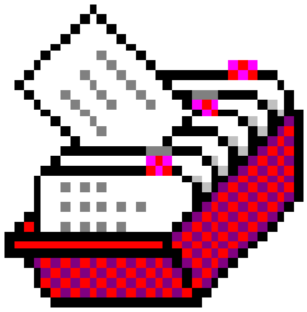
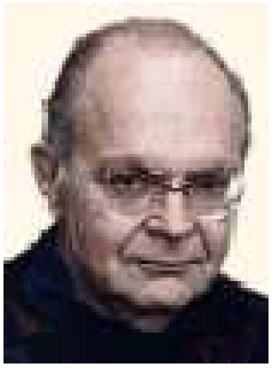
ON THE ANALYSIS OF LINEAR PROBING HASHING



Philippe Flajolet, INRIA Rocquencourt (France)


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

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



With P. POBLETE (Chili) and A. VIOLA (Uruguay)

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.

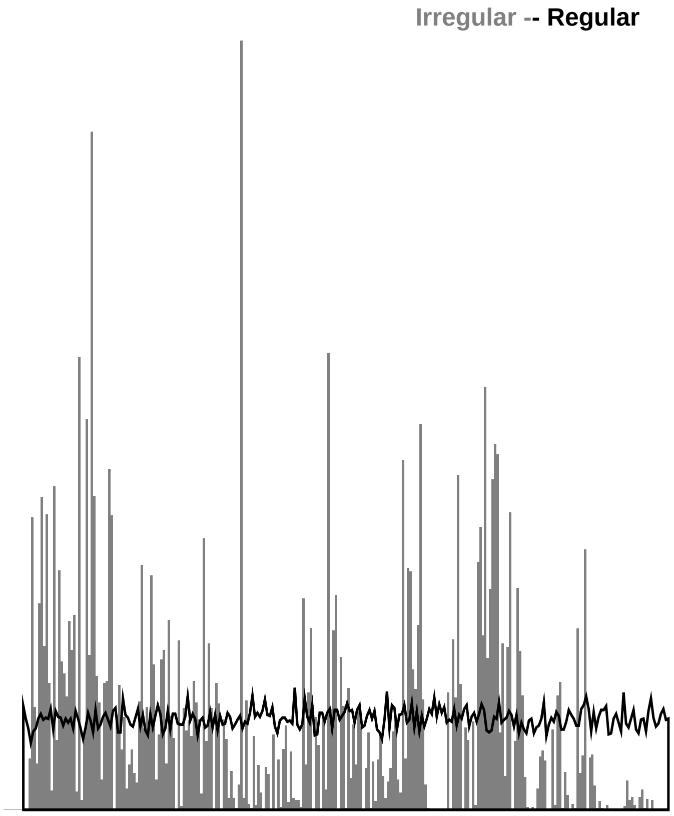
Works well for constrained data or random uniform data.

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



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

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).

0	1	2	3	4	5	6	7	8	9	
							ON			7
					THE		ON			5
		ANALYSIS			THE		ON			2
		ANALYSIS			THE		ON		OF	9
		ANALYSIS	LINEAR		THE		ON		OF	3
PROBING		ANALYSIS	LINEAR		THE		ON		OF	$\boxed{9}$
PROBING		ANALYSIS	LINEAR	HASHING	THE		ON		OF	4
PROBING		ANALYSIS	LINEAR	HASHING	THE	BY	ON		OF	5
PROBING		ANALYSIS	LINEAR	HASHING	THE	BY	ON	PHILIPPE	OF	3

http://orange.mcs.dundee.ac.uk:8080/CS202/tuts/mosaic/hashing/HashDemo/index.html

$h(x) =$ ASCII sum mod 10

2 Random Allocations

Collisions are barely avoidable, distrib. is irregular.

$\boxed{\text{m} = \# \text{ Urns}, \qquad\qquad \text{n} = \# \text{ Balls}}$

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:

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

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

3 Linear Probing Hashing

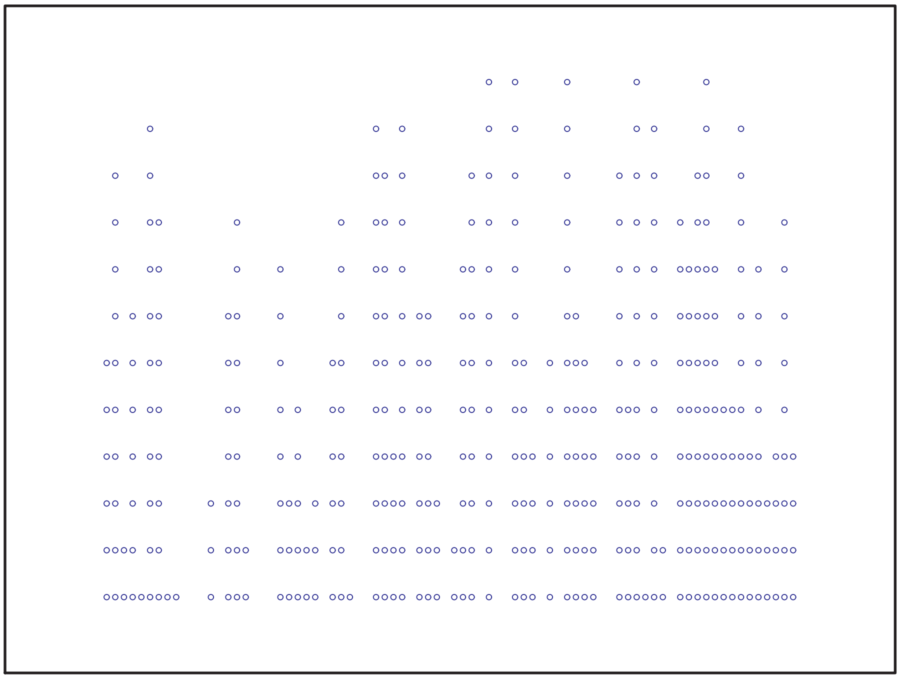
LINEAR PROBING: Physics of "What goes on"?

— 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 $\sim \dfrac{m}{2}$.



QUANTIFY?

Theorem K. • *α-sparse tables*, $\alpha = n/m$

$$\textit{L.P.H.} \quad C_n^+ \sim \frac{1}{2}\left(1 + \frac{1}{(1-\alpha)}\right) \quad C_n^- \sim \frac{1}{2}\left(1 + \frac{1}{(1-\alpha)^2}\right)$$

Compare

$$\text{S.C.} \qquad C^+ = 1 + \frac{\alpha}{2} \qquad C^- = 1 + \alpha$$

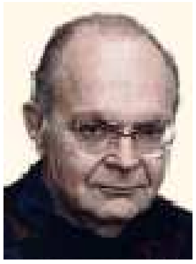
• *Full tables*, $n = m - 1$

$$\textit{L.P.H.} \qquad C_n^+ \sim \sqrt{\frac{\pi n}{2}} \qquad C_n^- \sim \frac{n}{2}.$$

♡ Fine for α small

♠ In a full table mean displacement is $O(\sqrt{n})$, distance to nearest gap is $O(n)$.

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

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

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

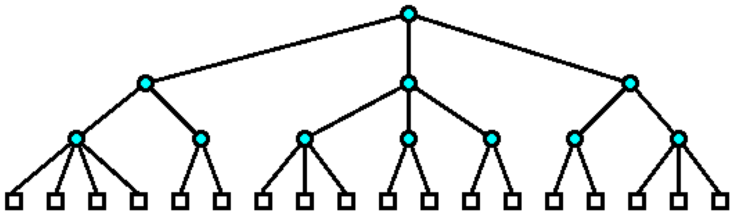
$$\left[{n \atop k}\right] = (n+1)^k \left(1 - \frac{k}{n+1}\right).$$

4 Analytic Combinatorics

Two basic principles $\mapsto$ "dictionaries"

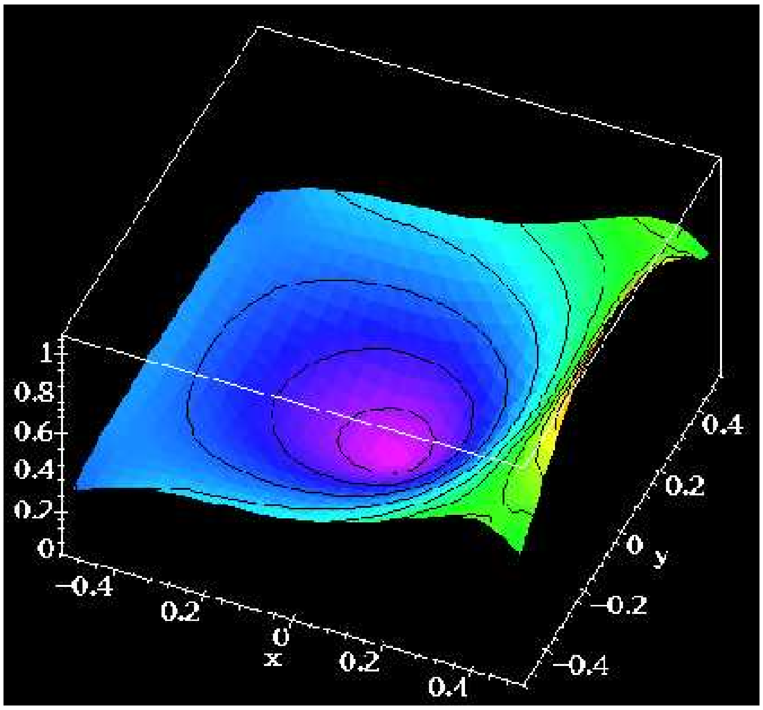
SYMBOLIC METHODS

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)

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

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

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

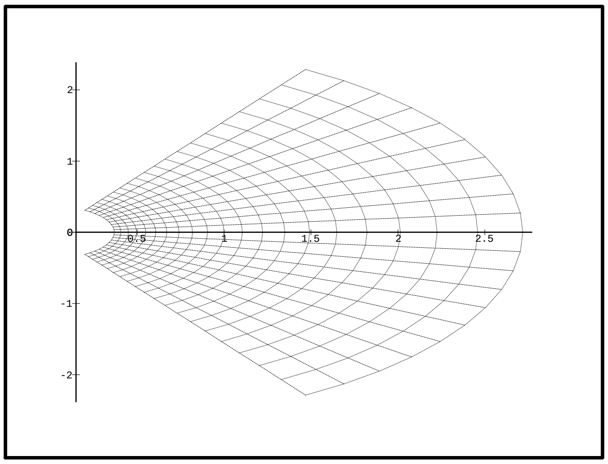
$$\begin{array}{lcl}
\mathbf{A} \cup \mathbf{B} & \mapsto & A(z) + B(z) \\
\mathbf{A} \times \mathbf{B} & \mapsto & A(z) \times B(z) \\
\mathsf{Seq}\ \mathbf{A} & \mapsto & \dfrac{1}{1 - A(z)} \\
\mathsf{Set}\ \mathbf{A} & \mapsto & \exp(A(z)) \\
\mathsf{Cycle}\ \mathbf{A} & \mapsto & \log \dfrac{1}{1 - A(z)}
\end{array}$$

COMPLEX ASYMPTOTICS

Dictionary (II)

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

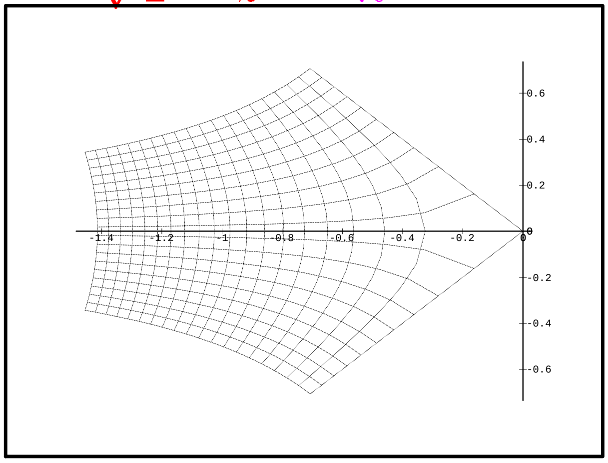
$\exp(z)$



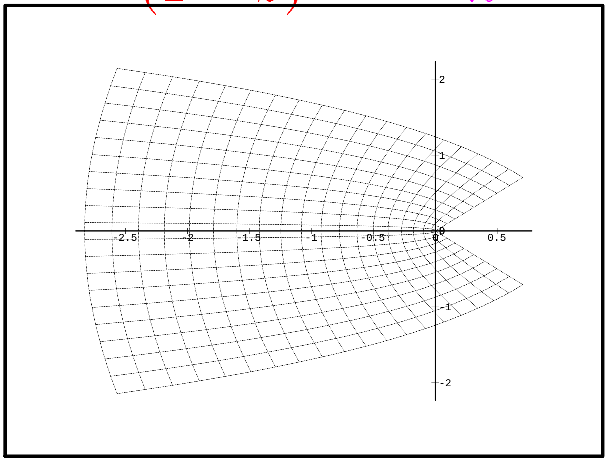
Point of singularity. $\neg\exists \left. \frac{d}{dz} f(z) \right|_{z_0}$

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

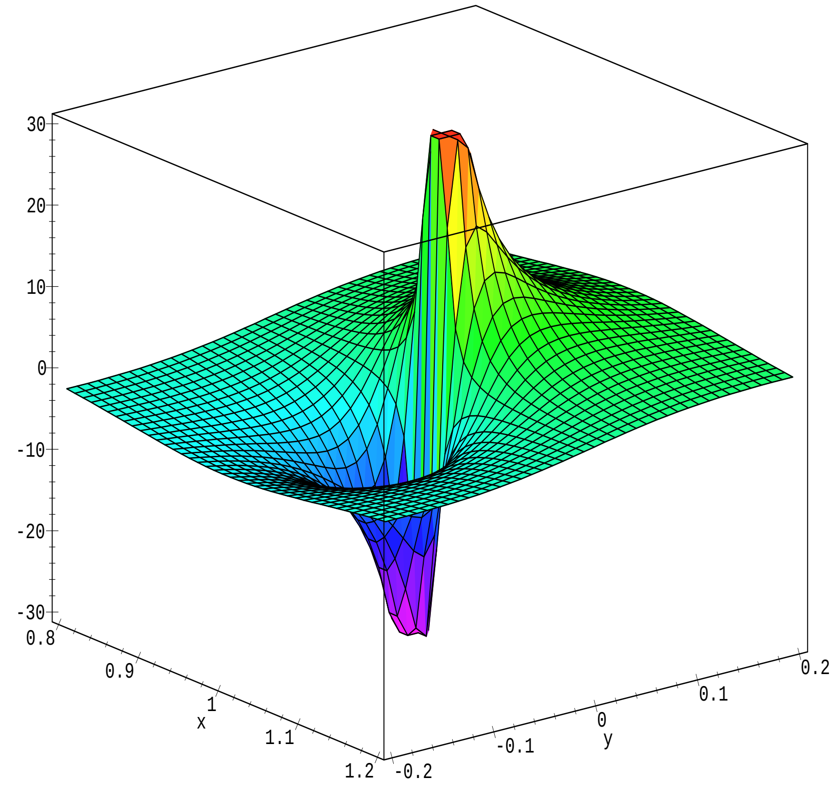
$-\sqrt{1-z} \mapsto n^{-3/2}$



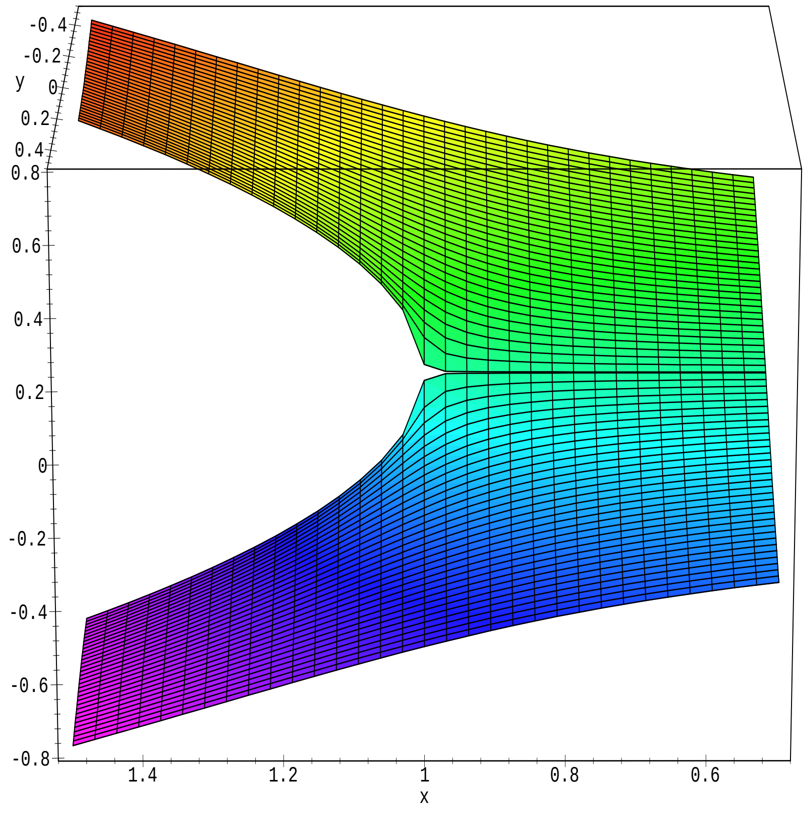
$-(1-z)^{3/2} \mapsto n^{-5/2}$



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



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



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 $\mapsto \int$, $\quad C_n = A_{n-1}$

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

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

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

A nonlinear ODE translates Linear Probing

$$\begin{array}{lcl} \mathit{Table} & = & \mathit{Table} \text{ `}\star\text{' } \mathit{Table} \\ F & = & \int (zF)'F \\ T'(1-\frac{1}{T}) & = & \frac{1}{z} \qquad \text{with } T = zF \end{array}$$

Lemma 1.

$$F(z) = \frac{1}{z}T(z) \quad \text{where} \quad T = ze^T$$

Lemma 2. [Lagrange + Eisenstein + Cayley]

$$T(z) = z + 2\frac{z^2}{2!} + 9\frac{z^3}{3!} + \cdots = \sum_n n^{n-1}\frac{z^n}{n!}$$

Defines the Tree function $T(z)$

```
Tree = Root * Set(Tree)
```

[Knuth63] The number of almost full tables (n keys)

$$F_n = (n+1)^{(n-1)}$$

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 $T(z)$

Distributional analysis (almost full tables)

- Record *construction cost* = *total displacement*

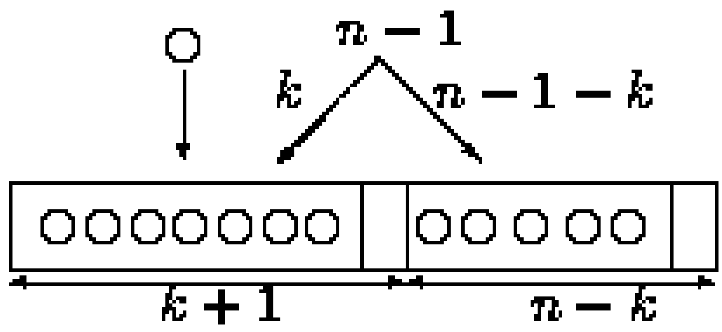


FIGURE 1. The binary tree decomposition of almost full tables.

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

- q–Calculus

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

The construction cost is worst-case quadratic

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] $(\exp(z))^{-1} = (\exp(-z))$

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

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

$$\partial_z F(z) = \partial_z(zF(z)) \cdot F(z)$$
$$\partial_z F(z,q) = \left(\frac{F(z,q) - qF(qz,q)}{1-q}\right) \cdot F(z,q)$$
$$\partial_z F = \mathrm{H}\, F \cdot F$$

Moments result from

— ∂_q, differentiation w.r.t. q

— U, setting $q = 1$

For the r-th moment, apply $U\,\partial_q^r$ to difference-differential equation

Lemma. Commutation rule with $\mathrm{Z}[f] = z \cdot f$

$$\begin{aligned} \mathrm{U\,H} &= \partial_z \mathrm{Z} \\ \mathrm{U}\,\partial_q\,\mathrm{H} &= \mathrm{U}\,\partial_z \mathrm{Z} \partial_q + \frac{1}{2}\mathrm{Z}\partial_z^2 Z, \\ \mathrm{U}\,\partial_q^2\,\mathrm{H} &= \partial_z \mathrm{Z}\,\mathrm{U}\,\partial_q^2 + \mathrm{Z}\partial_z^2\mathrm{Z}\,\mathrm{U}\,\partial_q + \tfrac{1}{3}\mathrm{Z}^2\partial_z^3\mathrm{Z}\,\mathrm{U} \end{aligned}$$

Proof. $\partial_q[n+1]$, Leibniz, &c.

Lemma. *The generating functions associated with moments* $f_r(z) = \mathrm{U}\,\partial_q^r F(z,q)$ *satisfy a linear ODE that involves* $T(z)$.

$$\begin{aligned} z f_1 &= \frac{1}{2}\frac{T^3}{(1-T)^2}, \\ z f_2 &= \frac{1}{12}\frac{T^4(24-11T+2T^2)}{(1-T)^5} \end{aligned}$$

Theorem. [Full tables, exact form of moments]

$$\begin{aligned} \mathbf{E}[d_{n,n}] &= \frac{n}{2}(Q(n)-1) \\ \mathbf{E}[d_{n,n}^2] &= \frac{n}{12}(5n^2+4n-1-8n\ Q(n)). \end{aligned}$$

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

Theorem. [Full tables, asymptotic form of moments]

$$\begin{aligned} \mathbf{E}[d_{n,n}] &= \frac{\sqrt{2\pi}}{4}n^{3/2} - \frac{2}{3}n + \frac{\sqrt{2\pi}}{48}n^{1/2} - \frac{2}{135} + O(n^{-1}), \\ \mathbf{Var}[d_{n,n}] &= \frac{10-3\pi}{24}n^3 + \frac{16-3\pi}{144}n^2 + \frac{\sqrt{2\pi}}{135}n^{3/2} - \cdots \end{aligned}$$

Knuth (1962-3); Flajolet-Poblete-Viola (1997); Knuth (1997)

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

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

— 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

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

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

Details.

$$\partial_z F = F \cdot H$$

$$\mathcal{L}\, Y(z) = R_r(z)$$

$$\mathcal{L}\, Y := Y' \cdot (1-T) - Y \cdot \frac{T(2-T)}{z(1-T)}$$

- Homogenous equation

$$\mathcal{L} Y = 0 \quad \Longrightarrow \quad Y = \frac{e^T}{1-T}$$

- Variation-of-constant method

$$Y(z) = \frac{e^{T(z)}}{1-T(z)} \int_0^z R_r(u) e^{-T(u)}\, du$$

$\boxed{\text{Pumping moments } \textit{ad libidinem}}$

Singularity analysis is enough.

++ [Wright, Janson et al.] Rational function in T.

Airy functions

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

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

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,

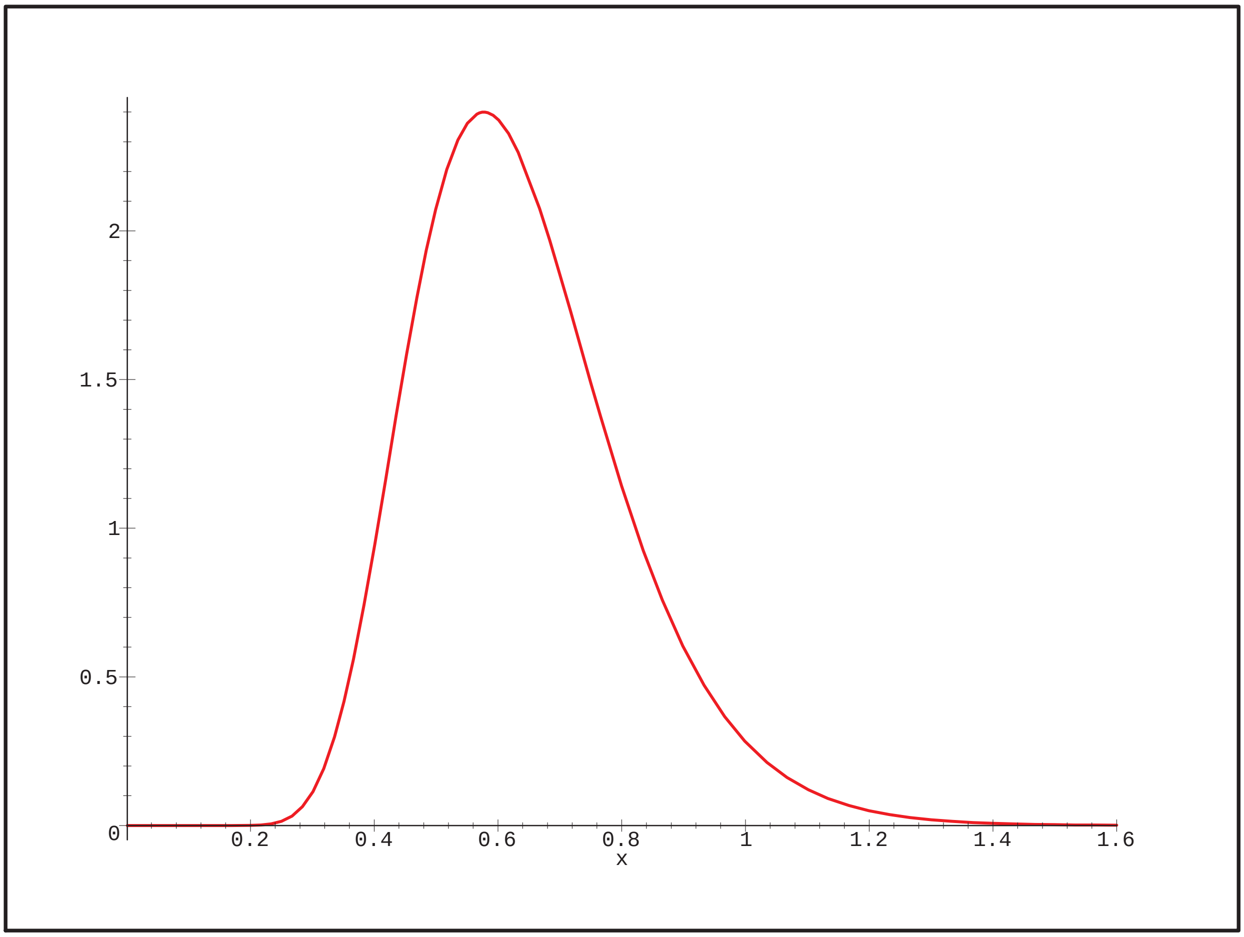
$$\Pr\{\frac{d_{n,n-1}}{(n/2)^{3/2}} \le x\} \to \Pr\{X \le x\} \qquad (n \to \infty),$$

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

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

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

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:

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

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

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

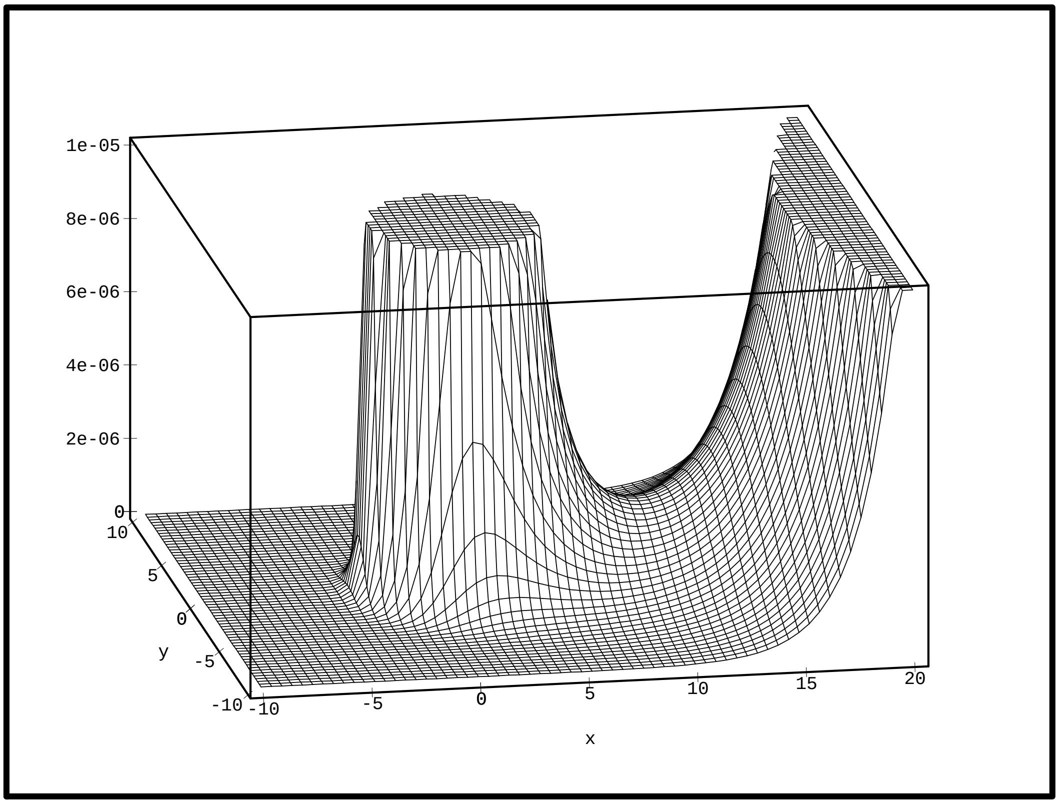
Flajolet-Poblete-Viola (1997); Knuth (1997)

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.

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.

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

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

$\boxed{\text{Hashing}} \mapsto$ 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!

$\boxed{\text{Hashing} \mapsto \text{Inversions}} \mapsto$ 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 $\mapsto$ $\boxed{\text{Inversions} \mapsto \text{Graphs}}$ $\mapsto$ 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 $\mapsto$ $\boxed{\text{Graphs} \mapsto \text{Paths}}$ $\mapsto$ 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 $\mapsto$ $\boxed{\text{Paths} \mapsto \text{Tree P.L.}}$

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.

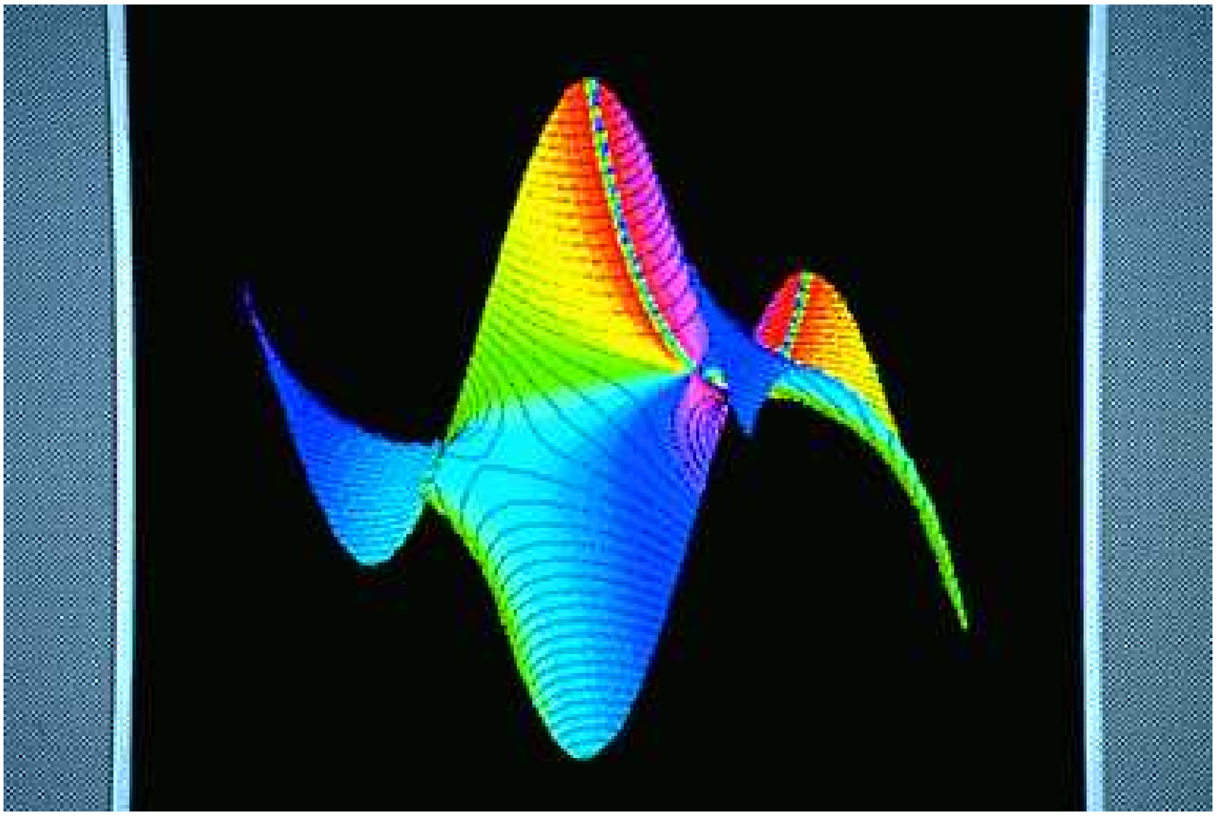
Airy phenomena have a large degree of universality

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.