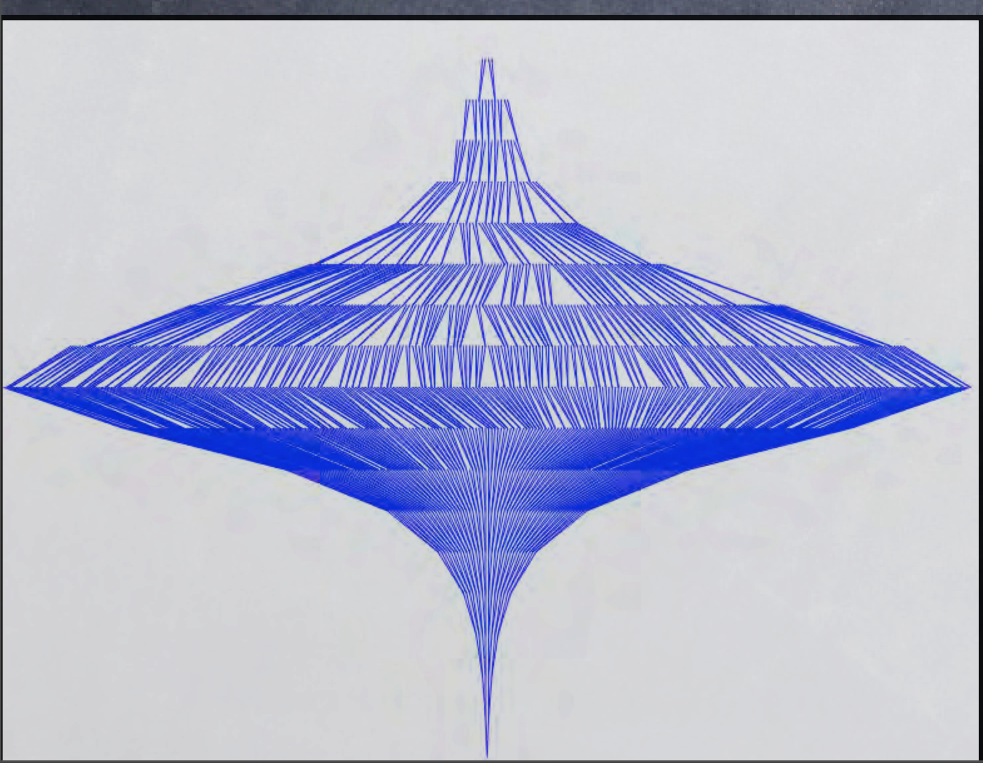
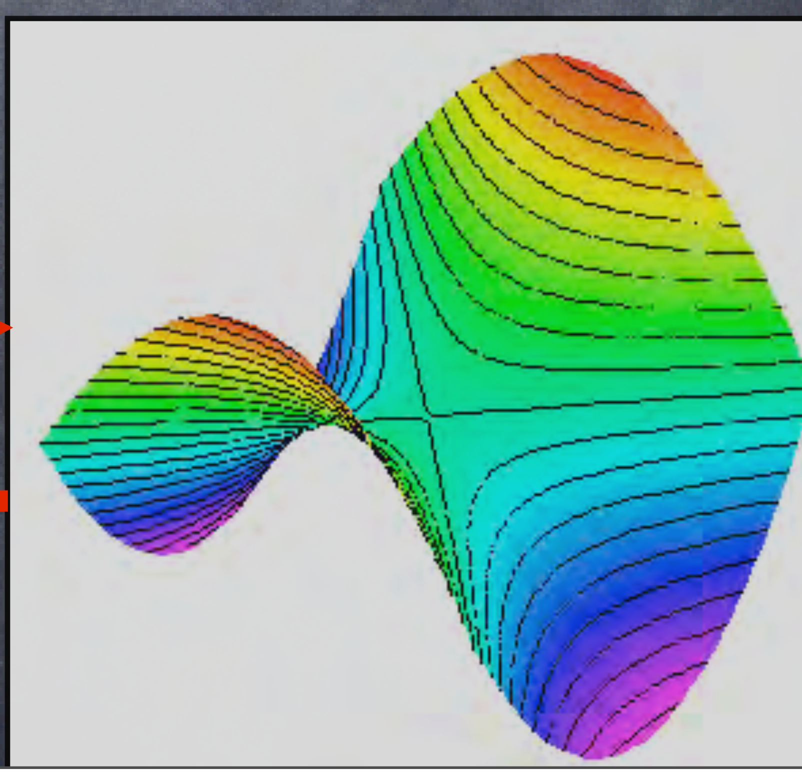
From the Discrete to the Continuous, and Back...

Philippe Flajolet
INRIA, France





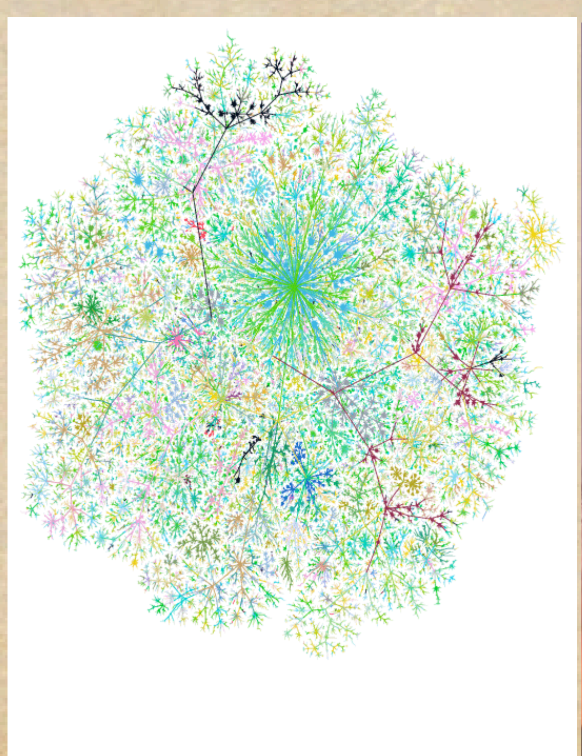
Discrete structures in ...

- Combinatorial mathematics
- Computer science:
  data structures & algorithms
- Information and communication theory
- Models of physical sciences

Discrete structures ...

require quantification...

- What is the "typical" profile of a tree?
- How well-connected is a graph?
- Which patterns are "expected"?



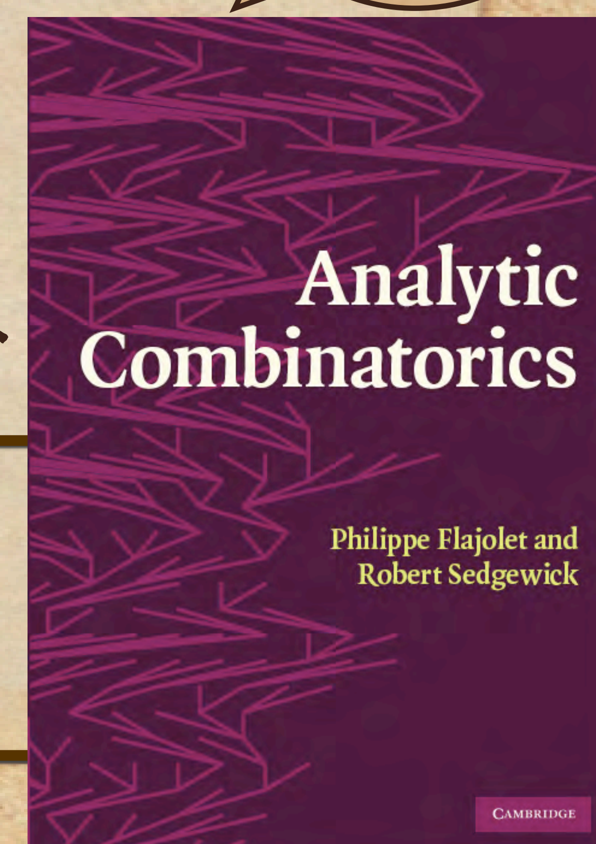

$$\frac{\sqrt{x}}{2}\left\{1+\left(\frac{1}{2}\right)^2\frac{x}{3}+\left(\frac{1.3}{2.4}\right)^2\frac{x^2}{5}+\left(\frac{1.3.5}{2.4.6}\right)^2\frac{x^3}{7}+\&c\right\}$$
$$=\log\frac{1+e^{-y/2}}{1-e^{-y/2}}-3\log\frac{1+e^{-3y/2}}{1-e^{-3y/2}}+5\log\frac{1+e^{-5y/2}}{1-e^{-5y/2}}-\&c$$

Continuous mathematics helps!

- Generating functions and singularities
- Functional equations
- Perturbative methods
- Limit theorems from probability, ...

Cf: Analytic Combinatorics, F. & Sedgewick C.U.P. 2008

Free!



A few snapshots:

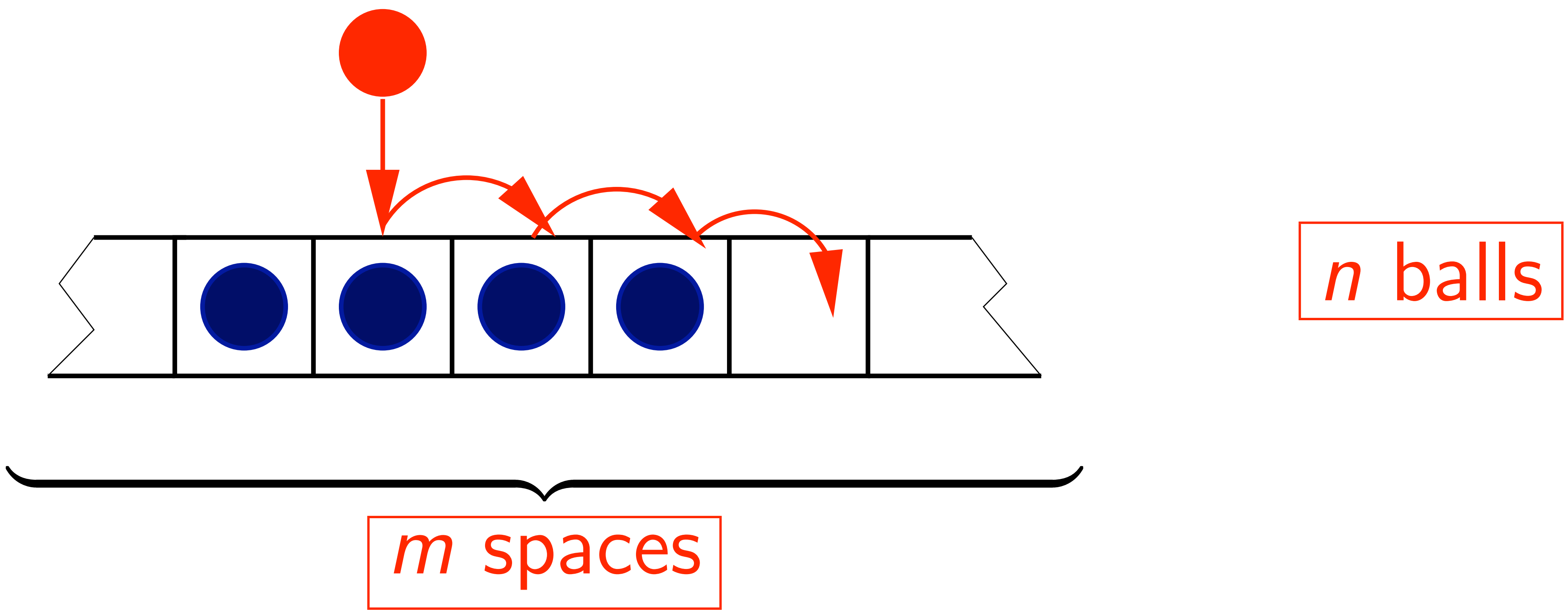
- 1. A parking problem and hashing
- 2. Digital trees (tries)
- 3. Data compression
- 4. Multidimensional search
- 5. Arithmetic computations and cont'd fractions

1. Parking & Hashing

The central rôle of generating functions

HASHING (1)

Knuth [1973, p. 545]: "A man and his dozing wife drive by, and suddenly she wakes up and orders him to park immediately. He dutifully parks at the first available space . . . ".

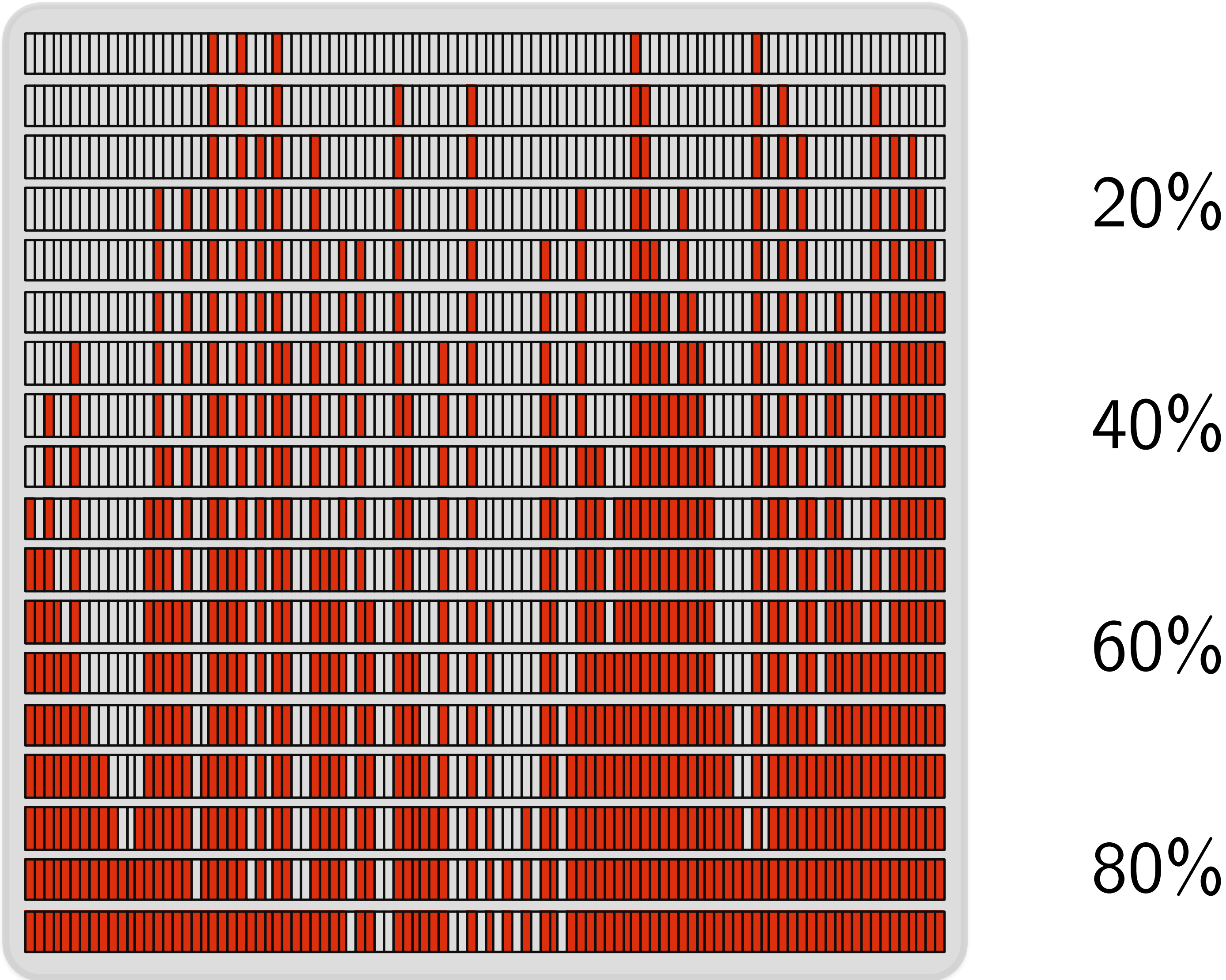


In computer science: place item x at a location $h(x)$ computed from x (or nearby).

HASHING (2)

The quantity $\alpha := \frac{n}{m}$ is the *filling ratio*.

As α increases, longer and longer "islands" get formed.

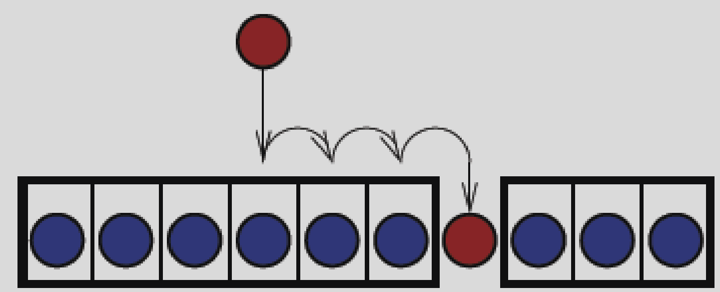


- Geometry of the randomly formed islands?
- Total displacement (dissatisfaction)?

HASHING (3): Generating functions

$F_n = \#$ ways to form an island with n balls and $m = n+1$ cells.

Generating function: $F(z) := \sum_{n\geq 0} F_n \frac{z^n}{n!}$

Island: $=$ $\square$ $+$ 

$$F(z) = 1 + \int_0^z \frac{d}{dz}(F(t)\, t) \cdot F(t) \quad dt$$

Solve: $F' = (zF)' \cdot F \Longrightarrow \frac{F'}{F} = (zF)' \Longrightarrow \log F = zF \Longrightarrow$ $F = e^{zF}$

- Hashing: [Knuth 1962–3, 1973] [F. Poblete Viola, *Algorithmica*, 1998]
- Symbolic methods in enumeration: Stanley, Goulden–Jackson, Wilf, . . .
- *Analytic Combinatorics* book by F. & Sedgewick, CUP, 2008;

`http://algo.inria.fr/flajolet/Publications/`

HASHING (4): More generating functions

- Use Lagrange Inversion Theorem to get $\boxed{F_n = (n+1)^{n-1}}$.
- Locate singularity at $z = e^{-1}$ to find (implicit functions):

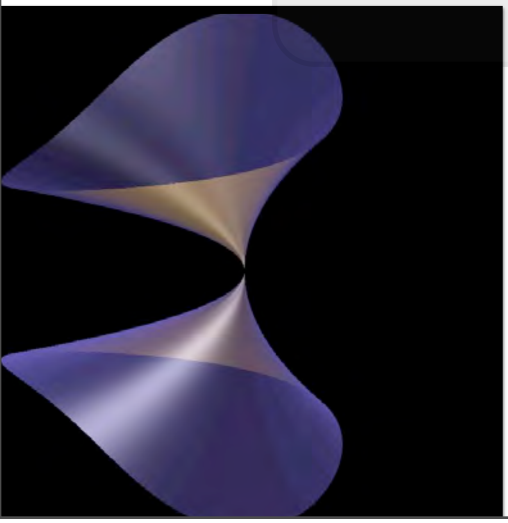
$$F(z) = e - e\sqrt{2}\sqrt{1 - ze} + \cdots .$$

Apply *singularity analysis* to get back asymptotics of F_n.

Singularity analysis ($\rightsquigarrow$ later):

$$f(z) \sim \left(1 - \frac{z}{\rho}\right)^{-\alpha} \quad \overset{\text{(Conditions)}}{\Longrightarrow} \quad \text{coeff}[z^n]f(z) \sim \rho^{-n}\frac{n^{\alpha-1}}{\Gamma(\alpha)}$$

FUNCTION COEFFICIENTS



HASHING (5): Results

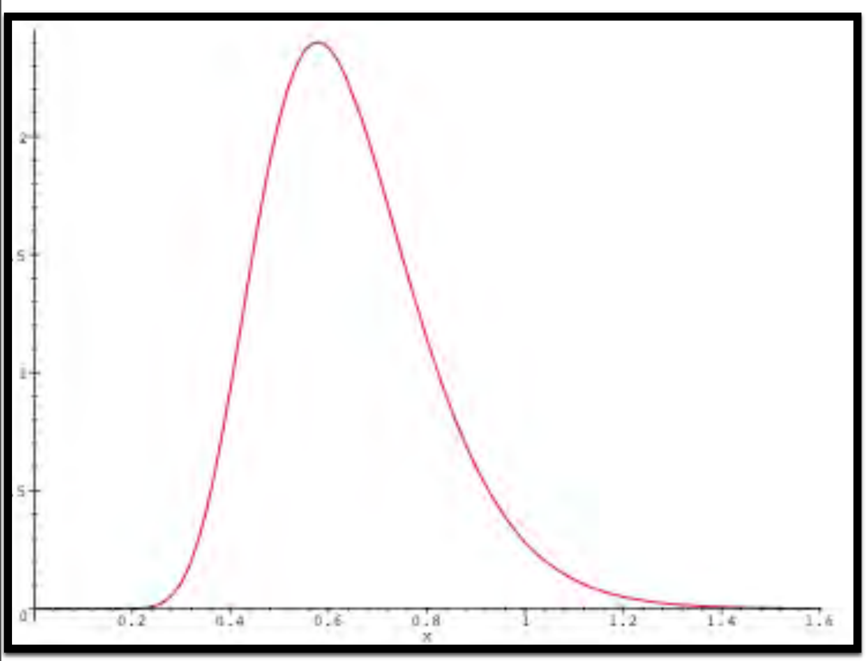
Theorem A. *The distribution of island size from random point is*

$$\mathbb{P}(k) = (1-\alpha)\frac{(k+1)^{k-1}}{(k-1)!}\alpha^{k-1}\left(e^{-\alpha}\right)^{k+1}.$$

$\alpha = 0.9 \Longrightarrow$ 54% of elements are in islands of size ≥ 50.

Theorem B. *The limit distribution of total displacement in full tables scales: $D_n \propto n^{3/2}$; it is characterized by its moments:*

$$\mathbb{E}\left[\mathcal{D}_\infty^r\right] = \frac{2\sqrt{\pi}}{\Gamma\left((3r-1)/2\right)}\Omega_r, \qquad \frac{\mathrm{Ai}'(z)}{\mathrm{Ai}(z)} \underset{z\to+\infty}{\sim} \sum_{r=0}^{\infty}\Omega_r(-1)^r\frac{z^{-(3r-1)}}{2^r\, r!}.$$



Ai(z) is the Airy function: $y'' - zy = 0$,

$$\mathrm{Ai}(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(zt+t^3/3)}\,dt.$$

"The limit distribution of total displacement in full tables scales: $D_n \propto n^{3/2}$."

- A bivariate generating function $F(z,q)$ with $q \leftrightarrow$total displacement:

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

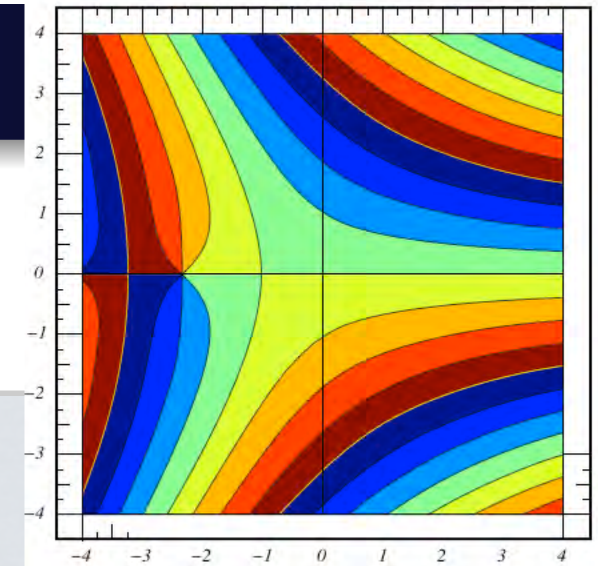
- Set up a *dedicated calculus* of commutations

$$\mathbf{U}h \equiv h(z,1), \quad \partial_z h \equiv \frac{\partial}{\partial z}h(z,q), \quad \partial_q h \equiv \frac{\partial}{\partial q}h(z,q),$$

- Moment pumping: reduce to the singular structure of $\mathbf{U}\partial_q^r$, for $r = 1, 2, \ldots$, via a *dedicated integral transform*.
- Apply singularity analysis to get (factorial) moments as $[z^n]\mathbf{U}\partial_q^r F(z,q)$. Airy function $\ll$ quadratic relations.
- Use: *"convergence of moments $\Longrightarrow$ convergence of distributions"*.

After [F.–Poblete–Viola, 1998]

The Airy world



Airy coeffs

$$\mathrm{Ai}(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(zt+t^3/3)}\,dt; \qquad \frac{\mathrm{Ai}'(z)}{\mathrm{Ai}(z)} \underset{z\to+\infty}{\rightsquigarrow} \sum_r \Omega_r z^{-r} \ldots$$

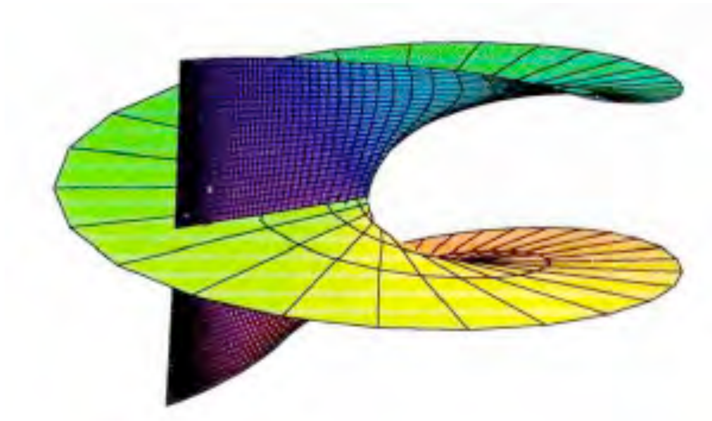
Graph connectivity

Pathlength in trees

Area in paths

[**Knuth**, *Algorithmica* 1998], [**Janson, Knuth, Łuczak, Pittel** 1993], Spencer, FPV98, Janson, **Louchard**, Takács, Aldous, [F., Salvy, Schaeffer 2004], . . .

Singularity Analysis: a technology

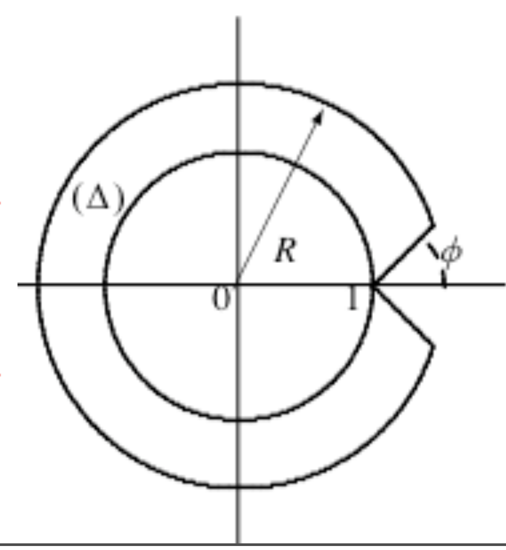


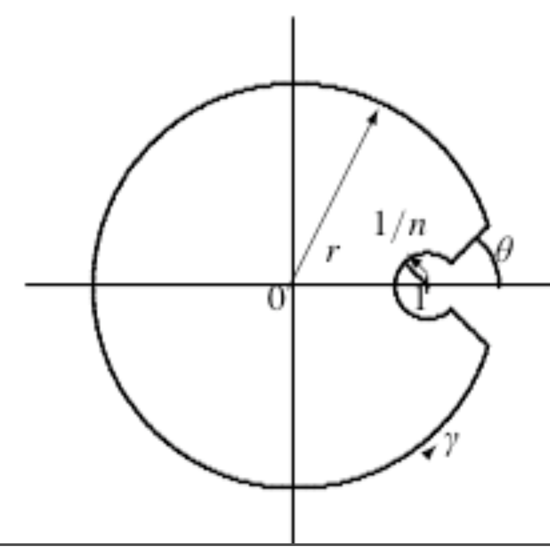
Function ⟶ Coefficients

- Under isolated singularities, transfer to coefficients is valid:
  - Base functions $Z^{-\alpha}(\log Z)^{\beta}$;
  - $O, o, \sim$–transfers, loglog's &c.
- Works for polylogarithms of various sorts $\sum \frac{\log n}{\sqrt{n}} z^n$ &c.
- Class closed under ∂ and $\int$; also Hadamard product $\odot$.
- Sometimes adapts to *natural boundaries*.

[**Wong–Wyman 1974**], [**F.–Odlyzko 1990**], [F. 1999], [Fill–F.–Kapur 05], [FFGPP, *Elec. J. Comb.* 2006], [**F-Sedgewick, Analytic Combinatorics 2008**]

$$\operatorname{coeff}[z^n]\, f(z) = \frac{1}{2i\pi} \int_{\gamma} f(z) \frac{dz}{z^{n+1}}$$

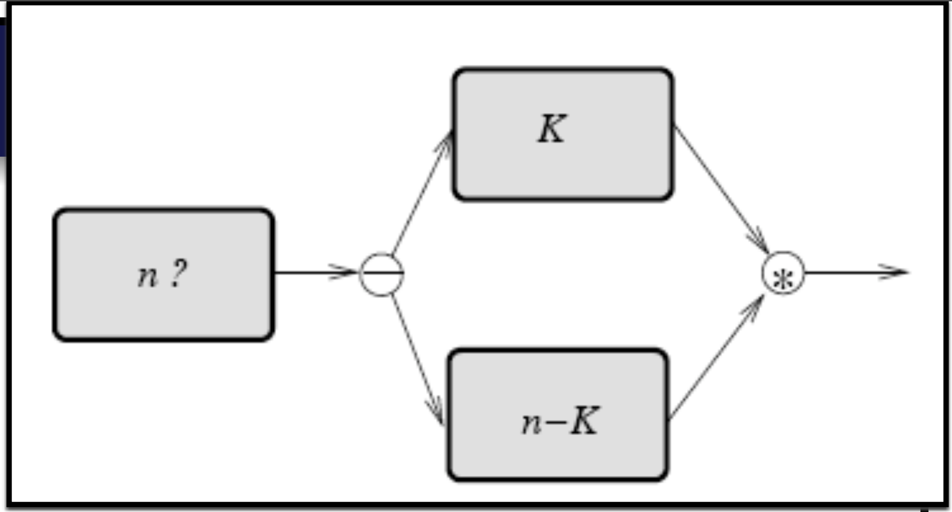




2. Digital Trees

Mellin transforms

Digital trees aka tries (1)



- The divide-and-conquer paradigm:

$$f_n = t_n + \sum_k \pi_{n,k} \left(f_k + f_{n-k} \right).$$

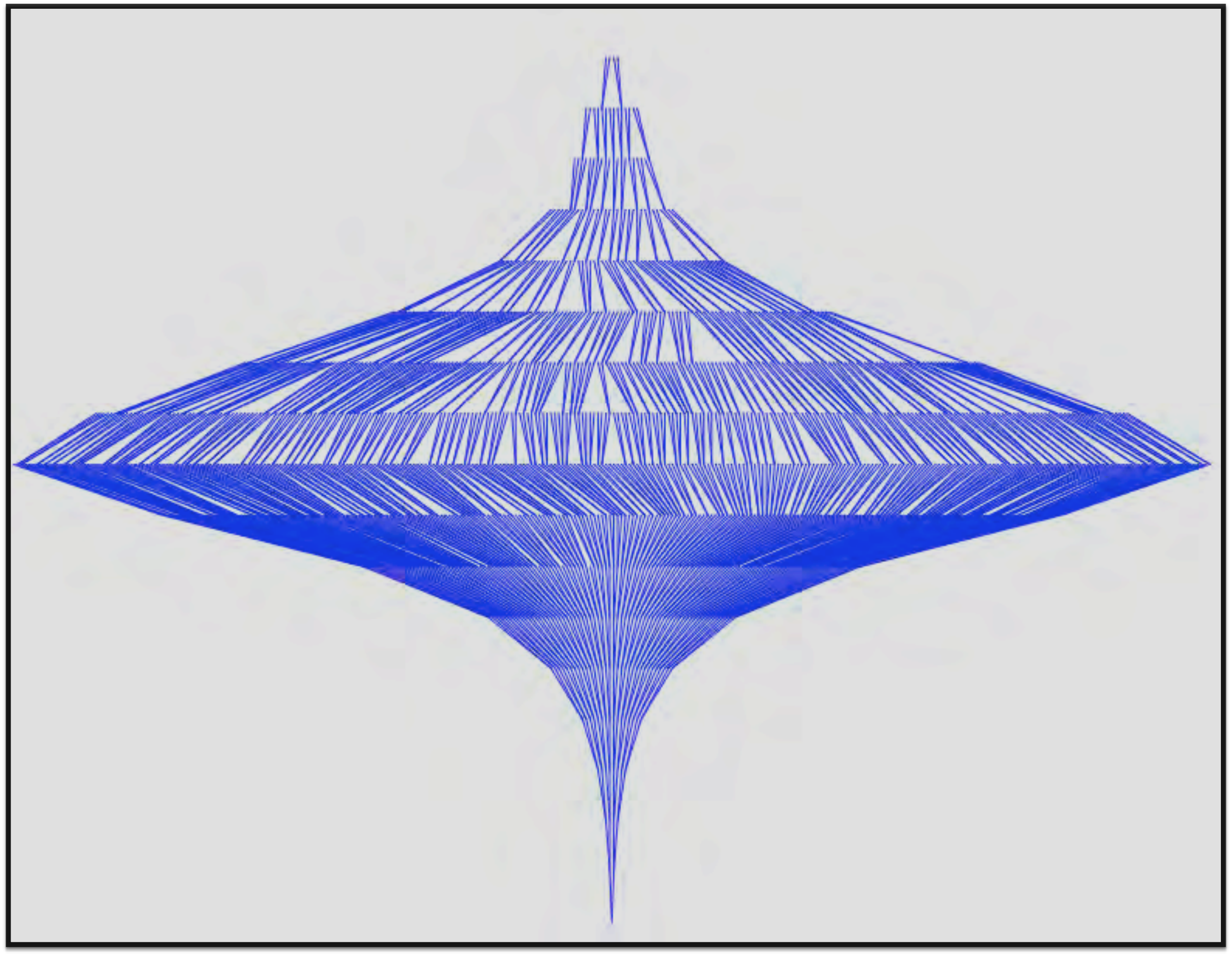
- The "trie" process = split on bits or on coin flips:

$$\pi_{n,k} = \frac{1}{2^n} \binom{n}{k}.$$

Applications: organising and sorting bit strings; electing a leader; communicating over a distributed channel, and many more.

What is the shape of a typical tree built on n items?

- Size is linear, $O(n)$, with high probability.
- Height is logarithmic, $O(\log n)$.



Luc Devroye

[**Knuth**, *TAOCP*, 1973], [**Vallée** 2001], [**Szpankowski** 2001], [F. 2006], . . .

Digital trees (2): Generating functions & Mellin transforms

To a sequence of tolls or costs associate a Poisson generating function:

$$f_n \quad \rightsquigarrow \quad \varphi(z) := e^{-z} \sum_{n=0}^{\infty} f_n \frac{z^n}{n!}.$$

The fundamental D&C recurrence becomes a functional equation:

$$\begin{array}{ccccc} \varphi(z) & = & \tau(z) & + & 2\varphi\left(\frac{z}{2}\right) \\ \text{cost} & & \text{toll} & + & \text{sub-costs.} \end{array}$$

Introduce Mellin transforms:

$$\varphi^\star(s) := \int_0^\infty \varphi(x) x^{s-1}\, dx$$

to get a solvable equation for tree-size ($\tau(x) = 1 - (1+x)e^{-x}$):

$$\varphi^\star(s) = -\frac{(s+1)\Gamma(s)}{1 - 2^{1+s}}.$$

Mellin transforms have two major properties (for us):

[Wong 1990]

- Harmonic sum property:

$$\sum_{(\lambda,\mu)} \lambda g(\mu x) \quad \xrightarrow{\mathcal{M}} \quad \Big(\sum \lambda \mu^{-s}\Big) \cdot g^{\star}(s).$$

- Mapping property:

Asymptotics of $f \longleftrightarrow$ Singularities of $f^{\star}$.

$$\varphi^{\star}(s) := \int_0^{\infty} \varphi(x) x^{s-1}\, dx$$

Mellin transforms have two major properties (for us):

- Harmonic sum property:

$$\sum_{(\lambda,\mu)} \lambda g(\mu x) \quad \xrightarrow{\mathcal{M}} \quad \left(\sum \lambda \mu^{-s}\right) \cdot g^{\star}(s).$$

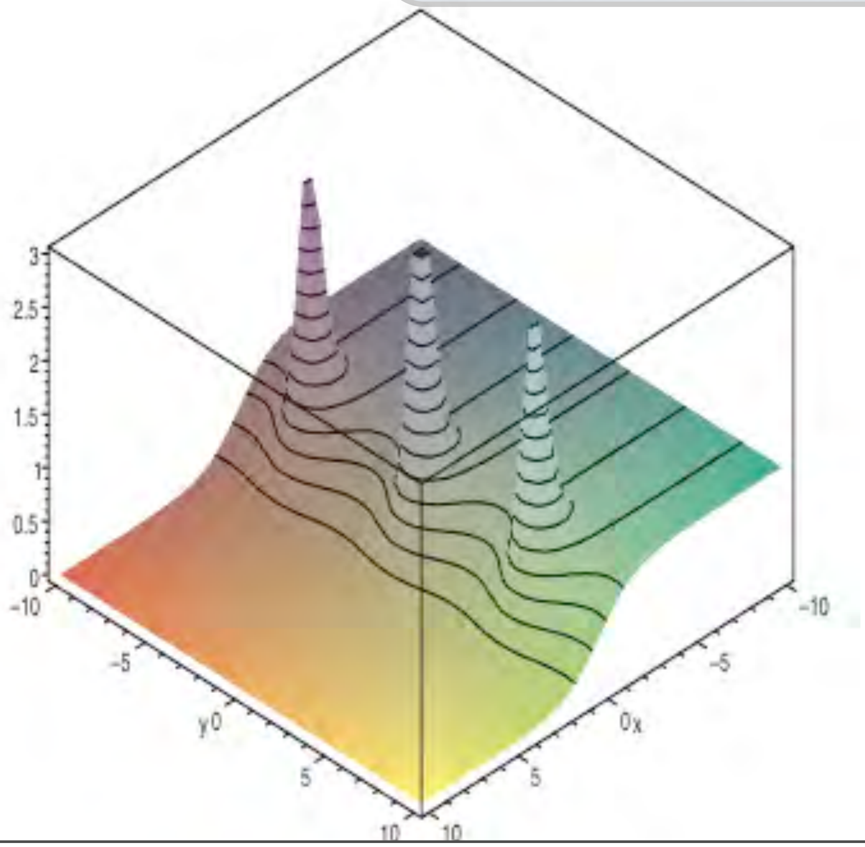
- Mapping property:

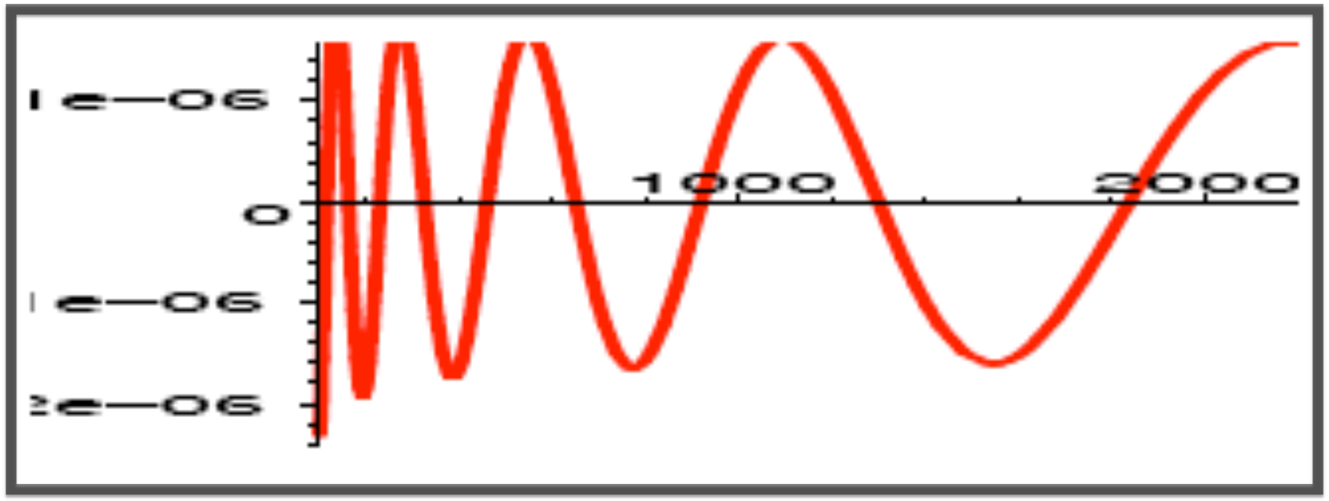
  Asymptotics of f $\longleftrightarrow$ Singularities of $f^{\star}$.

[Wong 1990]

Size : $\varphi^{\star}(s) = -\dfrac{(s+1)\Gamma(s)}{1-2^{1+s}}$ has complex poles, implies fluctuations.

Theorem: *Mean size is asymptotic to* $\dfrac{2n}{\log 2}(n + P(\log_2 n))$.





3. Data Compression

The Lempel-Ziv algorithm
& functional equations

Compression by dictionary (1)

Text:

"The square of the hypotenuse of a right triangle is equal to the sum of the squares on the other two sides."

Static dictionary:

1=the, 2=square, 3=of

Compressed text:

"1 2 3 1 hypotenuse 3 a right triangle is equal to 1 sum 3 1 2s on 1 o1r two sides."

Lempel-Ziv compression: a way to build a text-dependent dictionary on-line.

The "deja-vu" principle

Lempel-Ziv compression:

- Segment the text (from left to right) into "phrases".
- New phrase = *longest* earlier-seen phrase plus 1 character.
- Encode ~~rank of~~ earlier-seen phrase by its rank.

|a| |ab| |abr| |abra|abrac|

1 6 8 11 12

0a *1b* *6r* *8a* *11c*

The "deja-vu" principle

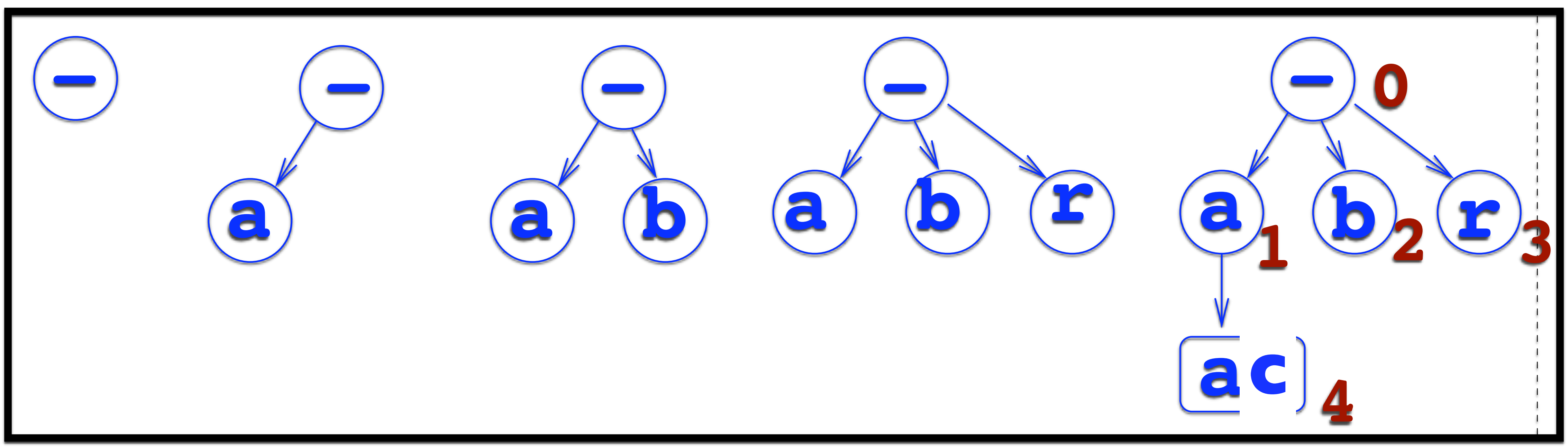
```
abracadabraabracadabraabracadabraa...

| |a|b|r|ac|ad|ab|ra|abr|aca|d|abra|abrac|ada|br|aa|br|ac|adab|raa|
(0) 1 2 3 4  5  6  7  8   9  10  11   12   13  14 15 16 17  18   19

0a0b0r1c1d1b3a6r4a0d8a11c5a2r1a2r1c13b7a
```

Compression (2): an implementation and a model

- Organize the "phrases" into a **tree**:
  - Follow branch and find longest matching phrase.
  - Occupy next vacant node.



Compression (3): tree analysis

- Take the **tree** from the root: split $n \to \langle 1,\ K,\ n-1-K \rangle$ with

$$\mathbb{P}_n(K=k) \equiv \pi_{n,k} = \frac{1}{2^{n-1}}\binom{n-1}{k}.$$

- Get fundamental generating function schema with $p+q=1$:

probabilities p, q

$$\frac{\partial}{\partial z}\varphi(z) = \tau(z) + \varphi(pz) + \varphi(qz).$$

Mellin transform gives:

Theorem. *Path length* $\sim \frac{n}{H}\log n$; $H = p\log\frac{1}{p} + q\log\frac{1}{q}$ *is entropy.*

[Knuth 1973], [F–Sedgewick 1986], [**Szpankowski** 2001]...

Compression (4): results

Theorem. *Lempel-Ziv achieves entropy; know redundancy & fluct.*

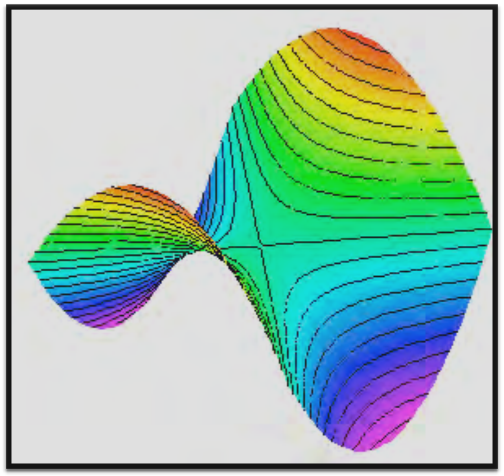
From [Jacquet, Szpankowski, and Louchard, 1995+]

- Study the bivariate differential–difference equation

$$\frac{\partial}{\partial z}\varphi(z,u) = \varphi(pz,u)\,\varphi(qz,u)\,.$$

  and get that PATH-LENGTH$_n$ is asymptotically normal.

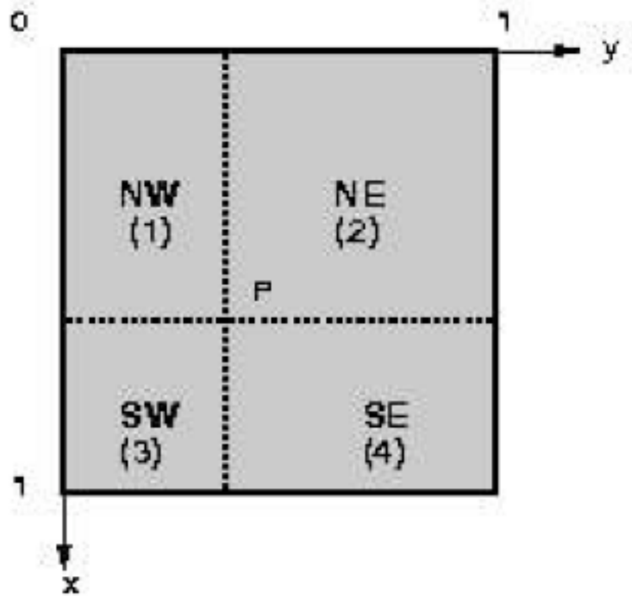
- "Invert" the relation TREE-SIZE $\leftrightarrow$ PATH-LENGTH.
- Saddle-point (analytic) depoissonization
- Newton series and Nörlund integrals in relation to Mellin

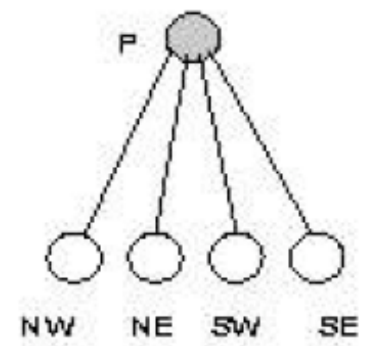


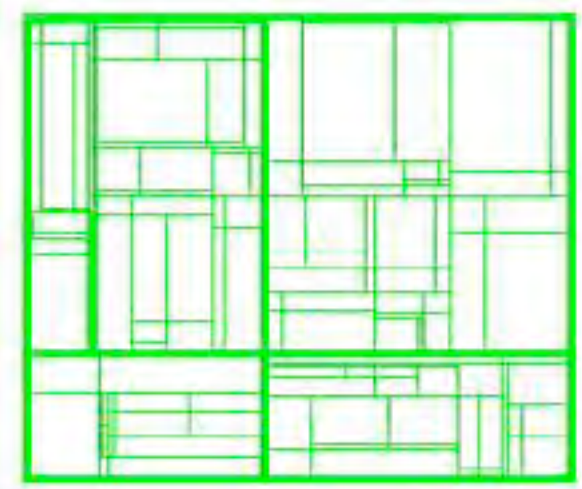
4. Multidimensional Search

Differential equations and singularities

Quadtrees (1)







Theorem A. *Cost a fully specified search is asymptoticaly normal with mean* $\sim \frac{2}{d} \log n$.

$$F(z,u) = 1 + 2^3 u \int_0^z \frac{dx_1}{x_1(1-x_1)} \int_0^{x_1} \frac{dx_2}{1-x_2} \int_0^{x_2} F(x_3,u) \frac{dx_3}{1-x_3}.$$

Theorem B. *Mean cost a partial match search is* $\asymp n^{\frac{1+\sqrt{17}}{2}}$.

- Singularities of linear ODEs and perturbation + singularity analysis.

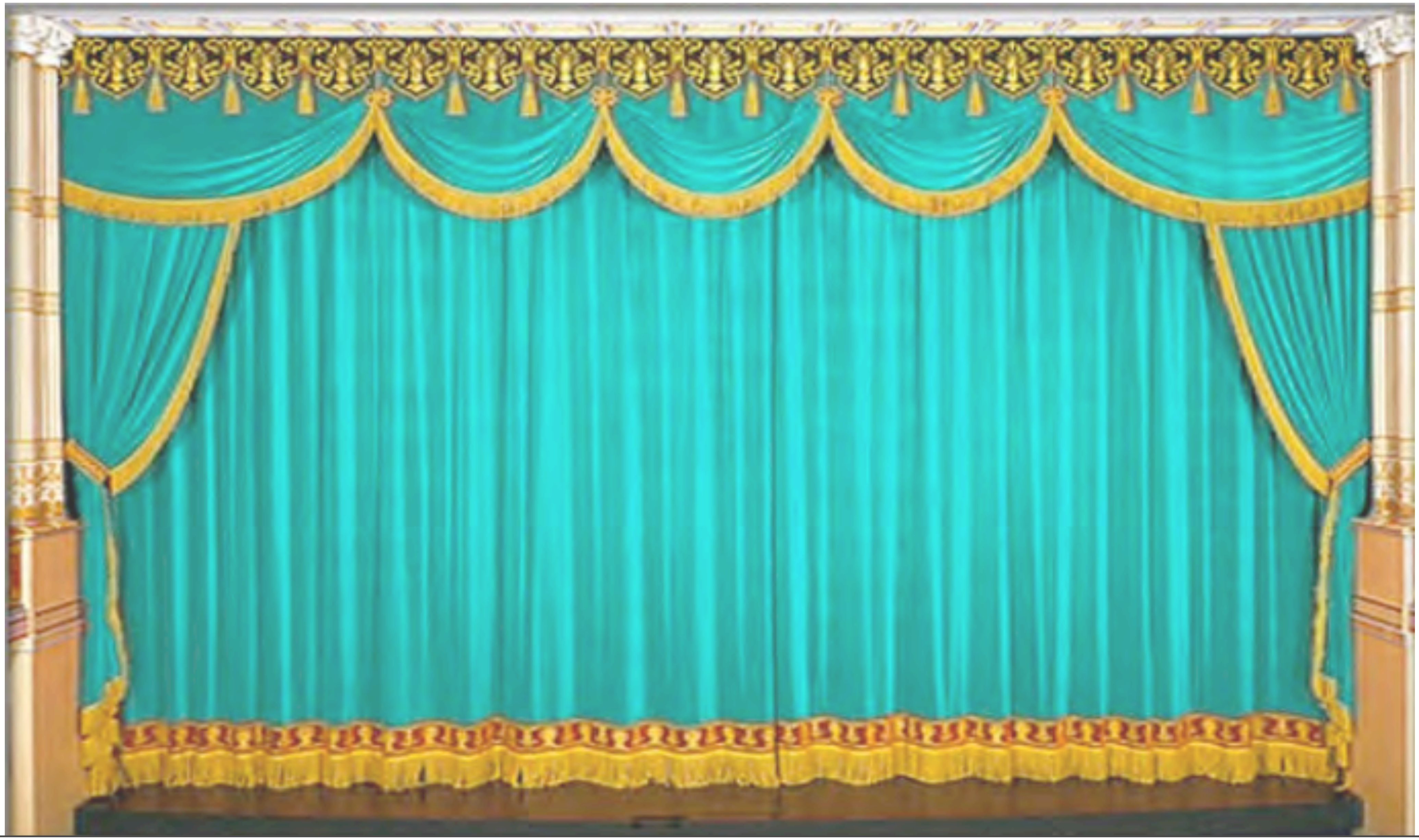
[F-Gonnet-Puech-Robson, 1993], [F-Lafforgue 94], [F-Labelle-Laforest-Salvy 1998], [Hwang* 2000+]

Gaston Gonnet

Quadtrees (2): the “holonomic” framework

Holonomic functions aka *differentiably finite* aka *D–finite*:

- satisfy linear differential equations with polynomial coefficientss;
- have coefficients satisfying P–recurrences;
- are s.t. the vector space of all partial derivatives is finite-dimensional over $\mathbb{C}(x, y, z, \ldots)$.

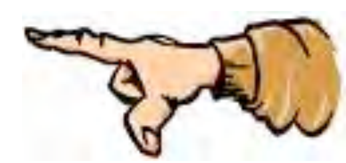


Quadtrees (2): the “holonomic” framework

Holonomic functions aka *differentiably finite* aka *D–finite*:

- satisfy linear differential equations with polynomial coefficientss;
- have coefficients satisfying *P*–recurrences;
- are s.t. the vector space of all partial derivatives is finite-dimensional over $\mathbb{C}(x, y, z, \ldots)$.

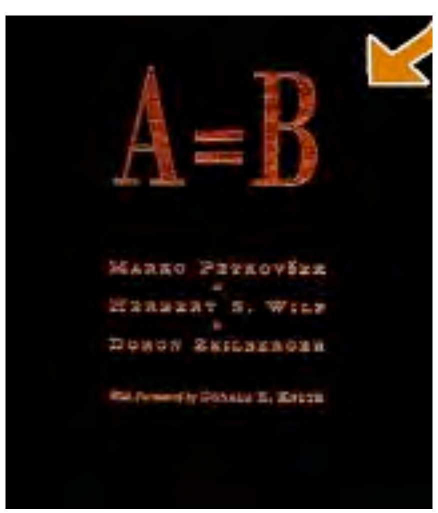
Stanley–Lipschitz–Zeilberger theory:



- contain algebraic and hypergeometric functions;
- closed under $+, \times, \int, \partial,$ Diag, algebraic substitution, &c
- Zeilberger: specifiable by a finite amount of information.

• Identity is decidable.

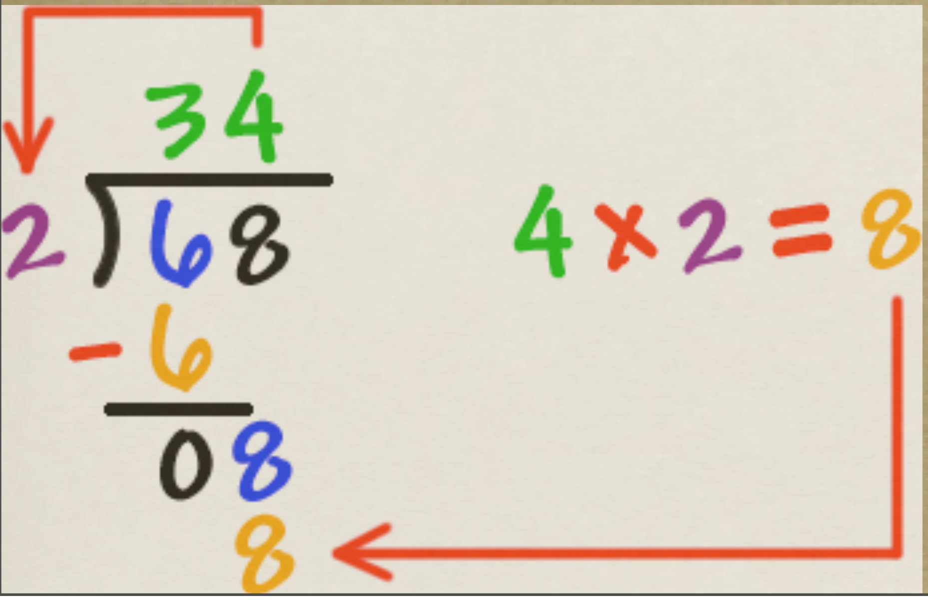
• Asymptotics of coefficients is (largely) decidable.



[Petkovšek, Wilf, **Zeilberger** $A = B$, 1996]

5. Arithmetic computations

A glimpse of transfer operators...



The Euclidean algorithm

Hensley, Baladi, and Vallée [1994+]:

- The dynamics of continued fractions depends on the Ruelle/transfer operator

$$\mathcal{G}_s[f](x) := \sum_{n \geq 1} \frac{1}{(n+x)^{2s}} f\left(\frac{1}{n+x}\right);$$

- gives also the dynamics of the (discrete) Euclidean algorithm.

Theorem A. *Euclid's algorithm is Gaussian!*

• A large number of variations: *binary GCD, nearest integer GCD, least/most significant bits*, ...

[D. Mayer 1971], [**Hensley** 1994], [Vallée 1998], [Dolgopyat 1998] [**Baladi–Vallée** 2005], [**Vallée** 2006],...

Comparing and sorting numbers with Euclid

Theorem B. *The average cost of the HAKMEM algorithm for comparing two numbers is*

$$-\frac{60}{\pi^4}\left(24\mathrm{Li}_4(\tfrac{1}{2}) - \pi^2(\log 2)^2 + 21\zeta(3)\log 2 + (\log 2)^4\right) + 17.$$

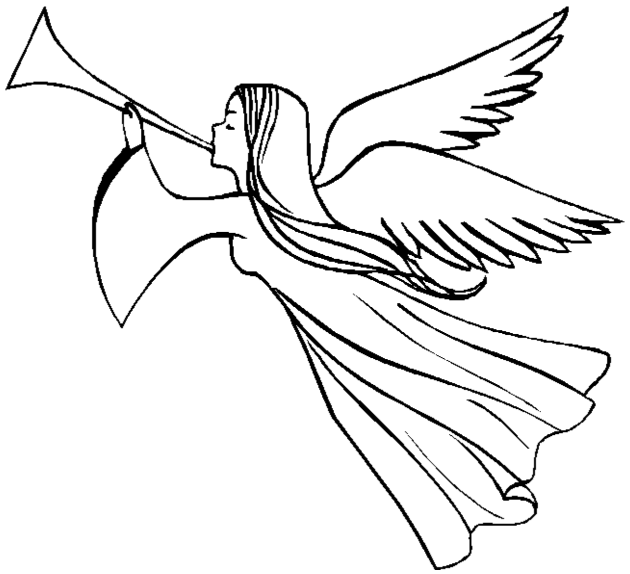
Theorem C. *The mean cost of sorting n numbers based on continued fractions is*

$$K_0 n \log n + K_1 n + Q(n) + K_2 + o(1),$$

$$K_0 = \frac{6\log 2}{\pi^2}, \qquad K_1 = 18\frac{\gamma \log 2}{\pi^2} + 9\frac{(\log 2)^2}{\pi^2} - 72\frac{\log 2\,\zeta'(2)}{\pi^4} - \frac{1}{2}.$$

[F. Vallée 2000]

The order of Q(n) depends on the Riemann hypothesis!



- Many algorithms and models of computational sciences a priori live in a discrete world;
- but continuous mathematics is highly relevant.
- Encounters with many old and some new maths: asymptotic and complex analysis...