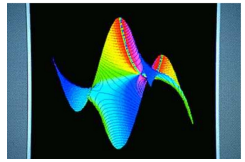


What do we learn from the analysis of hashing algorithms?

(guided tour by Philippe Flajolet to the mysteries of the sky, with the help of hashing).

Alfredo Viola

Universidad de la República, Uruguay

December 16th, 2011

Philippe Flajolet and Analytic Combinatorics

Conference in the memory of Philippe Flajolet

Other potential titles for this talk.

- Non-trivial exercises to illustrate the use of general methodological tools to find universal laws.
- Solution to hashing problems to be "solved" in future published papers.
- How to survive in a world full of (very few!) universal laws!: "the hashing experience, what is left to us to solve!".

Personal approach

- To give just a "taste" of Philippe Flajolet's approach to Analytic Combinatorics and motivate to take (or continue!) our personal discovery experience.
- Today, the main issue are not the specific results. This talk is about ideas, "personal experience" and "feelings"!
- Hope to reach a general audience. Philippe is a specialist in achieving this goal!
- Try to present Philippe Flajolet's own views, through the analysis of hashing problems.

The fingerprint of every Philippe's paper.

Our initial motivation when starting this project was to build a coherent set of methods useful in the analysis of algorithms, a domain of computer science now well-developed and presented in books by Knuth, Hofri, Mahmoud, and Szpankowski, in the survey by Vitter–Flajolet, as well as in our earlier *Introduction to the Analysis of Algorithms* published in 1996. This book, *Analytic Combinatorics*, can then be used as a systematic presentation of methods that have proved immensely useful in this area; see in particular the *Art of Computer Programming* by Knuth for background.

THIS BOOK is meant to be reader-friendly. Each major method is abundantly illustrated by means of concrete *Examples*[1] treated in detail—there are scores of them, spanning from a fraction of a page to several pages—offering a complete treatment of a specific problem. These are borrowed not only from combinatorics itself but also

Hashing and methodology in "Analytic Combinatorics".

Saddle Point and Hashing

mean, itself asymptotic to $\log n / \log\log n$. In addition, the saddle-point method may be used instead of crude bounds. These results, in the context of longest probe sequences in hashing, were obtained by Gonnet [301] under the Poisson model. Many key estimates regarding random allocations (including capacity) are to be found in the book by Kolchin *et al.* [388]. Analyses of this type are also useful in evaluating various dynamic hashing algorithms by means of saddle-point methods [217, 504]. ..... ■

Moment Pumping, Parking Problem and Hashing

▷ **VII.54.** *A parking problem II.* This continues Example II.19, p. 146. Consider m cars and condition by the fact that everybody eventually finds a parking space and the last space remains empty. Define *total displacement* as the sum of the distances (over all cars) between the initially intended parking location and the first available space. The analysis reduces to the difference-differential equation [249, 380], which generalizes (65), p. 146,

$$\frac{\partial}{\partial z} F(z,q) = F(z,q) \cdot \frac{F(z,q) - qF(qz,q)}{1-q}.$$

Moment pumping is applicable [249]: the limit distribution is once more an Airy (of area type). This problem arises in the analysis of the *linear probing hashing* algorithm [380, §6.4] and is of relevance as a discrete version of important coalescence models. It is also shown in [249] based on [285] that the number of inversions in a Cayley tree is asymptotically Airy. ◁

Hashing and other combinatorial problems.

Random Allocation and Hashing

II.3.2. Applications to words and random allocations. Numerous enumeration problems present themselves when analysing statistics on letters in words. They find applications in the study of *random allocations* [388] and the design of *hashing algorithms* in computer science [378, 538]. Fix an alphabet

▷ **III.11.** *Hashing and random allocations.* Random allocations of balls into bins are central in the understanding of a class of important algorithms of computer science known as *hashing* [378, 537, 538, 598]: given a universe $\mathcal{U}$ of data, set up a function (called a hashing func-

Birthday Paradox, Coupon Collector and Hashing

itself but also in probability and statistics. In particular, labelled constructions of words provide an elegant solution to two classical problems, the birthday problem and the coupon collector problem, as well as several of their variants that have numerous applications in other fields, including the analysis of hashing algorithms in computer science.

1 Introduction

Address calculation methods *may* provide direct access to data.

$$\boxed{\text{SOURCE}} \Longrightarrow \boxed{\text{ADDRESSES}}$$
$$x \mapsto h(x) = \lfloor m \cdot x \rfloor$$

— Place key x at location $h(x)$

— Resolve collisions

- by chaining
- by "linear probing", or other methods.

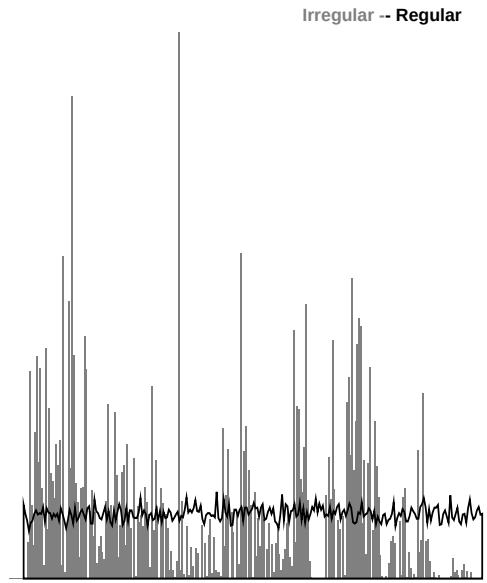
Works well for constrained data or random uniform data.

Example: Apple II BASIC: Use 2 letter identifiers and index in 26×26 table.

Natural data are often *not* uniformly random.

Reduction to randomness via

a HASHING FUNCTION $h(x) \ : \ \mathcal{U} \mapsto [0\,..\,m-1]$
(previously $h(x) = \lfloor m \cdot x \rfloor$)



Arithmetic hashing: take $h(x) = \texttt{value}\,(x)$ mod m

Others: binary, Xor, etc, etc.

Vincent Lum et alii, 1979: works well on real files (DB).

The distribution of natural digrams in a dictionary of the English language with about 25,000 words (in gray), and the smoothing that results form hashing (in black).

2 Random Allocations

Collisions are barely avoidable, distrib. is irregular.

m = # Urns,	n = # Balls

Theorem 0.

(*i*) *Collisions occur early* = *Birthday Paradox*

$$Ex\{\text{First collision } /m \text{ cells}\} \sim \sqrt{\frac{\pi m}{2}}$$

(*ii*) *Probability of no collisions in a full table is*

$$\Pr\{\text{No collision } n = m\} = \frac{n!}{n^n} \sim \frac{e^{-n}}{\sqrt{2\pi n}}.$$

(*iii*) *Empty cells disappear late* = *Coupon Collector*

$$Ex\{\text{All } m \text{ cells non-empty}\} \sim m \cdot \log m.$$

(*iv*) *Even in an* α*-sparse allocation,* $\alpha = \frac{n}{m}$

$$Ex\{\text{Max bucket occupancy}\} \sim \frac{\log n}{\log \log n}$$

Collision management is a necessity

The Poisson Law governs balls-in-urns models

Thow n balls into m buckets. Let

$$\boxed{\alpha = \frac{n}{m}}$$

be fixed, $0 < \alpha < \infty$. Then, asymptotically $(m, n \to +\infty)$

- the proportion of empty urns is $e^{-\alpha}$, 36% for $\alpha = 1$;
- the proportion of k-urns is a Poisson law of param. α

$$\text{Poisson}(\alpha, k) := e^{-\alpha} \frac{\alpha^k}{k\,!}.$$

Proof.

$$\begin{array}{lll} \text{empty urns} & m \times \left(\dfrac{m-1}{m}\right)^n & \sim m \times e^{-\alpha}. \\ k\text{–urns:} & m \times \dbinom{n}{k} \cdot \dfrac{(m-1)^{n-k}}{m^n} & \sim m \times e^{-\alpha} \dfrac{\alpha^k}{k!} \end{array}$$

$\Longrightarrow$ Analysis of separate chaining:

$$Ex\{C^-(m,n)\} \sim 1 + \sum_{k=1}^{\infty} k \cdot e^{-\alpha} \frac{\alpha^k}{k!} \equiv \boxed{1+\alpha}.$$

$$Ex\{C^+(m,n)\} \sim \boxed{1 + \tfrac{\alpha}{2}}$$

Trie searching, dynamic hashing and extendible hashing

A BRANCHING PROCESS ARISING IN DYNAMIC HASHING,
TRIE SEARCHING AND POLYNOMIAL FACTORIZATION

Philippe FLAJOLET
INRIA
78150 - Rocquencourt
(France)

Jean-Marc STEYAERT
Ecole Polytechnique
91128 - Palaiseau
(France)

On the Performance Evaluation of Extendible Hashing and Trie Searching

Philippe Flajolet*
INRIA, Rocquencourt BP105, F-78153 Le Chesnay Cedex, France

Abstract.

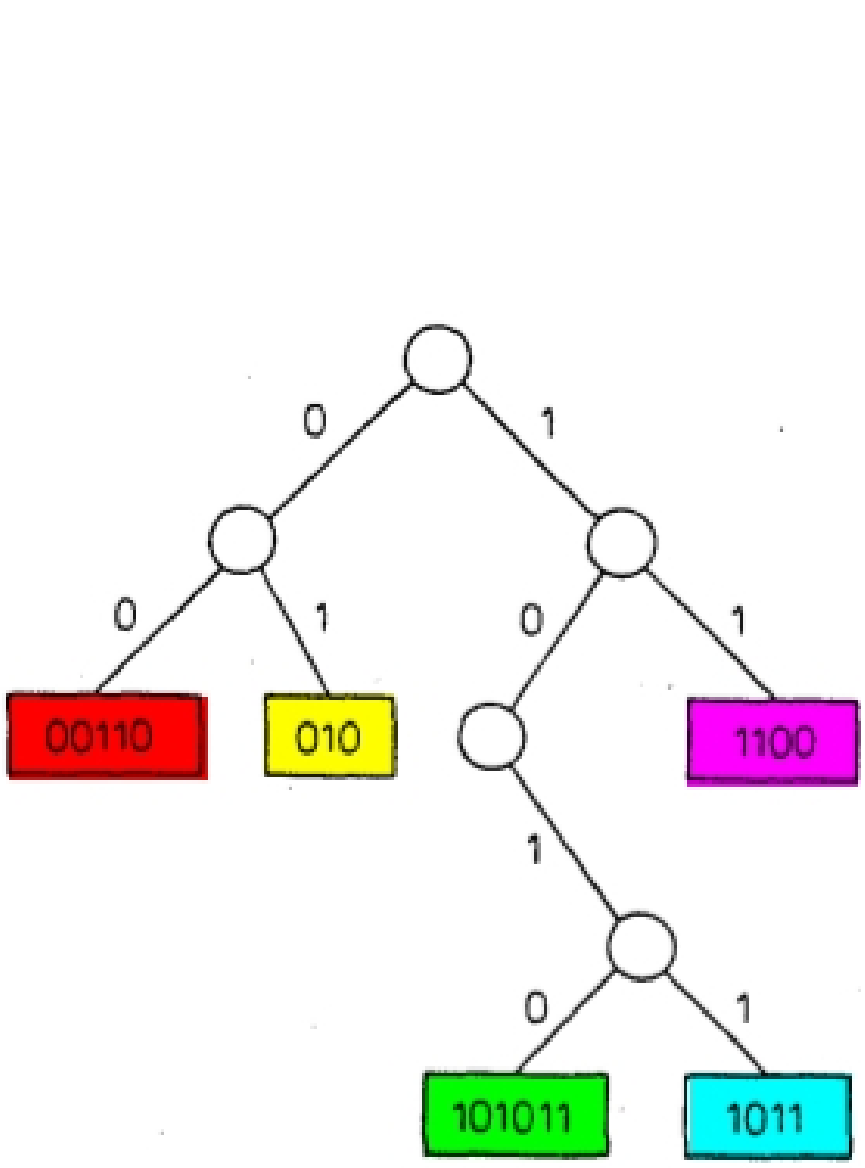
On the Performance Evaluation of Extendible Hashing and Trie Searching

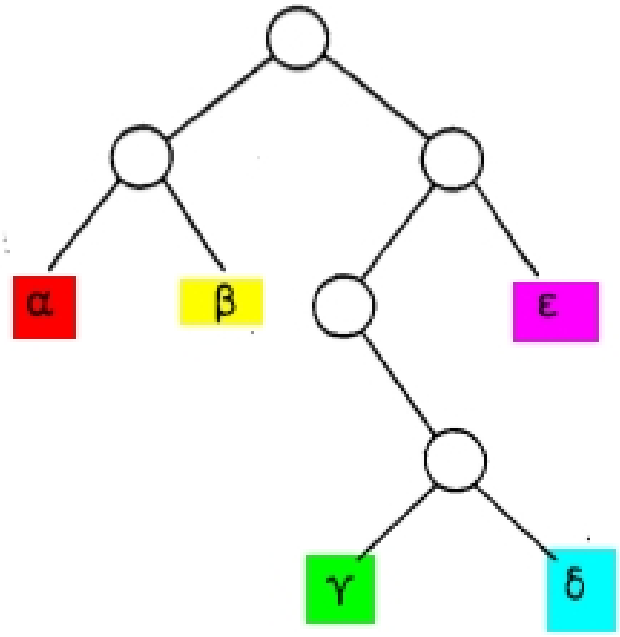
Philippe Flajolet*

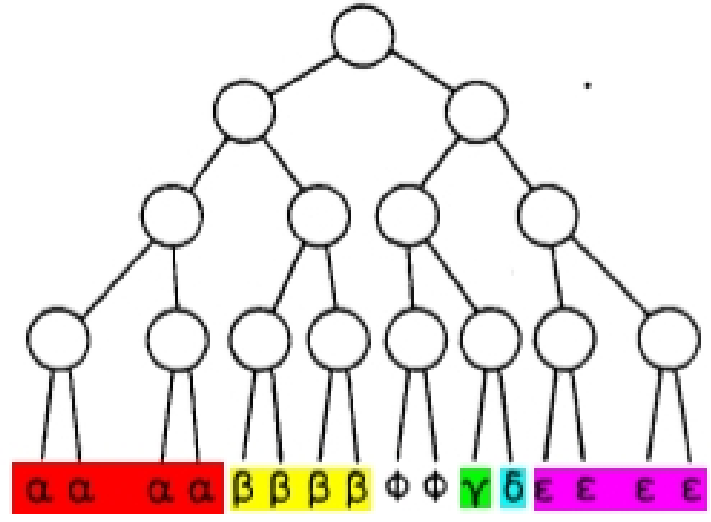
INRIA, Rocquencourt BP105, F-78153 Le Chesnay Cedex, France

Summary. A class of treés occurs both in digital searching and in schemes for maintaining dynamic hash tables. We study the distribution of height in these trees using the saddle point method of complex analysis. As a result, we derive a precise evaluation of the memory requirements of extendible hashing – a dynamic hashing scheme – and discuss some related implementation issues.

Directory trie in Extendible Hashing.







Height of the Directory in Extendible Hashing.

Theorem 2. *The average height of a b-trie constructed on n keys satisfies:*

$$\bar{H}_n = \left(1+\frac{1}{b}\right)\log_2 n + P\left(\left(1+\frac{1}{b}\right)\log_2 n\right) + o(1)$$

where P is a periodic and continuous function with period 1. Function P is representable by a Fourier series with coefficients

$$p_k = \int_0^1 P(u)\, e^{2ik\pi x}\, dx$$

given by

$$p_0 = \frac{1}{b\log 2}\left[\gamma - \log(b+1)!\right],$$

$$p_k = -\frac{1}{b\log 2}\,\beta^{-\chi_k}\,\Gamma(\chi_k)$$

where $\chi_k = \dfrac{2ik\pi}{b\log 2}$ *and* $\beta = 1/(b+1)!$.

Size of the Directory in Extendible Hashing.

Theorem 3. *The average size of the directory in extendible hashing when n keys are present in the file satisfies for $b>1$*

$$\bar{S}_n = Q\left(\left(1+\frac{1}{b}\right)\log_2 n\right) n^{1+1/b}\left[1+O\left(\frac{1}{(\log n)^{b-1}}\right)\right],$$

where Q is a continuous periodic function with period 1 and Fourier coefficients:

$$q_0 = \frac{1}{\log 2}[(b+1)!]^{-1/b}\Gamma\left(1-\frac{1}{b}\right),$$

$$q_k = \frac{-1}{b\log 2}[(b+1)!]^{\chi_k - 1/k}\Gamma\left(\chi_k - \frac{1}{b}\right) \qquad \chi_k = \frac{2ik\pi}{b\log 2}.$$

The Fourier series of Q is absolutely convergent, so that Q is representable as:

$$Q(u) = \sum_{k\in Z} q_k e^{-2k\pi u}.$$

As a technical comment on Theorems 2 to 3, it may be of interest to notice that similar Fourier series otherwise occur in the analysis of algorithms as a result of the use of Mellin transforms or in relation to certain regularities of arithmetic sequences. For a related problem relative to the expected length of the longest probe sequence in hashing with separate chaining, the reader can consult [5].

Bias probabilities and a factorization algorithm.

Finally, in Section 4, we indicate how to extend our methods to cope with the case of a biased distribution on bits of keys of hashed values. Consideration of this situation is motivated by some recent improvements on Berlekamp's factorization algorithm [Kn69], [CZ81], [La81] and we prove :

THEOREM 3 : *The expected height of a simple trie* (b=1) *of* n *leaves when bits in keys have a biased probability of* p *for zeros and* q *for ones satisfies*

$$\overline{H}_n = \frac{2 \log_2 n}{\log_2 (p^2+q^2)^{-1}} + O(1)$$

This theorem solves a problem left open in [La81] whose algorithm appears to factorize a plynomial with r factors in approximately $(2+\varepsilon) \log_2 r$ "main" steps.

Computer graphic tools in 1982.

We propose to take for contour Γ a contour that crosses the saddle point closest to the origin along the direction of decrease and hope for localization properties of the resulting integral (7).

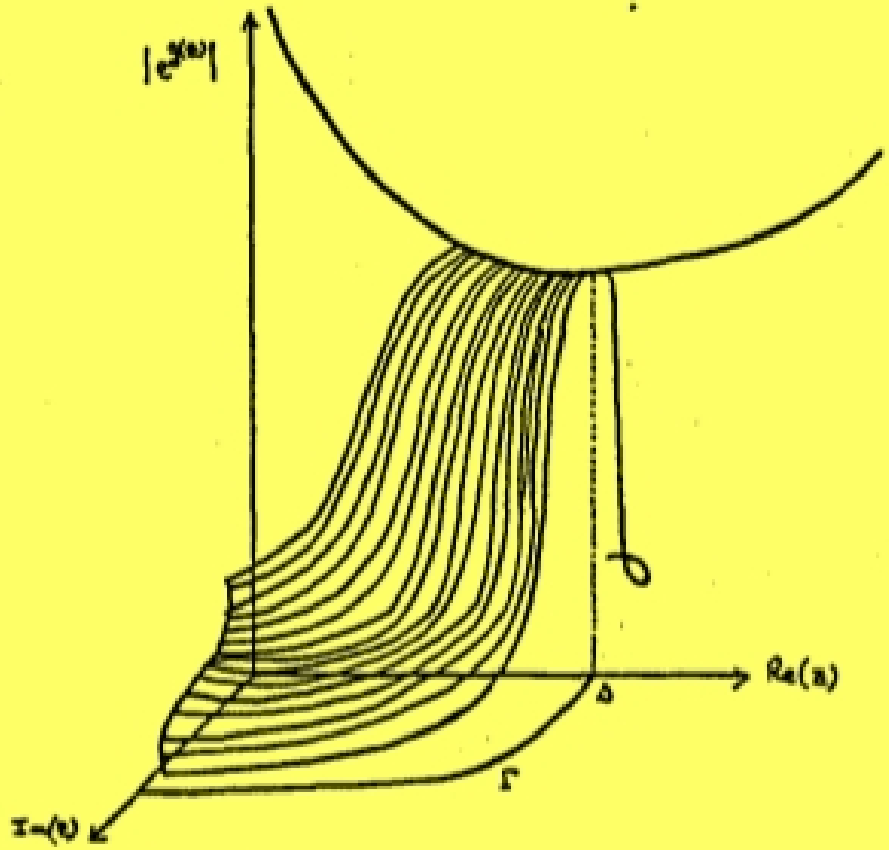


Figure 2 : The Topography of $|e^{\eta(z)}|$

Hashing, Random Allocation and Probabilistic Languages

RANDOM ALLOCATIONS

AND

PROBABILISTIC LANGUAGES

Philippe Flajolet
INRIA, Rocquencourt, 78150 Le Chesnay (France)

Danièle Gardy
LRI, Université Paris-Sud, 91405 Orsay (France)

Loÿs Thimonier
Université de Picardie, 33 rue Saint Leu, 80039 Amiens (France)
and LRI, Université Paris-Sud

Birthday paradox, coupon collectors, caching algorithms and self-organizing search

Philippe Flajolet
INRIA, Rocquencourt, 78150 Le Chesnay, France

Danièle Gardy
LRI, Université Paris-Sud, 91405 Orsay, France

Loÿs Thimonier
Université de Picardie, 33 rue Saint Leu, 80039 Amiens, France; and LRI, Université Paris-Sud, 91405 Orsay, France

Received 4 August 1987
Revised 3 December 1990

Abstract and Conclusions.

This paper introduces a unified framework for the analysis of a class of random allocation processes that include: (i) the birthday paradox; (ii) the coupon collector problem; (iii) least-recently-used (LRU) caching in memory management systems under the independent reference model; (iv) the move-to-front heuristic of self-organizing search. All analyses are relative to general nonuniform probability distributions.

Our approach to these problems comprises two stages. First, the probabilistic phenomena of interest are described by means of regular languages extended by addition of the shuffle product. Next, systematic translation mechanisms are used to derive integral representations for expectations and probability distributions.

7. Some conclusions

It is clear that our derivations are not "unique", and alternative combinatorial or probabilistic arguments could be (or have been) given for some of our results. Our goal has been to show how addition of the shuffle product to regular languages leads to direct analysis of a natural class of random allocation problems. That approach is of value in more complex situations. For instance, problems around multi-level caching are natural candidates and they are discussed in [16].

Birthday Paradox.

3. Birthday paradox

The alphabet $\mathcal{A}$ represents here the dates in a year with m days, and p_i is the probability of date $a_i \in \mathcal{A}$. We consider the following generalization of the birthday problem:

BP. Determine the expectation of the number B_j of elements that need to be drawn from $\mathcal{A}$ (with replacement) till we first encounter j distinct elements that are each repeated at least k times (i.e., the waiting time till the jth different letter occurrence of a k-hit).

The case $k=2$, $j=1$ is the classical birthday problem. Klamkin and Newman [23] have given an integral formula for $j=1$ (first hit) and general k in the *uniform case* where $p_i = 1/m$.

Theorem 3.1. *The expectation* $\mathrm{E}\{B_j\}$ *of the time for obtaining* j *different letter occurrences of a* k*-hit under a general probability distribution* $\{p_i\}_{i=1}^{m}$ *is given by*

$$\mathrm{E}\{B_j\} = \sum_{q=0}^{j-1} \int_0^{\infty} [u^q] \left(\prod_{i=1}^{m} (e_{k-1}(p_i t) + u(\mathrm{e}^{p_i t} - e_{k-1}(p_i t))) \right) \mathrm{e}^{-t}\, \mathrm{d}t \qquad (4)$$

where $e_k(t)$ *represents the truncated exponential*

$$e_k(t) = 1 + \frac{t}{1!} + \frac{t^2}{2!} + \cdots + \frac{t^k}{k!}.$$

Coupon Collector.

4. Coupon collector problem

The alphabet $\mathscr{A}$ now represents the set of coupons, with p_i being the probability that coupon i is issued. The general coupon collector problem is the following:

CCP. Determine the expectation of the number C_j of elements that need to be drawn from $\mathscr{A}$ (with replacement), till one first obtains a collection with j different coupons.

Theorem 4.1. *The expectation* $\mathrm{E}\{C_j\}$ *of the time necessary to gather a collection of* j *different items under a general probability distribution is given by*

$$\mathrm{E}\{C_j\} = \sum_{q=0}^{j-1} \int_0^{\infty} [u^q] \left(\prod_{i=1}^{m} (1 + u(\mathrm{e}^{p_i t} - 1)) \right) \mathrm{e}^{-t} \, \mathrm{d}t, \tag{13a}$$

and for a full collection,

$$\mathrm{E}\{C_m\} = \int_0^{\infty} \left(1 - \prod_{i=1}^{m} (1 - \mathrm{e}^{-p_i t}) \right) \mathrm{d}t. \tag{13b}$$

Proof. Form (13a) is just a specialization of formula (4) to the case $k = 1$, and

LRU Caching.

(3) *Least-recently-used caching* [LRU]: Caching algorithms aim at maintaining fast access to a large number of items by keeping a small "cache" that may be addressed quickly. The classical problem of cache analysis consists in determining the steady state probability of a cache "fault" when items are accessed with a fixed, not necessarily uniform, distribution. The LRU caching strategy consists in applying replacement, when needed, to the oldest element in the cache (the "least-recently-used" element).

Theorem 5.1. *The long run probability D of a cache fault in the LRU algorithm is given by*

$$1 - D = \sum_{q=0}^{k-1} [u^q] \int_0^{+\infty} \Phi(t,u)\Psi(t,u)e^{-t}\,dt, \tag{17}$$

where functions Φ and Ψ are

$$\Phi(t,u) = \prod_{j=1}^{m} (1 + u(e^{p_j t} - 1)) \quad \textit{and} \quad \Psi(t,u) = \sum_{i=1}^{m} \frac{p_i^2}{1 + u(e^{p_i t} - 1)}. \tag{18}$$

Move to Front.

MTF. When an element in position j is accessed, it is moved to the front of the file. Elements in position $1, 2, \ldots, j-1$ are shifted back by one position.

Formally, the MTF rule is exactly an LRU caching algorithm in which the cache size k is equal to the file size $m = \text{card}(\mathcal{A})$. Page faults disappear and the transition

Our purpose is only to show that the analysis of MTF can be cast in the framework of shuffles of languages. The theorem that follows is due to McCabe [32], useful references on the subject being [26, 34, 4, 18].

Theorem 6.1 [32]. *The expected cost of a search with the move-to-front heuristic applied to a file with access probabilities* $\{p_i\}_{i=1}^m$ *is*

$$E = -\frac{1}{2} + \sum_{1 \le i,j \le m} \frac{p_i p_j}{p_i + p_j}.$$

Our line of proof, admittedly not the simplest possible, "explains" the derivation of Theorem 6.1 that appears in [26, p. 403]. In essence Knuth's derivation amounts to operating with an ordinary generating function equivalent to $B(z; v)$ and computed directly by summing over all possible cases. Our proof also yields information

Bucket selection and sorting.




Acta Informatica 36, 735–760 (2000)

Analytic variations on bucket selection and sorting

Hosam Mahmoud[1], Philippe Flajolet[2], Philippe Jacquet[2], Mireille Régnier[2]

Abstract.

Abstract. We provide complete average-case as well as probabilistic analysis of the cost of bucket selection and sorting algorithms. Two variations of bucketing (and flavors therein) are considered: distributive bucketing (large number of buckets) and radix bucketing (recursive with a small number of buckets, suitable for digital computation). For Distributive Selection a compound Poisson limit is established. For all other flavors of bucket selection and sorting, central limit theorems underlying the cost are derived by asymptotic techniques involving perturbation of Rice's integral and contour integration (saddle point methods). In the case of radix bucketing, periodic fluctuations appear in the moments of both the selection and sorting algorithms.

Relation with Hashing (Methods!).

of Radix Sort. Although central limit theorems were not known before in the context of bucket sorting, the methods of analysis are connected to some classical as well recent analyses in tries (Jacquet and Régnier (1988)) and hashing with linear probing (Flajolet, Poblete and Viola (1998)). Therefore our arguments in Section 3 will be sketchy. Section 4 concludes with a discussion.

For this form of bucket sorting we consider the functional equation of Lemma 2 when specialized to the case $b = n$. Specialized to the case $b = n$, the right hand side of this equation can be viewed as the nth coefficient in a generating function. We can express the specialized equation in the form

$$\phi_n(u) = \frac{n!}{n^n}[z^n]\left(\sum_{j=0}^{\infty}\psi_j(u)\frac{z^j}{j!}\right)^n.$$

This representation admits the following central limit result. The theorem follows from a rather general result in Flajolet, Poblete and Viola (1998) for hashing with linear probing which broadly states that, under suitable conditions, coefficients of bivariate generating functions raised to large powers follow a Gaussian law. We work through some of the details to obtain the

Distributive selection.

2.1 Distributive selection

This version (with $b = n$) of bucket selection is easiest to analyze when the algorithm switches to a standard selection algorithm A within a bucket. Extracting the coefficient of u^k from the equality of Lemma 1 when specialized to the case $b = n$:

$$\mathbf{P}\{Z_n = k\} = \frac{1}{n^n} \sum_{j=1}^{\infty} j\mathbf{P}\{Y_j = k\}(n-1)^{n-j} \binom{n}{j}. \tag{4}$$

The term $n^{-n}(n-1)^{n-j}\binom{n}{j} = n^{-j}(1-1/n)^{n-j}\binom{n}{j}$ is the probability that $B(n, 1/n) = j$, which for any given j converges to $e^{-1}/(j-1)!$ by the standard approximation of the binomial distribution of $B(n, 1/n)$ to $\mathcal{P}(1)$. At any fixed k, passing to the limit (as $n \to \infty$) gives us

$$\lim_{n\to\infty} \mathbf{P}\{Z_n = k\} = \sum_{j=1}^{\infty} \mathbf{P}\{Y_j = k\} \frac{e^{-1}}{(j-1)!}. \tag{5}$$

Theorem 1 *In the selection of a randomly chosen rank from among n random keys by Distributive Selection, whose algorithm within a bucket makes Y_j operations for random selection in a file of size j, the extra cost after the first layer of bucketing satisfies a limiting compound Poisson law (in the sense of (5)):*

$$Z_n \xrightarrow{\mathcal{D}} Y_{\mathcal{P}(1)+1}.$$

Radix Selection.

Theorem 2 *Let C_n be the number of bucket operations (digit extractions) performed by Radix Select using B (fixed) buckets to find a randomly chosen order statistic among n keys. If we set $\lambda_B = B/(B-1)$, then as $n \to \infty$,*

$$\mathbf{E}[C_n] = \lambda_B n + \log_B n + P(\log_B n) + o(1),$$
$$\mathbf{Var}[C_n] = \frac{\lambda_B}{B-1} n + O(\ln^2 n),$$

where P is a smooth periodic function. The law of C_n is asymptotically normal:

$$\frac{C_n - \lambda_B n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\lambda_B}{B-1}\right).$$

Distributive sorting.

Theorem 3 *Let C_n be the cost of Distributive Sort to sort n random keys. Suppose for some fixed $\theta > 0$, the algorithm applied in the buckets uses $Y_j \leq j^\theta$ operations costing α units each (the unit being the cost of one bucketing operation). Then*

$$\frac{C_n - (1+\alpha\mu)n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \alpha^2\sigma^2),$$

where

$$\mu = e^{-1}\sum_{j=0}^{\infty}\frac{\mathbf{E}[Y_j]}{j!}, \qquad \sigma^2 = e^{-1}\sum_{j=0}^{\infty}\frac{\mathbf{E}[Y_j^2]}{j!} - \mu^2.$$

Radix Sorting.

Theorem 4 *Let C_n be the cost of Radix Sort (number of digit extractions) to sort n random keys. Then $(C_n - L_n)/\sqrt{V_n}$ tends in distribution and in moments to the standard normal variate $\mathcal{N}(0,1)$ with*

$$L_n = \left(\ln n + \gamma + \frac{\ln B}{2} + P_1(\ln n)\right)\frac{n}{\ln B} + O(\ln n),$$
$$V_n = n(C(B) + P_2(\ln n)) + O(\ln^2 n),$$

where

$$C(B) = \frac{1}{\ln B}\left(\frac{1}{4} + \ln 2 + 2\sum_{k=1}^{\infty}\left[\ln(1+B^{-k}) + (1+B^{-k})^{-2}\right]\right),$$

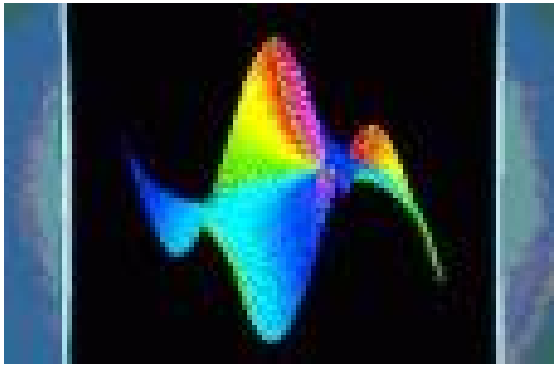
and the functions $P_1(x)$, and $P_2(x)$ are periodic functions with small amplitude and period $\ln B$.

Remark. The average was found by Knuth and De Bruijn; the analysis for the case $B = 2$ is presented in Knuth (1973; pp. 131–134). Related variance analyses appear in Kirschenhofer, Prodinger, and Szpankowski (1989).

General idea of the proof.

Proof of Theorem 4. The proof is divided into three parts: the mean value analysis, the variance analysis, and finally the limit distribution result. Mean value analysis is done by *Poissonization*, where one considers the number of keys to have come from the $\mathcal{P}(n)$ distribution instead of a fixed population model. Recurrence equations arise in a natural way for Poissonized averages. These equations are solved exactly by iteration and the resulting form is approximated via the Mellin transform and its inverse. The variance follows suit, only the exact computation and the approximation by Mellin transform are a lot more complicated. The limit distribution is obtained by a direct analysis of the recurrence at hand. The Poissonized form (14) is manipulated to show that a suitably normed version of the Poissonized number of comparisons follows a Gaussian law. The fixed population result is then extracted by depoissonization.

Linear Probing Hashing.



MATHÉMATIQUES ET INFORMATIQUE

Hachage, arbres, chemins & graphes

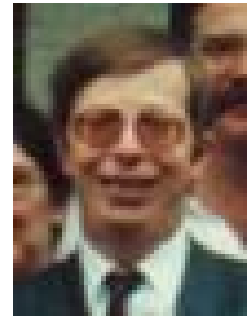
Philippe Chassaing[†] et Philippe Flajolet[‡]

The mathematical beauty of Linear Probing!

Mathématiques discrètes et continues se rencontrent et se complètent volontiers harmonieusement. C'est cette thèse que nous voudrions illustrer en discutant un problème classique aux ramifications nombreuses—l'analyse du hachage avec essais linéaires. L'exemple est issu de l'analyse d'algorithmes, domaine fondé par Knuth et qui se situe lui-même « à cheval » entre l'informatique, l'analyse combinatoire, et la théorie des probabilités. Lors de son traitement se croisent au fil du temps des approches très diverses, et l'on rencontrera des questions posées par Ramanujan à Hardy en 1913, un travail d'été de Knuth datant de 1962 et qui est à l'origine de l'analyse d'algorithmes en informatique, des recherches en analyse combinatoire du statisticien Kreweras, diverses rencontres avec les modèles de graphes aléatoires au sens d'Erdös et Rényi, un peu d'analyse complexe et d'analyse asymptotique, des arbres qu'on peut voir comme issus de processus de Galton-Watson particuliers, et, pour finir, un peu de processus, dont l'ineffable mouvement Brownien ! Tout ceci contribuant *in fine* à une compréhension très précise d'un modèle simple d'aléa discret.

Ramanujan's Q Function.






ELSEVIER

Journal of Computational and Applied Mathematics 58 (1995) 103–116

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

On Ramanujan's Q-function☆

Philippe Flajolet*,[a], Peter J. Grabner[b], Peter Kirschenhofer[c], Helmut Prodinger[c]

[a] *Algorithms Project, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France*
[b] *Technische Universität Graz, Steyrergasse 30, A-8010 Graz, Austria*
[c] *Technische Universität Wien, Wiedner Hauptstrasse 8–10, A-1040 Vienna, Austria*

Dedicated to D.E. Knuth, on the occasion of the 30th anniversary of his first analysis of an algorithm in 1962

Abstract.

Abstract

This study provides a detailed analysis of a function which Knuth discovered to play a central rôle in the analysis of hashing with linear probing. The function, named after Knuth $Q(n)$, is related to several of Ramanujan's investigations. It surfaces in the analysis of a variety of algorithms and discrete probability problems including hashing, the birthday paradox, random mapping statistics, the "rho" method for integer factorization, union-find algorithms, optimum caching, and the study of memory conflicts.

A process related to the complex asymptotic methods of singularity analysis and saddle point integrals permits to precisely quantify the behaviour of the $Q(n)$ function. In this way, tight bounds are derived. They answer a question of Knuth (*The Art of Computer Programming*, Vol. 1, 1968, [Ex. 1.2.11.3.13]), itself a rephrasing of earlier questions of Ramanujan in 1911–1913.

In the 1911 issue of the *J. Indian Math. Soc.*, Ramanujan [18] poses the following problem: Show that

$$\tfrac{1}{2}e^n = 1 + \frac{n}{1!} + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}\theta, \quad \text{where } \theta \text{ lies between } \tfrac{1}{2} \text{ and } \tfrac{1}{3}. \tag{1.1}$$

A solution was then outlined in [19]. Later in his first letter to Hardy dated 16 January 1913 (see [20, p. xxvi], [1, p. 181], [8]), Ramanujan makes a stronger assertion, namely that

$$\theta = \tfrac{1}{3} + \frac{4}{135(n+k)}, \quad \text{where } k \text{ lies between } \tfrac{8}{45} \text{ and } \tfrac{2}{21}. \tag{1.2}$$

A solution to the weaker inequality (1.1) was given by Szegő in 1928 [21], and almost simultaneously Watson [23] wrote a paper where he proved (1.1) and adds regarding (1.2): "*I shall also give reasons, which seem to me fairly convincing, for* believing *that k lies between* $\frac{8}{45}$ *and* $\frac{2}{21}$". Our purpose here is to finally provide a complete proof of Ramanujan's assertion (1.2).

Ramanujan's Q function into play.

The variant form used by Knuth introduces the two functions

$$Q(n) = 1 + \frac{n-1}{n} + \frac{(n-1)(n-2)}{n^2} + \cdots,$$
$$R(n) = 1 + \frac{n}{n+1} + \frac{n^2}{(n+1)(n+2)} + \cdots, \tag{1.3}$$

and one finds easily

$$Q(n) + R(n) = n!e^n/n^n.$$

$$\frac{n^n}{n!}Q(n) = \frac{n^{n-1}}{(n-1)!} + \frac{n^{n-2}}{(n-2)!} + \cdots + 1, \tag{1.4}$$

which entails

$$\frac{n^n}{n!}Q(n) = \tfrac{1}{2}e^n - \theta\frac{n^n}{n!} \quad \text{and} \quad \theta(n) = \tfrac{1}{2}(R(n) - Q(n)).$$

In this way, Ramanujan's problem can be rephrased as: "*Show that*

$$R(n) - Q(n) = \tfrac{2}{3} + \frac{8}{135(n+k)},$$

where $k \equiv k(n)$ *lies between* $\frac{2}{21}$ *and* $\frac{8}{45}$". Following Knuth, this is the form that we shall take as our starting point, setting $D(n) = R(n) - Q(n)$, so that $D(n) = 2\theta(n)$.

The approaches followed by Ramanujan himself and later authors all make use of *real* integral representations derived from

$$Q(n) = \int_0^\infty e^{-x}\left(1 + \frac{x}{n}\right)^{n-1} dx, \tag{1.5}$$

The Tree function y(z).

An important function in combinatorial analysis is the function $y(z)$ defined implicitly by the equation

$$y(z) = ze^{y(z)}, \tag{2.1}$$

with $y(z) = z + z^2 + 3z^3/2 + \cdots$. By the Lagrange inversion formula, we have[1] the following proposition.

Proposition 1. *The Taylor coefficients of* $y(z) = ze^{y(z)}$ *and its powers are given by*

$$[z^n]\, y(z) = \frac{n^{n-1}}{n!} \quad \text{and} \quad [z^n]\, y^k(z) = k\frac{n^{n-k-1}}{(n-k)!}. \tag{2.2}$$

Furthermore, a generating function of $Q(n)$ *is expressible in terms of* $y(z)$:

$$\sum_{n=1}^{\infty} Q(n)\, n^{n-1}\, \frac{z^n}{n!} = \log\frac{1}{1-y(z)}. \tag{2.3}$$

Theorem 2 (Ramanujan, Watson and Knuth). *The quantities* $Q(n)$, $R(n)$ *admit full asymptotic expansions in descending powers of* $\sqrt{n}$:

$$Q(n) \sim \sqrt{\frac{\pi n}{2}} - \frac{1}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2n}} - \frac{4}{135n} + \cdots,$$

$$R(n) \sim \sqrt{\frac{\pi n}{2}} + \frac{1}{3} + \frac{1}{12}\sqrt{\frac{\pi}{2n}} + \frac{4}{135n} + \cdots.$$

Proof. We sketch here the proof based on singularity analysis (see [22, 14, 5] for related develop-

Main result.

Theorem 7. *With the quantity* $\theta \equiv \theta(n)$ *being defined by*

$$\frac{1}{2}\,e^n = 1 + \frac{n}{1!} + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}\,\theta,$$

one has, for all integers $n \geqslant 0$,

$$\theta = \frac{1}{3} + \frac{4}{135(n+k)},$$

where $k \equiv k(n)$ *lies between* $\frac{8}{45}$ *and* $\frac{2}{21}$.

Methodology ...

5. Some conclusions

It may be of interest, at last, to reflect on the various alternatives that offer themselves in order to estimate asymptotically sequences like $Q(n)$ or $D(n)$.

(1) *Laplace method.* The Laplace method for integrals, based on the integral representation (1.5) was the starting point of earlier approaches. As Szegő and Watson show, it can be made "constructive" (instead of providing only O-bounds) but its operation becomes then somewhat intricate.

(2) *Singularity analysis.* This is the method that gave us here the expansion of Theorem 2. It is based on the fact that the implicitly defined function $y(z)$ has an algebraic singularity of the $\sqrt{}$-type, from which the singularity types of the generating functions associated with $Q(n)$ or $D(n)$ follow. The method can also accommodate constructive bounds on a function's coefficients [6]. Consequently, it might be applicable to derive Theorem 7, although, in this case, bounding $y(z)$ in the appropriate region would probably prove unwieldy.

. . . and more methodology!

(3) *Darboux's method.* Darboux's method also leads to a full asymptotic expansion by a route very similar to singularity analysis. However, it does not have the capacity to provide bounds since it is intrinsically based on a nonconstructive lemma on Fourier series.

(4) *Saddle point.* This is in essence the route that we took, after a suitable change of variable. Its application in the case of implicitly defined functions and Lagrange series is also to be traced in Darboux's works, an interesting combinatorial application occurring in [12]. By this method, we were able to reduce the problem to the task of finding simple bounds for elementary functions on circles and line segments. Interestingly enough, when considering the conformal mapping defined by $y^{(-1)}(z)$, it appears that the induced contour in the z-plane closely resembles the type of "Hankel" contour used in the z-plane under the singularity analysis approach. This establishes a perhaps unexpected relation between two seemingly unrelated methods — singularity analysis and saddle point — at least in the context of implicitly defined functions and Lagrange series.

Linear Probing and a box full of surprises!

Algorithmica (1998) 22: 490–515

Algorithmica

On the Analysis of Linear Probing Hashing[1]

P. Flajolet,[2] P. Poblete,[3] and A. Viola[4]

Dedicated to Don Knuth on the occasion of the 35th anniversary of his first analysis of an algorithm in 1962–1963.

Abstract

Abstract. This paper presents moment analyses and characterizations of limit distributions for the construction cost of hash tables under the linear probing strategy. Two models are considered, that of full tables and that of sparse tables with a fixed filling ratio strictly smaller than one. For full tables, the construction cost has expectation $O(n^{3/2})$, the standard deviation is of the same order, and a limit law of the Airy type holds. (The Airy distribution is a semiclassical distribution that is defined in terms of the usual Airy functions or equivalently in terms of Bessel functions of indices $-\frac{1}{3}, \frac{2}{3}$.) For sparse tables, the construction cost has expectation $O(n)$, standard deviation $O(\sqrt{n})$, and a limit law of the Gaussian type. Combinatorial relations with other problems leading to Airy phenomena (like graph connectivity, tree inversions, tree path length, or area under excursions) are also briefly discussed.

Companion paper in Special Issue (Flajolet's 50th birthday).

Linear Probing and Graphs

Donald E. Knuth, Stanford University

DEDICATED TO PHILIPPE PATRICK MICHEL FLAJOLET

Abstract. Mallows and Riordan showed in 1968 that labeled trees with a small number of inversions are related to labeled graphs that are connected and sparse. Wright enumerated sparse connected graphs in 1977, and Kreweras related the inversions of trees to the so-called "parking problem" in 1980. A combination of these three results leads to a surprisingly simple analysis of the behavior of hashing by linear probing, including higher moments of the cost of successful search.

Main part of one of several mails exchanged with D. Knuth.

Date: Mon, 29 Sep 1997 13:15:21 -0700 (PDT)

. . .

To: Philippe.Flajolet@inria.fr

Subject: note from Don Knuth

Dear Ph, Ordinarily I am not happy to receive email, but in this case it was very touching to learn that you had decided to dedicate such a nice paper to me, just after I had (secretly) decided to dedicate reference [22] to you!

But I haven't time to study it in detail now, as I'm working 150% time on the new edition of Volume 3...

. . .

. . .

. . .

Best regards, Don

Combinatorial approach to Linear Probing.

The purpose of this note is to exhibit a surprisingly simple solution to a problem that appears in a recent book by Sedgewick and Flajolet [9]:

> **Exercise 8.39** Use the symbolic method to derive the EGF of the number of probes required by linear probing in a successful search, for fixed M.

The authors admitted that they did not know how to solve the problem, in spite of the fact that a "symbolic method" was the key to the analysis of all the other algorithms in their book. Indeed, the second moment of the distribution of successful search by linear probing was unknown when [9] was published in 1996.

Combinatorial Analysis (FPV).

2.1. *Combinatorial Analysis.* We define $F_{n,k}$ as the number of ways of creating an almost full table with n elements and total displacement k. The corresponding bivariate generating function is then

$$F(z, q) = \sum_{n,\, k \geq 0} F_{n,k} q^k \frac{z^n}{n!},$$

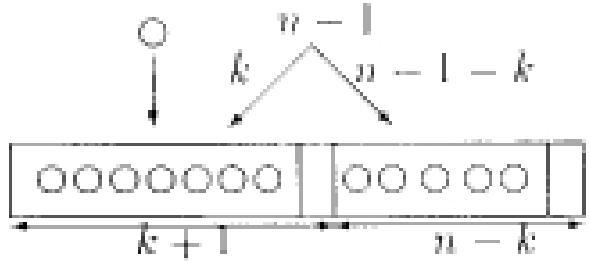


Fig. 1. The binary tree decomposition of almost full tables.

$$F_n(q) = \sum_{k=0}^{n-1} \binom{n-1}{k} F_k(q)(1 + q + \cdots + q^k) F_{n-1-k}(q).$$

Solution to the fundamental recurrence (Knuth).

$$\sum_{n=1}^{\infty}(x-1)^{n-1}\,F_{n-1}(x)\,\frac{z^n}{n!} = C(z) = \ln\sum_{n=0}^{\infty} x^{n(n-1)/2}\,\frac{z^n}{n!}\,. \tag{2.17}$$

3. Connected graphs. We are interested in the behavior of $F_n(x)$ near $x = 1$, so it is convenient to write $x = 1 + w$. Then (2.17) becomes

$$\sum_{n=1}^{\infty} w^{n-1}\,F_{n-1}(1+w)\,\frac{z^n}{n!} = \ln\sum_{n=0}^{\infty}(1+w)^{n(n-1)/2}\,\frac{z^n}{n!}\,. \tag{3.1}$$

Aha—the right side of this equation is well known as the exponential generating function for labeled connected graphs [8]. Thus we have

$$w^{n-1}\,F_{n-1}(1+w) = C_n(1+w) = \sum w^{\text{edges}(G)}\,, \tag{3.2}$$

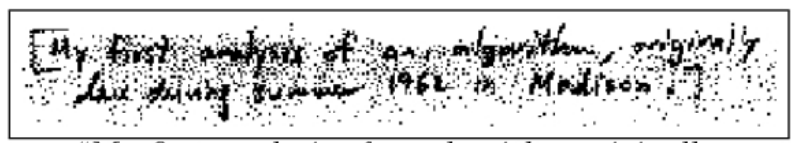
Conclusions (Knuth).

7. Personal remarks. The problem of linear probing is near and dear to my heart, because I found it immensely satisfying to deduce (5.4) when I first studied the problem in 1962. Linear probing was the first algorithm that I was able to analyze successfully, and the experience had a significant effect on my future career as a computer scientist. None of the methods available in 1962 were powerful enough to deduce the expected square displacement, much less the higher moments, so it is an even greater pleasure to be able to derive such results today from other work that has enriched the field of combinatorial mathematics during a period of 35 years.

The reader will note that Sedgewick and Flajolet's exercise 8.39 has not truly been solved, strictly speaking, because we have not found the EGF $\sum_{n=0}^{m-1} F_{mn}(x)\, z^n/n!$ as requested. However, Sedgewick and Flajolet should be happy with any analysis of linear probing that uses symbolic methods associated with generating functions in an informative way.

ON THE ANALYSIS OF LINEAR PROBING HASHING

Philippe Flajolet, INRIA Rocquencourt
(France)



"My first analysis of an algorithm originally done during Summer 1962 at Madison."

3 Linear Probing Hashing

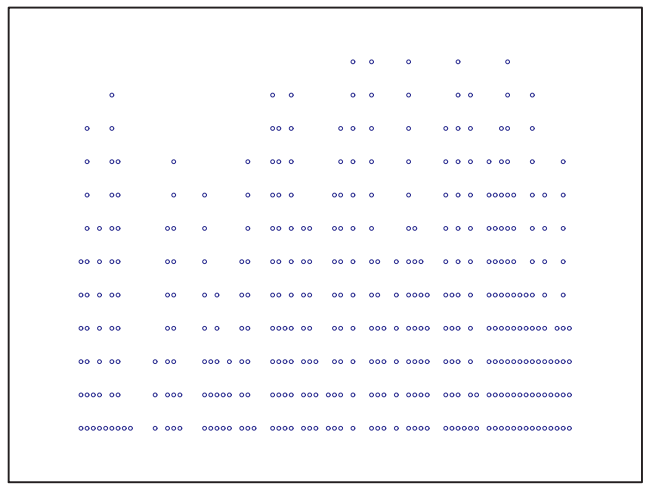
LINEAR PROBING: Physics of "What goes on"?

— For very small α, expect a behaviour like separate chaining (S.C.H.): isolated elements only.

— For moderate α, clusters start to form. These should be a bit larger than S.C.H.

— As $\alpha \to 1$, "clotting" takes place.

— Last element suffers from large displacement $\sim \frac{m}{2}$.



QUANTIFY?

A story that starts with Knuth in 1962

♠  A decomposition based on adding a new element

♠♠  Going through Abel identities

$$(x+y)(x+y+n)^{n-1} = xy \sum_{k=0}^{n} \binom{n}{k} (x+k)^{k-1}(y+n-k)^{n-k-1}.$$

♠♠♠  And concluding with Ramanujan's function

$$Q(n) := 1+\frac{n-1}{n}+\frac{(n-1)(n-2)}{n^2}+\frac{(n-1)(n-2)(n-3)}{n^3}+\cdots$$

[My first analysis of an algorithm, originally done during Summer 1962 in Madison.]

8354

NOTES ON "OPEN" ADDRESSING. D. Knuth 7/22/63

1. Introduction and Definitions. Open addressing is a widely-used technique for keeping "symbol tables." The method was first used in 1954 by Samuel, Amdahl, and Boehme in an assembly program for the IBM 701. An extensive discussion of the method was given by Peterson in 1957 [1], and frequent references have been made to it ever since (e.g. Schay and Spruth [2], Iverson [3]). However, the timing characteristics have apparently never been exactly established, and indeed the author has heard reports of several reputable mathematicians who failed to find the solution after some trial. Therefore it is the purpose of this note to indicate one way by which the solution can be obtained.

We will use the following abstract model to describe the method: N is a positive integer, and we have an array of N variables $x_1, x_2, \ldots, x_N$. At the beginning, $x_i = 0$, for $1 \leq i \leq N$.

To "enter the k-th item in the table," we mean that an integer a_k is calculated, $1 \leq a_k \leq N$, depending only on the item, and the following process is carried out:

1. Set $j = a_k$.
2. "The comparison step." If $x_j = 0$, set $x_j = 1$ and stop; we say "the k-th item has fallen into position x_j."
3. If $j = N$, go to step 5.
4. Increase j by 1 and return to step 2.
5. "The overflow step." If this step is entered twice, the table is full, i.e. $x_i = 1$ for $1 \leq i \leq N$. Otherwise set j to 1 and return to step 2.

Observe the cyclic character of this algorithm.

We are concerned with the statistics of this method, with respect to the number of times the comparison step must be executed. More precisely, we consider all of the N^k possible sequences $a_1, a_2, \ldots a_k$ to be equally probable, and we ask, "What is the probability that the comparison step is used precisely m times when the k-th item is placed?"

2. Non-overflow (self-contained) sequences.

Let $\left[{n \atop k} \right]$ denote the number of sequences $a_1, a_2, \ldots a_k$ $(1 \leq a_i \leq n)$ in which no overflow step occurs during the entire process of placing k items, if the algorithm is used for $N = n$. (By convention, we set

$$\left[{n \atop 0} \right] = 1.)$$

Lemma 1: If $0 \leq k \leq n+1$, then $\left[{n \atop k} \right] = (n+1)^k - k(n+1)^{k-1}$.

Proof: This proof is based on the fact that $\left[{n \atop k} \right]$ is precisely the number of sequences $b_1, b_2, \ldots b_k$ $(1 \leq b_i \leq n+1)$ in which, if the algorithm is carried out for $N = n+1$, then $x_{n+1} = 0$ at the end of the operation. This follows because every sequence of the former type is one of the latter, and conversely the condition implies in particular that $1 \leq b_i \leq n$, and that no overflow step occurs.

But sequences of the latter type are easily enumerated, because the algorithm has circular symmetry; of the $(n+1)^k$ possible sequences $b_1, b_2, \ldots b_k$, exactly $k/(n+1)$ of these leave $x_{n+1} \neq 0$. This shows that

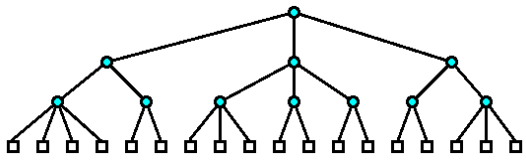
$$\left[{n \atop k} \right] = (n+1)^k \left(1 - \frac{k}{n+1} \right).$$

4 Analytic Combinatorics

Two basic principles $\mapsto$ "dictionaries"

SYMBOLIC METHODS

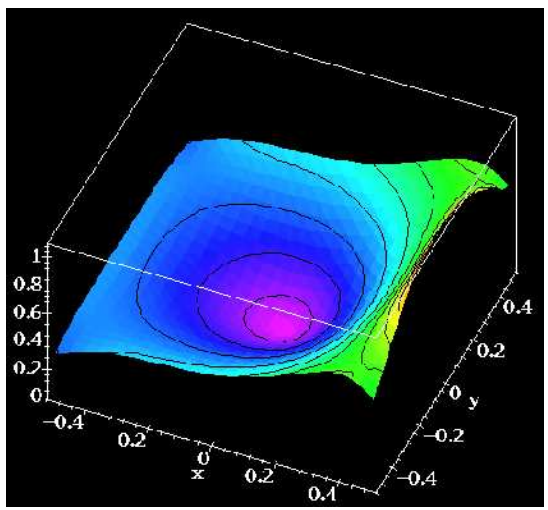
Generating functions



$\mapsto \quad \mathbf{z}^{11}$

$$\mathbf{z} + \mathbf{z}^2 + \mathbf{z}^3 + 2\,\mathbf{z}^4 + 2\,\mathbf{z}^5 + 4\,\mathbf{z}^6 + 5\,\mathbf{z}^7 + 9\,\mathbf{z}^8 + \cdots$$

Analytic functions and singularities



CONSTRUCTIONS

Dictionary (I)

$$\mathcal{F} \quad \mapsto \quad \{f_n\} \quad \mapsto \quad f(z) = \sum_n f_n \frac{z^n}{n!}.$$

$$\frac{1}{1-f} = 1 + f + f^2 + f^3 + \cdots$$

$$\exp(f) = 1 + f + \frac{1}{2!}f^2 + \frac{1}{3!}f^3 + \cdots$$

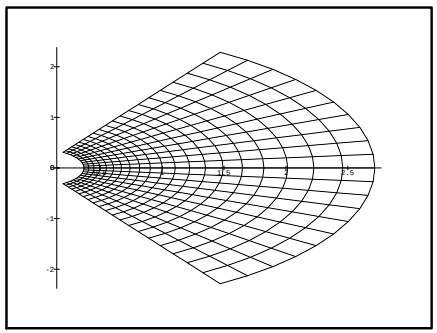
$$\begin{aligned}
\mathbf{A} \cup \mathbf{B} &\mapsto A(z) + B(z) \\
\mathbf{A} \times \mathbf{B} &\mapsto A(z) \times B(z) \\
\text{Seq } \mathbf{A} &\mapsto \frac{1}{1 - A(z)} \\
\text{Set } \mathbf{A} &\mapsto \exp(A(z)) \\
\text{Cycle } \mathbf{A} &\mapsto \log \frac{1}{1 - A(z)}
\end{aligned}$$

COMPLEX ASYMPTOTICS

Dictionary (II)

Point of regularity. $f(z) \approx f(z_0) + f'(z_0)(z - z_0)$

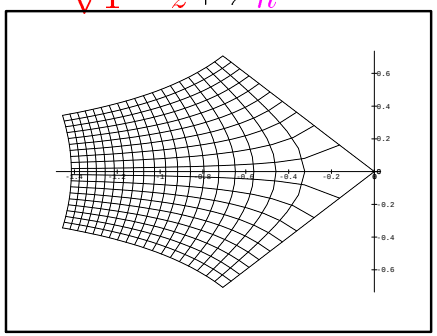
$\exp(z)$



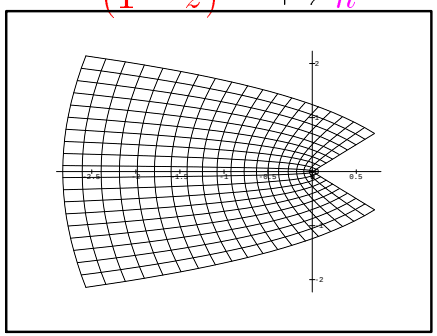
Point of singularity. $\neg\exists \; \frac{d}{dz} f(z)\Big|_{z_0}$

$$(1-z)^{\beta} \;\mapsto\; \frac{n^{\beta-1}}{\Gamma(\beta)}$$

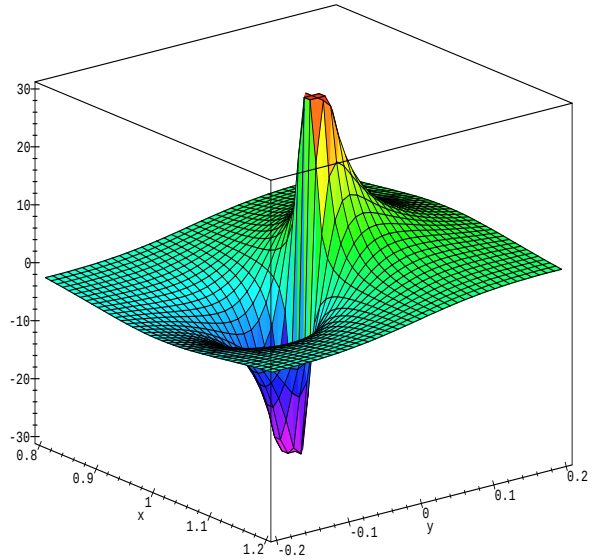
$-\sqrt{1-z} \mapsto n^{-3/2}$



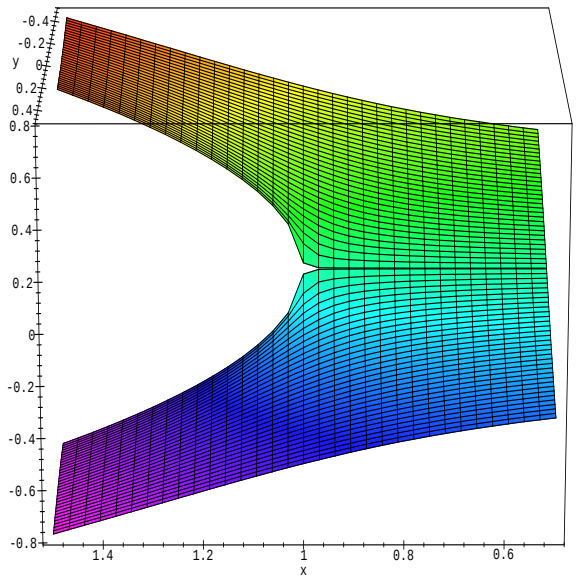
$-(1-z)^{3/2} \mapsto n^{-5/2}$



Permutations: $(1-z)^{-1}$



Trees: $1-\sqrt{1-z}$



Distributional analysis (almost full tables)

- Record *construction cost* = *total displacement*

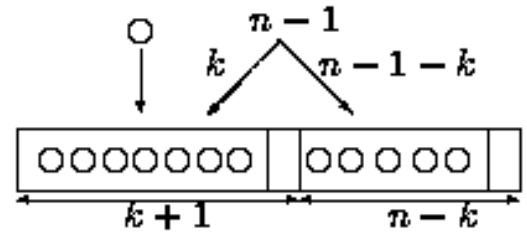


FIGURE 1. The binary tree decomposition of almost full tables.

$$F_n(q) = \sum_{k=0}^{n-1} \binom{n-1}{k} F_k(q)(1 + q + \cdots + q^k) F_{n-1-k}(q).$$

- q–Calculus

$$\sum_n q^{n^2} z^2$$

The construction cost is worst-case quadratic

5 Analysis of L.P.H

Almost full tables $n = m - 1$ have a **tree decomposition**.

```
<Full> := <Full> * <Last> * <Full>
          ^^^^^^  with Position
```

Dictionary: Products $\mapsto$ Products

$$\mathcal{C} = \mathcal{A} \star \mathcal{B} \qquad \mapsto \qquad C(z) = A(z) \cdot B(z)$$

$$C_n = \sum_k \binom{n}{k} A_k B_{n-k}$$

Dictionary: Adding an element $\mapsto \int$, $C_n = A_{n-1}$

$$\mathcal{C} = \mathsf{Add}(\mathcal{A}) \qquad \mapsto \qquad C(z) = \int_0^z A(w)\, dw.$$

Dictionary: Choosing a position $\mapsto \partial$, $C_n = (n+1)A_n$

$$\mathcal{C} = \mathsf{Pos}(\mathcal{A}) \qquad \mapsto \qquad \frac{\partial}{\partial z}\left(z\, A(z)\right).$$

Asymptotics

- Find the singularities of an Implicit Function

$$z - Te^{-T} = 0$$

— Implicit function theorem: OK if partial derivative is nonzero

— Singularity when partial derivative equals 0

$$(1 - T)e^{-T} = 0$$

The system gives $\boxed{T = 1,\ z = e^{-1}}$.

Singular dependence between z and T is locally quadratic

$$\boxed{T(z) \underset{z \to e^{-1}}{=} 1 - \sqrt{1}\,\sqrt{1 - ez} + O(1 - ez)}$$

$$\frac{T_n}{n!} \sim \sqrt{\frac{n}{2\pi}} e^n.$$

Gives back Stirling's formula since $T_n = n^{n-1}$.

$\Longrightarrow$ Analyse any expression involving $T(z)$

Q-calculus

$$
\begin{array}{lcl}
n & \mapsto & [n] = 1 + q + q^2 + \cdots q^{n-1} = \dfrac{1-q^n}{1-q} \\
n! & \mapsto & [n]! = [1] \cdot [2] \cdots [n] \\
\sum (n+1) f_n z^n & \mapsto & \sum [n+1] f_n z^n \\
\dfrac{\partial}{\partial z}(z f(z)) & \mapsto & \mathrm{H}[f(z)] = \dfrac{F(z) - qF(qz)}{1-q}
\end{array}
$$

Many "combinatorial identities" survive in the q-world, with $q^\star \mapsto 1$

EXAMPLE. [Euler] $(\exp(z))^{-1} = (\exp(-z))$

$$\left(\sum \frac{z^n}{n!} \right)^{-1} = \left(\sum \frac{(-z)^n}{n!} \right)$$

$$\left(\sum \frac{z^n}{[n]!} \right)^{-1} = \left(\sum q^{n(n-1)/2} \frac{(-z)^n}{[n]!} \right)$$

6 Limit distribution

A method of "pumping" moments

— Start from **nonlinear** combinatorial decomposition (BGF)

$$\Phi[F(z,q)] = 0$$

— Apply derivatives $U\partial_q^r$ to get rth moment.

— Expect **linear** operator $\mathcal{L}$ with

$$\mathcal{L} f_r = \Phi_r[f_0, f_1, \dots, f_{r-1}]$$

— Solve exactly and/or or asymptotically (singularities)

Method used on

- Quicksort, Hennequin 1989: 100 moments; nonGaussian law
- Path length in trees, Takacs 1990^+
- Area below walks, Louchard 1984
- In situ permutation, Knuth 1972, Prodinger *et alii*

Path length in Cayley trees: $F(z,q) - ze^{F(qz,q)} = 0$

Moments and reduced singular structure

Pumping moments *ad libidinem*

$$zf_1 \sim \frac{1}{2}\,\frac{1}{(1-T)^2},\ zf_2 \sim \frac{5}{4}\,\frac{1}{(1-T)^5},\ zf_3 \sim \frac{45}{4}\,\frac{1}{(1-T)^8}$$

Lemma.

$$\begin{cases} zf_r(z) \sim \dfrac{C_r}{(1-T(z))^{3r-1}} \\ 2C_r = (3r-4)rC_{r-1} + \displaystyle\sum_{j=1}^{r-1} \binom{r}{j} C_j C_{r-j} \quad r \geq 1. \end{cases}$$

Tree decomp. $\mapsto$ Functional Eq. $\mapsto$ Quadratic recurrence

By singularity analysis implies

$$\boxed{E\{(d_{n,n-1})^r\} \sim \frac{-\Gamma(\frac{-1}{2})}{\Gamma(\frac{3r-1}{2})} C_r \left(\frac{n}{2}\right)^{3r/2}}$$

$$\Gamma(s) := \int_0^\infty e^{-t} t^{s-1}\, dt$$

Airy functions

Quadratic recurrence $\mapsto$ Riccati ODE $\mapsto$ Linear ODE

$$y' = y^2 + by + c \quad \mapsto \quad y = -\frac{Y'}{Y}$$

The GF of moment coefficients diverges (taken in asymptotic sense) but it is expressed in terms of Airy function.

$$\sum_r C_r \frac{z^r}{r!} = -\frac{I_{2/3}(\frac{1}{3z})}{I_{2/3}(\frac{1}{3z})}$$

Airy solution to $Y'' - zY = 0$

$$\mathrm{Ai}(z) = \frac{1}{\pi}\int_0^\infty \cos\left(\frac{1}{3}t^3 + zt\right)\, dt$$

Use Airy–Bessel asymptotics to get all moments.

Moment problem

A classical theorem: *if the "Moment generating function"*

$$M(z) = \sum_r \mu_r \frac{z^r}{r!}$$

has nonzero radius of convergence, then the law is uniquely determined by its moments.

A corollary: Convergence of moments implies convergence of distributions

EXAMPLE. $w(x) = e^{-x}$, $\mu_r \equiv r! \implies M(z) = \frac{1}{1-z}$

Theorem. For almost full tables, convergence to the Airy distribution,

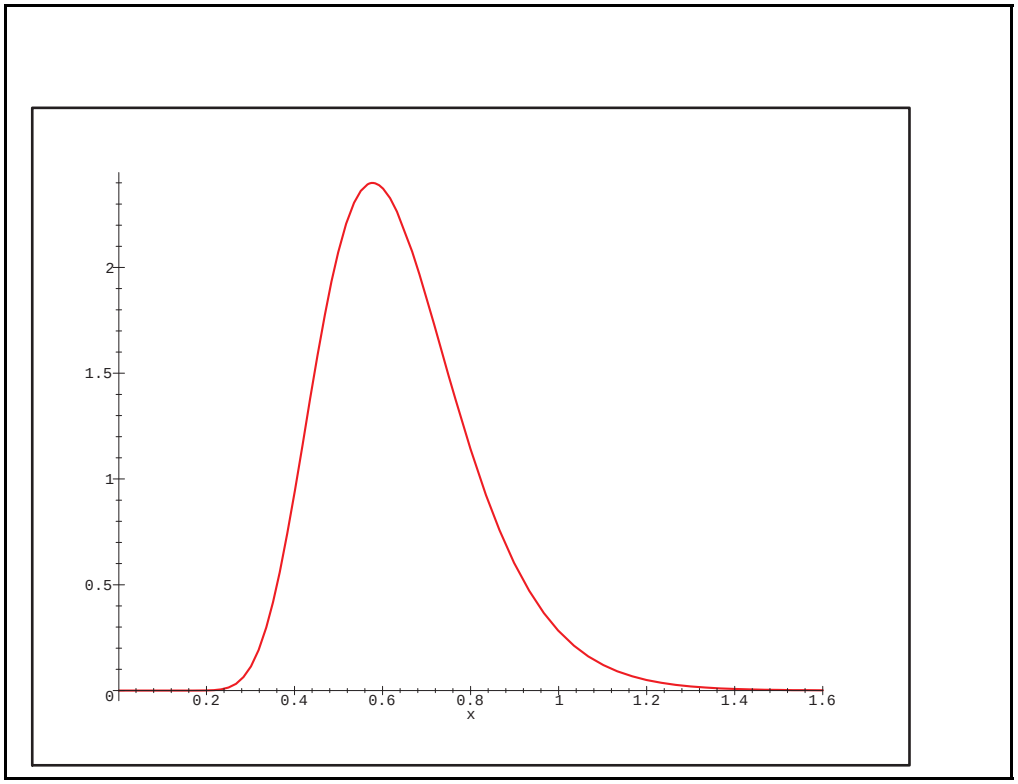
$$\Pr\{\frac{d_{n,n-1}}{(n/2)^{3/2}} \le x\} \to \Pr\{X \le x\} \qquad (n \to \infty),$$

where X is Airy distributed. $\mathbf{E}[X^r] = -\frac{\Gamma(-\frac{1}{2})}{\Gamma(\frac{3r-1}{2})}\,\Omega_r$.

$$\sum_{r\ge 0} \Omega_r \frac{w^r}{r!} = -\frac{\Phi_{2/3}(w)}{\Phi_{-1/3}(w)}$$

$$\begin{aligned}\Phi_\nu(w) \quad = \quad & 1 - (4\nu^2-1)\left(\frac{w}{24}\right) + \frac{(4\nu^2-1)(4\nu^2-9)}{2!}\left(\frac{w}{24}\right)^2 \\ & - \frac{(4\nu^2-1)(4\nu^2-9)(4\nu^2-25)}{3!}\left(\frac{w}{24}\right)^3 + \cdots .\end{aligned}$$

The Airy desnity



7 Sparse tables

A table with m cells and n elements has $g = m - n$ "gaps".

```
SparseTable := <Full> *  ...  * <Full>
               (----  g times ----)
```

Bivariate GF is: $(F(z,q))^g$

The analysis can be "recycled"

Theorem. For α-sparse tables, $\alpha = \dfrac{n}{m}$, mean and variance:

$$\begin{aligned}
&\mathbf{E}[d_{m,n}] = \tfrac{n}{2}(Q_0(m,n-1)-1),\\
&\mathbf{E}[d^2_{m,n}] = \tfrac{n}{12}\Big((m-n)^3+(n+3)(m-n)^2+(8n+1)(m-n)+5n^2+4n-1\\
&\quad -((m-n)^3+4(m-n)^2+(6n+3)(m-n)+8n)Q_0(m,n-1)\Big).
\end{aligned}$$

$$Q_0(m,n) := 1 + \frac{n-1}{m} + \frac{(n-1)(n-2)}{m^2} + \frac{(n-1)(n-2)(n-3)}{m^3} + \cdots$$

$$\begin{aligned}
\mathbf{E}[d_{m,n}] &= \frac{\alpha}{2(1-\alpha)}n - \frac{\alpha}{2(1-\alpha)^3} + O(n^{-1}),\\
\mathbf{Var}[d_{m,n}] &= \frac{6\alpha - 6\alpha^2 + 4\alpha^3 - \alpha^4}{12(1-\alpha)^4}n - \cdots
\end{aligned}$$

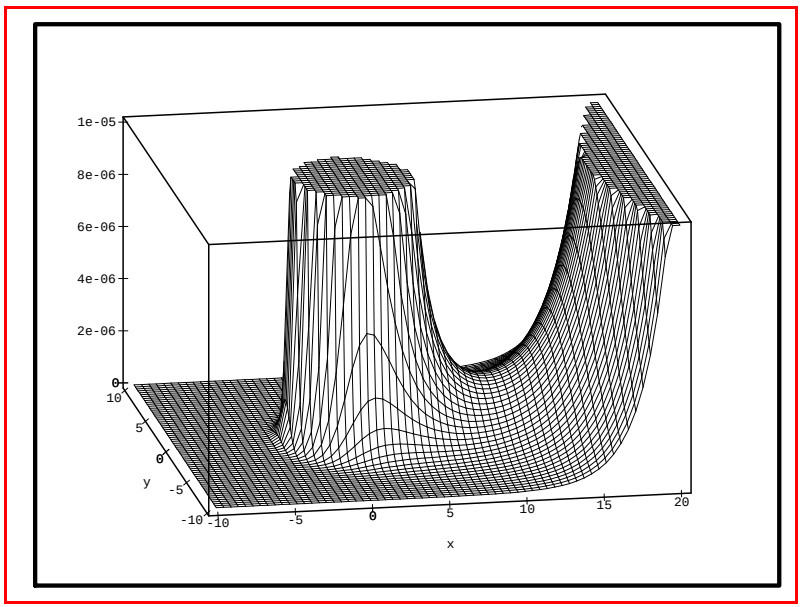
Flajolet-Poblete-Viola (1997); Knuth (1997)

The limit distribution

Theorem. A Gaussian law.

Proof. Integral of large powers by saddle point

$$[z^n](F(z,q))^{m-n} = \frac{1}{2i\pi}\oint (F(z,q))^{m-n}\,\frac{dz}{z^{n+1}}$$



Plus continuity theorem for characteristic functions $q = e^{i\theta}$

Corollary. Works for large assemblies!

E.g. [Mahmoud] Distribution sorts with $O(n)$ buckets.

Methodology again! Specific method is general!

The process used in the proof of the last theorem is in fact very general and we encapsulate it into a general statement.

COROLLARY 1. *A Gaussian limit law holds for the coefficients of any "large power,"*

$$[z^n]G(z,q)^{\beta n}, \qquad \beta > 0,$$

(β fixed, $n \to \infty$) provided the following conditions hold:

- (C_1) $G(z,q) = \sum_n g_n(q)z^n$ *has nonnegative coefficients and* $\deg g_n(q) = O(n^\kappa)$ *for some integer* κ.
- (C_2) *There exists some* r *with* $0 < r \leq +\infty$, *such that* $G(z,1)$ *is analytic in* $|z| < r$, *and* $G(0,1) \neq 0$, $G'_z(0,1) \neq 0$.
- (C_3) $\lim_{z \to r^-} zG'_z(z,1)/G(z,1) = +\infty$.
- (C_4) *There exists* n_1, n_2, k_1, k_2 *with* $k_1 \neq k_2$ *such that the coefficients* $[z^{n_1}q^{k_1}]G(z,q)$ *and* $[z^{n_2}q^{k_2}]G(z,q)$ *are nonzero.*

8 Combinatorics

The Airy distribution occurs in

— (1) Full tables for L.P.H.

— (2) Inversions in Trees

— (4) Area below random walks

— (5) Path length in random trees

The Airy coefficients occur in

— **(3)** Enumeration of connected graphs

Why?? Foata, Kreweras, Gessel, Knuth, Spencer, Louchard, Takacs, Wright, etc.

$$F(z,q) = \frac{\displaystyle\sum_{n=0}^{\infty} q^{n(n+1)/2} \frac{z^n (q-1)^{-n}}{n!}}{\displaystyle\sum_{n=0}^{\infty} q^{n(n-1)/2} \frac{z^n (q-1)^{-n}}{n!}}$$

$$F(z,q+1) = \sum_{n,t} \gamma(n, n+t-1) q^t \frac{z^{n-1}}{(n-1)!},$$

Hashing $\mapsto$ Inversions $\mapsto$ Graphs $\mapsto$ Paths $\mapsto$ Tree P.L.

LPH has Airy distrib.

Knuth 1997; Mallows–Riordan 1968

$$\partial_z F = F \cdot \mathrm{H}\, F$$

Set $z \mapsto z(1-q)$ and get connected graph GF.

$$F(z,q) \quad = \quad \frac{\displaystyle\sum_{n=0}^{\infty} q^{n(n+1)/2} \frac{z^n (q-1)^{-n}}{n!}}{\displaystyle\sum_{n=0}^{\infty} q^{n(n-1)/2} \frac{z^n (q-1)^{-n}}{n!}}$$

$\Longrightarrow$ "Closed form" but moments are still "hard" to find!

Hashing $\mapsto$ Inversions $\mapsto$ Graphs $\mapsto$ Paths $\mapsto$ Tree P.L.

Foata ca 1971; Knuth 1997, Kreweras 1980. Almost full table are combinatorially equivalent to trees.

Correspondence: Connect a key/car to (immediately before) where it wanted to land.

Inversions in trees have Airy distrib.

Hashing $\mapsto$ Inversions $\mapsto$ Graphs $\mapsto$ Paths $\mapsto$ Tree P.L.

Correspondence: Depth-first search as a combinatorial correspondence.
Gessel–Wang 1979. Wright 1977 Count connected graphs by excess. Janson et al. 1993 The Giant paper

Graphs by excess counted by Airy coeff.

Hashing $\mapsto$ Inversions $\mapsto$ Graphs $\mapsto$ Paths $\mapsto$ Tree P.L.

Correspondence: Breadth-first search as a combinatorial correspondence.

Spencer 1997. Leads to Poisson walks. Louchard 1984. Analyse area by moments and/or Brownian motion Connects to area of Dyck/Catalan paths by universality of Brownian motion

$$\cfrac{1}{1-\cfrac{qz}{1-\cfrac{q^2z}{1-\cfrac{q^3z}{\ddots}}}} = q\text{-exponentials}$$

Excursion area has an Airy distribution

Hashing $\mapsto$ Inversions $\mapsto$ Graphs $\mapsto$ Paths $\mapsto$ Tree P.L.

Correspondence: Catalan (Dyck) walks as traversal sequences of Catalan trees.

Takacs 1990-1994. Moment methods apply to simple families.

Path length in simple trees has Airy distrib.

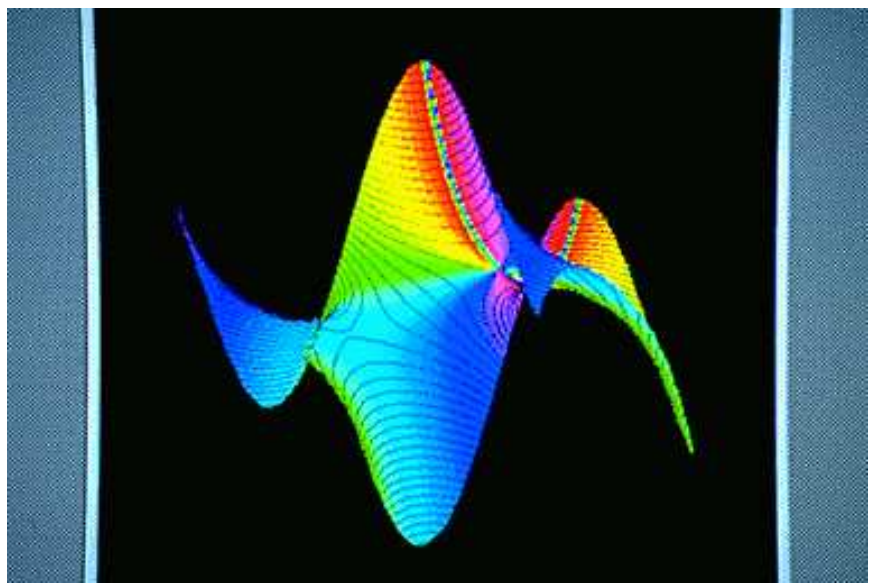
Airy phenomena have a large degree of **universality**

From Flajolet-Salvy (1995): analytic explanation

Lemma. Let $U(z) := \sum u_k z^k$. Then, $q = e^{-t}$,

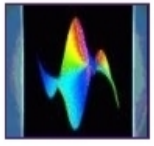
$$\sum_k q^{k^2/2} u_k = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} U(e^{x\sqrt{t}})\, dx$$

Combinatorics $\mapsto$ Coalescent saddles $\mapsto$ **Airy**



Cf. Prellberg 1993. Flajolet-Noy 1997 Crossing in chord systems.

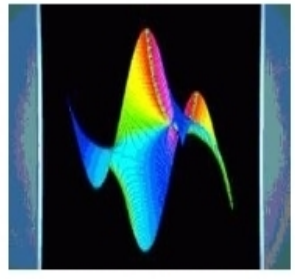
"Monkey Saddle" as the ALGO project's logo.



Algorithms Project's Logo

Welcome!	Research Topics	People	Publications	Seminars	Software	On-Line Applications	Jobs & Internships

Our logo shows the behaviour in the complex plane of the generating function of connected graphs counted according to number of nodes and edges. In critical regions, two saddle points coalesce giving rise to a so-called "monkey saddle" (a saddle that you'd use if you had three legs!)



The fine analysis of this coalescence is crucial to the understanding of connectivity in random graphs. This problem has applications in the design of communication networks and it relates to a famous series of problems initiated by Erdös and Renyi in the late 1950's. See the paper [Janson, Knuth, Luczak, Pittel: The birth of the giant component. Random Structures Algorithms 4 (1993), no. 3, 231--358]. As said by Alan Frieze in his review [**MR**94h:05070]:

> "*This paper and its predecessor [**MR**90d:05184] mark the entry of generating functions into the general theory of random graphs in a significant way. Previously, their use had mainly been restricted to the study of random trees and mappings. Most of the major results in the area, starting with the pioneering papers of P. Erdös and A. Renyi [**MR**22#10924] have been proved without significant use of generating functions. However, at the early stages of the evolution of a random graph we find that it is usually not too far from being a forest, and this allows them an entry...*"

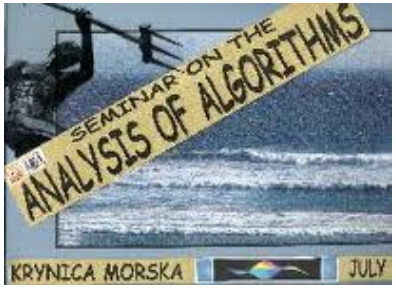
The icon was generated by Maple code like this:

```
f:=u^3/3-3*u+1;
plots[complexplot3d]([Re(f),argument(f)],
              u=-4-3*I..4+3*I,style=patchcontour,
                     contours=30,numpoints=50*50);
```

18 years have already passed

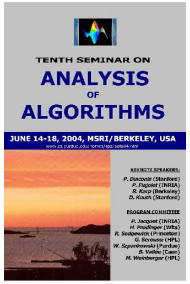
from our first scientific meeting!

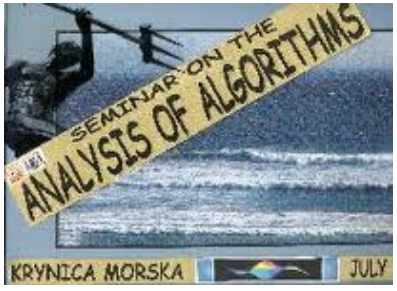
Since then


SEMINAR ON THE
ANALYSIS OF ALGORITHMS
KRYNICA MORSKA
JULY


TATIHOU FRANCE
SEVENTH SEMINAR ON
ANALYSIS OF ALGORITHMS
JULY 2 - 8, 2001

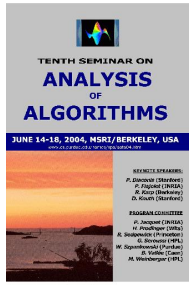
. . . Philippe dreamed with . . .

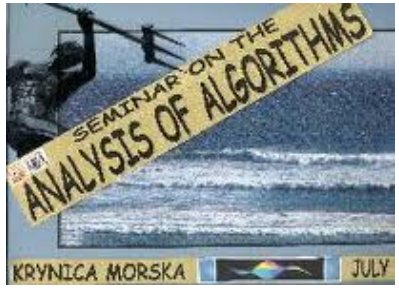


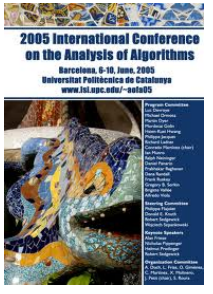


ANALCO05 Workshop on Analytic Algorithmics & Combinatorics
January 22, 2005
Sheraton Vancouver Wall Center
Vancouver, British Columbia, Canada

. . . and lead the construction of . . .

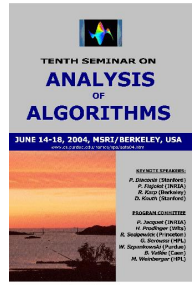


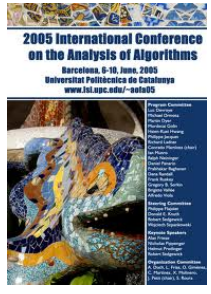




ANALCO05 Workshop on Analytic Algorithmics & Combinatorics
January 22, 2005
Sheraton Vancouver Wall Center
Vancouver, British Columbia, Canada

. . . a strong research group.





It is up to us to keep this dream alive.

AofA'11

22th International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms

Conference dedicated to the memory of Philippe Flajolet

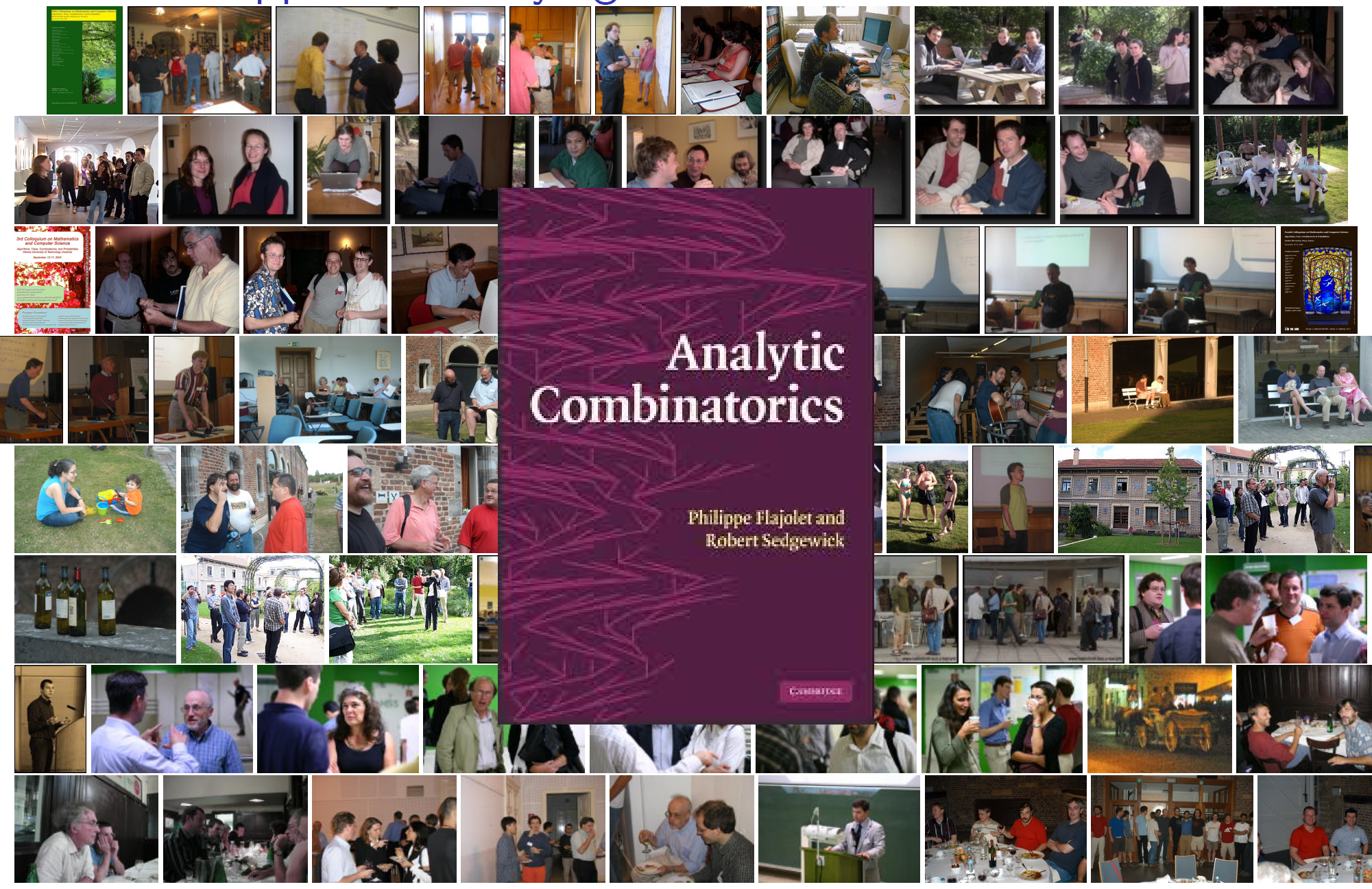
- Home
- Letter
- Obiutary
- Previous
- Lectures
- Plenary speakers
- Committees
- Travel
- Social events
- Conference fees
- Participants
- Participant Form

Participants

Our photo:

We have to keep working as we have always done.

Philippe will always guide us in difficult moments.



This is the way to honor him and his work.