INTRODUCTION 1

Introduction to Hashing Papers by Philippe Flajolet

Alfredo Viola

1. Some historical context

The idea of hashing seems to have been originated by Luhn, in an internal IBM memorandum in January 1953. The first major paper published in the area is the classic article by Peterson in 1956, where he defines open addressing in general and gives empirical statistics about linear probing hashing. He also notices the degradation in performance when lazy deletions are presented. Nevertheless, as noted by Knuth in [Knu98b], the word “hashing” to identify this technique appears for the first time in the literature in the survey of Morris in 1968, although it had been in common usage for several years.

On the other hand, one of the first mathematical challenges in the early 60’s of the nascent Computer Science was the design of models to understand and predict the practical behavior of access methods to data. Hashing methods had shown to have very good empirical complexity, although no mathematical proof of their behavior had been proposed. In 1962 Knuth presents a solution for linear probing hashing, and this milestone is considered to be the the first algorithm ever analyzed as well as the origin of the Analysis of Algorithms.

2. Linear probing and the Ramanujan’s Q-function

Philippe Flajolet was dazzled by the analysis of hashing algorithms. The mathematical properties presented in these problems have seemed to be emerged from a box full of surprises! In this regard, Flajolet was specially interested in the analysis of linear probing hashing. Most of Flajolet’s motivations in linear probing are clearly presented in the introduction of the wonderful survey together with Chassaing in 2003 [PF175] oriented to french students, whose introduction can be summarized as follows:

“Discrete and continuous mathematics willingly and harmoniously encounter and complement. We would like to illustrate this thesis by presenting a classical problem with several ramifications – the analysis of linear probing hashing. This example is typical of the analysis of algorithms, a topic pioneered by Knuth and which is at the intersection of computer science, combinatorics, and probability theory”. Historical and scientific motivations include “questions asked by Ramanujan to Hardy in 1913, a summer work in 1962 by Knuth that is at the origin of the analysis of algorithms in computer science, the research in combinatorics done by the statistician Kreweras,
several encounters with the model of random graphs by Erdős and Rényi, some complex and asymptotic analysis, trees generated by specific Galton-Watson processes, and, to conclude, a bit of processes like the ineffable Brownian motion!" All this contributes to a “very precise understanding of a very simple discrete random problem”.

His first approach to linear probing is paper [PF121], that presents a solution to the question asked by Ramanujan to Hardy in 1913. More specifically, the goal is to present a complete proof of the following assertion made by Ramanujan:

\[ \frac{1}{2} e^n = 1 + \frac{n}{1!} + \frac{n^2}{2!} + \ldots + \frac{n^n}{n!} \theta(n), \]

where

\[ \theta(n) = \frac{1}{3} + \frac{4}{135(n + k)}, \quad \frac{2}{21} \leq k \leq \frac{8}{45}. \]

Knuth in [Knu98b] presents his complete analysis of linear probing, where a key rôle is played by a function closely related to \( \theta(n) \), the Ramanujan’s \( Q(n) \)-function. In [Knu97] Knuth uses two variants (\( Q(n) \) and \( R(n) \)) to illustrate asymptotic analysis techniques. These two variants are:

\[ Q(n) = 1 + \frac{n - 1}{n} + \frac{(n - 1)(n - 2)}{n^2} \ldots \]
\[ R(n) = 1 + \frac{n}{n + 1} + \frac{n^2}{(n + 1)(n + 2)} \ldots \]

Since \( Q(n) + R(n) = n! e^n / n^n \) and \( \theta(n) = \frac{1}{2} (R(n) - Q(n)) \) then the problem to be solved can be rephrased in terms of the Ramanujan’s \( Q(n) \) function.

The approach followed by Ramanujan and later authors all make use of real integral representations of the \( Q(n) \) and proceed using the Laplace method for the asymptotic evaluation of integrals. To completely solve the problem, the authors “develop a complex integral representation based on a generating function of the \( Q(n) \) from which yet another proof of its asymptotic expansion follows”. Then, the approach “is in fact a hybrid of singularity analysis and saddle point in the following sense: It starts with an integral representation based on an expansion essentially dictated by the singularity analysis method; then a suitable change of variable is introduced that causes the integration contour to pass through a saddle point”. The analysis is then completed after presenting effective error bounds for the intervening integrals and then weaving together all lemmas and estimates to complete the proof of the main inequality.

The last section is devoted to reflections on the various methodological alternatives to estimate asymptotically these type of sequences. In this regard, a succinct and pedagogical comparison of the Laplace method, Singularity Analysis, the Darboux’s method and the saddle point methods is presented. Most of Flajolet’s papers have this special touch about discussing at depth the methodologies used to solve the problem at hand, and this paper is not an exception.

As a historical note, the paper is “Dedicated to D. E. Knuth on the occasion of the 30th anniversary of his first analysis of an algorithm in 1962”.
3. Linear probing, analytic combinatorics and the Airy distribution

Paper [PF142] presents key contributions to the analysis of linear probing, among them the formal definition of the Airy distribution that appears as a limit law in multiple combinatorial problems. At the time of working in [PF142], he had already started with Robert Sedgewick to compile and to set the foundations of Analytic Combinatorics [PF111; PF110; PF118; PF137] in preparation to his major scientific legacy [PF201]. It is clear that most of his scientific effort was devoted to develop general methodological contributions in Analytic Combinatorics, and that his papers can also be seen as non-trivial exercises that illustrate the use of these methods. His web page contains, in addition to the collection of his papers, a very impressive set of lectures that guide us in the understanding of their main methodological contributions. His approach to Analytic Combinatorics is clearly presented in the preface of [PF201]:

“Our initial motivation when starting this project was to build a coherent set of methods useful in the analysis of algorithms . . . This book, Analytic Combinatorics, can then be used as a systematic presentation of methods that have proved immensely useful in this area; . . . The present book expounds this view by means of a very large number of examples concerning classical objects of discrete mathematics and combinatorics”.

In this regard, one of his main personal challenges in relation with linear probing is explicitly stated in [Knu98a]:

“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 [PF130]:

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 [PF130] was published in 1996”.

In [Knu98a], Knuth’s approach to the problem (construction cost of hash tables) relates linear probing with Wright’s enumeration of sparse connected graphs. This relation was based on the work of Mallows and Riordan who showed that labeled trees with a small number of inversions are related to labeled graphs that are connected and sparse, and the work of Kreweras who related the inversions of trees to the so-called “parking problem”.

A complementary approach to the problem is presented independently, at the same time, in [PF142]. Similar to [Knu98a], for confined (“almost full”) tables, the same recurrence to the main generating function $F_n(q)$ is derived. More specifically, let $F_{n,k}$ be the number of ways of of creating an almost full table with $n$ elements and total displacement $k$. The corresponding bivariate generating function is then

\begin{equation}
F(z, q) = \sum_{n, k \geq 0} F_{n,k} q^k z^n/n!.
\end{equation}
A natural binary tree decomposition of almost full tables gives rise to a recurrence on the $F_n(q) = n! [z^n] F(z, q)$:

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

leading to the basic functional equation

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

Equation (2) is solved in [Knu98a] leading to the identity

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

where the right side of this equation is well known as the exponential generating function for labeled connected graphs.

In [PF142], on the other hand, from the functional equation (3) moments, in either exact or asymptotic form, are obtained by a “pumping” process. This approach leads to a limit law of the Airy type defined in [PF142]. Let $d_{m,n}$ be the random variable for the total displacement of a table of length $m$ with $n$ elements, under the assumption that all $m^n$ hash sequences are equally likely. Then, Theorem*For almost full tables, the distribution of the random variable $d_{m,n}/(n/2)^{3/2}$, converges to the Airy distribution, in the sense that, pointwise for each $x$,

$$\Pr \left\{ \frac{d_{m,n-1}}{(n/2)^{3/2}} \leq x \right\} \to \Pr X \leq x \quad (n \to \infty),$$

where $X$ is Airy distributed in the sense of Definition 1. The same property holds for completely full tables and the random variable $d_{n,n}/(n/2)^{3/2}$.

In both papers, the symbolic method handles the analysis of the general case ($n$ elements in a table of length $m$), since “such a table then decomposes as a labeled product of $m - n$ clusters (sometimes also figuratively called “islands”) that are, up to relabeling, almost full tables” [PF142]. The behavior of such tables turns out to be much more tame than that of full tables. In this sense, the limit law of the total displacement $d_{m,n}$ in tables with filling ratio $\alpha = n/m$ ($\alpha < 1$) is asymptotically Gaussian as $n \to \infty$.

Furthermore, and in consistency with Flajolet’s interest in presenting general methodological contributions, the “delicate” saddle point analysis used to prove this normal limit theorem gives rise to the following observation to justify Corollary 1 in [PF142]: “the process used in the proof of the last theorem is in fact very general and we encapsulate it into a general statement”. This result is then cited in [PF159] to assert that Theorem 3 “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.”
As a historical note, Flajolet knew identity (4) at the last session of the “Average Case Analysis of Algorithms” meeting held at Dagstuhl in July 1997, where the combinatorial solution proposed in [PF142] had been presented. He immediately realized about all the deep non-trivial relations between linear probing, tree inversions, tree path length, and area under excursions, most of them already known in the literature, although in independent work presented in several unrelated papers. He was dazzled by all these relations, that are briefly (although deeply!) presented in the last section of [PF142]. All these “unexpected” mathematical coincidences and the connections with the Airy distribution motivated him to work in other problems where the “Airy phenomena” appears (Volume II, Chapter 2). The Algorithms Project’s logo [Fla] is a first-class witness of this productive and moving stage in his scientific career. Both [PF142] and [Knu98a] appear, mutually dedicated, as companion papers in a special issue of Algorithmica dedicated to Flajolet’s 50th anniversary.

4. Other hashing problems

Philippe Flajolet has also published other papers with technical contributions related with hashing. Two pieces of work are briefly presented in this subsection.

The first one to mention is the the analysis of generalizations of the birthday paradox, coupon collectors, caching algorithms and self-organizing search presented in [PF073; PF100]. Working on very general probabilistic settings based on formal languages (presented in Volume V, Chapter 1), these papers present deep connections between hashing and random allocation problems.

The second piece of work is presented in [PF037; PF036]. Analytic Combinatorics, as presented in the Preface of [PF201], “aims at predicting precisely the properties of large structured combinatorial configurations, through an approach based extensively on analytic methods”. Moreover “Given its capacity of quantifying properties of large discrete structures, Analytic Combinatorics is susceptible to many applications . . . the analysis of algorithms and data structures in computer science has served and still serves as an important incentive for the development of the theory”. Even though these papers have been published long before the project leading to [PF201] started, one of the main goals of Flajolet’s work through his scientific career is the precise evaluation of important parameters used in “practical” algorithms. The summary of [PF037] is very clear in this regard:

“A class of trees 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”.

There is a strong connection between digital searching and schemes for dynamic hashing, since operations on b-tries are isomorphic to operations performed on a dynamic hashing table with page capacity equal to b. The part of the tree in dynamic hashing schemes obtained by pruning leaves (that contain record or page references) is called the directory. The “paper is relative to the performance evaluation of Dynamic Hashing andExtendible Hashing”. The main parameter studied is the depth of
The directory that is closely related with the most important parameter in Extendible Hashing - size of the representation of the directory. The paper proves that the expectation of the depth is $O(\log n)$ and that the expected size is $O(n^{1+1/b})$, where $n$ is the number of records in the hash table.

The probabilistic model considered is the Bernoulli model which intuitively corresponds to the uniform measure over $n$-tuples of infinite binary sequences. The problem is then translated into a “balls and urn” model on the interval $[0,1]$. The saddle point method is the main methodological tool to derive asymptotics, that includes very precise characterization of the oscillations. At the end, the paper presents a discussion of implementation issues based on these analytical results.

Every paper by Flajolet contains some “unexpected” relation that involves the problem at hand. In [PF036] this problem has been shown to be closed related with branching processes in trie searching and polynomial factorization (Lazard’s algorithm).

To conclude, it is worth mentioning that the height of general tries is thoroughly studied in [PF161], in the framework of dynamical systems (Volume II, Chapter 3).

5. The legacy of Philippe Flajolet

In this brief introduction I invite you to discover together part of the light that Flajolet gives us behind his deep and beautiful contributions presented in his analysis of hashing algorithms. The rest is up to each of us to discover. Every time we read again each of his papers (even if we are coauthors!), there is room to always learn something new, or better understand some important idea. Flajolet’s work will always be a source of inspiration and light in our everyday research.
Bibliography


