fast* random walk (e.g., β-random walk, Metropolis walk) is a challenge.

Acknowledgements

The author would like to thank Prof. Kijima for his comments on the manuscript. The author is also grateful to Prof. Sauerwald and Dr. Pajak for a discussion on the topic. This work is supported by JSPS KAKENHI Grant Number 15J03840. The author also gratefully acknowledge to the ELC project (Grant-in-Aid for Scientific Research on Innovative Areas MEXT Japan) for encouraging the research presented in this paper.

References

[1] H. Akbari and P. Berenbrink, Parallel rotor walks on finite graphs and applications in discrete load balancing, Proc. SPAA 2013, 186–195.

[2] D. Aldous and P. Diaconis, Strong uniform times and finite random walks, Advances in Applied Mathematics, **8** (1987), 69–97.

[3] R. Aleliunas, R. Karp, R. Lipton, L. Lovasz and C. Rackoff, Random walks, universal traversal sequences, and the complexity of maze problems, Proc. FOCS 1979, 218–223.

[4] N. Alon, C. Avin, M. Koucky, G. Kozma, Z. Lotker, M. R. Tuttle, Many random walks are faster than one, Combinatorics, Probability & Computing, **20** (2011), 481–502.

[5] O. Angel, A.E. Holroyd, J. Martin, and J. Propp, Discrete low discrepancy sequences, arXiv:0910.1077.

[6] E. Bampas, L. Gasieniec, N. Hanusse, D. Ilcinkas, R. Klasing, and A. Kosowski, Euler tour lock-in problem in the rotor-router model, Proc. DISC 2009, 423–435

[7] P. Berenbrink, R. Klasing, A. Kosowski, F. Mallmann-Trenn, and P. Uznanski, Improved analysis of deterministic load-balancing schemes, Proc. PODC 2015, 301–310.

[8] A. Broder, A. Karlin, P. Raghavan, E. Upfal, Trading space for time in undirected s–t connectivity, SIAM Journal on Computing **23** (1994) 324–334.

[9] J. Chalopin, S. Das, P. Gawrychowski, A. Kosowski, A. Labourel and P. Uznanski, Lock-in problem for parallel rotor-router walks, arXiv:1407.3200.

[10] J. Cooper, B. Doerr, T. Friedrich, and J. Spencer, Deterministic random walks on regular trees, Random Structures & Algorithms, **37** (2010), 353–366.

[11] J. Cooper, B. Doerr, J. Spencer, and G. Tardos, Deterministic random walks on the integers, European Journal of Combinatorics, **28** (2007), 2072–2090.

[12] J. Cooper and J. Spencer, Simulating a random walk with constant error, Combinatorics, Probability and Computing, **15** (2006), 815–822.

[13] D. Dereniowski, A. Kosowski, D. Pajak, and P. Uznanski, Bounds on the cover time of parallel rotor walks, LIPICS, **25** (STACS 2014), 263–275.

[14] B. Doerr and T. Friedrich, Deterministic random walks on the two-dimensional grid, Combinatorics, Probability and Computing, **18** (2009), 123–144.

[15] R. Elsasser and T. Sauerwald, Tight bounds for the cover time of multiple random walks, Theoretical Computer Science, **412** (2011), 2623–2641.

[16] U. Feige, A tight upper bound for the cover time of random walks on graphs, Random Structures and Algorithms, **6** (1995), 51–54.

[17] U. Feige, A tight lower bound for the cover time of random walks on graphs, Random Structures and Algorithms, **6** (1995), 433–438.

[18] T. Friedrich, M. Gairing, and T. Sauerwald, Quasirandom load balancing, SIAM Journal on Computing, **41** (2012), 747–771.

[19] T. Friedrich and T. Sauerwald, The cover time of deterministic random walks, The Electronic Journal of Combinatorics, **17** (2010), R167. Lecture Notes in Computer Science, **6196** (COCOON 2010), 130–139.

[20] A. E. Holroyd and J. Propp, Rotor walks and Markov chains, M. Lladser, R.S. Maier, M. Mishna, A. Rechnitzer, (eds.), Algorithmic Probability and Combinatorics, The American Mathematical Society, 2010, 105–126.

[21] S. Ikeda, I. Kubo, and M. Yamashita, The hitting and cover times of random walks on finite graphs using local degree information, Theoretical Computer Science, **410** (2009), 94–100.

[22] H. Kajino, S. Kijima, and K. Makino, Discrepancy analysis of deterministic random walks on finite irreducible digraphs, discussion paper.

[23] S. Kijima, K. Koga, and K. Makino, Deterministic random walks on finite graphs, Random Structures & Algorithms, **46** (2015), 739–761.

[24] A. Kosowski and D. Pajak, Does adding more agents make a difference? A case study of cover time for the rotor-router, Proc. ICALP 2014, 544–555.

[25] D. A. Levine, Y. Peres, and E. L. Wilmer, Markov Chain and Mixing Times, The American Mathematical Society, 2008.

[26] Y. Nonaka, H. Ono, K. Sadakane, M. Yamashita, The hitting and cover times of Metropolis walks, Theoretical Computer Science, **411** (2010), 1889–1894.

[27] Y. Rabani, A. Sinclair, and R. Wanka, Local divergence of Markov chains and analysis of iterative load balancing schemes, Proc. FOCS 1998, 694–705.

[28] T. Shiraga, Y. Yamauchi, S. Kijima, and M. Yamashita, Deterministic random walks for rapidly mixing chains, arXiv:1311.3749.

[29] T. Shiraga, Y. Yamauchi, S. Kijima, and M. Yamashita, Total variation discrepancy of deterministic random walks for ergodic Markov chains, Proc. ANALCO 2016, 138–148.

[30] R. Tijdeman, The chairman assignment problem, Discrete Mathematics. **32** (1980), 323–330.

[31] V. Yanovski, I.A. Wagner, and A.M. Bruckstein, A distributed ant algorithm for efficiently patrolling a network, Algorithmica, **37** (2003), 165–186.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

A half-normal distribution scheme for generating functions and the unexpected behavior of Motzkin paths

Michael Wallner[1]

[1] *Institute of Discrete Mathematics and Geometry, TU Wien, Wiedner Hauptstr. 8-10/104, A-1040 Wien, Austria*

We present an extension of a theorem by Michael Drmota and Michèle Soria [Images and Preimages in Random Mappings, 1997] that can be used to identify the limiting distribution for a class of combinatorial schemata. This is achieved by determining analytical and algebraic properties of the associated bivariate generating function. We give sufficient conditions implying a half-normal limiting distribution, extending the known conditions leading to either a Rayleigh, a Gaussian, or a convolution of the last two distributions. We conclude with three natural appearances of such a limiting distribution in the domain of Motzkin paths.

Keywords: Lattice path, analytic combinatorics, singularity analysis, limit laws

1 Introduction

Generating functions have proved very useful in the analysis of combinatorial questions. The approach builds on general principles of the correspondence between combinatorial constructions and functional operations. The symbolic method [14] provides a direct translation of the structural description of a class into an equation on generating functions. In [11], Drmota and Soria provided general methods for the analysis of bivariate generating functions $F(z,u)=\sum f_{nk}z^nu^k$. In general, n is the length or size, and k is the value of a "marked" parameter.

They continued their work in [12], wherein they derived three general theorems which identify the limiting distribution for a class of combinatorial schemata from certain properties of their associated bivariate generating function. These lead to a Rayleigh, a Gaussian, or a convolution of both distributions. Especially for a Gaussian limit distribution there are many schemata known: Hwang's quasi-powers theorem [16] or [14, Theorem IX.8], the supercritical composition scheme [14, Proposition IX.6], the algebraic singularity scheme [14, Theorem IX.12], an implicit function scheme for algebraic singularities [10, Theorem 2.23], or the limit law version of the Drmota-Lalley-Woods theorem [2, Theorem 8]. But such schemata also exist for other distributions, like e.g., the Airy distribution, see [4]. In general it was shown in [1] and [2, Theorem 10] that even in simple examples "any limit law", in the sense that the limit curve can be arbitrarily close to any càdlàg multi-valued curve in $[0,1]^2$, is possible.

In this paper we extend the work of [12], by providing an additional limit theorem, Theorem 2.1, which reveals a half-normal distribution. This distribution is generated by the absolute value $|X|$ of a normally distributed random variable X with mean 0. We will encounter several distributions, whose most important properties are summarized in Table 1.

We also present three natural appearances of this distribution in combinatorial constructions. In particular we consider Motzkin walks. Despite them being well-studied objects [7, 9, 17], they still hide some mysterious properties. Our applications extend some examples of random walks presented by Feller in [13, Chapter III] to Motzkin walks. We show that the same phenomena appear which, to quote Feller, "not only are unexpected but actually come as a shock to intuition and common sense".

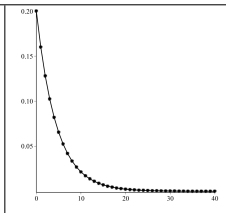
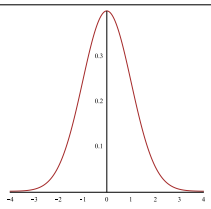
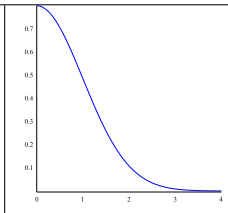
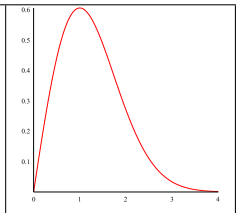
	Geometric Geom(p)	**Normal** $\mathcal{N}(\mu,\sigma)$	**Half-normal** $\mathcal{H}(\sigma)$	**Rayleigh** $\mathcal{R}(\sigma)$
Graph				
Support	$x \in \{0,1,\ldots\}$	$x \in \mathbb{R}$	$x \in \mathbb{R}_{\geq 0}$	$x \in \mathbb{R}_{\geq 0}$
PDF	$(1-p)^k p$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\sqrt{\frac{2}{\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$	$\frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$
Mean	$\frac{1-p}{p}$	μ	$\sigma\sqrt{\frac{2}{\pi}}$	$\sigma\sqrt{\frac{\pi}{2}}$
Variance	$\frac{1-p}{p^2}$	σ^2	$\sigma^2\left(1-\frac{2}{\pi}\right)$	$\sigma^2\left(2-\frac{\pi}{2}\right)$

Table 1: A comparision of the geometric, normal, half-normal, and Rayleigh distribution. We will encounter all four of them in the context of Motzkin walks.

Plan of this article. First, in Section 2, we present our main contribution: a scheme for bivariate generating functions leading to a half-normal distribution. In Section 3, we introduce Motzkin paths and establish the analytic framework which will be used in the subsequent sections. In Section 4, we apply our result to three properties of Motzkin walks: the number of sign changes, the number of returns to zero, and the height. In the case of zero drift a half-normal distribution appears. In Section 5, we give a summary of our results.

2 The half-normal theorem

Let $c(z) = \sum_n c_n z^n$ be the generating function of a combinatorial structure and $c(z,u) = \sum c_{nk} z^n u^k$ be the bivariate generating function where a parameter of interest has been marked, i.e., $c(z,1) = c(z)$. We introduce a sequence of random variables $X_n, n \geq 1$, defined by

$$\mathbb{P}[X_n = k] = \frac{c_{nk}}{c_n} = \frac{[z^n u^k]c(z,u)}{[z^n]c(z,1)},$$

where $\mathbb{P}$ denotes the probability. As we are interested in the asymptotic distribution of the marked parameter among objects of size n where n tends to infinity, the probabilistic point of view is given by finding the limiting distribution of X_n.

Important combinatorial constructions are "sequences" or "sets of cycles" (in the case of exponential generating functions) which imply the following decomposition

$$c(z,u) = \frac{1}{1-a(z,u)},$$

with a generating function $a(z,u)$ corresponding to the elements of the sequence, or the cycles, respectively. Another important and recurring phenomenon is the one of an algebraic singularity $\rho(u)$ of the square-root type such that $a(\rho(1),1) = 1$. According to further analytic properties of $a(z,u)$ the limiting distribution of X_n is shown to be either Gaussian, Rayleigh, the convolution of Gaussian and Rayleigh (see [12, Theorems 1-3]), or half-normal (see Theorem 2.1).

We start with the general form of the analytic scheme. In contrast to the original hypothesis [H] in [12] we call our hypothesis [H'] because we drop the condition that $h(\rho,1) > 0$ and we require it only for $\rho(u) = const.$

Hypothesis [H']: Let $c(z,u) = \sum_{n,k} c_{nk} z^n u^k$ be a power series in two variables with non-negative coefficients $c_{nk} \geq 0$ such that $c(z,1)$ has a radius of convergence of $\rho > 0$.

We suppose that $1/c(z,u)$ has the local representation

$$\frac{1}{c(z,u)} = g(z,u) + h(z,u)\sqrt{1-\frac{z}{\rho}}, \tag{1}$$

for $|u-1|<\varepsilon$ and $|z-\rho|<\varepsilon$, $\arg(z-\rho)\neq 0$, where $\varepsilon>0$ is some fixed real number, and $g(z,u)$, and $h(z,u)$ are analytic functions. Furthermore, these functions satisfy $g(\rho,1)=0$.

In addition, $z=\rho$ is the only singularity on the circle of convergence $|z|=|\rho|$, and $1/c(z,u)$, respectively $c(z,u)$, can be analytically continued to a region $|z|<\rho+\delta, |u|<1+\delta, |u-1|>\frac{\varepsilon}{2}$ for some $\delta>0$. ♢

Theorem 2.1 (Half-normal limit theorem) *Let $c(z,u)$ be a bivariate generating function satisfying [H']. If $g_z(\rho,1)\neq 0$, $h_u(\rho,1)\neq 0$, and $h(\rho,1)=g_u(\rho,1)=g_{uu}(\rho,1)=0$, then the sequence of random variables X_n defined by*

$$\mathbb{P}[X_n=k] = \frac{[z^n u^k]c(z,u)}{[z^n]c(z,1)},$$

has a half-normal limiting distribution, i.e.,

$$\frac{X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{H}(\sigma),$$

where $\sigma = \sqrt{2}\frac{h_u(\rho,1)}{\rho g_z(\rho,1)}$, and $\mathcal{H}(\sigma)$ has density $\frac{\sqrt{2}}{\sqrt{\pi\sigma^2}}\exp\left(-\frac{z^2}{2\sigma^2}\right)$ for $z\geq 0$. Expected value and variance are given by

$$\mathbb{E}[X_n] = \sigma\sqrt{\frac{2}{\pi}}\sqrt{n}+\mathcal{O}(1) \qquad \textit{and} \qquad \mathbb{V}[X_n] = \sigma^2\left(1-\frac{2}{\pi}\right)n+\mathcal{O}(\sqrt{n}).$$

Moreover, we have the local law

$$\mathbb{P}[X_n=k] = \frac{1}{\sigma}\sqrt{\frac{2}{\pi n}}\exp\left(-\frac{k^2/n}{2\sigma^2}\right)+\mathcal{O}\left(kn^{-3/2}\right)+\mathcal{O}\left(n^{-1}\right),$$

uniformly for all $k\geq 0$.

Remark 2.2 (Non-trivial dependency of ρ on u) The assumption of a constant singularity in z given by ρ can be weakened to a singularity $\rho(u)=\rho(1)+\mathcal{O}((u-1)^3)$, i.e., $\rho'(1)=\rho''(1)=0$. However, no example is known where $\rho(u)$ is not constant in a neighborhood of $u\sim 1$. ■

Proof (Sketch): The proof ideas are similar to the ones of [12, Theorem 1]. For details on the half-normal distribution we refer to [19], but all we need is the characteristic function. The main idea is to derive the asymptotic form of the characteristic function of $X_n/\sqrt{n}$. Since

$$\mathbb{E}[e^{itX_n/\sqrt{n}}] = \frac{[z^n]c(z,e^{\frac{it}{\sqrt{n}}})}{[z^n]c(z,1)},$$

we need to expand $[z^n]c(z,u)$ for $u=e^{it/\sqrt{n}}=1+\frac{it}{\sqrt{n}}+\mathcal{O}(n^{-1})$. To achieve this, we will apply Cauchy's integral formula for the following path of integration $\Gamma=\Gamma_1\cup\Gamma_2$:

$$\Gamma_1 = \left\{z=\rho\left(1+\frac{s}{n}\right) : s\in\gamma'\right\},$$
$$\Gamma_2 = \left\{z=Re^{i\vartheta} : R=\rho\left|1+\frac{\log^2 n+i}{n}\right|,\ \arg\left(1+\frac{\log^2 n+i}{n}\right)\leq|\vartheta|\leq\pi\right\},$$

where $\gamma'=\{s:|s|=1,\ \Re s\leq 0\}\cup\{s:0<\Re s<\log^2 n,\ \Im s=\pm 1\}$ is the major part of a Hankel contour γ, see Figure 1.

What remains is to investigate the parts separately: The first part gives the claimed result, whereas the second one is asymptotically negligible. Note that the changes in the hypothesis [H] are responsible for the appearance of characteristic function of the half-normal distribution in the limit. We omit these technical steps. □

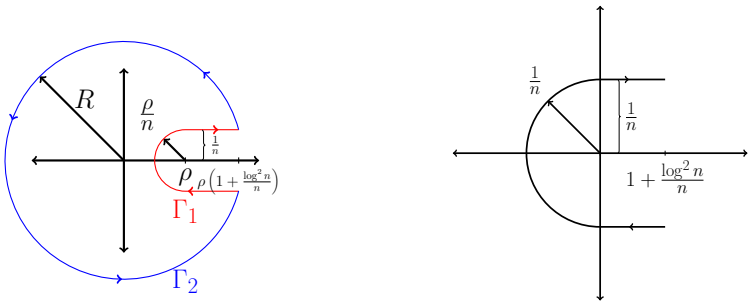


Figure 1: Hankel contour decomposition (left), and contour of γ' (right).

3 Motzkin paths

In this section we present needed, known results on directed lattice paths. Readers familiar with the exposition of Banderier and Flajolet [3] or related results may skip this section.

Definition 3.1 (Lattice paths) A *step set* $\mathcal{S} \subset \mathbb{Z}^2$ is a fixed, finite set of vectors $\{(a_1, b_1), \dots, (a_m, b_m)\}$. An n-step *lattice path* or *walk* is a sequence $(v_1, \dots, v_n)$ of vectors, such that v_j is in $\mathcal{S}$. Geometrically, it is a set of points $\{\omega_0, \omega_1, \dots, \omega_n\}$ where $\omega_i \in \mathbb{Z}^2, \omega_0 = (0,0)$ and $\omega_i - \omega_{i-1} = v_i$ for $i = 1, \dots, n$. The elements of $\mathcal{S}$ are called *steps* or *jumps*. The *length* $|\omega|$ of a lattice path is its number n of jumps. ◊

We restrict our attention to *simple directed paths* for which every element in the step set $\mathcal{S}$ is of the form $(1, b)$. In other words, these walks constantly move one step to the right. We introduce the abbreviation $\mathcal{S} = \{b_1, \dots, b_m\}$ in this case.

Along these restrictions, we introduce the following classes (see Table 2): A *bridge* is a path whose end-point ω_n lies on the x-axis. A *meander* is a path that lies in the quarter plane $\mathbb{Z}_+^2$. An *excursion* is a path that is at the same time a meander and a bridge. Their generating functions have been fully characterized in [3] by means of analytic combinatorics, see [14].

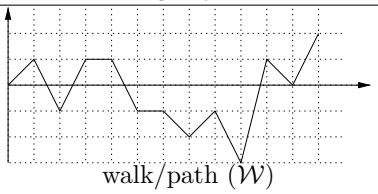
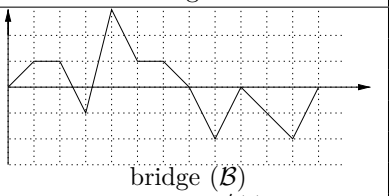
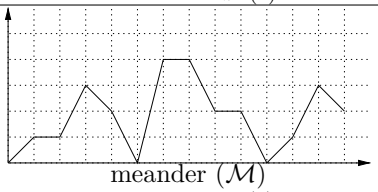
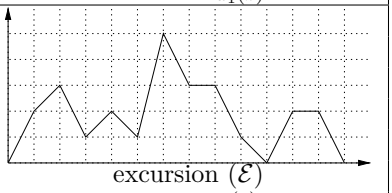
	ending anywhere	ending at 0
unconstrained (on $\mathbb{Z}$)	 walk/path ($\mathcal{W}$) $W(z) = \frac{1}{1-zP(1)}$	 bridge ($\mathcal{B}$) $B(z) = z\frac{u_1'(z)}{u_1(z)}$
constrained (on $\mathbb{Z}_+$)	 meander ($\mathcal{M}$) $M(z) = \frac{1-u_1(z)}{1-zP(1)}$	 excursion ($\mathcal{E}$) $E(z) = \frac{u_1(z)}{p_{-1}z}$

Table 2: The four types of paths: walks, bridges, meanders and excursions, and the corresponding generating functions for Motzkin paths [3, Fig. 1].

Definition 3.2 (Motzkin paths) A Motzkin path is a path that starts at the origin and is given by the step set $\mathcal{S} = \{-1, 0, +1\}$. ◊

We will refer to Motzkin walks/meanders/bridges/excursions depending on the different restrictions. In the literature Motzkin paths are often defined as Motzkin excursions, e.g. in [9].

In many situations it is useful to associate weights to single steps.

Definition 3.3 (Weights) For a given step set $\mathcal{S}$, we define the respective *system of weights* as $\{p_s \mid s \in \mathcal{S}\}$ where $p_s > 0$ is the associated weight to step $s \in \mathcal{S}$. The *weight of a path* is defined as the product of the weights of its individual steps. ◊

A typical weighted lattice path model is $p_s = 1$ (enumeration of paths), or $\sum_s p_s = 1$ (probabilistic model of paths, i.e., step s is chosen with probability p_s).

The following definition is the algebraic link between weights and steps. It is given only for the case of Motzkin paths, which is sufficient for our purpose.

Definition 3.4 (Jump polynomial of Motzkin paths) The *jump polynomial* is defined as the polynomial in u, u^{-1} (a Laurent polynomial)

$$P(u) := p_{-1}u^{-1} + p_0 + p_1 u, \qquad \text{with} \qquad p_{-1}, p_0, p_1 > 0.$$

The *kernel equation* is defined by

$$1 - zP(u) = 0, \qquad \text{or equivalently} \qquad u - z(uP(u)) = 0.$$

The quantity $K(z, u) := u - zuP(u)$ is called *kernel*. ◊

A walk is called *periodic* with period p if there exists a polynomial $H(u)$ and integers $b \in \mathbb{Z}$ and $p \in \mathbb{N}$, $p > 1$ such that $P(u) = u^b H(u^p)$. Otherwise its called *aperiodic*. The condition $p_0 > 0$ implies aperiodicity for Motzkin paths. Note that generating functions of aperiodic walks possess a unique singularity on the positive real axis [3].

The kernel plays a crucial rôle and is name-giving for the *kernel method*, which is the key tool characterizing this family of lattice paths. The interested reader is referred to [3, Chapter 2]. In the heart of this method lies the observation that the kernel equation is of degree 2 in u, and therefore has generically 2 roots. These correspond to branches of an algebraic curve given by the kernel equation.

Proposition 3.5 (Roots of the kernel) *The kernel equation* $1 - zP(u) = 0$ *has* 2 *solutions:*

$$u_{1,2}(z) = \frac{1 - p_0 z \mp \sqrt{(1 - p_0 z)^2 - 4p_{-1}p_1 z^2}}{2p_1 z}.$$

It holds that $\lim_{z\to 0} u_1(z) = 0$, and $\lim_{z\to 0} u_2(z) = \infty$. Because of that, we call $u_1(z)$ the small branch, and $u_2(z)$ the large branch.

Banderier and Flajolet showed that the generating functions of bridges, excursions and meanders can be expressed in terms of the small branch(es) and the jump polynomial, see Table 2. The branch $u_1(z)$ is real positive near 0. It is responsible for the asymptotic behavior of bridges, excursions and meanders, compare [3, Theorem 3 and 4].

In order to understand their behavior we need the following constants:

Lemma 3.6 (Structural constants) *The* structural constant τ*, which is the unique positive solution of* $P'(u) = 0$*, is* $\tau = \sqrt{\frac{p_{-1}}{p_1}}$*. The* structural radius *is* $\rho = \frac{1}{P(\tau)} = \frac{1}{p_0 + 2\sqrt{p_{-1}p_1}}$.

The theory of Newton-Puiseux series implies that the small branch $u_1(z)$ is analytic on the open interval $(0, \rho)$, and satisfies the singular expansion

$$u_1(z) = \tau - C\sqrt{1 - \frac{z}{\rho}} + \mathcal{O}\left(1 - \frac{z}{\rho}\right), \tag{2}$$

for $z \to \rho^-$, where $C = \sqrt{2\frac{P(\tau)}{P''(\tau)}}$. This is a direct consequence of the implicit function theorem.

Proposition 3.7 (Square-root singularity) *There exists a neighborhood* $\Omega \setminus (\rho, \infty)$ *such that for* $z \to \rho$ *in* $\Omega \setminus (\rho, \infty)$ $u_1(z)$ *has a local representation of the kind*

$$u_1(z) = a(z) + b(z)\sqrt{1 - z/\rho}, \qquad \text{with} \qquad a(\rho) = \tau, \text{ and } b(\rho) = -C,$$

where $a(z)$ *and* $b(z)$ *are analytic functions for every point* $z \in \Omega \setminus (\rho, \infty)$, $z \neq z_0$.

Proof: This is a direct consequence of the explicit structure of $u_1(z)$ from Proposition 3.5. □

4 Properties of Motzkin paths

The following examples are motivated by the very nice presentation of Feller [13, Chapter III] about one-dimensional symmetric, simple random walks. Therein, the discrete time stochastic process $(S_n)_{n\geq 0}$ is defined by $S_0 = 0$ and $S_n = \sum_{j=1}^{n} X_j$, $n \geq 1$, where the $(X_i)_{i\geq 1}$ are iid Bernoulli random variables with $\mathbb{P}[X_i = 1] = \mathbb{P}[X_i = -1] = \frac{1}{2}$. These results are generalized to the case of Motzkin paths. In particular compare [13, Problems 9-10] and [18, Remark of Barton] for returns to zero of symmetric and asymmetric random walks, respectively. Furthermore, see [13, Chapter III.5] for sign changes, and [13, Chapter III.7] for the height. See also the recent paper of Döbler [8] on Stein's method for this questions in which he derives bounds for the convergence rate in the Kolmogorov and the Wasserstein metric.

Let us now analyze these properties in the case of Motzkin walks. For the sake of brevity we will only mention the weak convergence law. However, in all cases the local law and the asymptotic expansions for mean and variance hold as well.

4.1 Returns to zero

A *return to zero* is a point of a walk of altitude 0, except for the starting point; in other words a return to the x-axis, see Figure 3. In order to count them we consider "minimal" bridges, in the sense that the bridges touch the x-axis only at the beginning and at the end. We call them *arches*. As a bridge is a sequence of such arches, we get their generating function in the form of $A(z) = 1 - \frac{1}{B(z)}$.

Lemma 4.1 *The generating function of arches $A(z)$ is for $z \to \rho$ of the kind*

$$A(z) = a(z) + b(z)\sqrt{1 - z/\rho},$$

where $a(z)$ and $b(z)$ are analytic functions in a neighborhood $\Omega\setminus(\rho,\infty)$ of ρ (i.e., for $z \in \Omega\setminus(\rho,\infty)$ it holds that $z \notin (\rho,\infty)$).

Proof: We know that $B(z) = z\frac{u_1'(z)}{u_1(z)}$ is analytic for $|z| < \rho$, see [3, Theorem 3]. Due to $p_0 > 0$ (aperiodicity) ρ is the only singular point on the circle of convergence. Hence,

$$B(z) = \frac{C_1}{\sqrt{1 - z/\rho}} + \mathcal{O}(1), \qquad C_1 = \frac{C}{2\tau}, \tag{3}$$

by (2) for $z \to \rho$. Proposition 3.7 together with (3) implies the desired decomposition. □

Here, we are interested in the number of returns to zero of walks which are unconstrained by definition. Every walk can be decomposed into a maximal initial bridge, and a walk that never returns to the x-axis, see Figure 2 . Let us denote the generating function of this *tail* by $T(z)$.

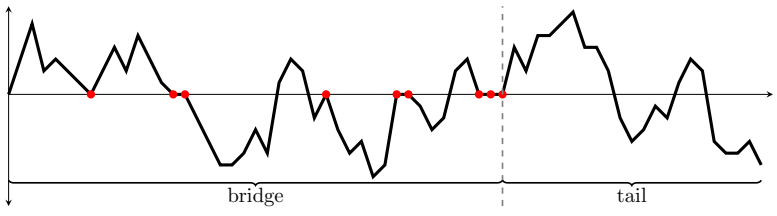


Figure 2: A walk with 9 returns to zero decomposed into a bridge and a tail.

As we want to count the number of returns to zero, we mark each arch by an additional parameter u and reconstruct the generating function of walks. This gives

$$W(z,u) = \frac{1}{1 - uA(z)}T(z) = \frac{W(z)}{u + (1-u)B(z)}, \qquad \text{with} \qquad T(z) = \frac{W(z)}{B(z)}.$$

Let us define the random variable X_n which stands for the number of returns to zero of a random meander of length n. Thus, $\mathbb{P}[X_n = k] = \frac{[u^k z^n]W(z,u)}{[z^n]W(z,1)}$.

Theorem 4.2 (Limit law for returns to zero) *Let X_n denote the number of returns to zero of a walk of length n. Let $\delta = P'(1)$ be the drift.*

1. *For $\delta \neq 0$ we get convergence to a geometric distribution:*

$$X_n \xrightarrow{d} \mathrm{Geom}\left(\frac{|p_1 - p_{-1}|}{P(1)}\right).$$

2. *For $\delta = 0$ we get convergence to a half-normal distribution:*

$$\frac{X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{H}\left(\sqrt{\frac{P(1)}{P''(1)}}\right).$$

Proof: First of all, we see that $[z^n]W(z,1) = [z^n]W(z) = P(1)^n$. Note that because of $p_0 > 0$ (aperiodicity) $B(z)$ is singular only at ρ. Obviously, $W(z)$ is singular at $\rho_1 := \frac{1}{P(1)}$.

Note that $P(\tau)$ is the unique minimum of $P(u)$ on the positive real axis. Hence, only two cases are possible: $\rho_1 < \rho$, if $\tau \neq 1$; or $\rho_1 = \rho$, if $\tau = 1$. These cases are equivalent to $\delta \neq 0$ and $\delta = 0$, respectively. In the first case $W(z)$ is responsible for the dominant singularity. Then we get ($B(z)$ is analytic for $|z| < \rho$)

$$[z^n]W(z,u) = \frac{1}{B(\rho_1)} \frac{P(1)^n}{1 - u\left(1 - \frac{1}{B(\rho_1)}\right)} + o(P(1)^n).$$

Thus, the limit distribution is a geometric distribution with parameter $\lambda = \frac{1}{B(\rho_1)}$. Distinguishing between a positive and a negative drift, and some tedious calculations with the help of relations implied by the kernel equation, give the final result for $\delta \neq 0$.

In the second case $\tau = 1$ or $\delta = 0$, we apply Theorem 2.1. By Lemma 4.1 it holds that $1/W(z,u)$ has a decomposition of the kind (1). In particular, from (3) we directly get that

$$\frac{1}{W(z,u)} = \left(1 - \frac{z}{\rho}\right) u + \frac{C}{2}(1-u)\sqrt{1 - \frac{z}{\rho}} + \mathcal{O}\left(\left(1 - \frac{z}{\rho}\right)(1-u)\right),$$

for $z \to \rho$ and $u \to 1$, with $g(\rho,1) = h(\rho,1) = g_u(\rho,1) = g_{uu}(\rho,1) = 0$; and $g_z(\rho,1) = -P(1)$ and $h_u(\rho,1) = -\sqrt{\frac{P(1)}{2P''(1)}}$. Hence, Theorem 2.1 yields the result. □

4.2 Sign changes of Motzkin walks

We say that nodes which are strictly above the x-axis have a *positive sign* denoted by "+", whereas nodes which are strictly below the x-axis have a *negative sign* denoted by "−", and nodes on the x-axis are *neutral* denoted by "0". This notion easily transforms a walk $\omega = (\omega_n)_{n\geq 0}$ into a sequence of signs. In such a sequence a *sign change* is defined by either the pattern $+(0)-$ or $-(0)+$, where (0) denotes a non-empty sequence of 0's, see Figure 3.

The main observation in this context is the non-emptiness of the sequence of 0's. Geometrically this means that it has to touch the x-axis when passing through it. This means that we can count the number of sign changes by counting the number of maximal parts above or below the x-axis. The idea is to decompose a walk into an alternating sequence of positive (above the x-axis) and negative (below) excursions terminated by a positive or negative meander.

We introduce two new terms: *positive excursions* are "traditional" excursions, i.e., they are required to stay above the x-axis, whereas *negative excursions* are walks which start at zero, end on the x-axis, but are required to stay below the x-axis.

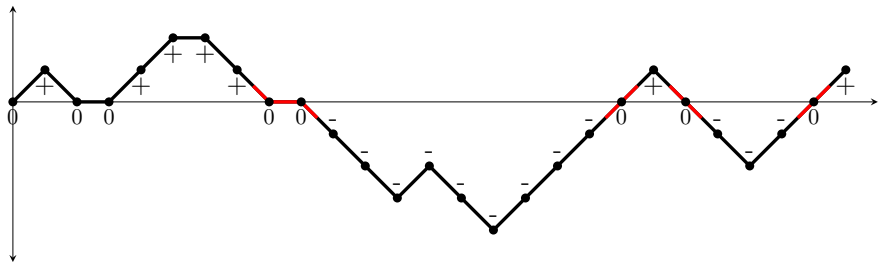


Figure 3: A Motzkin walk with 7 returns to zero and 4 sign changes. The positive, neutral or negative signs of the walks are indicated by $+, 0$, or $-$, respectively.

Lemma 4.3 *Among all walks of length n, the number of positive excursions is equal to the number of negative excursions.*

Proof: Mirroring bijectively maps positive excursions to negative ones. □

We define the bivariate generating function $B(z,u) = b_{n,k} z^n u^k$, where $b_{n,k}$ denotes the number of bridges of size n having k sign changes. Furthermore, we define

$$C(z) = \frac{1}{1-p_0 z},$$

as the generating function of *chains*, which are walks constructed solely from the jumps of height 0. Then the generating function of excursions starting with a $+1$ jump is given by

$$E_1(z) = \frac{E(z)}{C(z)} - 1,$$

because we need to exclude all excursions which start with a chain or are a chain. Due to Lemma 4.3 this is also the generating function for excursions starting with a -1 jump.

Theorem 4.4 *The bivariate generating function of bridges (where z marks the length, and u marks the number of sign changes of the walk) is given by*

$$B(z,u) = C(z)\left(1 + \frac{2E_1(z)}{1-uE_1(z)}\right).$$

Proof: A bridge is either a chain, which has zero sign changes, or it is not a chain. In the latter it is an alternating sequence of positive and negative excursions, starting with either of them. We decompose it uniquely into such excursions, by requiring that all except the first one start with a non-zero jump. Therefore the first excursion is counted by $E(z) - C(z)$, whereas all others are counted by $E_1(z)$. The decomposition is shown in Figure 4. □

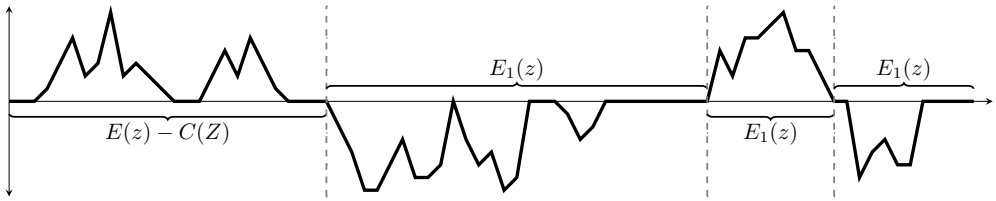


Figure 4: A bridge is an alternating sequence of positive and negative excursions. Here, it starts with a positive excursion, followed by excursions starting with a non-zero jump.

Let X_n be the random variable for the number of sign changes of a random bridge of length n. Thus, $\mathbb{P}[X_n = k] = \frac{[u^k z^n]B(z,u)}{[z^n]B(z,1)}$.

Theorem 4.5 (Limit law for returns to zero for bridges) *Let X_n denote the number of sign changes of a Motzkin bridge of length n. Then for $n \to \infty$ the normalized random variable has a Rayleigh$^{(\mathrm{i})}$ limit distribution*

$$\frac{X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{R}(\sigma) \qquad \text{and} \qquad \sigma = \frac{\tau}{2}\sqrt{\frac{P''(\tau)}{P(\tau)}},$$

where $\tau = \sqrt{\frac{p_{-1}}{p_1}}$ and $\mathcal{R}(\sigma)$ has the density $\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)$ for $x \geq 0$.

Proof (Sketch): We apply the first limit theorem of Drmota and Soria, [12, Theorem 1]. Proposition 3.7 implies that $E_1(z)$ and therefore $B(z,u)$ has a decomposition of the desired form (1). Checking the other conditions with the help of Lemma 3.6 yields the result. □

Finally, we consider sign changes of walks. Since we want to count the number of sign changes we need to know whether a bridge ended with a positive or negative sign. Let *positive bridges* be bridges whose last non-zero signed node was positive, and *negative bridges* be bridges whose last non-zero signed node was negative. Their generating functions are denoted by $B_+(z,u)$ and $B_-(z,u)$, respectively. Figure 4 shows a negative bridge.

Lemma 4.6 *The number of positive and negative bridges is the same and given by*

$$B_+(z,u) = \frac{B(z,u) - C(z)}{2} = \frac{E(z) - C(z)}{1 - uE_1(z)}.$$

Proof: The result is a direct consequence of Lemma 4.3, because a positive bridge is either a non trivial excursion or a negative bridge where an additional excursion starting with a $+1$ jump was appended. For negative bridges an analogous construction holds. □

Proposition 4.7 *The bivariate generating function of walks $W(z,u) = \sum_{n,k\geq 0} w_{nk} z^n u^k$ where w_{nk} is the number of all walks of length n with k sign changes, is given by*

$$W(z,u) = B(z,u)\frac{W(z)}{B(z)} + B_+(z,u)\left(\frac{W(z)}{B(z)} - 1\right)(u-1),$$

where $W(z) = \frac{1}{1 - zP(1)}$ is the generating function of walks.

Proof: Combinatorially, a walk is either a bridge or a bridge concatenated with a meander that does not return to the x-axis again. In the second case an additional sign change appears if the bridge ends with a positive sign and continues with a meander always staying above the x-axis, or vice versa. By Lemma 4.6 the desired form follows. □

The next theorem concludes this discussion. Its proof is similar to the one of Theorem 4.2.

Theorem 4.8 (Limit law for sign changes) *Let X_n denote the number of sign changes of Motzkin walks of length n. Let $\delta = P'(1)$ be the drift.*

1. *For $\delta \neq 0$ we get convergence to a geometric distribution:*

$$X_n \xrightarrow{d} \mathrm{Geom}(\lambda), \qquad \text{with} \qquad \lambda = \begin{cases} \frac{p_1}{p_{-1}}, & \text{for } \delta < 0, \\ \frac{p_{-1}}{p_1}, & \text{for } \delta > 0. \end{cases}$$

2. *For $\delta = 0$ we get convergence to a half-normal distribution:*

$$\frac{X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{H}\left(\frac{1}{2}\sqrt{\frac{P''(1)}{P(1)}}\right).$$

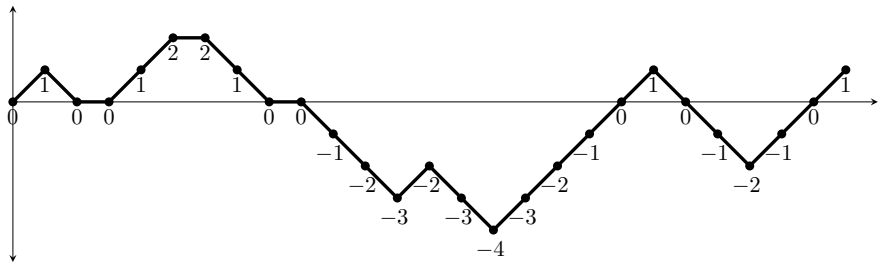


Figure 5: A Motzkin walk of height 2. The relative heights are given at every node.

4.3 Height of Motzkin walks

For a path of length n we define the *height* as its maximally attained y-coordinate, see Figure 5. Formally, let $\omega = (\omega_k)_{k=0}^n$ be a walk. Then its height is given by $\max_{k\in\{0,\dots,n\}} \omega_k$.

In order to analyze the distribution of heights, we define the bivariate generating function $F(z,u) = \sum_{n,h\geq 0} f_{nh} z^n u^h$. The coefficient f_{nh} represents the number of walks of height h among walks of length n. First we need a relation between the branches of the kernel equation:

Lemma 4.9 *Let $P(u) = p_{-1}u^{-1} + p_0 + p_1 u$. Then the small branch $u_1(z)$ and the large branch $u_2(z)$ of the kernel equation $1 - zP(u) = 0$ fulfil*

$$u_1(z)u_2(z) = \frac{p_{-1}}{p_1} \qquad \text{and} \qquad u_1(z) + u_2(z) = \frac{1 - zp_0}{zp_1}.$$

Proof: The kernel equation factorizes into $u(1-zP(u)) = -zp_1(u-u_1(z))(u-u_2(z))$. Comparing the coefficients gives the results. □

This relation gives us an explicit expression of $F(z,u)$ in terms of the large and small branch. For the final analysis we will use the latter.

Theorem 4.10 *The bivariate generating function of Motzkin walks (where z marks the length, and u marks the height of the walk) is given by*

$$F(z,u) = \frac{1}{1-zP(1)} \frac{u_2(z)-1}{u_2(z)-u} = \frac{1}{1-zP(1)} \frac{1 - \frac{p_1}{p_{-1}} u_1(z)}{1 - u\frac{p_1}{p_{-1}} u_1(z)}.$$

Proof: Banderier and Nicodème derived in [5, Theorem 2] the generating function $F^{[-\infty,h]}(z)$ for walks staying always below a wall $y = h$. For the case of Motzkin walks we get $F^{[-\infty,h]}(z) = \frac{1-\left(\frac{1}{u_2(z)}\right)^{h+1}}{1-zP(1)}$, where $u_2(z)$ is the large branch of the kernel equation. From this we directly get the generating function $F^{[h]}(z)$ for walks that have height exactly h. For $h \geq 1$ it equals

$$F^{[h]}(z) = F^{[-\infty,h]}(z) - F^{[-\infty,h-1]}(z) = \frac{u_2(z)-1}{1-zP(1)} \left(\frac{1}{u_2(z)}\right)^h.$$

The last formula also holds for $h = 0$. Finally, marking the heights by u and summing over all possibilities yields the result. The second formula is a consequence of Lemma 4.9. □

Let X_n be the random variable for the height of a random walk of length n. Thus, $\mathbb{P}[X_n = k] = \frac{[u^k z^n]F(z,u)}{[z^n]F(z,1)} = \frac{[u^k z^n]F(z,u)}{P(1)^n}$. This time the behavior will be different for $\delta < 0$ and $\delta > 0$. We omit its proof, however the ideas are again similar to the ones of Theorem 4.2.

(i) The parameter $\lambda = \sigma^{-2}$ was used in [12, Theorem 1].

Theorem 4.11 (Limit law for the height) *Let X_n denote the height of a Motzkin walk of length n. Let $\delta = P'(1)$ be the drift.*

1. *For $\delta < 0$ we get convergence to a geometric distribution:*

$$X_n \xrightarrow{d} \mathrm{Geom}\left(\frac{p_1}{p_{-1}}\right).$$

2. *For $\delta = 0$ the standardized random variable converges to a half-normal distribution:*

$$\frac{X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{H}\left(\sqrt{\frac{P''(1)}{P(1)}}\right).$$

3. *For $\delta > 0$ the standardized random variable converges to a normal distribution:*

$$\frac{X_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1), \qquad \mu = \frac{\delta}{P(1)}, \qquad \sigma^2 = 1 - \frac{p_0}{P(1)} - \left(\frac{\delta}{P(1)}\right)^2.$$

5 Conclusion

Drmota and Soria [12] presented three schemata leading to three different limiting distributions: Rayleigh, normal, and a convolution of both. This paper can be seen as an extension, by adding Theorem 2.1 yielding a half-normal distribution to this family. Other popular limit theorems are Hwang's quasi-powers theorem [16], and (implied by it) the supercritical composition scheme [14, Proposition IX.6]. These lead to a normal distribution.

The question may arise, how Theorem 2.1 behaves in the situation of a singularity $\rho(u)$ with $\rho'(1) \neq 0$ and $\rho''(1) \neq 0$, compare Remark 2.2. This remains an object for future research.

However, the more interesting question is if more "natural" appearances of such situations exist. Another known example is the limit law of the final altitude of meanders with zero drift in the reflection-absorption model in [6]. Chronologically, this was the starting point for the research of this paper. But this distribution also appears in number theory, see [15].

Yet another question is how the zero drift behavior of the analyzed parameters generalizes to other lattice path models. We will comment on these questions in the full version of this work.

Summing up, the applications to Motzkin paths show that intuition might lead you into the wrong direction. In Table 3 we see a comparison of the parameters. Obviously, the situation depends strongly on the drift. The critical case of a 0 drift seems to be the most delicate one, as it changes the nature of the law. In this case the limiting probability functions are concentrated at 0. In particular the expected value for $\Theta(n)$ trials grows like $\Theta(\sqrt{n})$ and not linearly. Equipped with the presented tools they might still be a "shock to intuition and common sense" but should not come "unexpected" anymore.

drift	returns to zero	sign changes	height
$\delta < 0$	$\mathrm{Geom}\left(\frac{p_{-1}-p_1}{P(1)}\right)$	$\mathrm{Geom}\left(\frac{p_1}{p_{-1}}\right)$	$\mathrm{Geom}\left(\frac{p_1}{p_{-1}}\right)$
$\delta = 0$	$\mathcal{H}\left(\sqrt{\frac{P(1)}{P''(1)}}\right)$	$\mathcal{H}\left(\frac{1}{2}\sqrt{\frac{P''(1)}{P(1)}}\right)$	$\mathcal{H}\left(\sqrt{\frac{P''(1)}{P(1)}}\right)$
$\delta > 0$	$\mathrm{Geom}\left(\frac{p_1-p_{-1}}{P(1)}\right)$	$\mathrm{Geom}\left(\frac{p_{-1}}{p_1}\right)$	Normal distribution

Table 3: Summary of the limit laws for Motzkin paths.

Acknowledgments

The author is grateful to Cyril Banderier, Bernhard Gittenberger, and Michael Drmota for insightful discussions. The author also thanks the anonymous referees for their suggested improvements. This work was supported by the Austrian Science Fund (FWF) grant SFB F50-03.

References

[1] Cyril Banderier, Olivier Bodini, Yann Ponty, and Hanane Tafat Bouzid. On the diversity of pattern distributions in rational language. In *ANALCO12—Meeting on Analytic Algorithmics and Combinatorics*, pages 107–115. SIAM, Philadelphia, PA, 2012.

[2] Cyril Banderier and Michael Drmota. Formulae and asymptotics for coefficients of algebraic functions. *Combin. Probab. Comput.*, 24(1):1–53, 2015.

[3] Cyril Banderier and Philippe Flajolet. Basic analytic combinatorics of directed lattice paths. *Theoret. Comput. Sci.*, 281(1-2):37–80, 2002. Selected papers in honour of Maurice Nivat.

[4] Cyril Banderier, Philippe Flajolet, Gilles Schaeffer, and Michèle Soria. Random maps, coalescing saddles, singularity analysis, and Airy phenomena. *Random Structures Algorithms*, 19(3-4):194–246, 2001.

[5] Cyril Banderier and Pierre Nicodème. Bounded discrete walks. *Discrete Math. Theor. Comput. Sci.*, AM:35–48, 2010.

[6] Cyril Banderier and Michael Wallner. Some reflections on directed lattice paths. *Prob., Comb. and Asymp. Methods for the Analysis of Algorithms*, BA:25–36, 2014. http://arxiv.org/abs/1605.01687 (Corrected drift 0 case in Table 4).

[7] Frank R. Bernhart. Catalan, Motzkin, and Riordan numbers. *Discrete Math.*, 204(1-3):73–112, 1999.

[8] Christian Döbler. Stein's method for the half-normal distribution with applications to limit theorems related to the simple symmetric random walk. *ALEA*, 12(1):171–191, 2015.

[9] Robert Donaghey and Louis W. Shapiro. Motzkin numbers. *J. Combinatorial Theory Ser. A*, 23(3):291–301, 1977.

[10] Michael Drmota. *Random trees: an interplay between combinatorics and probability*. Springer Science & Business Media, 2009.

[11] Michael Drmota and Michèle Soria. Marking in combinatorial constructions: Generating functions and limiting distributions. *Theoretical Computer Science*, 144(1–2):67–99, 1995.

[12] Michael Drmota and Michèle Soria. Images and preimages in random mappings. *SIAM J. Discrete Math.*, 10(2):246–269, 1997.

[13] William Feller. *An Introduction to Probability Theory and its Applications*. Wiley & Sons, New York, 1968.

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

[15] Virginija Garbaliauskienė, Jonas Genys, and Ruta Ivanauskaitė. A limit theorem for the arguments of zeta-functions of certain cusp forms. *Acta Appl. Math.*, 96(1-3):221–231, 2007.

[16] Hsien-Kuei Hwang. On convergence rates in the central limit theorems for combinatorial structures. *European J. Combin.*, 19(3):329–343, 1998.

[17] Theodore Motzkin. Relations between hypersurface cross ratios, and a combinatorial formula for partitions of a polygon, for permanent preponderance, and for non-associative products. *Bull. Amer. Math. Soc.*, 54:352–360, 1948.

[18] John G. Skellam and Leonard R. Shenton. Distributions associated with random walk and recurrent events. *J. Roy. Statist. Soc. Ser. B.*, 19:64–111 (discussion 111–118), 1957.

[19] Michail Tsagris, Christina Beneki, and Hossein Hassani. On the folded normal distribution. *Mathematics*, 2(1):12, 2014.

Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

On the exponential decay of the characteristic function of the quicksort distribution

Vytas Zacharovas

Department of Mathematics and Informatics
Vilnius University
Naugarduko 24
Vilnius, Lithuania
E-mail: vytas.zacharovas@mif.vu.lt

Abstract. We prove that the characteristic function of the quicksort distribution is exponentially decreasing at infinity. As a consequence it follows that the density of the quicksort distribution can be analytically extended to the vicinity of the real line.

Keywords: Quicksort, characteristic function, density, Laplace transform, analytic continuation

1 Introduction

Let X_n be the number of steps required by Quicksort algorithm to sort the list of values $\sigma(1), \sigma(2), \dots, \sigma(n)$ where σ is a random permutation chosen with uniform probability from the set of all permutations S_n of order n. It has been proven by Régnier (1989) and Rösler (1991) that the appropriately scaled distribution of X_n converges to some limit law

$$\frac{X_n - \mathbb{E}X_n}{n} \to^d Y$$

as $n \to \infty$. Let us denote as $f(t)$ the characteristic function of the limiting distribution

$$f(t) = \mathbb{E}e^{itY}$$

Tan and Hadjicostas (1995) proved that the characteristic function $f(t)$ has a density $p(x)$. Knessl and Szpankowski (1999) using heuristic approach established a number of very precise estimates for the behavior of $p(x)$ at infinity. Later Fill and Janson (2000) showed that the characteristic function $f(t)$ of the limit quicksort distribution together with its all derivatives decrease faster than any polynomial at infinity. More precisely they showed that for all real $p > 0$ there is such a constant c_p that

$$|f(t)| \leqslant \frac{c_p}{|t|^p}, \quad \text{for all} \quad t \in \mathbb{R}.$$

They also proved that

$$c_p \leqslant 2^{p^2+6p}.$$

Hence

$$|f(t)| \leqslant \inf_{p>0} \frac{2^{p^2+6p}}{|t|^p}.$$

The infimum in the above inequality can be evaluated as

$$|f(t)| \leqslant \inf_{p>0} \frac{2^{p^2+6p}}{|t|^p} \leqslant |t|^3 e^{-\frac{\log^2|t|}{4\log 2}}.$$

The main result of this paper is the following theorem stating that the characteristic function $f(t)$ of limiting Quicksort distribution decreases exponentially at infinity.

Theorem 1 *There is a constant $\eta > 0$ such that*

$$f(t) = O(e^{-\eta|t|})$$

as $|t| \to \infty$.

Corollary 2 *Quicksort distribution has a bounded density that can be extended analytically to the vicinity of the real line $|\Im(s)| < \eta$. Where η is the same positive number as in the formulation of Theorem 1.*

2 Proofs

It has been shown in Rösler (1991) that the characteristic function $f(t)$ satisfies the functional equation

$$f(t) = e^{it} \int_0^1 f(tx) f\big(t(1-x)\big) e^{2itx\log x + 2it(1-x)\log(1-x)}\, dx$$

which after a change of variables $x \to y/t$ becomes

$$tf(t)e^{2it\log t} = e^{it} \int_0^t f(y) f(t-y) e^{2iy\log y + 2i(t-y)\log(t-y)}\, dy$$

It follows hence by taking Laplace transform of the both sides that function

$$\psi(s) = \int_0^\infty f(t) e^{2it\log t} e^{-st}\, dt$$

satisfies an equation

$$-\psi'(s) = \psi^2(s-i). \tag{1}$$

The Laplace transform $\psi(s)$ together with the above differential equation will be the main tool of proving the result stated in the introduction.

It is well known that the quicksort distribution has finite moments of all orders. In the following analysis we will only need the fact that it has finite first moment, which implies that $|f'(t)|$ is bounded. Thus integrating by parts we conclude that

$$\begin{aligned}
\psi(s) &= \int_0^\infty f(t) e^{2it\log t} e^{-st}\, dt \\
&= \frac{1}{s} + \frac{1}{s} \int_0^\infty \big(f'(t) e^{2it\log t} + f(t) e^{2it\log t}(2i\log t + 2i)\big) e^{-st}\, dt \qquad (2) \\
&\leqslant \frac{A}{|s|}\left(1 + \frac{|\log \Re s|}{\Re s}\right),
\end{aligned}$$

for all s lying in the right half-plane $\Re s > 0$ and $A > 0$ being some positive absolute constant.

Lemma 3 *For all s lying in the right half-plane $\Re s > 0$ and all integer $n \geqslant 0$ holds the inequality*

$$\left|\psi^{(n)}(s)\right| \leqslant n! \left(\max_{r\in\{0,1,\ldots,n\}} \left|\psi(s-ir)\right|\right)^{n+1}$$

Proof: The proof is done by applying mathematical induction on n and using the fact that the differential equation for $\psi(s)$ allows us to express the derivatives $\psi^{(n)}(s)$ as a polynomial function of $\psi(s-ik)$ with $0 \leqslant k \leqslant n$.

Indeed, for $n = 0$ the above inequality becomes an identity. Suppose this identity holds for all n not exceeding m. Let us consider now $n = m + 1$. Replacing the first derivative of $\psi(s)$ by $-\psi^2(s-i)$ we obtain

$$\begin{aligned}\psi^{(m+1)}(s) &= \big(\psi'(s)\big)^{(m)} = -\big(\psi^2(s-i)\big)^{(m)}\\ &= -\sum_{k=0}^{m}\binom{m}{k}\psi^{(k)}(s-i)\psi^{(m-k)}(s-i).\end{aligned}$$

Thus applying the inductive hypothesis to the derivatives of $\psi(s-i)$ we get

$$\begin{aligned}\left|\psi^{(m+1)}(s)\right| &\leqslant \sum_{k=0}^{m}\binom{m}{k}k!\left(\max_{r\in\{0,1,\ldots,k\}}\left|\psi(s-i-ir)\right|\right)^{k+1}(m-k)!\left(\max_{r\in\{0,1,\ldots,m-k\}}\left|\psi(s-i-ir)\right|\right)^{m-k+1}\\ &\leqslant (m+1)!\left(\max_{r\in\{0,1,\ldots,n\}}\left|\psi(s-ir)\right|\right)^{m+2}.\end{aligned}$$

The last inequality is the same as stated in the lemma with $n = m + 1$. This completes the proof of the lemma □

Lemma 4 *For all s lying in the lower part of the right half-plane $\Re s > 0$ and $\Im s < 0$ holds the inequality*

$$\left|\psi^{(n)}(s)\right| \leqslant n!\left(\frac{C(\sigma)}{|s|}\right)^{n+1}$$

Where $\sigma = \Re s$ and

$$C(\sigma) = A\left(1 + \frac{|\log\sigma|}{\sigma}\right)$$

with some absolute constant $A > 0$.

Proof: Our upper bound (2) for $\psi(s)$ implies that for $\Re s > 0$ and $\Im s < 0$ we have

$$\max_{r\in\{0,1,\ldots,n\}}\left|\psi(s-ir)\right| \leqslant \max_{r\in\{0,1,\ldots,n\}}\frac{C(\sigma)}{|s-ir|} \leqslant \frac{C(\sigma)}{|s|}.$$

Since imaginary part of s is negative so $|s-ir| \geqslant |s|$. Using this inequality to evaluate the right hand side of the inequality of Lemma 3 we complete the proof of the lemma.

□

Proposition 5 *The function $\psi(s)$ can be continued analytically to the whole complex plane. Moreover, for all s belonging to the lower half-plane $\Im(s) < 0$ and $\Re s \geqslant -B$ with any fixed $B > 0$ holds the estimate*

$$\psi(s) = O_B(1/|s|)$$

Proof: For $\Re(s) \geqslant 1$ the estimate of the proposition already follows from (2). By this estimate of Lemma 4 we have that the Taylor series

$$\psi(s) = \sum_{j=0}^{\infty} \frac{\psi^{(j)}(1-iK)}{j!}\big(s-(1-iK)\big)^j$$

converges in the circle $|1-iK-s| < |1-iK|/C(1)$ and moreover in this circle holds the estimate

$$|\psi(s)| \leqslant \sum_{j=0}^{\infty} \left(\frac{C(1)}{|1-iK|}\right)^{j+1} \big|s-(1-iK)\big|^j = \frac{C(1)}{|1-iK|}\frac{1}{1-\frac{C(1)}{|1-iK|}|1-iK-s|}.$$

This means that $\psi(s)$ can be analytically continued to the region of complex plane that consists of such s that are contained in any of the circles of radius $|1-iK|/C(1)$ with center at $1-iK$ with some $K > 0$. Note that all complex number s with negative imaginary part such that $1 + \frac{\Im(s)}{C(1)} \leqslant \Re(s)$ satisfy such condition. See the figure 1.

Note that $\psi(s)$ satisfies a shift-differential equation (1) which is by integrating its both sides yields the identity

$$\psi(s) = \psi(s-i) + i\int_0^1 \psi(s-i-it)^2\,dt.$$

The repeated application of the above identity allows us to continue $\psi(s)$ analytically to the whole complex plane.

□

We have already proven that for $\Im(s) \leqslant 0$ we have

$$\psi(s) = O\left(\frac{1}{|s|}\right)$$

when $\Re(s) \geqslant -H$ with an arbitrary fixed $H > 0$. Let us now try to obtain a similar estimate for the values of s lying in the upper half-plane.

Lemma 6 *For all $\sigma > 0$ we have*

$$\sup_{y\in\mathbb{R}} |\psi(\sigma+iy)| < \frac{1}{\sigma}.$$

Proof: The proof of the lemma relies on a standard trick that is used to prove that if a modulus of a characteristic function of a random variable reaches 1 at some point other than 0 then the random variable has a lattice distribution. We have

$$|\psi(\sigma+iy)| = \left|\int_0^\infty f(t)e^{2it\log t}e^{-(\sigma+it)t}\,dt\right| \leqslant \left|\int_0^\infty e^{-\sigma t}\,dt\right| \leqslant \frac{1}{\sigma},$$

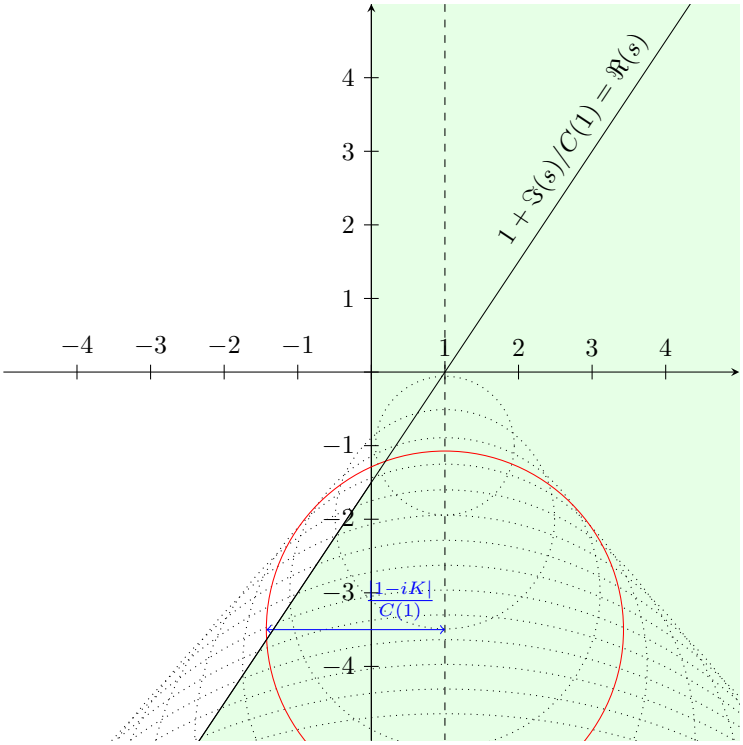


Fig. 1: The continuation of $\psi(s)$ to the left half-plane

for $\sigma > 0$. Note that the estimate 2 for fixed $\sigma > 0$ implies that $\psi(\sigma + iy) = O(1/|y|)$ as $|y| \to \infty$ which means that the supremum of $|\psi(\sigma + iy)|$ will be reached on some finite point $y_0 = y_0(\sigma)$. It remains to prove that this supremum cannot be equal to $1/\sigma$. Indeed if

$$|\psi(\sigma + iy_0)| = \frac{1}{\sigma},$$

then recalling the definition of ψ we can rewrite this identity as

$$\left| \int_0^\infty f(t) e^{2it \log t} e^{-\sigma t} e^{-iy_0 t}\, dt \right| = \int_0^\infty e^{-\sigma t}\, dt$$

or equivalently

$$e^{i\theta} \int_0^\infty f(t) e^{2it \log t} e^{-\sigma t} e^{-iy_0 t}\, dt = \int_0^\infty e^{-\sigma t}\, dt$$

for some real θ. Since $|f(t)| \leqslant 1$ taking the real part of the above equation we have

$$\Re\big(e^{i\theta} f(t) e^{2it \log t} e^{-iy_0 t}\big) \equiv 1.$$

The above identity together with the fact that $\big|e^{i\theta} f(t) e^{2it \log t} e^{-iy_0 t}\big| \leqslant 1$ implies that $\Im\big(e^{i\theta} f(t) e^{2it \log t} e^{-iy_0 t}\big) \equiv 0$ and thus

$$e^{i\theta} f(t) e^{it \log t} e^{-iy_0 t} \equiv 1.$$

Which means that

$$\psi(s) = \int_0^\infty f(t) e^{2it \log t} e^{-st}\, dt = e^{-i\theta} \int_0^\infty e^{-st} e^{iy_0 t}\, dt = \frac{e^{-i\theta}}{s - iy_0}.$$

However such function does not satisfy the equation $-\psi'(s) = \psi^2(s - i)$. □ With the help of the just proven lemma we can obtain an upper bound for $\psi(s)$ in the vicinity of the imaginary line $\Im(s) = 0$.

Lemma 7 *We have*

$$|\psi(s)| \leqslant \frac{1 - \varepsilon}{1 - |\Re(s) - 1|(1 - \varepsilon)},$$

for s belonging to the vertical strip $-\frac{\varepsilon}{1-\varepsilon} < \Re(s) < \frac{2-\varepsilon}{1-\varepsilon}$, where ε is such that $\sup_{y \in \mathbb{R}} |\psi(1+iy)| = 1 - \varepsilon$.

Proof: Applying the inequality of Lemma 6 with $\sigma = 1$ we have

$$\psi(1 + iy) \leqslant 1 - \varepsilon$$

for all $y \in \mathbb{R}$ and some fixed $\varepsilon > 0$. Hence inequality of Lemma 3 yields that

$$\psi^{(k)}(1 + iy) \leqslant k!(1 - \varepsilon)^{k+1} \tag{3}$$

uniformly for $y \in \mathbb{R}$. This implies that $\psi(s)$ is bounded in the vicinity of the imaginary line $\Re(s) \geqslant -\varepsilon'$ where $\varepsilon' < \varepsilon$. Indeed by Taylor expansion

$$\psi(s) = \sum_{k=0}^{\infty} \frac{\psi^{(k)}(1 + iy)}{k!} (s - 1 - iy)^k$$

Thus

$$|\psi(s)| \leqslant \sum_{k=0}^{\infty}(1-\varepsilon)^{k+1}|s-1-iy|^k = \frac{1-\varepsilon}{1-|s-1-iy|(1-\varepsilon)}$$

for $|s-1-iy| < \frac{1}{1-\varepsilon}$. Suppose $|\Re(s)-1| < \frac{1}{1-\varepsilon}$ then taking $y = \Im(s)$ we get

$$|\psi(s)| \leqslant \frac{1-\varepsilon}{1-|\Re(s)-1|(1-\varepsilon)},$$

for all s lying in the strip $|\Re(s)-1| < \frac{1}{1-\varepsilon}$. □

A more precise estimate can be obtained combining the obtained two upper bounds for derivatives of $\psi(s)$.

Lemma 8 *We have an upper bound*

$$|\psi(s)| = O\left(\frac{1}{|s|}\right)$$

in the region $\Re(s) > -\frac{\varepsilon'}{1-\varepsilon'}$. *Where* ε' *is a fixed number that* $0 < \varepsilon' < \varepsilon = 1 - \sup_{y\in\mathbb{R}}|\psi(1+iy)|$, *the constant in the symbol depends on* ε' *only.*

Proof: Putting $\sigma = 1$ in our non-uniform bound (2) for $\psi(s)$ we have

$$|\psi(1+iy)| \leqslant D/|y|$$

for some fixed $D > 0$. Again by induction for $k \leqslant |y|/2$ we have

$$|\psi^{(k)}(1+iy)| \leqslant k!\left(\frac{2D}{|y|}\right)^{k+1}.$$

Suppose $|\Re(s)-1| < \frac{1}{1-\varepsilon'}$. Let us take $y = \Im(s)$. Combining the above upper bound with our previous uniform estimate (3) for the derivatives of $\psi^{(j)}(1+iy)$ we get

$$\begin{aligned}
|\psi(s)| &\leqslant \sum_{k\leqslant|y|/2}\left(\frac{2D}{|y|}\right)^{k+1}|s-1-iy|^k + \sum_{k>|y|/2}|s-1-iy|^{k+1}(1-\varepsilon)^k \\
&\leqslant \frac{2D}{|y|}\frac{1}{1-\frac{2D}{|y|}|s-1-iy|} + \frac{|s-1-iy|\big(|s-1-iy|(1-\varepsilon)\big)^{|y|/2}}{1-|s-1-iy|(1-\varepsilon)} \\
&\leqslant \frac{2D}{|y|-\frac{2D}{(1-\varepsilon')}} + \frac{\left(\frac{1-\varepsilon}{1-\varepsilon'}\right)^{|y|/2}}{(1-\varepsilon')\left(1-\frac{1-\varepsilon'}{1-\varepsilon}\right)},
\end{aligned}$$

for $|\Re(s)-1| < \frac{1}{1-\varepsilon'}$ and $|y| > \frac{2D}{1-\varepsilon'}$. Since $\frac{1-\varepsilon}{1-\varepsilon'} < 1$ we have

$$|\psi(s)| = O\left(\frac{1}{|s|}\right).$$

□ A number of conclusions can be drawn from the estimate of the just proven lemma.

Proof of Theorem 1: The Laplace transform of $tf(t)e^{2it\log t}$ is $-\psi'(s)$ so, by inversion formula we have

$$-f(t)e^{2it\log t} = \frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \psi'(s)e^{ts}\,ds = \frac{-1}{2\pi it}\int_{\sigma-i\infty}^{\sigma+i\infty} \psi^2(s-i)e^{ts}\,ds$$

and taking into account that $|\psi(s-i)| \ll 1/|s|$ in the region $\Re(s) \geqslant -2\eta$ for some fixed $\eta > 0$ we can shift the integration line to the left and obtain

$$f(t)e^{2it\log t} = \frac{1}{2\pi it}\int_{-\eta-i\infty}^{-\eta+i\infty} \psi^2(s-i)e^{ts}\,ds \ll e^{-\eta t}.$$

□

Proof of Corollary 2: The density is given by formula

$$p(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-ixt}\,dt.$$

The fact that $f(t)$ is exponentially decreasing $|f(t)| \ll e^{-\eta|t|}$ at infinity $|t| \to \infty$ immediately implies that the integral

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-ist}\,dt.$$

is absolutely convergent in the vicinity of the real line $|\Im(s)| < \eta$ where it defines an analytic function that coincides with the density of the quicksort distribution $p(x)$ on the real line $s = x \in \mathbb{R}$. □

Corollary 9 *The density function $p(x)$ of the quicksort distribution can have only finite number of zeros in any finite interval. The same is true for the derivatives of $p(x)$ of all orders.*

Proof: Since an analytic function that is not identically equal to zero can have only finite number of zeros in any closed circle $|s-x| \leqslant r/2$ for any $x \in \mathbb{R}$, so the density $p(x)$ can have only finite number of zeros in any finite interval $[x-r/2, x+r/2]$ with all $x \in \mathbb{R}$. □

Acknowledgements

The author sincerely thanks Prof. Hsien-Kuei Hwang for numerous discussions on the topic of the paper as well as for his hospitality during the author's visits to Academia Sinica (Taiwan). The author also thanks the anonymous referee for pointing out numerous errors in the previous draft of the paper and his suggestions that lead to considerable improvement in the exposition of the results of the paper.

References

J. A. Fill and S. Janson. Smoothness and decay properties of the limiting Quicksort density function. In *Mathematics and computer science (Versailles, 2000)*, Trends Math., pages 53–64. Birkhäuser, Basel, 2000.

C. Knessl and W. Szpankowski. Quicksort algorithm again revisited. *Discrete Math. Theor. Comput. Sci.*, 3(2):43–64 (electronic), 1999. ISSN 1365-8050.

M. Régnier. A limiting distribution for quicksort. *RAIRO Inform. Théor. Appl.*, 23(3):335–343, 1989. ISSN 0988-3754.

U. Rösler. A limit theorem for "Quicksort". *RAIRO Inform. Théor. Appl.*, 25(1):85–100, 1991. ISSN 0988-3754.

K. H. Tan and P. Hadjicostas. Some properties of a limiting distribution in Quicksort. *Statist. Probab. Lett.*, 25(1):87–94, 1995. ISSN 0167-7152. doi: 10.1016/0167-7152(94)00209-Q. URL http://dx.doi.org/10.1016/0167-7152(94)00209-Q.