Mathematisches Forschungsinstitut Oberwolfach

Report No. 23/2011

DOI: 10.4171/OWR/2011/23

Mini-Workshop: Random Trees, Information and Algorithms

Organised by
Ralph Neininger (Frankfurt)
Wojciech Szpanowski (West Lafayette)

April 24th – April 30th, 2011

Abstract. The subject of this Mini-Workshop is the probabilistic analysis of random tree models that originate from applications in Computer Science. Emphasis is put on their connections to algorithms and information theory. Trees with a stochastic growth dynamic appear in Computer Science as data structures, in the context of coding schemes as well as connected to fundamental algorithms such as sorting, searching and selecting. The focus of this Mini-Workshop is on probabilistic and analytic techniques that have been developed recently in the asymptotic analysis of random trees such as martingale methods, connections to branching random walks, the contraction method, the method of moments as well as various techniques based on generating functions.

Mathematics Subject Classification (2000): 60F05, 68W40.

Introduction by the Organisers

In the asymptotic analysis of random tree models that originate from Computer Science diverse probabilistic and analytic techniques have been developed in the last decades with a strongly increasing interest during the last years. The techniques being developed include methods based on martingales, connections to branching random walks, the contraction method, techniques using generating functions, and the method of moments. Classically, in Computer Science random trees appear in the performance analysis of data structures, in the context of coding schemes as well as connected to fundamental algorithms such as sorting, searching and selecting. However, in the last years also fascinating connections of these random tree models to coalescent processes, fragmentation theory and

other combinatorial stochastic processes have been found. The aims of this Mini-Workshop are a deeper understanding and advances in the probabilistic analysis of random tree models as well as to discover their connections beyond the realm of models motivated from Computer Science.

The topics discussed at the workshop are briefly summarized as follows: For the model of simply generated trees Svante Janson discussed convergence of the size n trees to limit objects in a topology corresponding to convergence of all out-degrees. In particular, he discussed the cases where there is a representation of the simply generated tree as conditioned subcritical Galton-Watson tree or there is no representation as a conditioned Galton-Watson tree, the generating function associated having zero radius of convergence. Louigi Addario-Berry considered the problem of cutting down such simply generated trees (for finite variance offspring distributions) and its connection to the distance between two independent, uniformly chosen nodes in the tree by a coupling method. Here, the Rayleigh distribution appears in a $\sqrt{n}$ scaling and connections to Brownian excursions are found. Christina Goldschmidt also discussed the cutting of random trees to isolate the root. However, in contrast to the $\sqrt{n}$-height simply generated trees she considered a $\log n$-height tree, the random recursive tree. The stochastic process describing the cutting procedure on the set of partitions of $\{1, \dots, n\}$ turns out to be the Bolthausen-Sznitman coalescent. Also asymptotic frequencies of blocks were discussed and the open problem was raised to find other coalescents that can be represented by cutting down a combinatorial tree.

An important tree structure to store bit-strings are tries. In the talk of Wojciech Szpankowski tries were considered under the symmetric and asymmetric Bernoulli model (which describe memoryless sources). The quantities under consideration are the internal and external profile of tries. Results on asymptotic expectations with their phase changes were presented together with asymptotic variances and limit laws. In the talk of Mark Ward the internal profile of tries was considered under a different probabilistic model that makes the trie a suffix tree, another important data structure in applications. A related quantity is the subword complexity of the given string; the approach is based on generating functions.

A couple of talks discussed the analysis of search trees, mainly by the use of the contraction method. Ludger Rüschendorf presented results on depths and various distance measures of the weighted b-ary tree together with applications to special kinds of random trees that are covered by the class of weighted b-ary trees. Uwe Rösler discussed a functional limit law for a process version associated to the Quicksort algorithm: the algorithm on size n input always first recursively sorts the left sublist generated and stops once the smallest $\ell \in \{1, \dots, n\}$ items are sorted. The number of key comparisons as a process in ℓ is considered asymptotically in n. Ralph Neininger developed the contraction method on the spaces $C[0,1]$ and $D[0,1]$ of continuous resp. cádlág functions with the supremum norm by use of the Zolotarev metric. This leads to a framework that covers as an application Donsker's invariance principle. Another application was presented by

Henning Sulzbach. He showed a functional limit law for the complexity of partial match queries of the form $(s, *)$, $s \in [0, 1]$ in random two-dimensional (point) quadtrees, the process being in s. As corollaries, open questions regarding the variance and limit law for uniform queries and the order of worst case queries could be solved. The approach also covers the k-d trees for $k = 2$. Related to the contraction method Gerold Alsmeyer characterized the set of solutions of general smoothing transforms. By asking for fixed points of a functional equation under various constraints on the solutions he covers recursive distributional equations of sum and max type. The relation between solutions of homogeneous and inhomogeneous equations was also discussed as well as the phenomenon of endogeneity that plays a role in cases also important in applications. Rudolf Grübel viewed the stochastic evolution of random (search) trees as a transient Markov chain and used connections to discrete potential theory. He discussed almost sure convergence of the normalized trees to limiting random measures. Also a metric on trees based on subtree sizes was introduced and the resulting limiting metric space was explored.

Nicolas Broutin considered algorithms to resolve collisions in communication of multiple users on one broadcast channel. He studied protocols for which the execution of the algorithms can be represented by a tree. The main focus was on the stability of the protocols, analyzed via the long-term behavior of an associated conditioned Markov chain.

Talks with geometrically motivated topics started with Hsien-Kuei Hwang discussing different notions of dominance in random point sets in space. He studied threshold phenomena and uniform estimates for the expected number of such points among n iid. points in d-dimensional cubes and simplices in various asymptotic settings. Yuliy Baryshnikov discussed bounding the unimodal category of functions, in particular addressed the case of a simple random function in dimension 1, the uniformly chosen random Dyck path of length $2n$, for which a limit law was derived. Gábor Lugosi presented a random geometric graph model on high-dimensional spheres that is motivated from a statistical hypothesis testing problem. In particular the clique number and its dependence on the dimension of the spheres were addressed.

There was a special session on Wednesday morning dedicated to the memory of Philippe Flajolet who passed away about a month before the workshop. Hsien-Kuei Hwang gave a survey on the subjects of Philippe Flajolet's research continued by the talk of Wojciech Szpankowski.

Mini-Workshop: Random Trees, Information and Algorithms

Table of Contents

Hsien-Kuei Hwang (joint with Wei-Mei Chen and Tsung-Hsi Tsai)
Threshold phenomena in k-dominant skylines of random samples 1247

Svante Janson
Simply generated trees and conditioned Galton–Watson trees 1249

Yuliy Baryshnikov
Unimodal category of random univariate functions 1252

Louigi Addario-Berry (joint with Nicolas Broutin, Cecilia Holmgren)
Cutting down trees with a Markov chainsaw 1252

Christina Goldschmidt (joint with James Martin)
Cutting random recursive trees, and the Bolthausen–Sznitman coalescent 1255

Hsien-Kuei Hwang
The works of Philippe Flajolet 1258

Wojciech Szpankowski (joint with Gahyun Park, Hsien-Kuei Hwang, Pierre Nicodème)
Profile of Tries 1259

Ludger Rüschendorf (joint with G. Olaf Munsonius)
On depths and distances in random weighted $\boldsymbol{b}$*-ary trees* 1264

Gerold Alsmeyer (joint with John D. Biggins, Matthias Meiners)
The Functional Equation of the Smoothing Transform 1265

Mark Daniel Ward (joint with Pierre Nicodème)
Towards the Variance of the Profile of Suffix Trees 1269

Uwe Rösler
The Quicksort Process 1272

Ralph Neininger (joint with Henning Sulzbach)
On a Functional Contraction Method 1274

Henning Sulzbach (joint with Nicolas Broutin and Ralph Neininger)
A Process Convergence Result for Partial Match Queries in Random Quadtrees 1276

Rudolf Grübel
Metric aspects of binary search trees 1279

Nicolas Broutin (joint with C. Holmgren)
Behaviour of tree-based contention algorithms 1281

Gábor Lugosi (joint with Luc Devroye, András György, Frederic Udina)
Random geometric graphs in high dimension 1284

Abstracts

Threshold phenomena in k-dominant skylines of random samples

Hsien-Kuei Hwang
(joint work with Wei-Mei Chen and Tsung-Hsi Tsai)

Skylines of multivariate data sample were introduced for selecting representative groups in the database query literature by Börzsönyi et al. (see [4]) and had appeared in diverse areas under several different guises and names: *Pareto optimality*, *efficiency*, *maxima*, *admissibility*, *elite*, *sink*, etc.; see [6, 7] and the references therein for more information. These diverse terms reveal the importance of the use of skyline in practice. Many different notions and variants of skylines have been proposed in the literature, following the original paper [4]. In particular, the k-dominant skylines were introduced by Chan et al. (see [5]) in situations when the skylines are abundant and have received widespread discussions since. We focus in this paper on the asymptotic estimates of such skylines and prove several types of results under different probability assumptions of the input samples, which are believed to be useful for practitioners.

The definitions of skyline and many of its variants are based on the notion of dominance. Given a d-dimensional dataset $\mathcal{D}$, a point $\mathbf{p} \in \mathcal{D}$ is said to *dominate* another point $\mathbf{q} \in \mathcal{D}$ if $p_j \le q_j$ for $1 \le j \le d$, where $\mathbf{p} = (p_1, \dots, p_n)$ and $\mathbf{q} = (q_1, \dots, q_n)$, and is less than in at least one dimension. The non-dominated points in $\mathcal{D}$ are called the *skyline* (or *skyline points*) of $\mathcal{D}$. By relaxing the dominance definition to partial dominance, we say that a point $\mathbf{p} \in \mathcal{D}$ *k-dominates* another point $\mathbf{q} \in \mathcal{D}$ if there are k dimensions in which p_j is not greater than q_j and is less than in at least one of these k dimensions. The points in $\mathcal{D}$ that are not k-dominated by any other points are defined to be the *k-dominant skyline* of $\mathcal{D}$. The definition of k-dominant skyline implies that for a fixed dataset the number of k-dominant skylines decreases as k becomes smaller.

The number of skyline points is a key issue in their use and usefulness. This quantity under suitable random assumptions of the input is also important for practical modeling or reference purposes, as well as for the analysis of skyline-finding algorithms. The two major simple, representative random models are *hypercubes* and *simplices*. Assuming that the input dataset $\mathcal{D} = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ is taken uniformly and independently from the hypercube $[0,1]^d$, then it has been known since the 1960's (see [1]) that the expected number of skyline points of $\mathcal{D}$ is asymptotic to $(\log n)^{d-1}/(d-1)!$ for large n and finite d, exhibiting roughly the independence of the coordinates. On the other hand, if we assume that the input points are uniformly sampled from the d-dimensional simplex $\{|x_1| + \cdots + |x_d| \le 1, x_j \in (-1, 0]\}$, then the expected number of skyline points is asymptotic to $\Gamma(1/d)n^{1-1/d}$, reflecting obviously a stronger negative correlation of the coordinates; see [3] and the references cited there. Here Γ denotes Euler's Gamma function. For the number of skyline points under other models, see [2, 8, 11] and the references therein.

On the other hand, in contrast to the recent growing high-p-low-n trend (p being our d, the dimensionality), not much is known for the expected number of skyline points when d is large compared with n. The only exception is the uniform estimates given in [9] (see also [3]) for the expected number of skyline points $\mu_{n,d}$ in a random uniform samples of n points from the hypercube $[0,1]^d$. While the order $(\log n)^{d-1}/(d-1)!$ may seem slowly growing as d increases, it soon reaches the order n when d is around $\log n$, which is relatively small for moderate values of n. Consequently, the skyline points become too numerous to be of direct use. The growth of skyline points in the random d-dimensional simplex model is even faster and we can show that almost all points are skylines when d roughly exceeds $(\log n)/(\log\log n)$.

Since k-dominant skyline were proposed (see [5]) to resolve the abundance problem of skyline, it is of interest to know their quantity under suitable random models. A critical step in applying k-dominant skyline is to identify an appropriate k such that the size of the k-dominant skyline is within the acceptable ranges. But this may not be always feasible. Consider the 5-dimensional dataset $\mathcal{D}$ given in Table 1. The six points are all skyline points, one ($\mathbf{p}_6$) is the 4-dominant skyline point and no point is in the 3-dominant skyline. Clearly, $\mathbf{p}_6$ is to some extent better than the other points since it contains two components with the lowest value 1. However, it was already mentioned in [5] that some k-dominant skylines may be empty. For example, if we drop $\mathbf{p}_6$ from $\mathcal{D}$, then the five points are all skyline points but all k-dominant skylines are empty for $1 \le k \le 4$. In this example, other alternatives to k-dominant skylines have to be used. Unfortunately, such a property of *excessive skylines but few k-dominant skylines* is not uncommon, and we show in this paper that, under the hypercube and the simplex random models, the expected number of k-dominant skylines both tends to zero for large n and $1 \le k \le d-1$.

point	skyline	4-dominant skyline	3-dominant skyline
$\mathbf{p}_1$ $(1,2,2,3,3)$	✔	-	-
$\mathbf{p}_2$ $(3,1,2,2,3)$	✔	-	-
$\mathbf{p}_3$ $(3,3,1,2,2)$	✔	-	-
$\mathbf{p}_4$ $(2,3,3,1,2)$	✔	-	-
$\mathbf{p}_5$ $(2,2,3,3,1)$	✔	-	-
$\mathbf{p}_6$ $(2,3,1,1,3)$	✔	✔	-

Table 1: An example showing the property of many skylines but few k-dominant skylines.

More precisely, We present first an asymptotic vanishing property for the number of k-dominant skyline points under a common hypercube model when the dimensionality is fixed. The extension to include more points in the partial dominant skyline is showed to suffer from a similar drawback. We then prove that changing the underlying model from hypercube to simplex does not change the asymptotic vanishing property. Switching from continuous model to a categorical model also does not help and we have too many skyline points. Roughly, as the

total number of sample points are finite in this model, the expected number of k-dominant skylines will be asymptotically linear, meaning too many choices for ranking or selection purposes. All these results point to the negative side for the use of k-dominant skylines under similar data situations. We then address the positive side by considering again the hypercubes but with growing dimensionality. A sharp threshold phenomenon is discovered for $\mathbb{E}[M_{d-1}(n)]$ when $d \to \infty$ with n, which says that if d is less than the threshold, then $\mathbb{E}[M_{d-1}(n)] \to 0$, while if d is larger than the threshold, then $\mathbb{E}[M_{d-1}(n)] \to \infty$, at the threshold $\mathbb{E}[M_{d-1}(n)]$ taking either the value 0 or 1.

References

[1] O. Barndorff-Nielsen and M. Sobel (1966). On the distribution of the number of admissible points in a vector random sample. *Theory of Probability and its Applications*, **11** 249–269.

[2] Y. Baryshnikov, On expected number of maximal points in polytopes. *2007 Conference on Analysis of Algorithms*, AofA 07, pp. 227–236, *Discrete Math. Theor. Comput. Sci. Proc.*, Nancy, 2007.

[3] Z.-D. Bai, L. Devroye, H.-K. Hwang and T.-H. Tsai, Maxima in hypercubes, *Random Structures and Algorithms*, **27** (2005), 290–309.

[4] S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator, *Proceedings 17th International Conference on Data Engineering*, 421–430, 2001.

[5] C. Y. Chan, H. V. Jagadish, K.-L. Tan, A. K. H. Tung, and Z. Zhang, Finding k-dominant skylines in high dimensional space, *Proceedings of the 2006 ACM SIGMOD international conference on management of data*, 503–514, 2006.

[6] W.-M. Chen, H.-K. Hwang, and T.-H. Tsai, Efficient maxima-finding algorithms for random planar samples, *Discrete Mathematics and Theoretical Computer Science*, **6**:1 (2003), 107–122.

[7] W.-M. Chen, H.-K. Hwang, and T.-H. Tsai, Simple, efficient maxima-finding algorithms for multidimensional samples, submitted. http://arxiv.org/abs/0910.1392

[8] L. Devroye, Records, the maximal layer, and uniform distributions in monotone sets. *Comput. Math. Appl.* **25** (1993), 19–31.

[9] H.-K. Hwang, Phase changes in random recursive structures and algorithms, in *Probability, Finance and Insurance*, pp. 82–97, World Sci. Publ., River Edge, NJ, 2004,

[10] H.-K. Hwang and T.-H. Tsai, Multivariate records based on dominance, *Electronic Journal of Probability*, **15** (2010), 1863–1892.

[11] T. Schreiber and J. E. Yukich, Variance asymptotics and central limit theorems for generalized growth processes with applications to convex hulls and maximal points. *Ann. Probab.* **36** (2008), 363–396.

Simply generated trees and conditioned Galton–Watson trees

Svante Janson

The trees that we consider are rooted and ordered (= plane); thus each node v has a number of children, ordered in a sequence $v_1, \dots, v_d$, where $d = d(v) \geq 0$ is the *outdegree* of v. (See [1] for more information on these and other types of trees; the trees we consider are there called *planted plane trees*.) We let $\mathfrak{T}_n$ denote the set of all ordered rooted trees with n nodes (including the root) and let $\mathfrak{T}_{\mathrm{f}} := \bigcup_{n=0}^{\infty} \mathfrak{T}_n$ be the set of all finite ordered rooted trees.

Let $(w_k)_{k\ge0}$ be a fixed *weight sequence* of non-negative real numbers. We then define the *weight* of a tree $T \in \mathfrak{T}_{\mathrm{f}}$ by

$$w(T) := \prod_{v\in T} w_{d(v)},$$

taking the product over all nodes v in T. Trees with such weights are called *simply generated trees* and were introduced by Meir and Moon [5]. To avoid trivialities, we assume that $w_0 > 0$ and that there exists some $k \ge 2$ with $w_k > 0$.

We let $\mathcal{T}_n$ be the random tree obtained by picking an element of $\mathfrak{T}_n$ at random with probability proportional to its weight, i.e.,

$$\mathbb{P}(\mathcal{T}_n = T) = \frac{w(T)}{Z_n}, \qquad T \in \mathfrak{T}_n,$$

where the normalizing factor Z_n, known as the *partition function*, is given by

$$Z_n := \sum_{T\in\mathfrak{T}_n} w(T).$$

We consider only n such that $Z_n > 0$.

One particularly important case is when $\sum_{k=0}^\infty w_k = 1$, so the weight sequence (w_k) is a probability distribution on $\mathbb{Z}_{\ge0}$. In this case, the random tree $\mathcal{T}_n$ is the same as the random Galton–Watson tree $\mathcal{T}$ with offspring distribution (w_k) conditioned on $|\mathcal{T}| = n$. In this case the random tree $\mathcal{T}_n$ is thus called a *conditioned Galton–Watson tree*.

The distribution of the tree $\mathcal{T}_n$ does not change if w_k is replaced by $\widetilde{w}_k := ab^k w_k$ for some $a, b > 0$. Using this, we can always reduce to one of the three following cases, where $\rho \in [0,\infty]$ is the radius of convergence of the generating function $\Phi(x) := \sum_{k=0}^\infty w_k x^k$ and $\mu := \sum_{k=0}^\infty k w_k = \Phi'(1)$:

(i) Critical Galton–Watson: (w_k) a probability distribution with mean $\mu = 1$. (In this case $\rho \ge 1$.)
(ii) Subcritical Galton–Watson: (w_k) a probability distribution with mean $\mu < 1$ and $\rho = 1$.
(iii) Not Galton–Watson: $\rho = 0$.

Case (i) is the standard case, and most work has been done for this case only (often with additional conditions like $\operatorname{Var}\xi < \infty$).

Probabilists, including myself, have often dismissed the remaining cases as uninteresting exceptional cases. However, some researchers, including mathematical physicists, have studied such cases and found a condensation, showing that there are interesting phenomena in the exceptional cases as well. The purpose of this talk is to give a unified limit theorem of $\mathcal{T}_n$ as $n \to \infty$ for all simply generated trees, extending the well-known result in the standard case (i), and to encourage further research in the other cases too.

A limit theorem

In cases (i) and (ii), let ξ be an integer-valued random variable with distribution (w_k); thus $0 < \mathbb{E}\,\xi \le 1$. In case (iii), let $\xi = 0$, so $\mathbb{E}\,\xi = 0$. In all cases, let $\widehat{\xi}$ be a

random variable with values in $\{0,1,\dots,\infty\}$ with the distribution

$$\mathbb{P}(\widehat{\xi}=k) := \begin{cases} k\,\mathbb{P}(\xi=k), & k=0,1,2,\dots,\\ 1-\mathbb{E}\,\xi, & k=\infty. \end{cases}$$

In case (i), this is the usual size-biased transformation of ξ.

We define the modified Galton–Watson tree $\widehat{\mathcal{T}}$ as follows: There are two types of nodes: *normal* and *special*, with the root being special. Normal nodes have offspring (outdegree) according to independent copies of ξ, while special nodes have offspring according to independent copies of $\widehat{\xi}$. Moreover, all children of a normal node are normal; when a special node gets an infinite number of children, all are normal; when a special node gets a finite number of children, one of its children is selected uniformly at random and is special, while all other children are normal.

The special nodes form a path from the root; we call this path the *spine* of $\widehat{\mathcal{T}}$. $\widehat{\mathcal{T}}$ behaves differently in our three different cases:

(i) In the critical Galton–Watson case, the spine is an infinite path. Each outdegree $d(v)$ in $\widehat{\mathcal{T}}$ is finite, so the tree is infinite but locally finite. This is the size-biased Galton–Watson tree defined by Lyons, Pemantle and Peres [4].

(ii) In the subcritical Galton–Watson case, the spine is a.s. finite with a number L of vertices that has a (shifted) Geometric distribution $\mathrm{Ge}(1-\mu)$:

$$\mathbb{P}(L=\ell) = (1-\mu)\mu^{\ell-1}, \qquad \ell=1,2,\dots.$$

(iii) In the non-Galton–Watson case, the spine consists of the root only; the root has infinitely many children, and all its children are leaves. $\widehat{\mathcal{T}}$ is thus an infinite star. (This is the limiting case $\mu=0$ of case (ii).)

In case (i), all vertices have finite degree, while in cases (ii) and (iii), the tree has (a.s.) exactly one node with infinite outdegree, viz. the top of the spine.

Our main theorem extends a result by Lyons, Pemantle and Peres to cases (ii) and (iii) in complete generality. For special cases, see [3] and [2].

Theorem 1. *In all three cases, T_n converges in distribution to $\widehat{\mathcal{T}}$ as $n\to\infty$, in the topology defined by convergence of all finite parts of the tree.*

The topology can, equivalently, be defined as convergence of each outdegree.

Acknowledgement. This research was started during a visit to NORDITA, Stockholm, during the program *Random Geometry and Applications*, 2010. I thank the participants, in particular Bergfinnur Durhuus, Thordur Jonsson and Sigurður Stefánsson, for stimulating discussions.

References

[1] M. Drmota, *Random Trees*, Springer, Vienna, 2009.

[2] S. Janson, T. Jonsson and S. Ö. Stefánsson, Random trees with superexponential branching weights. Preprint, 2011. `arXiv:1104.2810.`

[3] T. Jonsson and S. Ö. Stefánsson, Condensation in nongeneric trees. *Journal of Statistical Physics*, **142** (2011), no. 2, 277–313.
[4] R. Lyons, R. Pemantle and Y. Peres, Conceptual proofs of $L \log L$ criteria for mean behavior of branching processes. *Annals of Probability* **23** (1995), no. 3, 1125–1138.
[5] A. Meir and J. W. Moon, On the altitude of nodes in random trees. *Canad. J. Math.*, **30** (1978), 997–1015.

Unimodal category of random univariate functions

Yuliy Baryshnikov

Let M be a smooth manifold. We call a function $f : M \to \mathbb{R}_+$ *unimodal* if the excursion sets $M_f(c) := \{f^{-1}(c, \infty)\}, c > 0$ are *contractible*.

For a non-negative function $g : M \to \mathbb{R}_+$, a decomposition $f = \sum_\alpha f_\alpha$ is called *unimodal*, if all the summands are unimodal. The minimal number of unimodal summands in a unimodal decomposition of f is called the *unimodal category* (the term comes from a formal analogy with the *Lyusternik-Schnirelman category*) of f, and denoted as $\mathsf{UCat}(f)$. Unimodal category is an important "fully covariant" invariant of a probability distribution, giving a lower bound on the *effects* that result in this distribution. An upper bound on UCat is the number of *modes* (i.e. local maxima) of the distribution which is rarely exact.

The question of finding UCat for an arbitrary function is hard and answered satisfactory only in dimension 1. This talk addressed the question of finding unimodal category for some simplest *random* functions. (We remark that UCat for Brownian motions, for example for the Brownian excursion, is always infinity.)

Let $f : [0, 2n] \to \mathbb{R}_+$ be the (linearly interpolated) simple random walk with n steps 1 and n steps -1 conditioned to stay positive ("Dyck paths with uniform measure"). Our main result states that $\mathsf{UCat}(f)$ scaled by $\sqrt{n}$ converges in distribution to

$$\int_0^1 1/e(s)ds,$$

where $e(s) : [0, 1] \to \mathbb{R}_+$ is the standard Brownian excursion on $[0, 1]$. In particular, the expectation of $\mathsf{UCat}(f)$ grows as $\sqrt{2\pi n}$.

Cutting down trees with a Markov chainsaw

Louigi Addario-Berry
(joint work with Nicolas Broutin, Cecilia Holmgren)

The subject of cutting down trees has been introduced by [28, 29]. One is given a rooted tree T which is pruned by random removal of edges. At each step, only the portion containing the root is kept (we refer to the portion not containing the root as the *pruned* portion) and the process continues until eventually the root has been isolated. The main parameter of interest is the number of cuts necessary to isolate the root. (The dual problem of isolating a leaf or a node with a specific label has been considered by [24, 23].)

The procedure has been studied on different deterministic and random trees. Essentially two kinds of random models have been considered for the tree: *recursive trees* with typical inter-node distances of the order of $\log n$ [30, 19, 14, 18], and trees arising from critical branching processes conditioned on their size, with typical distances of order $\sqrt{n}$ [21, 16, 33, 32, 20, 33]. We are interested in the latter family, and refer to such trees as *conditioned trees* for short.

The original analyses by [28] include asymptotics for the mean and variance of the number of cuts. In recent years, the subject has regained interest. [32] and [16] have studied the somewhat simpler case where, conditional on its size, the distribution of the remaining tree is left unchanged by a cut; this naturally simplifies greatly the recursive treatment. The class of random trees which satisfy this property include the important example of rooted Cayley trees (uniformly random labelled rooted trees), or, equivalently, Poisson Galton–Watson trees. For this class, they obtained the limiting distribution of the number of cuts using the method of moments and an analytic treatment of the recursive equation describing the cutting procedure. [21, 20] used a representation of the number of cuts in terms of generalized records in a labelled tree to extend these results to all the family trees of critical branching processes with offspring distribution having a finite variance. His method is also based on the calculation of moments.

At this point, it is important to mention that once divided by $\sigma\sqrt{n}$, the number of cuts required to isolate the root of a tree of size n converges in distribution to a Rayleigh random variable with density $xe^{-x^2/2}$ on $[0, \infty)$. The fact that the Rayleigh distribution appears here with a $\sqrt{n}$ scaling in a setting involving conditioned trees struck us. The Rayleigh distribution also arises as the limiting distribution of the length of a path between two uniformly random nodes in a conditioned tree, after appropriate rescaling. We show that the existence of a Rayleigh limit in both cases is not fortuitous. We prove using a coupling method that the number of cuts and the distance between two random vertices are asymptotically equal in distribution (modulo a constant factor σ^2). This approach yields as a by-product a very simple proofs of the results concerning the distribution of the number of cuts obtained by [32, 16, 20, 21]. The connection is most striking in the case of rooted Cayley trees. In this case, for any finite n we exhibit a coupling which shows that the number of nodes on a path between two random nodes of a Cayley tree has *exactly* the same distribution as the number of cuts required to isolate the root. Our approach also allows us to describe the joint distribution of the sequence of pruned trees.

Using a classical bijection, it is possible to re-express the cutting down procedure as acting on excursions rather than trees. Considering the limiting version of this procedure allows us to define a cutting down procedure for the continuum random trees of [4, 2, 5], where the trees are cut according to a Poisson point process on the skeleton. This continuous version of the procedure yields a construction of the random variable mentioned by [21, Remark 1.11]. As a by-product, we obtain a novel random transformation of a Brownian excursion into a Brownian bridge.

References

[1] D. Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, 3:450–465, 1990.

[2] D. Aldous. The continuum random tree II: an overview. In M.T. Barlow and N.H. Bingham, editors, *Stochastic Analysis*, pages 23–70. Cambridge University Press, 1991.

[3] D. Aldous. Asymptotic fringe distibutions for general families of random trees. *The Annals of Applied Probability*, 1:228–266, 1991.

[4] D. Aldous. The continuum random tree. I. *The Annals of Probability*, 19:1–28, 1991.

[5] D. Aldous. The continuum random tree III. *The Annals of Probability*, 21:248–289, 1993.

[6] D. Aldous and J. Pitman. Brownian bridge asymptotics for random mappings. *Random Structures and Algorithms*, 5:487–512, 1994.

[7] D. Aldous and J. Pitman. The standart additive coalescent. *The Annals of Probability*, 26: 1703–1726, 1998.

[8] D. Aldous and J. M. Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In H. Kesten, editor, *Discrete and Combinatorial Probability*, pages 1–72. Springer Verlag, 2003.

[9] V. Anantharam and P. Tsoucas. A proof of the Markov chain tree theorem. *Statistics & Probability Letters*, 8(2):189–192, 1989.

[10] J. Bertoin. A fragmentation process connected to Brownian motion. *Probability Theory and Related Fields*, 117:289–301, 2000.

[11] P. Biane and M. Yor. Valeurs principales associées aux temps locaux Browniens. *Bull. Sci. Math. (2)*, 111:23–101, 1987.

[12] A. Broder. Generating random spanning trees. In *30th Annual Symposium on Foundations of Computer Science,*, pages 442–447, 1989.

[13] P. Chassaing and R. Marchand. Personal Communication, 2009.

[14] M. Drmota, A. Iksanov, M. Moehle, and U. Roesler. A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Structures and Algorithms*, 34:319–336, 2009.

[15] S.N. Evans. *Probability and real trees, École d'Été de Probabilités de Saint-Flour XXXV-2005*, volume 1920 of *Lecture Notes in Mathematics*. Springer, 2005.

[16] J.A. Fill, N. Kapur, and A. Panholzer. Destruction of very simple trees. *Algorithmica*, 46: 345–366, 2006.

[17] B. Haas and G. Miermont. Scaling limits of Markov branching trees with applications to Galton–Watson and random unordered trees. 2010.

[18] C. Holmgren. Random records and cuttings in split trees: extended abstract. In *Fifth Colloquium on Mathematics and Computer Science*, volume AI of *DMTCS Proceedings*, pages 269–282. Discrete Mathematics and Theoretical Computer Science, 2008.

[19] A. Iksanov and M. Möhle. A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree. *Electronic Communications in Probability*, 12:28–35, 2007.

[20] S. Janson. Random records and cuttings in complete binary trees. In M. Drmota, P. Flajolet, D. Gardy, and B. Gittenberger, editors, *Mathematics and Computer Science III: Algorithms, Trees, Combinatorics and Probability (Vienna)*, pages 242–253, Basel, 2004. Birkhäuser.

[21] S. Janson. Random cuttings and records in deterministic and random trees. *Random Structures and Algorithms*, 29:139–179, 2006.

[22] V. F. Kolchin. *Random Mappings*. Optimization Software, New York, 1986.

[23] M. Kuba and A. Panholzer. Isolating a leaf in rooted trees via random cuttings. *Annals of Combinatorics*, 12:81–99, 2008.

[24] M. Kuba and A. Panholzer. Isolating nodes in recursive trees. *Aequationes Mathematicae*, 76:258–280, 2008.

[25] J.-F. Le Gall. The uniform random tree in a Brownian excursion. *Probability Theory and Related Fields*, 96:369–383, 1993.

[26] J.-F. Le Gall. Random trees and applications. *Probability Surveys*, 2:245–311, 2005.
[27] J.-F. Marckert and A. Mokkadem. The depth first processes of Galton–watson trees converge to the same Brownian excursion. *The Annals of Probability*, 31:1655–1678, 2003.
[28] A. Meir and JW Moon. Cutting down random trees. *Journal of the Australian Mathematical Society*, 11:313–324, 1970.
[29] A. Meir and JW Moon. Cutting down recursive trees. *Mathematical Biosciences*, 21:173–181, 1974.
[30] A. Meir and J.W. Moon. On the altitude of nodes in random trees. *Canadian Journal of Mathematics*, 30:997–1015, 1978.
[31] G. Miermont. Tesselation of random maps of arbitrary genus. *Ann. Sci. ENS*, 42:725–781, 2009.
[32] A. Panholzer. Non-crossing trees revisited: cutting down and spanning subtrees. In *Discrete random walks (Paris 2003)*, volume AC of *DMTCS*, pages 265–276, 2003.
[33] A. Panholzer. Cutting down very simple trees. *Quaestiones Mathematicae*, 29:211–228, 2006.
[34] Y.L. Pavlov. *Random Forests*. VSP, Utrecht, 2000.
[35] J. Pitman. Coalescent random forests. *Journal of Combinatorial Theory, Series A*, 85:165–193, 1999.
[36] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer, Berlin, 2006.
[37] D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Springer, Berlin, 3rd edition, 2004.
[38] L.C.G. Rogers. A guided tour through excursions. *Bulletin of the London Mathematical Society*, 21:305–341, 1989.
[39] A.C. Yao. On the average behavior of set merging algorithms. In *STOC '76: Proceedings of the 8th Annual ACM Symposium on Theory of Computing*, pages 192–195, 1976.

Cutting random recursive trees, and the Bolthausen–Sznitman coalescent

Christina Goldschmidt
(joint work with James Martin)

The Bolthausen–Sznitman coalescent was introduced in the context of spin glasses in [1]. These days, it is usually thought of as a special case of a more general class of coalescent processes introduced by Pitman [5] and Sagitov [7] and usually referred to as the Λ-*coalescents*. These are Markov processes taking values in the space $\mathcal{P}_\infty$ of partitions of $\mathbb{N}$, or the space $\mathcal{P}_n$ of partitions of $\{1, 2, \ldots, n\}$, where the blocks of the partition represent particles which gradually coalesce over the course of time. The dynamics of the Bolthausen–Sznitman coalescent on $\mathcal{P}_n$ are very simple. Suppose that we have an initial state $\pi \in \mathcal{P}_n$ consisting of b blocks. Then any k of them coalesce at rate

$$\lambda_{b,k} = \frac{(k-2)!(b-k)!}{(b-1)!}, \quad 2 \leq k \leq b \leq n,$$

regardless of block sizes or which integers the blocks contain. Since the state-space $\mathcal{P}_n$ is finite, the distribution of the coalescent is entirely specified by its initial distribution and these transition rates.

A random partition Π of $[n]$ is *exchangeable* if

$$\mathbb{P}(\Pi = \pi) = \mathbb{P}(\Pi = \sigma(\pi))$$

for any $\pi \in \mathcal{P}_n$ and any permutation σ of $[n]$. A random partition of $\mathbb{N}$ is exchangeable if its restriction to $[n]$ is exchangeable in the above sense for all $n \geq 1$. The above dynamics preserve the property of exchangeability: if the initial state of the Bolthausen–Sznitman coalescent is exchangeable, then the state remains exchangeable for all times. Moreover, the rates are such that the restriction of the coalescent evolving in $\mathcal{P}_{n+1}$ to $[n]$ evolves exactly as the coalescent evolving in $\mathcal{P}_n$; in other words, we have *consistency* for each $n \geq 1$. This means that we can define the coalescent evolving in $\mathcal{P}_\infty$ simply as a projective limit.

Let $(\Pi(t), t \geq 0)$ be the Bolthausen–Sznitman coalescent in $\mathcal{P}_\infty$. An important consequence of the exchangeability of $\Pi(t)$ for all $t \geq 0$ is that its blocks possess asymptotic frequencies i.e. if B is a block of $\Pi(t)$ then

$$\lim_{n\to\infty} \frac{|B \cap [n]|}{n}$$

exists almost surely.

We now turn to random recursive trees. A *recursive tree* is a labelled, unordered tree which is rooted at its vertex of smallest label and has the property that its labels increase along non-backtracking paths away from the root. We allow any partition of $[n]$ to be a label-set for such a tree, where we order blocks according to their least elements; the canonical label-set is the partition into singletons $(\{1\}, \{2\}, \ldots, \{n\})$ for some $n \geq 1$. A *random recursive tree* on label-set $L = (\ell_1, \ell_2, \ldots, \ell_b)$ with $\ell_1 \preceq \ell_2 \preceq \ldots \preceq \ell_b$ is simply chosen uniformly at random from the $(b-1)!$ recursive trees with those labels. It is more easily constructed via a recursive procedure:

- start from a single vertex labelled by ℓ_1;
- for $k \geq 2$, attach a vertex labelled by ℓ_k to one of the vertices labelled by $\ell_1, \ldots, \ell_{k-1}$ chosen uniformly at random.

(Note that this procedure does not, in fact, require finiteness of the label-set.) We now consider a variant of a cutting procedure which was first introduced by Meir and Moon [3, 4] and has been subsequently much studied in the combinatorics literature. Pick an edge uniformly at random. Cut it, and combine all of the labels below the cut edge with the label of the vertex just above. Repeat, until only the root remains (necessarily labelled by $[n]$). If we start with a partition of $[n]$ then, at every subsequent step, we clearly obtain a coarser partition of $[n]$. We can easily put this procedure into continuous time by associating an independent standard exponential random variable with each edge: this random variable gives the time at which that edge will be cut, if it still exists in the tree at that time.

Suppose that the tree has the partition of $[n]$ into singletons as its initial label-set. Let $\Gamma^{[n]}(t)$ be the partition obtained by running the cutting procedure for time t.

Theorem 1. *The process* $(\Gamma^{[n]}(t), t \geq 0)$ *is the Bolthausen–Sznitman coalescent on* $[n]$.

The proof is straightforward and relies on the fact that a random recursive tree cut at a uniformly-chosen edge is again a random recursive tree on its new

label-set. Moreover, the rate at which a cut results in a coalescence of k labels is

$$\frac{(k-2)!(b-k)!}{(b-1)!}, \quad 2 \leq k \leq b \leq n.$$

We immediately recognise the rates of the Bolthausen–Sznitman coalescent. See [2] for the details of the proof.

Note that, because of the recursive way in which the tree is built, we have consistency in n and so, in fact, we can define $(\Gamma(t), t \geq 0)$ evolving in $\mathcal{P}_\infty$ by means of the cutting procedure applied to a random recursive tree labelled by $\mathbb{N}$.

The representation given by Theorem 1 is somewhat surprising. It splits the randomness of the coalescent into two parts: the randomness used to build the tree, and the randomness used to cut it. A particular realisation of the tree corresponds to a particular conditioning of the path of the coalescent. For example, if $\{2\}$ and $\{5\}$ are both children of $\{1\}$ then we condition 2 and 5 only to be in the same block once they have both coalesced with 1. The tree representation gives a size-biased viewpoint rather than the usual exchangeable one. The block containing 1 (which is always the label of the root) is a size-biased pick from amongst the blocks and so tends to be large. We can think of it as a tagged particle, and we watch the coalescent evolve from its point of view. This leads to a rather nice way to prove the following properties of the coalescent, originally due to Bolthausen and Sznitman [1] and Pitman [5] respectively. Write $\mathrm{PD}(\alpha, \theta)$ for the Poisson–Dirichlet distribution with parameters $0 < \alpha < 1$ and $\theta > -\alpha$ (see Pitman and Yor [6]).

Theorem 2. (1) *Write $F(t)$ for the asymptotic frequencies of the blocks of $(\Pi(t), t \geq 0)$, where the frequencies are listed in decreasing order of size. Then*

$$F(t) \sim \mathrm{PD}(e^{-t}, 0).$$

(2) *Write $F_*(t)$ for the frequency of the block containing 1 at time t. Then $(F_*(t), t \geq 0)$ is Markovian, with the same distribution as the process $(\gamma(1-e^{-t})/\gamma(1), t \geq 0)$, where $(\gamma(s), s \geq 0)$ is a Gamma subordinator. This entails that $F_*(t) \sim \mathrm{Beta}(1-e^{-t}, e^{-t})$. Moreover, if $J_1 \geq J_2 \geq \ldots \geq 0$ is the ranked sequence of jumps of $(F_*(t), t \geq 0)$ then $(J_1, J_2, \ldots) \sim \mathrm{PD}(0, 1)$.*

We refer the reader to [2] for the proofs and for further development.

Open problem. Find another exchangeable coalescent which may be represented by cutting down a combinatorial tree.

References

[1] E. Bolthausen and A.-S. Sznitman, *On Ruelle's probability cascades and an abstract cavity method*, Comm. Math. Phys. **197**(2) (1998), 247–276.

[2] C. Goldschmidt and J.B. Martin, *Random recursive trees and the Bolthausen–Sznitman coalescent*, Electron. J. Probab. **10** (2005), Paper no. 21, 718–745.

[3] A. Meir and J. W. Moon. *Cutting down random trees*, J. Austral. Math. Soc. **11** (1970), 313–324.

[4] A. Meir and J. W. Moon. *Cutting down recursive trees*, Math. Bioscience **21** (1974), 173–181.

[5] J. Pitman, *Coalescents with multiple collisions*, Ann. Probab. **27**(4) (1999), 1870–1902.
[6] J. Pitman and M. Yor, *The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator*, Ann. Probab. **25** (1997), 855–900.
[7] S. Sagitov, *The general coalescent with asynchronous mergers of ancestral lines*, J. Appl. Probab. **36**(4) (1999), 1116–1125.

The works of Philippe Flajolet

Hsien-Kuei Hwang

Philippe Flajolet was born in 1948 in Lyon and passed away on March 22, 2011. He was one of the most influential figures in several scientific fields, notably in analysis of algorithms and in analytic combinatorics, for each of which he published a book jointly with Robert Sedgewick). He was elected member of the French Academy of Science (l'Académie des Sciences) in 2003. The major events of his life are briefly summarized as follows.

1948 **Born in Lyon**

1966 **Baccalauréat**
1968
Ecole polytechnique
1970
1971 **DEA; Research Assistant at INRIA**
1973 **Thèse de 3e cycle, Université Paris 7**

Docteur ès Science, Université Paris 11
1979 **Research Director**

1981 **Head of Projet ALGO**

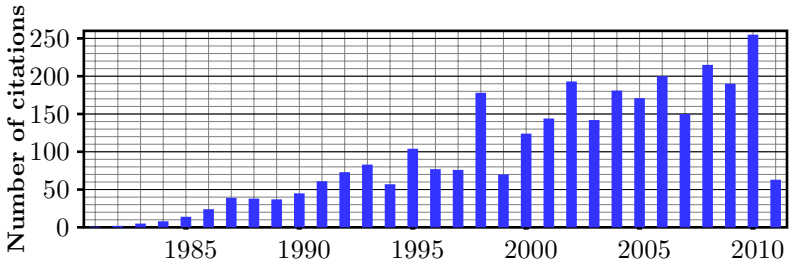
1986 **Prix Scientifique (Union des Assurances de Paris)**

Prix Michel Montpetit
1994 **corresponding member (Académie des Sciences)**
Doctorate Honoris Causa (U. Libre Bruxelles)
1996 **Analysis of Algorithms book published**
Fellow, Academia Europaea

2003 **Full member Académie des Sciences**
2004 **Médaille d'argent CNRS**

2007 **Simulation Prize of INFORMS**
2009 **Analytic Combinatorics book published**
2011 **Rest in peace**

We first give a more outsider's view of his major works through several figures and tables (for example, the number of citations per year is shown below), and then provide a more thorough "guided tour" for almost all of his publications, indicating important ideas, original developments, philosophical thoughts, interdisciplinary connections, and linguistic-complexity synthesis.



One can readily grasp an idea of his most popular works through a simple search on Google Scholar.

Quoted from the webpage[1] of EATCS (European Association for Theoretical Computer Science):

"Philippe Flajolet (1948–2011) passed away on Tuesday, March 22. He was a larger-than-life theorist, the kind of person who "makes" an institution and becomes one himself.

Philippe Flajolet (1 December 1948–22 March 2011) was a French computer scientist. A former student of École Polytechnique, Philippe Flajolet got a Ph.D. in computer science from University Paris Diderot in 1977 and a doctorate of state in 1979. Most of Philippe Flajolet's research work was dedicated towards generic methods for analyzing the computational complexity of algorithms, including the theory of average-case complexity. He introduced the theory of analytic combinatorics. With Robert Sedgewick of Princeton, he wrote the first book-length treatment of the topic, the 2009 book entitled Analytic Combinatorics. A summary of his research up to 1998 can be found in the article "Philippe Flajolet's research in Combinatorics and Analysis of Algorithms" by H. Prodinger and W. Szpankowski, Algorithmica 22 (1998), 366-387. At the time of his death from a serious illness, Philippe Flajolet was a research director (senior research scientist) at INRIA in Rocquencourt. From 1994 to 2003 he was a corresponding member of the French Academy of Sciences, and was a full member from 2003 on.

He was also a member of the Academia Europaea."

Profile of Tries

Wojciech Szpankowski

(joint work with Gahyun Park, Hsien-Kuei Hwang, Pierre Nicodème)

Tries are prototype data structures useful for many indexing and retrieval purposes. They were first proposed by de la Briandais [1] in the late 1950's for information processing; Fredkin [5] suggested the current name as it being part of re*trie*val. Tries are multiway trees whose nodes are vectors of characters or digits. Due to their simplicity and efficiency, tries found widespread use in diverse applications ranging from document taxonomy to IP addresses lookup, from data

[1] www.eatcs.org/index.php/component/content/article/1-news/922-in-memoriam-of-philippe-flajolet-19482011

compression to dynamic hashing, from partial-match queries to speech recognition, from leader election algorithms to distributed hashing tables (see [10, 11, 15]). Here, we are concerned with probabilistic properties of the profiles of tries, where the *profile* of a tree is the sequence of numbers each counting the number of nodes with the same distance from the root. We discover several new phenomena in the profiles of tries built over strings generated by a random memoryless source, and develop asymptotic tools to describe them.

Tries are natural choice of data structures when the input records involve a notion of alphabets or digits. They are often used to store such data so that future retrieval can be made efficient. Given a sequence of n words over the alphabet $\{a_1, \ldots, a_m\}$, $m \geq 2$, we can construct a trie as follows. If $n = 0$, then the trie is empty. If $n = 1$, then a single (external) node holding the word is allocated. If $n \geq 1$, then the trie consists of a root (internal) node directing words to the m subtrees according to the first alphabet of each word, and words directed to the same subtree are themselves tries (see [10, 11, 15] for more details).

Throughout, we write $B_{n,k}$ to denote the number of external nodes (leaves) at distance k from the root; the number of internal nodes at distance k from the root is denoted by $I_{n,k}$. For simplicity, we will refer to $B_{n,k}$ as the *external profile* and $I_{n,k}$ the *internal profile*. Figure 1 shows a trie and its profiles. Here we study the profiles of a trie built over n binary strings generated by a memoryless source. More precisely, we assume that the input is a sequence of n independent and identically distributed random variables, each being composed of an infinite sequence of Bernoulli random variables with mean p, where $0 < p < 1$ is the probability of a "1" and $q := 1 - p$ is the probability of a "0". The corresponding trie constructed from these n bit-strings is called a *random trie*.

Summary of main results

We summarize here our main results proved in [13]. Crucial to our analysis of the profiles is the asymptotics of the expected profiles. Not only are the results fundamental and highly interesting, but also the analytic methods we used are of certain generality.

The expected external profile $\mu_{n,k} := \mathbb{E}(B_{n,k})$ satisfies the following recurrence

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

for $n \geq 2$ and $k \geq 1$ with the initial values $\mu_{n,0} = 0$ for all $n \neq 1$ and 1 for $n = 1$. Furthermore, $\mu_{0,k} = 0, k \geq 0$ and $\mu_{1,k} = 0$ for $k \geq 1$ and equal to 1 when $k = 0$. Throughout we assume that $p > q = 1 - p$ unless stated otherwise.

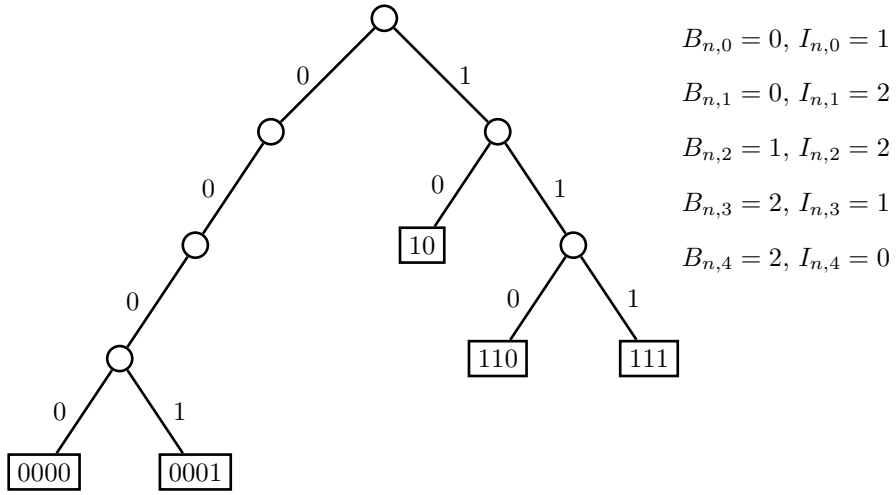


FIGURE 1. *A trie of $n = 5$ records and its profiles: the circles represent internal nodes and rectangles holding the records are external nodes.*

We solve asymptotically (1) for various ranges of k when $p \neq q$; a crude description of the asymptotics of $\mu_{n,k}$ is as follows.

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

where

(3)

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

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

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

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

To derive more precise asymptotics of $\mu_{n,k}$ than the phase transitions (2) of the polynomial order of $\mu_{n,k}$, we divide all possible values of k into four overlapping ranges.

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

(II) *Saddle-point range*: $\alpha_1(\log n - \log\log\log n + K_n) \le k \le \alpha_2(\log n - K_n\sqrt{\log n})$;
(III) *Gaussian transitional range*: $k = \alpha_2 \log n + o((\log n)^{2/3})$;
(IV) *Polar singularity range*: $k \ge \alpha_2 \log n + K_n\sqrt{\log n}$,

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

More precisely, in [13] we prove that for k lying in range (I) the expected external profile $\mu_{n,k}$ decays first exponentially fast (asymptotic to $q^k n(1-q^k)^{n-1}$). Then, when k is around $\alpha_1(\log n - \log\log\log n + \log(p/q - 1) + m\log(p/q))$ for some integer $m \ge 0$,

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

which is of order

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

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

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

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

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

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

and $G_1(\rho; x)$ is a periodic function. *Analytically*, these oscillations are consequences of an infinite number of saddle-points appearing in the integrand of the associated Mellin transform of the expected profile, but *visually* they look like certain sine waves due to the fact that the corresponding Fourier expansions involve Gamma function with increasing parameters, which decreases very fast along fixed vertical line for increasing imaginary part, so that only a few terms dominate.

Finally, in range (IV) we prove that

$$\mu_{n,k} \sim 2pqn^2(p^2+q^2)^{k-1} = \frac{2pq}{p^2+q^2}\, n^{\upsilon_2},$$

where $\upsilon_2 = 2 + \alpha_{n,k}\log(p^2+q^2)$, and the periodic function disappears. Here, the asymptotic behavior of the expected profile is dictated by the expected number of pairs (of input-strings) having common prefixes of length at least k. This property is analytically reflected by a polar singularity in the associated Mellin transform. Asymptotics of $\mu_{n,k}$ in range (III) for $k = \alpha_2\log n + o(\log^{2/3} n)$ is presented in [13]. In this transitional range, the saddle-point coalesces with the polar singularity, so we use the Gaussian integral to describe the behavior of $\mu_{n,k}$.

The expected value of the internal profile $\mathbb{E}(I_{n,k})$ is also discussed in [13]. In particular, the expected internal profile is asymptotically equivalent to 2^k for $k \leq \alpha_0(\log n - K_n\sqrt{\log n})$, where $\alpha_0 := 2/(\log(1/p) + \log(1/q))$. When $k \geq \alpha_2(\log n + K_n\sqrt{\log n})$, then $\mathbb{E}(I_{n,k}) \sim (p^2+q^2)\,\mathbb{E}(B_{n,k})/pq$. Between these two ranges, it is again the infinite number of saddle-points that yield the dominant asymptotic approximation. Unlike $\mu_{n,k}$, an additional phase transition appears in the asymptotics of the $\mathbb{E}(I_{n,k})$ when $k = \alpha_0 \log n + O(\sqrt{\log n})$, reflecting the structural change of the internal nodes from being asymptotically full to being of the same order as the number of external nodes.

In [13] we also deal with the variance of the profile. In particular, we derive asymptotic approximations to the variance of the profile, which asymptotically turns out to be of the same order as the expected value for all ranges of $k \geq 1$, namely, $\mathbb{V}(B_{n,k}) = \Theta(\mathbb{E}(B_{n,k}))$. In fact, we show that $\mathbb{V}(B_{n,k}) \sim \mathbb{E}(B_{n,k})$ in range (I), for range (IV) $\mathbb{V}(B_{n,k}) \sim 2\,\mathbb{E}(B_{n,k})$, while in range (II) (polynomial growth) the variance and the expected profile differ only by the oscillating functions. The variance of the internal profile behaves almost identically to the variance of the external profile; roughly, $\mathbb{V}(I_{n,k}) = \Theta(\mathbb{V}(B_{n,k}))$ for all k.

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

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

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

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

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

In passing, we should point out that recently Drmota and Szpankowski [2] extended the above analysis to the expected profile of digital search tree (see also [9]).

References

[1] R. de la Briandais, File searching using variable length keys, in *Proceedings of the AFIPS Spring Joint Computer Conference*. AFIPS Press, Reston, Va., (1959), pp. 295–298.

[2] M. Drmota and W. Szpankowski, The Expected Profile of Digital Search Trees *J. Combin. Theory, Ser. A*, 118, 1939-1965, 2011.

[3] P. Flajolet, X. Gourdon, and P. Dumas, Mellin transforms and asymptotics: harmonic sums, *Theoretical Computer Science* **144** (1995) 3–58.

[4] P. Flajolet and R. Sedgewick, Mellin transforms and asymptotics: finite differences and Rice's integrals, *Theoretical Computer Science* **144** (1995) 101–124.

[5] E. Fredkin, Trie memory, *Communications of the ACM*, **3** (1960) 490–499.

[6] P. Jacquet and M. Régnier, Trie partitioning process: limiting distributions, in *Lecture Notes in Computer Science*, **214** (1986) 196–210.

[7] P. Jacquet, and W. Szpankowski, Analysis of digital tries with Markovian dependency, *IEEE Transactions on Information Theory*, **37** (1991) 1470–1475.
[8] P. Jacquet and W. Szpankowski, Analytical depoissonization and its applications, *Theoretical Computer Science*, **201** (1998) 1–62.
[9] C. Knessl and W. Szpankowski, On the Average Profile of Symmetric Digital Search Trees, *Analytic Combinatorics*, 4, article #6, 2009.
[10] D. E. Knuth, *The Art of Computer Programming, Volume III: Sorting and Searching*, Second edition, Addison Wesley, Reading, MA, 1998.
[11] H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley & Sons, New York, 1992.
[12] G. Park, *Profile of Tries*, Ph.D. Thesis, Purdue University, 2006.
[13] G. Park, H-K. Hwang, P. Nicodeme, and W. Szpankowski, Profile of Tries, *SIAM J. Computing*, 38, 5, 1821-1880, 2009.
[14] W. Schachinger, Asymptotic normality of recursive algorithms via martingale difference arrays, *Discrete Mathematics and Theoretical Computer Science*, **4** (2001) 363–397.
[15] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001.

On depths and distances in random weighted b-ary trees

Ludger Rüschendorf
(joint work with G. Olaf Munsonius)

Weighted b-ary recursive trees are a combination of recursive trees as introduced in Szymański (1987) resp. of b-ary increasing trees as introduced in Bergeron, Flajolet, and Salvy (1992) with the random weighted b-ary trees, a continuous time tree model introduced in Broutin and Devroye (2006). The simplest description of the model is given by a recursive construction. If τ_n denotes the random weighted b-ary recursive tree with n nodes then an external node is chosen randomly with uniform distribution. This node is transformed into an internal node. It gets b external children which get their labels according to the appearance in the construction. Also a vector of random weights is attached to these children independent of the weights of the other nodes.

We establish that these trees belong to the class of well balanced $\log n$-trees. Central limit theorems are established for the depth D_n of the n-th node, for D_{U_n} the depth of a random node as well as for the distance Δ_{U_n,V_n} of two randomly chosen nodes. For the proof of this last result we establish that the (random) distance R_n of the least common ancestor of two randomly chosen nodes to the root is small in the sense that the sequence (R_n) is stochastically bounded. This allows to make use of the relationship

$$\Delta_{U_n,V_n} = D_{U_n} + D_{V_n} - 2R_n.$$

For the internal path length P_n and for the Wiener index W_n we give a recursive representation of the form

$$\begin{pmatrix} W_n \\ P_n \end{pmatrix} \stackrel{d}{=} \sum_{i=1}^{b} \begin{pmatrix} 1 & n - I_{n,i} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} W_{I_{n,i}}^{(i)} \\ P_{I_{n,i}}^{(i)} \end{pmatrix} + b_n$$

with subgroup sizes $I_{n,i}$, $(W_k^{(i)})$, $(P_k^{(i)})$ denoting independent copies and b_n a random toll term.

Based on this recursive structure we establish a second order asymptotic expansion for EW_n by means of Roura's (2001) Theorem. By means of the contraction method we obtain a central limit theorem

$$\left(\frac{W_n - EW_n}{n^2}, \frac{P_n - EP_n}{n}\right) \xrightarrow{d} (W, P),$$

where the limit (W, P) in the unique solution of a fixed point equation.

By a suitable choice of weighting vectors and embeddings these limit results can be extended to further classes of random trees including also models with unbounded degree. We give applications to the class of linear recursive trees which include the recursive tree and the plane oriented recursive tree (PORT) as special cases.

The distributional limit results for P_n and W_n obtained by this transference are new.

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, Springer (1992), 24–48.

[2] N. Broutin and L. Devroye, *Large deviations for the weighted height of an extended class of trees*, Algorithmica **46** (2006), 271–297.

[3] G. O. Munsonius, *Limit Theorems for Functionals of Recursive Trees*, PhD thesis, University of Freiburg (Germany) (2010).

[4] G. O. Munsonius and L. Rüschendorf, *Limit theorems for depths and distances in weighted random b-ary trees*, Preprint, University of Freiburg (Germany) (2010).

[5] R. Neininger and L. Rüschendorf, *A survey of multivariate aspects of the contraction method*, Discrete Math. Theor. Comput. Sci. **8** (2006), 31–56 (electronic).

[6] S. Roura, *Improved master theorems for dived-and-conquer recurrences*, J. ACM, **48** (2001), 170–205.

[7] J. Szymański, *On a nonuniform random recursive tree*, in: *Random graphs '85 (Poznań, 1985)*, volume **144** of North-Holland Math. Stud., North-Holland (Amsterdam), (1987), 297–306.

The Functional Equation of the Smoothing Transform

Gerold Alsmeyer

(joint work with John D. Biggins, Matthias Meiners)

Let $T = (T_j)_{j\geq 0}$ be a sequence of nonnegative random variables. Using T, a function f, defined on $\mathbb{R}$ or $\mathbb{R}_+$, can be transformed as follows:

$$f(t) \;\mapsto\; \mathbb{E}\prod_{k\geq 1} f(tT_k).$$

Call f a fixed point if

$$f(t) \;=\; \mathbb{E}\prod_{k\geq 1} f(tT_k). \tag{4}$$

Given a further random variable C and a possibly complex-valued function g, one may further consider the nonhomogeneous version of the above transform, viz.

$$f(t) \;\mapsto\; \mathbb{E}g(tC)\prod_{k\geq 1} f(tT_k),$$

and its fixed points satisfying

$$f(t) \;=\; \mathbb{E}g(tC)\prod_{k\geq 1} f(tT_k). \tag{5}$$

Here we are interested in fixed points f in the classes $\mathcal{L}$ of Laplace transforms, $\mathcal{F}$ of Fourier transforms, and $\mathcal{M}$ of survival functions of distributions on $\mathbb{R}_{\geq}$. The corresponding set of solutions (fixed points) are denoted $\mathcal{S}(\mathcal{L}), \mathcal{S}(\mathcal{F})$ and $\mathcal{S}(\mathcal{M})$, respectively. Note that $\mathcal{S}(\mathcal{L}) \subset \mathcal{S}(\mathcal{M})$. If X has Laplace (Fourier) transform f and $g(t) = e^{-t}$ $(= e^{it})$, then (5) in terms of random variables turns into the stochastic fixed point equation

$$X \;\stackrel{d}{=}\; \sum_{k\geq 1} T_k X_k + C, \tag{6}$$

where $X_1, X_2, ...$ are i.i.d. copies of X and independent of T, C and where $\stackrel{d}{=}$ means equality in distribution. If f is the survival function of X, i.e. $f(x) = \mathbb{P}(X \geq x)$, and $C = 0$, then (5) corresponds to the min-type equation

$$X \;\stackrel{d}{=}\; \inf_{k\geq 1:\, T_k>0} \frac{X_k}{T_k}, \tag{7}$$

where the infimum over the empty set is defined to be ∞. Examples of the above fixed point equations abound in the literature, for instance in the study of branching processes, random trees or divide and conquer algorithms.

We first look at the homogeneous case ($C = 0$) imposing the following conditions on T: For $\theta \geq 0$ define

$$m(\theta) \;:=\; \mathbb{E}\sum_{k\geq 1} T_k^{\theta}.$$

If α is the minimal positive real satisfying $m(\alpha) = 1$, then α will be called *characteristic exponent of* T. Now suppose that

$$m(0) \;=\; \mathbb{E}N \;>\; 1, \quad \text{where } N := \sum_{k\geq 1} \mathbf{1}_{\{T_k>0\}},$$

$$T \text{ has characteristic exponent } \alpha > 0,$$

$$\text{the closed subgroup generated by the positive } T_k \text{ is } \mathbb{R}^+,$$

and furthermore

$$m(\theta) < \infty \text{ for some } \theta < \alpha,$$

or

$$\mathbb{E}\sum_{i\geq 1} T_i^\alpha \log T_i \;\in\; (-\infty, 0)$$

$$\text{and} \qquad \mathbb{E}\left(\sum_{i\geq 1} T_i^\alpha\right)\log^+\left(\sum_{i\geq 1} T_i^\alpha\right) \;<\; \infty.$$

If one of the last two alternatives hold, we say that (A) is satisfied. Under these assumptions, the following two results settle the homogeneous case within class $\mathcal{M}$:

Theorem 1. *If (A) holds, there exists a unique (up to scaling) random variable W solving*

$$W \;\stackrel{d}{=}\; \sum_{k\geq 1} T_k^\alpha W_k, \tag{8}$$

such that all $f \in \mathcal{S}(\mathcal{M})$ are given by the family, parametrized by $h \in \mathbb{R}^+$,

$$f(t) \;=\; \mathbb{E}\exp(-Wht^\alpha).$$

Theorem 2 (Representation Theorem)**.** *Suppose that (A) holds. Then there exists a unique (up to a positive scaling factor) random variable W satisfying*

$$W \stackrel{d}{=} \sum_{i\geq 1} T_i^\alpha W_i$$

such that any disintegration M of a solution $f \in \mathcal{S}(\mathcal{M})$ has the following representation:

$$M(t) \;=\; \exp(-Wht^\alpha) \quad a.s. \quad (t>0). \tag{9}$$

In particular, any $f \in \mathcal{S}(\mathcal{M})$ is of the conjectured form.

Turning to two-sided solutions in the homogeneous case, which amounts to a study of Fourier transforms, we make the additional assumption that N is a.s. finite. Then the result for $\alpha \neq 1$ is as follows:

Theorem 3. *Suppose that (A) and $\alpha \in (0,2]\setminus\{1\}$ hold true. Then $\mathcal{S}(\mathcal{F})$, the set of two-sided solutions in terms of Fourier transforms, is given by*

$$\phi(t) \;=\; \begin{cases} \mathbb{E}\exp\left(-\sigma^\alpha W|t|^\alpha\left[1 - i\beta\frac{t}{|t|}\tan\left(\frac{\pi\alpha}{2}\right)\right]\right), & \text{if } \alpha \neq 2, \\ \mathbb{E}\exp(-\sigma^2 W t^2), & \text{if } \alpha = 2. \end{cases}$$

The range of the parameters is given by $\sigma > 0$, $\beta \in [-1,1]$ if $\alpha \neq 2$, and $\sigma > 0$ if $\alpha = 2$.

The case $\alpha = 1$ is more involved than the case $\alpha \neq 1$ due to a phenomenon called *endogeneity*, a notion coined by Aldous and Bandyopadhyay [1]. It means that a solution can be represented by a random variable that is a function of the weighted branching process. The random variable W appearing in the previous results, which is the unique endogeneous nonnegative solution (up to scaling), but

in order to rule out the existence of a second one within the class of real-valued variables we need the additional assumption

$$\text{(A+)} \qquad \mathbb{E}\sum_{j=1}^{N} T_j^{\alpha}(\log^{-} T_j)^2 \;<\; \infty.$$

Theorem 4. *Suppose that (A), (A+) and $\alpha = 1$ hold true. Then $\mathcal{S}(\mathcal{F})$ is given by the family*

$$\phi(t) \;=\; \mathbb{E}\exp\left(i\mu W t - \sigma W|t|\right), \tag{10}$$

where $\mu \in \mathbb{R}$, $\sigma \geq 0$ and $(\mu, \sigma) \neq (0, 0)$.

Finally regarding the non-homogeneous case ($C \neq 0$), we first have to state the following two conditions before stating our final result:

(C1) $m(1) < \infty$, $\mathbb{E}|C| < \infty$, and W_n^* is $\mathcal{L}^p$-bounded for some $p \geq 1$.

(C2) $m(\beta) < 1$ and $\mathbb{E}|C|^{\beta} < \infty$ for some $0 < \beta \leq 1$.

Theorem 5. *Suppose that (A) and one of* (C1) *or* (C2) *hold true. Additionally assume (A+) in the case $\alpha = 1$. Then there exists a coupling (W^*, W) of rv's such that W^* solves* (6)*, $W \geq 0$ solves* (8)*, and the Fourier transforms of solutions to* (6) *are*

$$\phi(t) \;=\; \begin{cases} \mathbb{E}\exp\left(iW^*t - \sigma^{\alpha}W|t|^{\alpha}\left[1 - i\beta\frac{t}{|t|}\tan\left(\frac{\pi\alpha}{2}\right)\right]\right), & \text{if } \alpha \notin \{1,2\}, \\ \mathbb{E}\exp\left(i(W^* + \mu W)t - \sigma W|t|\right), & \text{if } \alpha = 1, \\ \mathbb{E}\exp(iW^*t - \sigma^2 W\, t^2), & \text{if } \alpha = 2. \end{cases}$$

where $\sigma \geq 0$, $\beta \in [-1, 1]$ if $\alpha \notin \{1, 2\}$, $\mu \in \mathbb{R}$, $\sigma \geq 0$ if $\alpha = 1$, and $\sigma \geq 0$ if $\alpha = 2$.

Finally, we mention that there is a one-to-one correspondence between homogenous solutions and non-homogeneous ones which may be stated in terms of the disintegration of a solution. We refrain from giving details but mention that the result has been obtained under slightly stronger conditions by Rüschendorf [8].

A list of references mentioned in the talk is given below.

References

[1] Aldous, D.J., and Bandyopadhyay, A. *A survey of max-type recursive distributional equations.* Ann. Appl. Probab. **15** (2005), 1047–1110.

[2] Alsmeyer, G., Biggins, J.D., and Meiners, M. *The functional equation of the smoothing transform.* Preprint available at www.arxiv.org: 0906.3133v2

[3] Alsmeyer, G. and Meiners, M. *Fixed points of inhomogeneous smoothing transforms.* Preprint available at www.arxiv.org: 1007.4509v1

[4] Alsmeyer, G. and Meiners, M. *Fixed points of the smoothing transform: two-sided solutions.* Preprint available at www.arxiv.org: 1009.2412v1

[5] Caliebe, A. *Symmetric Fixed Points of a Smoothing Transformation.* Adv. Appl. Probab. **35** (2003), 377–394.
[6] Caliebe, A. *Representation of fixed points of a smoothing transformation.* Mathematics and computer science. III (2004), 311–324, Trends Math., Birkhäuser, Basel.
[7] Caliebe, A. and Rösler, U. *Fixed points with finite variance of a smoothing transformation.* Stochastic Process. Appl. **109** (2003), 105–129.
[8] Rüschendorf, L. *On stochastic recursive equations of sum and max type.* J. Appl. Probab. **43** (2006), 678-703.
[9] Spitzmann, J. *Lösungen inhomogener stochastischer Fixpunktgleichungen (Solutions of inhomogeneous fixed-point equations).* PhD Dissertation, Christian-Albrechts-Universität Kiel.

Towards the Variance of the Profile of Suffix Trees

Mark Daniel Ward
(joint work with Pierre Nicodème)

We consider randomly generated strings from which we (1) determine the profile of the analogous suffix tree, or (2) determine the subword complexity. A suffix tree is a retrieval tree (trie) built from the unique (occurring only once) prefixes of the suffixes of a string. E.g., if $S = 0101100111100001000111000\ldots$, and if we build a suffix tree from the first 12 strings of S, the 10th suffix has a unique prefix 11000, so it gets inserted as the leaf S_{10} in Figure 2. The suffix tree has "myriad" applications [1].

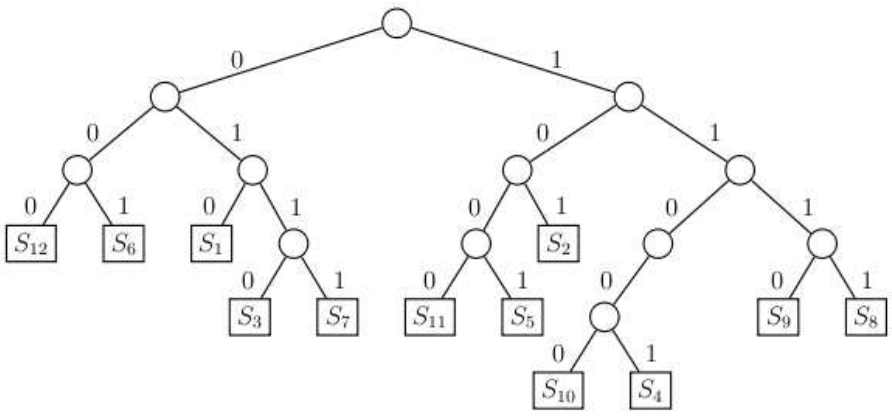


Figure 2. A suffix tree built from string $S = 0101100111100001000111000\ldots$

The (internal) **profile** of a suffix tree at level k is the number of (internal) nodes located on level k. Our goal is to make precise comparisons of the profile of a suffix tree versus the profile of a trie built over independent strings. When the underlying strings all derive from a Bernoulli source, a comparison of the average profile of a suffix tree versus the average profile of a trie built over independent strings was

made in [7]. Empirical evidence has been given, however, that the variance of the profile of a suffix tree at level k has asymptotically different behavior than the profile of a trie built over independent strings; see [5]. A recent, comprehensive study of the distribution of the profile of a trie built over independent strings appears in [6].

We use the following **notations**.

- $|S|_w$ is the number of occurrences of the word w in the string S.
- For a set of words $\mathcal{W}_n$ of cardinality n, we write

$$|\mathcal{W}_n|_w = |\{u \in \mathcal{W}_n;\ u = w\}|, \text{ the number of words of } \mathcal{W}_n \text{ equal to } w.$$

We generate strings randomly over an alphabet $\mathcal{A} = \{a, b\}$ according to a Bernoulli source. In other words, assuming that there are probabilities p and $q = 1 - p$ associated with letters a and b, the probability that a string of length n has exactly j occurrences of a is $\binom{n}{j} p^j q^{n-j}$.

Generating a random string S of length $n + k - 1$ and a set $\mathcal{T}_n$ of n random strings of length k, we consider the boolean indicators

- $I_{n,w}^{(d)} = 1$ if $|S|_w \geq d$ and $I_{n,w}^{(d)} = 0$ elsewhere,
- $J_{n,w}^{(d)} = 1$ if $|\mathcal{T}_n|_w \geq d$ and $J_{n,w}^{(d)} = 0$ elsewhere.

If a suffix tree is built from such a string S, then the profile $X_{n,k}^{(\text{prof})}$ of such a suffix tree is equal to the number of words of length k that occur two or more times as subwords in S. In other words, we observe

$$X_{n,k}^{(\text{prof})} = \sum_{w \in \mathcal{A}^k} I_{n,w}^{(2)},$$

where $\mathcal{A}^k$ is the collection of all binary words of length k.

Similarly, we define

$$Y_{n,k}^{(\text{prof})} = \sum_{w \in \mathcal{A}^k} J_{n,w}^{(2)},$$

which corresponds to the profile of a trie built upon n random strings of length k.

Then [7] proves $X_{n,k}^{(\text{prof})} - Y_{n,k}^{(\text{prof})} = O(n^{-\epsilon}\mu^k)$ for $\epsilon > 0$ and $\mu < 1$, but [5] gives empirical evidence that the variances are asymptotically different.

The k**th subword complexity** $X_{n,k}^{(\text{sub})}$ of S (of length $n+k-1$) is the number of distinct subwords of length k that occur at least once as a subword of S. We therefore have

$$X_{n,k}^{(\text{sub})} = \sum_{w \in \mathcal{A}^k} I_{n,w}^{(1)}.$$

Finally, we define

$$Y_{n,k}^{(\text{sub})} = \sum_{w \in \mathcal{A}^k} J_{n,w}^{(1)},$$

where the "sub" is just meant to remind us that $Y_{n,k}^{(\text{sub})}$ is defined similarly to the subword complexity $X_{n,k}^{(\text{sub})}$ above. Then [3] proves $X_{n,k}^{(\text{sub})} - Y_{n,k}^{(\text{sub})} = O(n^{-\epsilon}\mu^k)$

for $\epsilon > 0$ and $\mu < 1$, but empirical evidence (unpublished) also shows that the variances are asymptotically different.

The **correlation set** of a pair of words u, v (here, of the same length) is $\mathcal{C}_{u,v} = \{h \mid u.h = y.v,\ |y| < |u|\}$. The correlation polynomial is the relevant generating function. For example, $u = ababa$ and $v = abaab$ have correlation polynomial $C_{u,v}(z) = P(ab)z^2 + P(baab)z^4$. Previous approaches to problems of this nature use methods by Jacquet, Régnier, Szpankowski, and many others, tracing back to Goulden and Jackson, and Guibas and Odlyzko; here, we use the "cluster" approach that has been initially defined by Goulden and Jackson (see [2] for citations and recent discussion); we also do not consider the relevant complex analysis (this will follow in a longer treatment), but use the following intuitive approach: the primary results will be derived from noting that an autocorrelation polynomial is 1 plus much smaller terms, with high probability, and a correlation polynomial of two distinct words is 0 plus much smaller terms, with high probability; see [4]. Briefly, we have

$$\sum_{n\geq 0} E[Y_{n,k}^{(\mathrm{sub})}]z^n = \sum_{w\in\mathcal{A}^k} (1-(1-P(w)))z^n = \sum_{w\in\mathcal{A}^k} \frac{P(w)z}{(1-z)(1-(1-P(w))z)}.$$

To determine $\sum_{n\geq 0} E[X_{n,k}^{(\mathrm{sub})}]z^n$, we use the cluster approach. The probability generating function for the cluster of a word w is

$$\xi_w(z,t) = \frac{tP(w)z^{|w|}}{1-t(C_w(z)-1)}.$$

The probability generating function for the set of all words, with some of the w's distinguished, is $T_w(z,t) = 1/(1-z-\xi_w(z,t))$. Thus, the probability generating function for the set of words with no occurrences of w is

$$\frac{1}{1-z-\xi(z,-1)} = \frac{C_w(z)}{D_w(z)},$$

where $D_w(z) = (1-z)C_w(z) + P(w)z^{|w|}$. It follows that

$$\sum_{n\geq 0} E[X_{n,k}^{(\mathrm{sub})}]z^n = \sum_{w\in\mathcal{A}^k} \frac{P(w)z}{(1-z)D_w(z)}.$$

These results were first derived in [3], but the cluster approach allows a much more straightforward proof. Clusters allow quick verification of the probability generating functions from [7], and clusters allow the new derivation of the relevant probability generating functions for the variance of the profile of suffix trees and the variance of the subword complexity. MDW has derived several more results in this direction but did not have time to present these derivations during the relatively short talk at MFO. These results will be presented in a longer version of this paper in the near future, and we will complete the analysis using bootstrapping for complex-valued singularities and then using residue analysis.

References

[1] A. Apostolico. The myriad virtues of subword trees. In A. Apostolico and Z. Galil, editors, *Combinatorial Algorithms on Words*, pages 85–95, Berlin, 1985. Springer Verlag.

[2] F. Bassino, J. Clément, and P. Nicodème. Counting occurrences for a finite set of words: combinatorial methods. *ACM Transactions on Algorithms*. To appear.

[3] I. Gheorghiciuc and M. D. Ward. On correlation polynomials and subword complexity. *Discrete Mathematics and Theoretical Computer Science*, AH:1–18, 2007.

[4] P. Jacquet and W. Szpankowski. Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach. *Journal of Combinatorial Theory*, A66:237–269, 1994.

[5] P. Nicodème. q-gram analysis and urn models. *Discrete Mathematics and Theoretical Computer Science*, AC:243–258, 2003.

[6] G. Park, H.-K. Hwang, P. Nicodème, and W. Szpankowski. Profiles of tries. *SIAM Journal on Computing*, 38:1821–1880, 2009.

[7] M. D. Ward. The average profile of suffix trees. In *The Fourth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 183–193, 2007.

The Quicksort Process

Uwe Rösler

The sorting algorithm Quicksort, invented by Hoare '61, sorts a given list of n different reals. By now, we have a complete analysis of the running time, including the distribution and large deviation results. Is there an online version of Quicksort in the sense, that given the input of n different numbers, the online version provides first the smallest number, then the second smallest and so on during time. That is very easy to obtain, if we recall Quicksort every time for the list with the smallest numbers. But what about a limit as n tends to infinity as a process?

The answer to this question will be yes, the details will be given in a forthcoming paper by my PhD-student Mohammed Ragab. In this talk we discuss some related problems and technics via the Weighted Branching Process.

Let $X^n(l)$ denote the number of comparisons until the l-th smallest element appears for the online version of Quicksort. Mathematically we can describe the distribution of $X_n(l)$ recursively by

$$X^n(l) \stackrel{\mathcal{D}}{=} n - 1 + \mathbb{1}_{I^n \le l}(X_1^{I^n-1}(I^n - 1) + X_2^{n-I^n}(l - I^n)) + \mathbb{1}_{I^n > l} X_1^{I^n-1}(l)$$

Here I^n, X_i^k for $i = 1, 2$ and $0 \le k < n$ are independent. The distributions of X_1^k, X_2^k, X^k are the same and I^n has the uniform distribution on $\{1, 2, \ldots, n\}$.

By a result of Martínez, [1], the expectation $a^n(l) = E(X^n(l))$ can be explicitly calculated via the recursion and is

$$a^n(l) = 2n + 2(n+1)H_n - 2(n+3-l)H_{n+1-l} - 6l + 6$$

where H_n denotes the n-th harmonic number.

The natural normalization

$$Y^n\left(\frac{l}{n}\right) = \frac{X^n(l) - a^n(l)}{n+1}$$

provides the recursion

$$Y^n\left(\frac{l}{n}\right) \stackrel{\mathcal{D}}{=} \mathbb{1}_{I^n\le l}\left(\frac{I^n}{n+1}Y_1^{I^n-1}(1)+\frac{n-I^n+1}{n+1}Y_2^{n-I^n}\left(\frac{l-I^n}{n-I^n}\right)\right) + \mathbb{1}_{I^n>l}\frac{I^n}{n+1}Y_1^{I^n-1}\left(\frac{l}{I^n-1}\right)+C^n\left(\frac{l}{n}\right)$$

with C^n some toll term.

If Y^n converges as a process to some Y then $Y=(Y(t))_{t\in[0,1]}$ should satisfy the fixed point equation

$$Y \stackrel{\mathcal{D}}{=} \left(\mathbb{1}_{U\le t}\left(UY_1(1)+(1-U)Y_2\left(\frac{t-U}{1-U}\right)\right)+\mathbb{1}_{U>t}UY_1\left(\frac{t}{U}\right)+C(t)\right)_t$$

where $C=C(U)$ is a known function. The distribution of U is uniform.

The general approach as given in Knof and Rösler [2] for recursions

$$Y \stackrel{\mathcal{D}}{=} \sum_i A_iY_i\circ B_i + C$$

would not work here, since the contraction constant is 1.

We suggest to consider the following approach via the 'right' random variables defined via the Weighted Branching Process. Consider the binary tree $V=\{1,2\}^*$, edge weights $A_1^v, A_2^v, v\in V$ and vertex weights C_v as suggested above. Let L_v denote the path weight. Then the process

$$R_n=\sum_{|v|<n} L_n\circ C_v$$

converges to a limiting process R in terms of convergence of finite dimensional distributions. For example, $R_n(t)$ for fixed t is an L_2 martingale and converges a.e. to $R(t)$.

The work in progress and main forthcoming result is, the limit R satisfies the fixed point equation.

References

[1] Martínez, C. and Rösler, U. *Partial Quicksort and Quickpartitionsort.* DMTCS proc. **AM**, 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10) (2010), 505–512.

[2] Diether Knof and Uwe Rösler *The Analysis of Find or Perpetuities on Cadlag Functions.* DMTCS, accepted (2010).

On a Functional Contraction Method

Ralph Neininger
(joint work with Henning Sulzbach)

The contraction method was invented by Rösler [7] to derive a limit law for the normalized number of comparisons needed by the Quicksort algorithm to sort n randomly permuted numbers. The method is based on a recursive structure for the distributions of the random variables under consideration which often results from the recursive nature of the algorithms or from a decomposition of random trees into subtrees. In the last 20 years the method has been extended and applied to the asymptotic distributional analysis of numerous parameters, in particular in Rachev and Rüschendorf [6], Rösler [8], Neininger [3], and Neininger and Rüschendorf [4], where also many applications to random trees and recursive algorithms are discussed. Whereas all these studies consider quantities in $\mathbb{R}$ or $\mathbb{R}^n$, more recently, also systematic versions of the method on functional spaces have been developed: for quantities in separable Hilbert spaces in Drmota, Neininger and Janson [1], for quantities in the Banach space $L^p[0,1]$ of L^p-integrable functions on $[0,1]$ equipped with the L^p norm in Eickmeyer and Rüschendorf [2].

In this talk, based on Neininger and Sulzbach [5], the contraction method is developed for the Banach spaces $C[0,1]$ and $D[0,1]$ of continuous respectively cádlág functions each equipped with the supremum norm $\|\cdot\|_\infty$. It is exemplified at a short proof of Donsker's invariance principle. An algorithmic application to the probabilitic analysis of partial match queries is discussed in the subsequent talk by Henning Sulzbach.

Let $(V_j)_{j\geq 1}$ be a sequence of independent, identically distributed real random variables with $\mathbb{E}[V_1]=0$, $\mathrm{Var}(V_1)=1$ and $\mathbb{E}[|V_1|^{2+\varepsilon}]<\infty$ for some $\varepsilon>0$. We consider the normalized, linearly interpolated process $S^n=(S^n_t)_{t\in[0,1]}$ of the partial sums

$$S^n_t := \frac{1}{\sqrt{n}}\left(\sum_{j=1}^{\lfloor nt\rfloor} V_j + (nt-\lfloor nt\rfloor)V_{\lfloor nt\rfloor+1}\right), \quad t\in[0,1].$$

The idea in the context of the contraction method is that we have similar recursive decompositions for S^n as well as for Brownian motion: For $\beta>1$ we define operators

$$\varphi_\beta : C[0,1]\to C[0,1], \quad \varphi_\beta(f)(t)=\mathbf{1}_{\{t\leq 1/\beta\}}f(\beta t)+\mathbf{1}_{\{t>1/\beta\}}f(1),$$
$$\psi_\beta : C[0,1]\to C[0,1], \quad \psi_\beta(f)(t)=\mathbf{1}_{\{t\leq 1/\beta\}}f(0)+\mathbf{1}_{\{t>1/\beta\}}f\left(\tfrac{\beta t-1}{\beta-1}\right).$$

Both operators φ_β and ψ_β are linear, continuous and have operator norms $\|\varphi_\beta\|=\|\psi_\beta\|=1$. By construction we have for all $n\geq 2$,

$$S^n \stackrel{d}{=} \sqrt{\frac{\lceil n/2\rceil}{n}}\varphi_{\frac{n}{\lceil n/2\rceil}}\left(S^{\lceil n/2\rceil}\right)+\sqrt{\frac{\lfloor n/2\rfloor}{n}}\psi_{\frac{n}{\lceil n/2\rceil}}\left(\widehat{S}^{\lfloor n/2\rfloor}\right),$$

where $\stackrel{d}{=}$ denotes equality in distribution, $(S^j)_{j\geq 1}$ and $(\widehat{S}^j)_{j\geq 1}$ are independent and S^j and $\widehat{S}^j$ are identically distributed for all $j \geq 1$. Let $B = (B_t)_{t\in[0,1]}$ and $\widehat{B} = (\widehat{B}_t)_{t\in[0,1]}$ be independent standard Brownian motions. Properties of the Brownian motion imply

$$B \stackrel{d}{=} \sqrt{\frac{1}{\beta}}\varphi_\beta(B) + \sqrt{\frac{\beta-1}{\beta}}\psi_\beta(\widehat{B}),$$

for any $\beta > 1$. This implies that distances between $\mathcal{L}(S^n)$ and $\mathcal{L}(B)$ (more precisely a discretized version of B) can recursively been bounded.

For this we work with the Zolotarev distance. For an arbitrary Banach space $(\mathbb{B}, \|\cdot\|)$, $\mathcal{B}$ its Borel σ-algebra and $\mathcal{M}(\mathbb{B})$ the set of probability measures on $\mathcal{B}$ the Zolotarev metrics are defined as follows: For $s > 0$ fixed and $m := \lceil s \rceil - 1$, $\alpha := s - m$ we define

$$\mathcal{F}_s = \{f : \mathbb{B} \to \mathbb{R} : \|D^m f(x) - D^m f(y)\| \leq \|x - y\|^\alpha \ \forall\, x, y \in \mathbb{B}\},$$

where $D^m f$ denotes the m-th (Fréchet) derivative of f. For $\mu, \nu \in \mathcal{M}(\mathbb{B})$ the Zolotarev distance between μ and ν is defined by

$$\zeta_s(\mu, \nu) = \sup_{f\in\mathcal{F}_s} |\,\mathbb{E}\, f(X) - f(Y)|,$$

where X and Y are $\mathbb{B}$-valued random variables with $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$. Key issues for the contraction method to be developed in the Zolotarev metric on $\mathbb{B} = C[0,1]$ and $\mathbb{B} = D[0,1]$ with the uniform topology are

- finiteness of ζ_s on appropriate subspaces of $\mathcal{M}(C[0,1])$, $\mathcal{M}(D[0,1])$,
- completeness of ζ_s on appropriate subspaces of $\mathcal{M}(C[0,1])$, $\mathcal{M}(D[0,1])$,
- conditions under which convergence in ζ_s implies weak convergence on $\mathcal{M}(C[0,1])$, $\mathcal{M}(D[0,1])$,
- tightness criteria on these spaces in terms of the Zolotarev metric.

References

[1] Drmota, M., Janson, S. and Neininger, R. *A functional limit theorem for the profile of search trees.* Ann. Appl. Probab. **18** (2008), 288-333.

[2] Eickmeyer, K. and Rüschendorf, L. *A limit theorem for recursively defined processes in L^p.* Statistics and Decisions **25** (2007), 217-236.

[3] Neininger, R. *On a multivariate contraction method for random recursive structures with applications to quicksort.* Random Structures Algorithms **19** (2001), 498-524.

[4] Neininger, R. and Rüschendorf, L. *A general limit theorem for recursive algorithms and combinatorial structures.* Ann. Appl. Probab. **14** (2004), 378–418.

[5] Neininger, R. and Sulzbach, H. *On a functional contraction method*, in preparation.

[6] Rachev, S. T. and Rüschendorf, L. *Probability metrics and recursive algorithms.* Adv. in Appl. Probab. **27** (1995), 770–799.

[7] Rösler, U. *A limit theorem for "Quicksort".* RAIRO Inform. Théor. Appl. **25** (1991), 85–100.

[8] Rösler, U. *On the analysis of stochastic divide and conquer algorithms.* Algorithmica **29** (2001), 238–261.

A Process Convergence Result for Partial Match Queries in Random Quadtrees

Henning Sulzbach
(joint work with Nicolas Broutin and Ralph Neininger)

The quadtree is a data structure introduced by Finkel and Bentley [5] to store multidimensional data. For general references on multidimensional data structures and more details about their various applications, see the series of monographs by Samet [9, 10, 11]. For more information on the analysis of such tree data structure, we refer to [7, 4, 6].

The problem of partial match retrieval consists in reporting all the data with some specified values for some of their attributes. It is important for multidimensional databases which among others may arise in the management of geographical data and graphics algorithms.

In this paper, we focus on two-dimensional quadtrees. A quadtree is constructed by inserting data points into a tree data structure. For our model, we will assume the data to attain values in the unit square which is justified by the representation of elements by long binary strings. Consider a point sequence $p_1, p_2, \dots, p_n \in [0,1]^2$. As we build the tree, regions of the unit square are associated to the nodes where the points are stored. Initially, the root is associated with the region $[0,1]^2$ and the data structure is empty. The first point p_1 is stored at the root, and divides the unit square into four regions $Q_1, \dots, Q_4$. Each region is assigned to a child of the root. More generally, when i points have already been inserted, we have a set of $1 + 3^i$ (lower-level) regions that cover the unit square. The point p_{i+1} is stored in the node (say u) that corresponds to the region it falls in, divides it into four new regions that are assigned to the children of u. A partial match query for $(s, *), s \in [0,1]$, asks for all points whose first coordinate is s, where the second coordinate can be arbitrary. The complexity of the query is the number of nodes visited in the tree performing the partial search.

We are interested in the model of random quadtrees, where the data points are independent and uniformly distributed in the unit square. Let $C_n(s)$ denote the complexity of the partial match query for $(s, *)$ which coincides with the number of horizontal lines that insersect a vertical line at s in the unit square. It was conjectured in the 1970s that the order of $\mathbb{E}[C_n(U)]$, where the query U itself is uniform on $[0,1]$ and independent of the process, is $\sqrt{n}$. This was based on a approximation by a fully-balanced tree, i.e., all subtree sizes are concentrated around $n/4$. Examining singularities of generating functions Flajolet et. al [3] disprove this conjecture by showing

$$\mathbb{E}[C_n(U)] \sim \kappa\, n^{\beta} \qquad \text{where} \qquad \kappa = \frac{\Gamma(2\beta+2)}{2(\Gamma(\beta+1))^3}, \quad \beta = \frac{\sqrt{17}-3}{2},$$

where $\Gamma(x)$ denotes the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. They also give the order of growth in all dimensions. These results have since been strengthened

by Chern and Hwang [1], who provided the order of the error term together with the values of the leading constant in higher dimensions.

Recently, Curien and Joseph [2] were the first to give results for fixed $s \in [0,1]$. Using a continuous-time embedding they prove

$$\mathbb{E}[C_n(s)] \sim K(s(1-s))^{\beta/2} n^{\beta}, \qquad K = \frac{\kappa}{B\left(\frac{\beta}{2}+1, \frac{\beta}{2}+1\right)}. \tag{11}$$

Here, $B(a,b)$ denotes the Beta function $B(a,b) := \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Since subtrees behave independently of each other, given their sizes, the underlying structure of the problem proposes a recursive approach. For the cost within a subregion or subtree, what matters is the location of the query line *relative* to the region. Hence, decomposing the tree at the root with value (U,V) yields

$$\begin{aligned} C_n(s) \quad \stackrel{d}{=} \quad & 1 + \mathbf{1}_{\{s<U\}} \left[C^{(1)}_{I_1^{(n)}}\left(\frac{s}{U}\right) + C^{(2)}_{I_2^{(n)}}\left(\frac{s}{U}\right)\right] \\ & + \mathbf{1}_{\{s\geq U\}} \left[C^{(3)}_{I_3^{(n)}}\left(\frac{s-U}{1-U}\right) + C^{(4)}_{I_4^{(n)}}\left(\frac{s-U}{1-U}\right)\right], \end{aligned} \tag{12}$$

where $I_1^{(n)}, \ldots, I_4^{(n)}$ denote the sizes of the subtrees, i.e. the number of points falling in the four subregions. $(C_n^{(1)}), \ldots, (C_n^{(4)})$ are independent copies of (C_n), independent of $\left(U, V, I_1^{(n)}, \ldots, I_4^{(n)}\right)$. This does not imply a recurrence for the one-dimensional distributions expect for the case $s = 0$ which had already been studied in [3] and [2]. There, the number of nodes traversed, is of smaller order, $\mathbb{E}[C_n(0)] = \Theta\left(n^{\sqrt{2}-1}\right)$. However, $(C_n(s))_{s\in[0,1]}$ is a random stepfunction, hence the recurrence (12) can also be regarded in the space of cádlàg functions on the unit interval which is crucial for us. Then, if $n^{-\beta}C_n(s)$ converges to some random process $Z(s)$ uniformly, the limit is likely to satisfy

$$\begin{aligned} (Z(s))_{s\in[0,1]} \stackrel{d}{=} \Big(& \mathbf{1}_{\{s<U\}} \left[(UV)^{\beta} Z^{(1)}\left(\frac{s}{U}\right) + (U(1-V))^{\beta} Z^{(2)}\left(\frac{s}{U}\right)\right] \\ & + \mathbf{1}_{\{s\geq U\}} \left[((1-U)V)^{\beta} Z^{(3)}\left(\frac{s-U}{1-U}\right) \right. \\ & \left. \left. + ((1-U)(1-V))^{\beta} Z^{(4)}\left(\frac{s-U}{1-U}\right)\right]\right)_{s\in[0,1]}, \end{aligned} \tag{13}$$

where U and V are independent $[0,1]$-uniform random variables and $Z^{(i)}$, $i = 1, \ldots, 4$ are independent copies of the process Z, which are also independent of U and V. Our first result is

Theorem 1. *Subject to* $\mathbb{E}[Z(s)] = (s(1-s))^{\beta/2}$ *and* $\mathbb{E}[\|Z\|^2] < \infty$ *there exists a unique continuous solution of* (13).

The theorem is shown by constructing a sequence of random continuous functions satisfying a discrete recurrence approximating (13), that converges uniformly. The proof uses Chernoff-type concentration inequalities and tail bounds for the saturation level of random quadtrees.

Our main result is a functional limit law for $(C_n(s))$, accompanied by finer asymptotic properties of the one-dimensional marginals. In particular, it solves the open problems of the asymptotic variance and a distributional limit law for $C_n(U)$.

Theorem 2. *Let Z be as in Theorem 1. Then*

$$\left(\frac{C_n(s)}{Kn^\beta}\right)_{s\in[0,1]} \to (Z(s))_{s\in[0,1]}, \quad n\to\infty,$$

in distribution in $(\mathcal{D}[0,1], \|\cdot\|_\infty)$, the space of cádlág functions on the unit interval endowed with the supremum norm. Here K is defined in (11). For the marginals, we have

$$\frac{C_n(s)}{Kn^\beta} \xrightarrow{d,m} Z(s),$$

where $\xrightarrow{m}$ means convergence of all moments. If U is uniformly distributed on $[0,1]$, independent of (C_n) and Z, then

$$\frac{C_n(U)}{K_1 n^\beta} \xrightarrow{d,m} Z(U).$$

More precisely,

$$\mathbf{Var}[C_n(U)] \sim \bar{K} n^{2\beta},$$

with

$$\bar{K} = K^2\left(\frac{2(2\beta+1)}{3(1-\beta)}(B(\beta+1,\beta+1))^2 - \left(B\left(\frac{\beta}{2}+1,\frac{\beta}{2}+1\right)\right)^2\right) \approx 0.44736.$$

The proof of the result relies on the contraction method in Banach spaces, here $(\mathcal{D}[0,1], \|\cdot\|_\infty)$, as discussed in the previous talk by Ralph Neininger, see [8]. It is heavily based on a refinement of (11), towards a uniform polynomial rate of convergence.

As Svante Janson pointed out to us at the end of the talk, our method also implies that the distribution of $Z(s)$ is easily described by a single distribution on $\mathbb{R}^+$. More precisely, there exists a random variable $Z^* \geq 0$ such that for all $s \in [0,1]$ we have

$$Z(s) \stackrel{d}{=} Z^* \cdot (s(1-s))^{\beta/2}.$$

Analogous results can also be derived for random $2d$-trees which are closely related to quadtrees.

References

[1] Chern, H.H. and Hwang, H.K., *Partial match queries in random quadtrees*, SIAM Journal on Computing, **32** (2003), 904–915.

[2] Curien, N. and Joseph, A., *Partial match queries in two-dimensional quadtrees: a probabilistic approach*, Adv. in Appl. Probab. **43(1)** (2011), 178-194.

[3] Flajolet, P. and Gonnet, G. H. and Puech, C. and Robson, J. M., *Analytic variations on quadtrees*, Algorithmica **10** (1993), 473–500.

[4] Flajolet, P. and Sedgewick, R., *Analytic Combinatorics*, Cambridge University Press, (2009)

[5] Finkel R. A. and Bentley, J. L., *Quad trees, a data structure for retrieval on composite keys*, Acta Informatica **4** (1974), 1–19.
[6] Knuth, D. E., *The Art of Computer Programming: Sorting and Searching*, Addison-Wesley, (1998)
[7] Mahmoud, H., *Evolution of Random Search Trees*, Wiley, (1992)
[8] Neininger, R. and Sulzbach, H. *On a functional contraction method*, in preparation.
[9] Samet, H., *The Design and Analysis of Spatial Data Structures*, Addison-Wesley, (1990)
[10] Samet, H., *Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS*, Addison-Wesley, (1990)
[11] Samet, H., *Foundations of multidimensional and metric data structures*, Morgan Kaufmann, (2006)

Metric aspects of binary search trees

Rudolf Grübel

Binary trees

Let $\mathcal{V} := \{0,1\}^\star$ be the set of all finite sequences of 0's and 1's. By a *binary tree* we mean a prefix-stable subset x of $\mathcal{V}$. For $u = (u_1, \dots, u_k) \in \mathcal{V}$ let $\tilde{u} := (u_1, \dots, u_{k-1})$ if $k \geq 1$ and $u0 := (u_1, \dots, u_k, 0)$, $u1 := (u_1, \dots, u_k, 1)$ be its direct ancestor and its left and right direct descendant; $|u| := k$. Let $\mathbb{B}$ be the set of finite binary trees x, $\mathbb{B}_n$ those of size $|x| = n$. We write $\sigma(x,u)$ for the number of descendants of u in x, including u.

Our basic objects are Markov chains with state space $\mathbb{B}$: For the *BST chain* $X = (X_n)_{n\in\mathbb{N}}$ the next value X_{n+1}, given $X_n = x$, is uniformly distributed on those $y \in \mathbb{B}$ with $x \subset y$ and $|y| = |x| + 1$. For the *DST chain* $X^\mu = (X_n^\mu)_{n\in\mathbb{N}}$ driven by a measure μ on $\{0,1\}^\infty$ the transition from x to y, where again $|y| = |x| + 1$, happens with probability $\mu(A_{ui})$ if $u \in x$ and $ui \in y \setminus x$; here A_u is the set of 0-1 sequences that have u as a prefix. We always start with $x = \{\emptyset\}$ at time $n = 1$.

These are transient Markov chains that have the space-time property, meaning that the nth variable has all its values in $\mathbb{B}_n$. For these, discrete potential theory can be applied to obtain a strong law that is in a certain sense optimal. Convergences are understood to hold with probability 1 and refer to $n \to \infty$.

Theorem 1 ([2])**.** (a) *$X_n \to X_\infty$, a random measure on $\{0,1\}^\infty$, in the sense of*

$$\frac{1}{n}\,\sigma(X_n, u) \to X_\infty(A_u) \quad \textit{for all } u \in \mathcal{V}.$$

(b) *X_∞ generates the tail σ-field associated with X.*
(c) *The conditional distribution of X, given $X_\infty = \mu$, is the same as the distribution of the DST chain X^μ driven by μ.*

Moreover, an explicit expression for the distribution of X_∞ is available.

We sketch a proof of part (a) that makes use of the binary search tree algorithm and that also displays X_∞ as a function of its input sequence. (This algorithm, and the digital search tree algorithm, are responsible for the acronyms.)

Metric aspects

A metric d on a tree x is specified by the edge lengths $d(\tilde{u}, u)$, $u \in \mathcal{V}$, $u \neq \emptyset$. For simply generated trees the canonical tree distance, with $d(\tilde{u}, u) \equiv 1$, leads to a theory that is one of the highlights of modern probability; see e.g. [1] (the situation here turns out to be technically simpler). Metric trees (x, d) can be rescaled in the sense that αx refers to $(x, \alpha d)$. Theorem 1 suggests the use of a tree dependent metric that is based on subtree sizes: For $\rho > 0$ let

$$d_{x,\rho}(\tilde{u}, u) := \rho^{|u|}\sigma(x, u) \quad \text{for all } u \in \mathcal{V},\ u \neq \emptyset.$$

Theorem 1 then implies pointwise convergence of $n^{-1}X_n$ to X_∞, where the metric on the infinite limit tree is given by

$$d_{X_\infty,\rho}(\tilde{u}, u) := \rho^{|u|}X_\infty(A_u) \quad \text{for all } u \in \mathcal{V},\ u \neq \emptyset.$$

The next result implies that there is a transition where the limit space changes from being totally bounded to having infinite diameter.

Theorem 2. *Let $\rho_0 = 1.2617\ldots$ be the smaller root of the equation $2e\log(\rho) = \rho$. Then, with probability 1, (X_∞, d_{X_∞}) is compact for $\rho < \rho_0$. Further, for $\rho > \rho_0$,*

$$\sup\{d_{X_\infty,\rho}(\tilde{u}, u) :\ u \in \mathcal{V},\ u \neq \emptyset\} \ = \ \infty.$$

Applications

Over the years, many tree functionals have been studied, by different authors and with different techniques. For binary search trees perhaps the best known strong limit theorem refers to the *internal path length* $\Psi(x) := \sum_{u\in x} |u|$, where Régnier [4] showed that $Z_n := n^{-1}\Psi(X_n) - 2\log n$ converges almost surely (and in quadratic mean) to a limit variable Z_∞. In view of part (b) of Theorem 1 this limit must be a function of X_∞. Indeed, it turns out that

$$Z_\infty \ = \ \sum_{u\in\mathcal{V}} X_\infty(A_u)\, C\!\left(\frac{X_\infty(A_{u0})}{X_\infty(A_u)}\right) \quad \text{almost surely,}$$

with $C(s) = 1 + 2s\log(s) + 2(1-s)\log(1-s)$. Convergence on the level of metric trees with a suitable topology suggests an approach that puts such results into a general framework.

Of special interest are functionals that map a combinatorial structure x to a function $f(x, \cdot) : [0, 1] \to \mathbb{R}$, such as the Harris correspondence for simply generated trees. In [3] such a function was obtained by identifying $t \in [0, 1]$ with its binary expansion $(b_j)_{j\in\mathbb{N}} \in \{0, 1\}^\infty$ (in what follows, the binary rationals $\mathbb{Q}_b$ do not matter) and letting $f(x, t)$ be the depth of the first node along t that is outside of x. Replacing the canonical tree distance that is inherent in this definition by the subtree size metric $d_x = d_{x,1}$ we arrive at the *metric silhouette*,

$$f(x, t) \ := \ \sum_{k=1}^{\infty} \sigma(x, u(t, k)),$$

where $u(t,k) = (b_1, \dots, b_k)$ if $t = \sum_{j=1}^{\infty} b_j 2^{-j}$. For $s, t \in [0,1]$ let $d_0(s,t) := 2^{-k}$ where k denotes the length of the common prefix in the binary expansions of s and t. For tree nodes u and v the value k is the length of the last common ancestor of u and v. Again, convergence means almost sure convergence and refers to $n \to \infty$.

Theorem 3. *For all $t \in [0,1] \setminus \mathbb{Q}_b$,*

$$Y_n(t) := f(n^{-1}X_n, t) \;\to\; Y_\infty(t) = \int \big(-\log d_0(s,t)\big)\, X_\infty(ds).$$

Thus, the limit process Y_∞ for the metric silhouette Y_n of the BST chain is the logarithmic potential of the random limit measure X_∞ on the compact metric space $(\{0,1\}^\infty, d_0)$.

References

[1] S. Evans, *Probability and real trees*. Lecture Notes in Mathematics **1920** (2008), Springer, Berlin.

[2] S. Evans, R. Grübel, A. Wakolbinger, *Trickle-down processes and their boundaries* (2010), submitted.

[3] R. Grübel, *On the silhouette of binary search trees*, Ann. Appl. Probab. **19** (2009), 1781–1802.

[4] M. Régnier, *A limiting distribution for quicksort*, RAIRO Inform. Théor. Appl. **23** (1989), 335–343.

Behaviour of tree-based contention algorithms

Nicolas Broutin

(joint work with C. Holmgren)

Introduction

Consider the following general model of communication using a single broadcast channel (e.g., cable, radio, satellite channel, internet, mobile networks etc.) shared by many users (or sources). Suppose that the channel is in *free access*: every user transmits on the channel as soon as it has a message to send. When a source sends a message, it is picked up by the destination unless some other source also attempted to send its message, in which case the messages are corrupted and need to be resent. We assume that every message sent without interference reaches its destination, and that the corresponding source then quits the system. A strategy to resolve the collisions and (try to) ensure that each source eventually sends its message successfully is called a protocol.

Here, we are interested in a specific algorithm designed by [1] and [5] based on the divide-and-conquer paradigm: Using coin flips, any set of sources that collide is split into two subgroups; the users of the first group immediately try to send their message again, while those in the second group wait for the entire first group to be fully resolved. This strategy is used recursively to resolve the two groups, when their time has come.

A tree representation

The recursive splitting that underlies the protocol yields an elegant representation of the execution by a tree (For the sake of space, I will not give to much detail). The nodes in the tree represent time slots, and every node v carries the number of sources N_v that are allowed to emit during the corresponding time slot (sources at level 0). From now on, we identify nodes and time slots. If $N_v \geq 2$, the messages interfere and the sources are split into two subgroups that are associated with the children v_1 and v_2 of v. In general, the number of sources N_{v_1} and N_{v_2} that try to transmit during v_1 and v_2 are given by

$$(N_{v_1}, N_{v_2}) = \mathrm{Mult}(N_v; p, 1-p) + (A_{v_1}, A_{v_2}) \tag{14}$$

where A_{v_1} and A_{v_2} are the random numbers of additional sources that joined the system during time slots immediately before v_1 and v_2. Since priority is given to the first group, the dates (or time slots) associated to the nodes is recovered using a depth-first traversal of the tree. A branch of the tree is killed when the value of a node drops below one (successful transmission).

In general, we can consider splitting the sources allowed to transmit into b groups, according a the proportions given by a random split vector $(V_1, \ldots, V_b)$, $V_i \geq 0$, $V_1 + V_2 + \cdots + V_b = 1$. One is then lead to study the b-ary tree with the splitting rule:

$$(N_{v_1}, N_{v_2}, \ldots, N_{v_b}) = \mathrm{Mult}(N_v; V_1, V_2, \ldots, V_b) + (A_{v_1}, A_{v_2}, \ldots, A_{v_b}).$$

The branches are still killed when the value of the node drops below one (one can generalize to channel with capacity $s \geq 1$). In the following, we write T^n for an instance of this killed branching Markov chain, started from the value n at the root. Quite naturally, the stability of the protocol is related to the finiteness of the trees T^n, for $n \geq 1$.

Stability of the algorithm and conditioned Markov chains

Let (V_i, A_i) be i.i.d. copies of (V, A), where V is the distribution of a uniform random component of $(V_1, \ldots, V_b)$. Along any branch of the tree, one sees a Markov chain with transitions given by $N_{i+1} = \mathrm{Bin}(N_i; V_i) + A_i$, It is possible to make a strong connection between the stability of the algorithm and the long term behaviour of the Markov chain seen along a branch, conditioned on non-absorption (when it has a value less than two).

Under some mild conditions that are satisfied here provided $\mathbf{P}(V = 0) = 0$, one can show that there exists a random variable D that is quasi-stationary for the conditioned process. Here we mean that conditioned on $\mathrm{Bin}(D; V) + A > 1$, $\mathrm{Bin}(D; V) + A$ is distributed as D. Then, by definition of the distribution D, the tree T^D is a Galton–Watson tree. It is easily seen that D must charge all the natural numbers $i \geq 2$, so that the stability finiteness of the tree T^n corresponds to subcriticality for the Galton–Watson process. Writing $\rho^{-1} = \mathbf{P}(\mathrm{Bin}(D; V) + A > 1)$, the process is stable ($|T^n| < \infty$ a.s. for every n) precisely when $\rho^{-1} b \leq 1$.

Fix the distribution for V. Consider the distribution for D as a function (if there is only one such distribution) of that of A. By proving the uniqueness and the continuity of the function $D = D(A)$, as $A \to 0$ in probability, extend the result of [4], showing that the stability region is never empty:

Theorem 1. *Let $V \in [0,1]$ be a random variable such that $\mathbf{E}[V] = 1/b$ and $V > 0$ almost surely. There exists $\epsilon > 0$ (depending only on V) such that if $\mathbf{P}(A > 0) < \epsilon$ then $\mathbf{E}|T^n| < \infty$ for all $n \in \mathbb{N}$. In particular by Markov's inequality, if $\mathbf{E}[A] < \epsilon$ then the protocol is stable.*

Unfortunately, the result is based on a continuity argument, and does not give access to estimates for ϵ. To go further, we consider the more specific case of fair splits. Note that the result gives access to a universal stability region (given the split), regardless of the arrival distribution.

The case of fair splits

In the special case when $(V_1, \dots, V_b) = (1/b, \dots, 1/b)$, we can go further and identify the quasi-stationary distribution at the point that is critical for the stability. This allows us to pin down the precise value of the stability threshold as the root of certain series equations. We can express the stability threshold in terms of $a(z) = \mathbf{E}[z^A]$.

Theorem 2. *Assume that $V = 1/b$, $s = 1$ and that $\mathbf{E}[A^2] < \infty$. Then, the system is critical if and only if*

$$1 = \sum_{i\geq 0} b^{-i} \frac{a'(1-b^{-i})}{a(1-b^{-i})} + b \sum_{i\geq 0} b^i \prod_{j=0}^{i-1} a(1-b^{-j}) \Bigg[1 - a(1-b^{-i}) - b^{-i} a'(1-b^{-i}) + \left\{ a(1-b^{-i}) - 1 + b^{-i}\mathbf{E}A \right\} \sum_{k\geq i} b^{-k} \frac{a'(1-b^{-k})}{a(1-b^{-k})} \Bigg].$$

This representation is more than just theoretical, since it permits to compute effectively the stability threshold in concrete examples. For instance, we can recover the results of [3] and [2]

Corollary 1. *Suppose that the immigration A is Poisson(λ) and that $V = 1/b$ almost surely. Then, the process is stable if and only if $\lambda < \lambda_c$, where λ_c is the smallest positive root of*

$$1 + \frac{b(b-1)e^{-\frac{b\lambda}{b-1}}}{b-1-b\lambda} \sum_{i\geq 0} b^i e^{\frac{b\lambda b^{-i}}{b-1}} \left[e^{-\lambda b^{-i}} \left(1 - \frac{\lambda b^{-i}}{b-1} \right) - 1 + b \frac{\lambda b^{-i}}{b-1} \left(1 - \frac{\lambda}{b^i} \right) \right] = 0.$$

We can also obtain many more examples (previous results were only about Poisson arrivals). For instance:

Corollary 2. *Suppose that $V = 1/b$ almost surely and that A is Bernoulli(p). Then, the process is stable if and only if $p < p_c$, where p_c is the root of*

$$1 = \sum_{i\geq 0} \frac{pb^{-i}}{1-pb^{-i}}.$$

References

[1] J. Capetanakis. Tree algorithms for packet broadcast channels. *IEEE Transactions on Information Theory*, 25(5):505–515, 1979.

[2] G. Fayolle, P. Flajolet, and M. Hofri. On a functional equation arising in the analysis of a protocol for multi-access broadcast channel. *Advanced in Applied Probability*, 18:441–472, 1986.

[3] P. Mathys and P. Flajolet. Q-ary collision resolution algorithms in random-access systems with free or blocked channel access. *IEEE Transaction on Information Theory*, 31:217–243, 1985.

[4] H. Mohamed and P. Robert. Dynamic tree algorithms. *The Annals of Applied Probability*, 20:26–51, 2010.

[5] B.S. Tsybakov and V.A. Mikhailov. Free synchronous packet access in a broadcast channel with feedback. *Problemy Peredachi Informatsii*, 14:32–59, 1978.

Random geometric graphs in high dimension

Gábor Lugosi

(joint work with Luc Devroye, András György, Frederic Udina)

Motivated by a statistical hypothesis testing problem of detecting small correlations in Gaussian data, we introduce a model of random geometric graphs on high-dimensional spheres. We show that as the dimension grows, the graph becomes similar to and Erdős-Rényi random graph. We pay particular attention to the clique numbers of such graphs and show that the size of the dimension plays an important role in their behavior. In particular, we show that the clique number is very close to that of the corresponding Erdős-Rényi graph when the dimension is larger than $\log^3 n$ where n is the number of vertices.

Reporter: Ralph Neininger

Participants

Prof. Dr. Louigi Addario-Berry
Dept. of Mathematics and Statistics
McGill University
Burnside Hall
805 Sherbrooke Street West
Montreal QC, H3A 2K6
CANADA

Prof. Dr. Gerold Alsmeyer
Institut f. Mathematische Statistik
Universität Münster
Einsteinstr. 62
48149 Münster

Dr. Yuliy Baryshnikov
AT & T Bell Laboratories
P.O. Box 636
600 Mountain Avenue
Murray Hill , NJ 07974-0636
USA

Dr. Nicolas Broutin
INRIA Rocquencourt
Domaine de Voluceau
B. P. 105
F-78153 Le Chesnay Cedex

Prof. Dr. Luc Devroye
School of Computer Science
McGill University
Montreal Quebec H3A 2K6
CANADA

Dr. Christina Goldschmidt
Department of Statistics
University of Warwick
GB-Coventry CV4 7AL

Prof. Dr. Rudolf Grübel
Institut f. Mathematische Stochastik
Universität Hannover
Welfengarten 1
30167 Hannover

Prof. Dr. Hsien-Kuei Hwang
Institute of Statistical Science
Academia Sinica
Taipei , 11529
TAIWAN

Prof. Dr. Svante Janson
Matematiska Institutionen
Uppsala Universitet
Box 480
S-751 06 Uppsala

Prof. Dr. Gerhard Kramer
Lehrstuhl für Nachrichtentechnik
Technische Universität München
80290 München

Prof. Dr. Gabor Lugosi
Department of Economics
Pompeu Fabra University
Ramon Trias Fargas 25-27
E-08005 Barcelona

Prof. Dr. Ralph Neininger
Institut für Mathematik
J.W.Goethe-Universität
Robert-Mayer-Str. 10
60054 Frankfurt

Prof. Dr. Uwe Rösler
Mathematisches Seminar
Christian-Albrechts-Universität Kiel
Ludewig-Meyn-Str. 4
24118 Kiel

Prof. Dr. Ludger Rüschendorf
Institut f. Mathematische Stochastik
Universität Freiburg
Eckerstr. 1
79104 Freiburg

Henning Sulzbach
Institut für Mathematik
J.W.Goethe-Universität
Robert-Mayer-Str. 10
60054 Frankfurt

Prof. Dr. Wojciech Szpankowski
Department of Computer Sciences
Computer Science Building
Purdue University
West Lafayette , IN 47907-1398
USA

Prof. Dr. Mark Daniel Ward
Department of Statistics
Purdue University
150 North University Street
West Lafayette , IN 47907-2067
USA