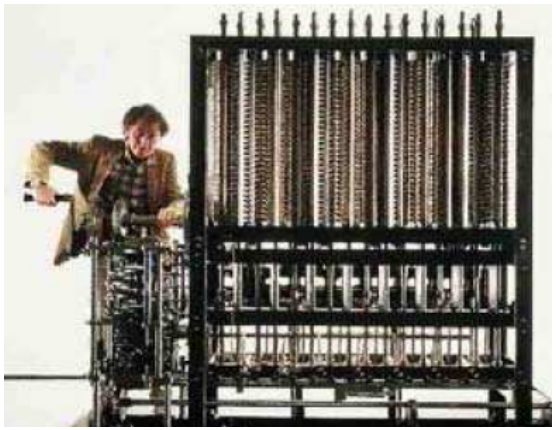
Analytic Combinatorics— A Calculus of Discrete Structures

Philippe Flajolet

INRIA Rocquencourt, France

SODA07, New Orleans, January 2007

Analysis of algorithms: *What is the cost of a computational task?*



Babbage (1837):

number of turns of the crank

On a data ensemble, as a function of size n?

- in the worst case
- typically: *on average*; *in probability in distribution*.

Also vital for randomized algorithms.

SURPRISE (1960-1970s): A large body of classical maths is adequate for many average-case analyses.

— Von Neuman 1946+Knuth 1978: adders=carry riples.
— Hoare 1960: Quicksort and Quickselect
— Knuth 1968–1973$^+$: *The Art of Computer Programming*.
— Sedgewick: median of three, halting on small subfiles, etc

"The Unreasonable Effectiveness of Mathematics" [E. Wigner]

. . . BUT . . . :

In the 1970s and 1980s, culmination of recurrences and real analysis ($\sum \to \int$) techniques.

— Limitations for richer data structures and algorithms
— analyses become more and more technical.

No clear relationship

Algorithmic structures $\longrightarrow$ Complexity structures.

+ Explosion in difficulty: average-case $\rightsquigarrow$ variance $\rightsquigarrow$ distribution

ADVANCES (1990–2007)

Synthetic approaches emerge based on generating functions.

- **A.** Combinatorial enumeration: Symbolic methods.
  Joyal's theory of species [Bergeron-Labelle-Leroux 1998];
  Rota–Stanley [books]; Goulden & Jackson's formal methods;
  Bender-Goldman's theory of "prefabs"; Russian school.
- **B.** Asymptotic analysis: Complex methods.
  Bender *et al.*. F-Odlyzko, 1990+: singularity analysis; Odlyzko's survey 1995; uses of saddle points and Mellin transform.
- **C.** Distributional properties: Perturbation theory.
  Bender, F-Soria; H.K. Hwang's Quasipowers, 1998;
  Drmota-Lalley-Woods. . . .

AofA Books: Hofri (1995), Mahmoud (1993); Szpankowski (2001).
+ *Analytic Combinatorics*, by F. & Sedgewick (2007).

PART A. SYMBOLIC METHODS

How to enumerate a combinatorial class $\mathcal{C}$?

$C_n = \#$ objects of size n

♡ *Generating function*: $C(z) := \sum_n z^n$.

Symbolic approach

- An object of size n is viewed as composed of *n atoms* (with additional structure): words, trees, graphs, permutations, etc.
- Replace each atom by symbolic weight z:

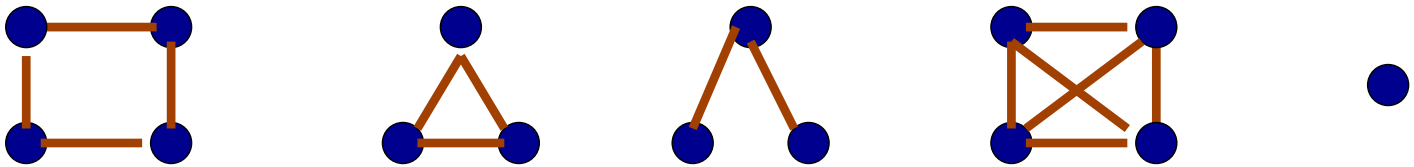
— Class: $\sum$ objects. Object: $\gamma \rightsquigarrow z^{|\gamma|}$.

Gives the **Ordinary Generating Function (OGF)**:

$$\mathcal{C} \qquad \rightsquigarrow \qquad C(z) := \sum_{\gamma \in \mathcal{C}} z^{\gamma} \equiv \sum_{n} C_n z^n.$$

Mathematician: "To count sheep, count legs and divide by 4."

E.g.: a class of graphs enumerated by # vertices

$\mathcal{C}$ = 

$C(z) = z\,z\,z\,z + z\,z\,z + z\,z\,z + z\,z\,z\,z + z$

$= \mathbf{1} \cdot z + \mathbf{2} \cdot z^3 + \mathbf{2} \cdot z^4$

$(C_n) = (0, 1, 0, 2, 2).$

Principle (Symbolic method)

The OGF of a class: (i) encodes the counting sequence; (ii) is nothing but a reduced form of the class itself.

Several set-theoretic constructions translate into GFs.

$$\text{disjoint union} \qquad \sum_{\mathcal{A}\oplus\mathcal{B}} = \sum_{\mathcal{A}} + \sum_{\mathcal{B}}$$

$$\text{cartesian product} \qquad \sum_{\mathcal{A}\times\mathcal{B}} = \sum_{\mathcal{A}} \cdot \sum_{\mathcal{B}}$$

There is a micro-dictionary:

$$\text{disjoint union} \qquad \mathcal{C} = \mathcal{A} \cup \mathcal{B} \quad \Longrightarrow \quad C(z) = A(z) + B(z)$$

$$\text{cartesian product} \qquad \mathcal{C} = \mathcal{A} \times \mathcal{B} \quad \Longrightarrow \quad C(z) = A(z) \cdot B(z)$$

Theorem (Symbolic method)

A dictionary translates constructions into generating functions:

Union	$+$
Product	$\times$
Sequence	$\frac{1}{1-\cdots}$
Set	Exp
Cycle	Log

♣ $\mathcal{C} = \textsc{Seq}(\mathcal{A}) \equiv \{\epsilon\} + \mathcal{A} + (\mathcal{A} \times \mathcal{A}) + \cdots$.

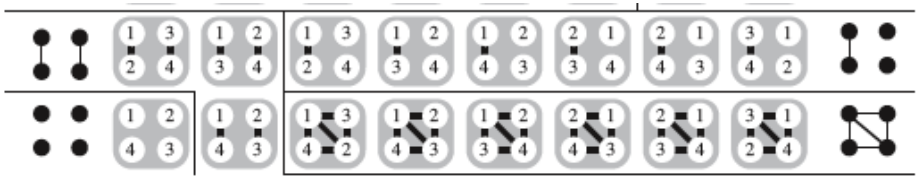
Thus $C = 1 + A + A^2 + A^3 = \dfrac{1}{1-A}$.

♣ $\mathcal{C} = \textsc{MSet}(\mathcal{A}) \equiv \prod_{\alpha \in \mathcal{A}} \textsc{Seq}(\alpha) \rightsquigarrow C = \mathrm{Exp}[A]$,

with $\mathrm{Exp}[f] := e^{f(z)+\frac{1}{2}f(z^2)+\cdots}$

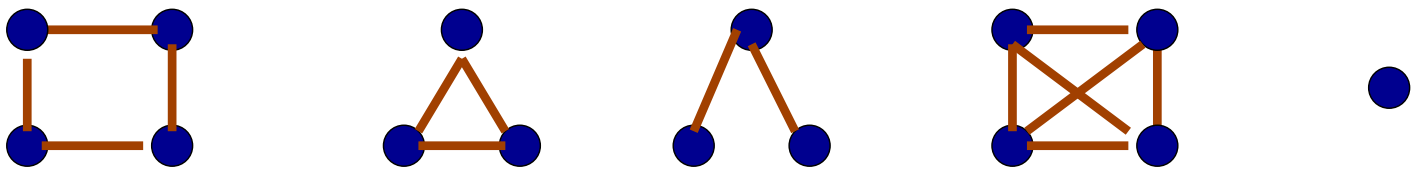
More generating functions . . .

- Labelled classes: via *exponential* GF (EGF) $\sum C_n \frac{z^n}{n!}$.



- Parameters: via *multivariate* GFs.

$\mathcal{C}$ =

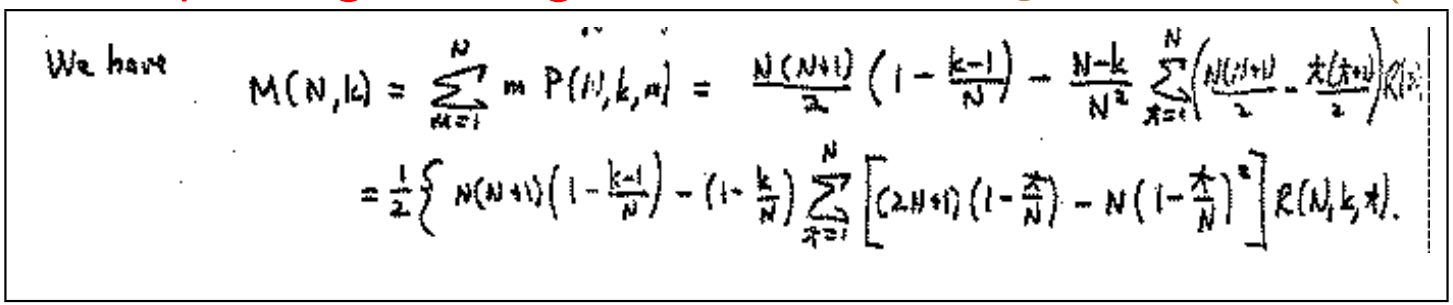


$$C(z,u) = \frac{z\,z\,z\,z}{u\,u\,u\,u} + \frac{z\,z\,z}{u\,u\,u} + \frac{z\,z\,z}{u\,u} + \frac{z\,z\,z\,z}{u\,u\,u\,u\,u\,u\,u} + \frac{z}{u^0}$$

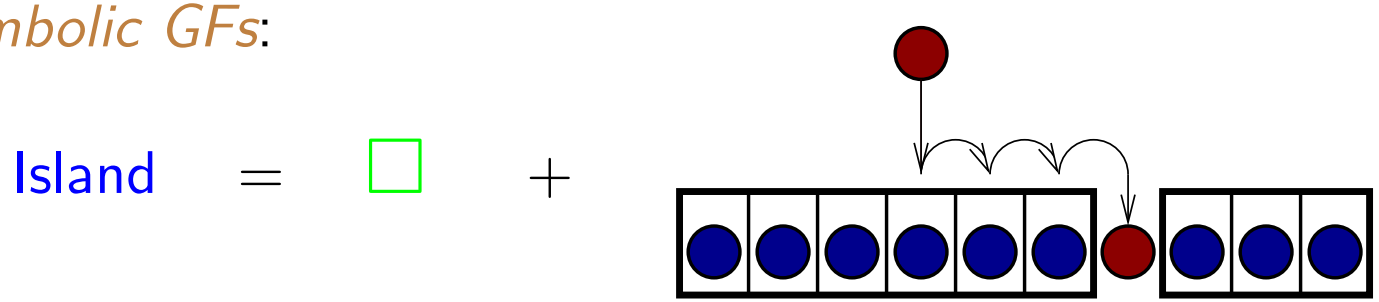
- Additional constructions: *substitution, pointing, order constraints*:

$$f \circ g, \qquad \partial f, \qquad \int f.$$

Linear probing hashing: From Knuth's original derivation (rec.):



to *symbolic GFs*:



$$I(z) = \mathbf{1} + \int \frac{\partial}{\partial z}(zI(z)) \times I(z)$$

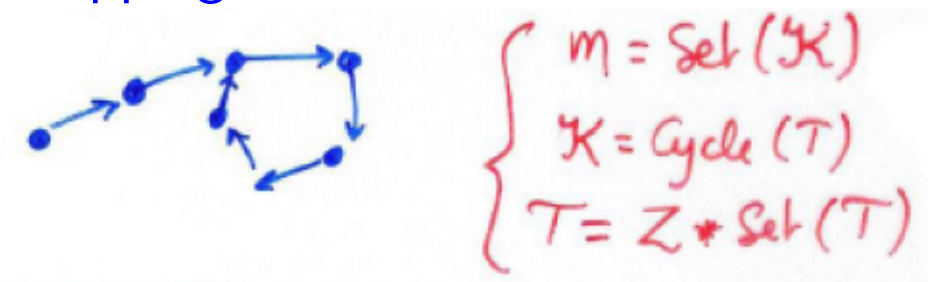
Get nonempty island by joining two islands by means of a gluing element.

$\rightsquigarrow$ wide encompassing extensions of original analyses [F-Poblete-Viola, Pittel, Knuth 1998, Janson, Chassaing-Marckert, . . .].

Some constructible families

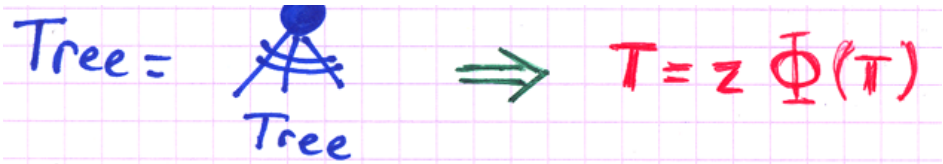
- Regular languages, FA, paths in graphs
- Unambiguous context-free languages
- Terms trees

- Increasing trees
- Mappings

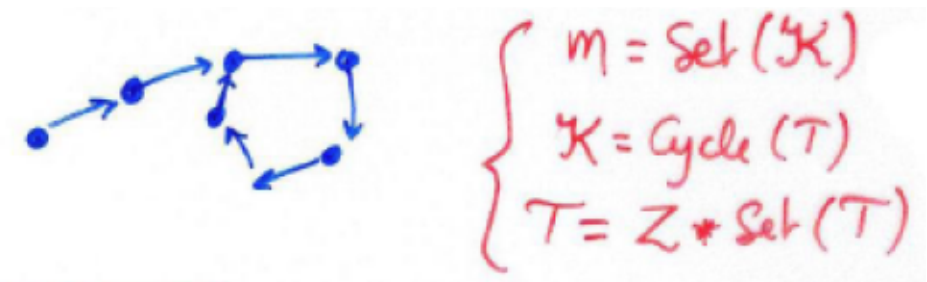


Some constructible families and generating fuctions

- Regular languages, FA, paths in graphs: $\rightsquigarrow$ rational fns
- Unambiguous context-free languages $\rightsquigarrow$ algebraic functions.
- Terms trees $\rightsquigarrow$ [+Pólya operators] implicit functions



- Increasing trees $\rightsquigarrow$ $Y = \int \Phi(Y)$ differential equation
- Mappings $\rightsquigarrow$ exp ∘ log ∘ implicit



$$\begin{cases} M = \exp(K) \\ K = \log(1-T)^{-1} \\ T = z\exp(T) \end{cases}.$$

PART B. COMPLEX ASYMPTOTICS

- The continuous [=analysis] helps understand the discrete.
- The complex domain has powerful properties.

"The shortest path between two truths on the real line goes through the complex plane."
— Jacques Hadamard

Erdös' *proofs from the Book* [cf Aigner-Ziegler]

Why are there infinitely many primes?

- Combinatorial proof ©Euclid: $n! + 1$ is divisible by a prime $> n$.
- Analytic proof ©Euler: consider a (Dirichlet) generating function

$$\begin{aligned} \zeta(s) &= \sum_{n \geq 1} \frac{1}{n^s} \\ &= \prod_{p \text{ Prime}} \frac{1}{1 - 1/p^s}. \end{aligned}$$

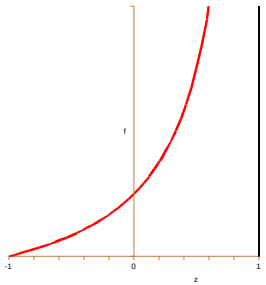
We have $\zeta(1^+) = +\infty$ while the finiteness of primes would imply $\zeta(1^+) < \infty$, a contradiction.

$\rightsquigarrow$ Riemann, Hadamard, de la Vallée-Poussin: Prime Number Theorem.

Complex asymptotics and GFs

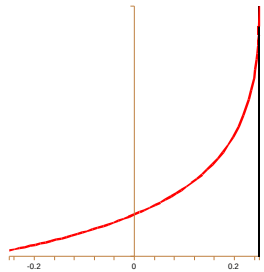
- formal z yields formal generating function as "power series";
- real z gives us a real function with convergence interval;

EGF of perms $\frac{1}{1-z}$



;

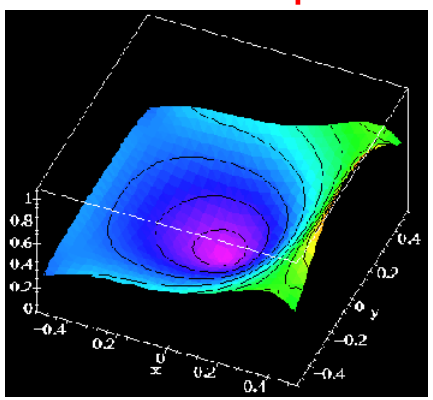
OGF of bin trees $\frac{1-\sqrt{1-4z}}{2z}$



- complex z gives us a function of a complex variable.

Surface

(here: *modulus* of OGF of balanced trees)



in $\mathbb{R}^4$ with $\langle \Re, \Im \rangle$.

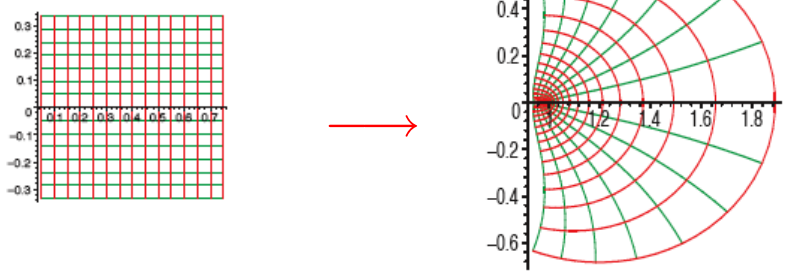
Analytic function := *smooth* transformation of the complex plane.

Definition

$f(z)$ is analytic (holomorphic, regular) if $\exists$: $\lim \frac{\Delta f}{\Delta z}$.

$\Longrightarrow$ Analytic functions satisfy rich closure properties.

(conformal mapping)



Definition

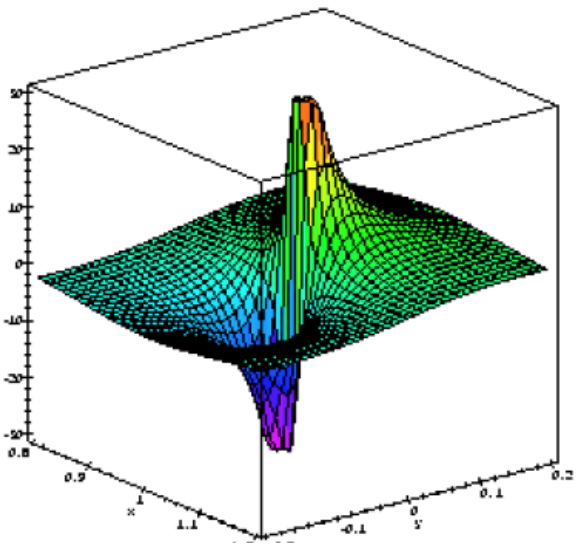
$f(z)$ has singularity at boundary point ζ if it cannot be made analytic around ζ.

E.g.: f discontinuous, infinite, oscillating, derivative blows up, etc.

Permutations

EGF: $P(z) = \dfrac{1}{1-z}$

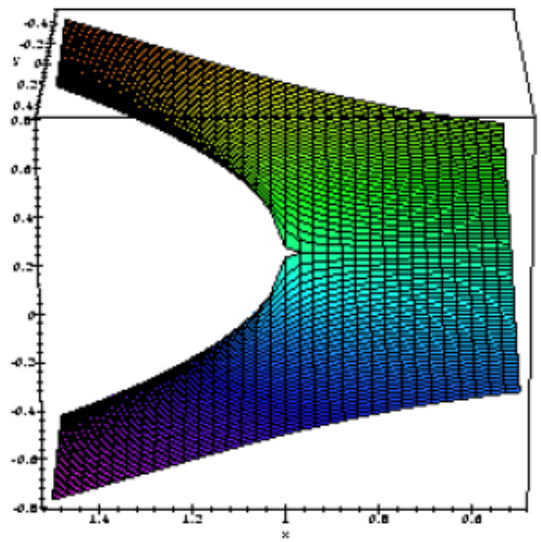
$\dfrac{P_n}{n!} \sim 1$



Bin. trees

OGF: $B(z) = \dfrac{1-\sqrt{1-4z}}{2z}$

$B_n \sim \dfrac{4^n}{\sqrt{\pi n^3}}$



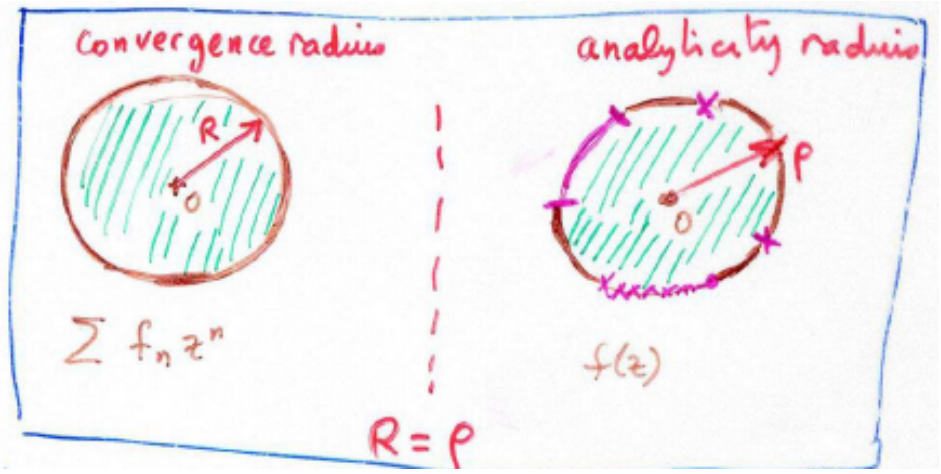
(Imaginary parts $\Im(f(z))$)

♡ Analytic properties of GF provide coefficients' asymptotics.

Principle (Singularity Analysis)

Singularities determine asymptotics of coefficients.
A singularity at ζ of $f(z)$ implies a contribution to f_n like $\zeta^{-n}\vartheta(n)$, where $\vartheta(n)$ is subexponential.

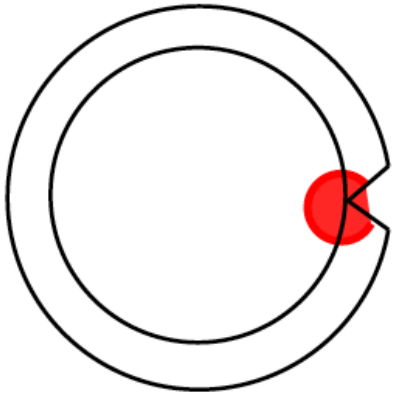
Theorem: $R_{\text{conv}} = \rho_{\text{sing}}$



LOCATION of SINGULARITY: by rescaling, $f(z/\zeta)$ is singular at 1. A factor of ζ^{-n} corresponds to a singularity at ζ.

NATURE of SINGULARITY: examine simple functions singular at 1:

Function	$\longrightarrow$	*Coefficient*	
$\dfrac{1}{(1-z)^2}$		$n+1$	$\sim n$
$\dfrac{1}{1-z}\log\dfrac{1}{1-z}$		$H_n \equiv 1+\frac{1}{2}+\cdots$	$\sim \log n$
$\dfrac{1}{1-z}$		1	~ 1
$\dfrac{1}{\sqrt{1-z}}$		$4^{-n}\binom{2n}{n}$	$\sim \frac{1}{\sqrt{\pi n}}$.



Let L be a slowly varying function, like $\log^{\beta} \log \log^{\delta}$.

Theorem (Singularity analysis)

Under a Camembert condition, the following implication is valid

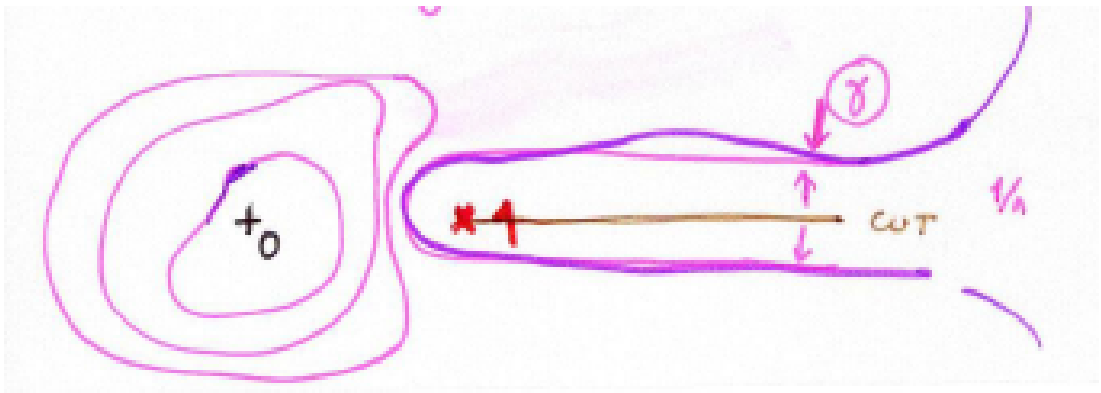
$$f(z) \approx \frac{1}{(1-z)^{\alpha}} L\left(\frac{1}{1-z}\right) \quad \longrightarrow \quad [z^n] f(z) \approx n^{\alpha-1} L(n).$$

Works for equality ($=$) with full asymptotic expansions; for $\mathcal{O}(.)$, $o(.)$, hence $\sim$.

[F., Odlyzko 1990]; closures $\partial, \int, \odot$ [Fill, F., Kapur 2005]; [F., Sedgewick 2007]

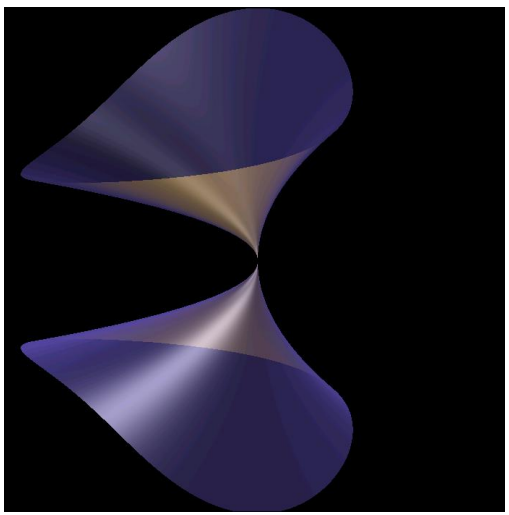
Proof of Singularity Analysis Theorems:

Cauchy's coefficient formula: $[z^n]f(z) = \frac{1}{2i\pi}\oint f(z)\frac{dz}{z^{n+1}}$.



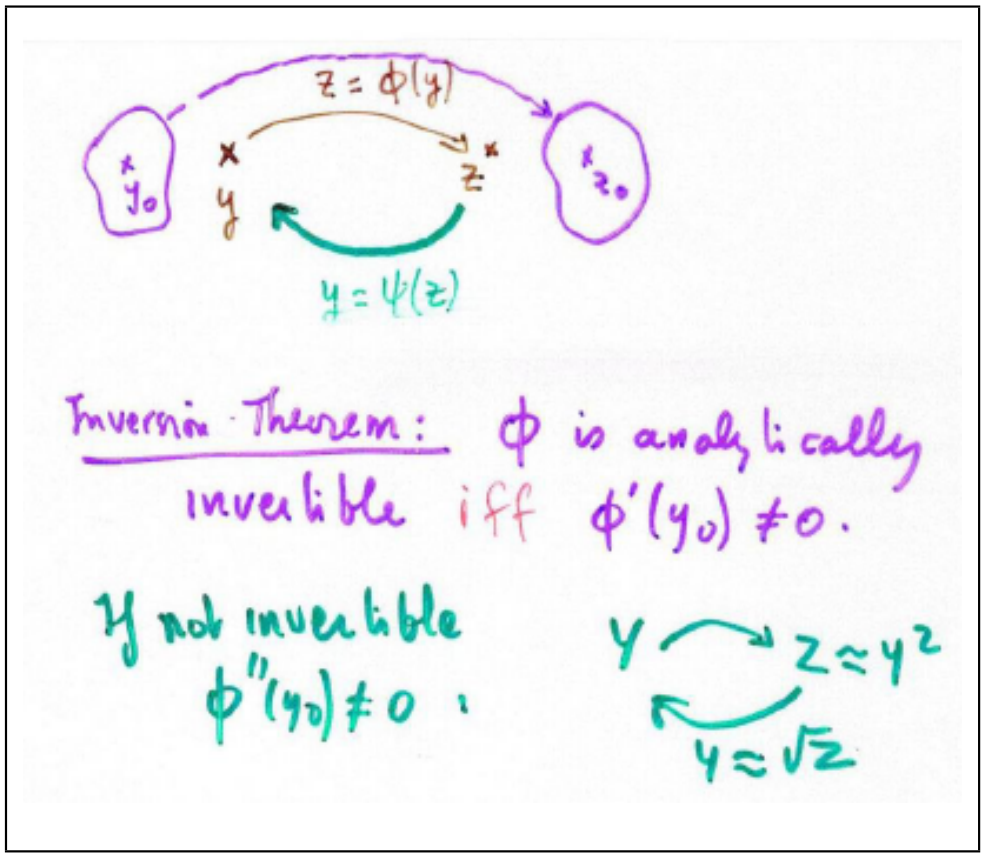
$z \to 1 + \frac{t}{n}$ $z^{-n} \to e^{-t}$; $dz \to \frac{dt}{n}$; $(1-z)^{-\alpha} \to (-t/n)^{-\alpha}$.

Singularity analysis works *automatically* for wide classes of generating functions.



— Rational [Perron-Frobenius] $\longrightarrow \zeta^{-n} n^k$
— Implicit $\longrightarrow \zeta^{-n} n^{-3/2}$
— Algebraic [Newton-Puiseux] $\longrightarrow \zeta^{-n} n^{p/q}$
— Holonomic [linear ODEs] $\longrightarrow \zeta^{-n} n^{\alpha} (\log n)^k$

Universality in trees and maps

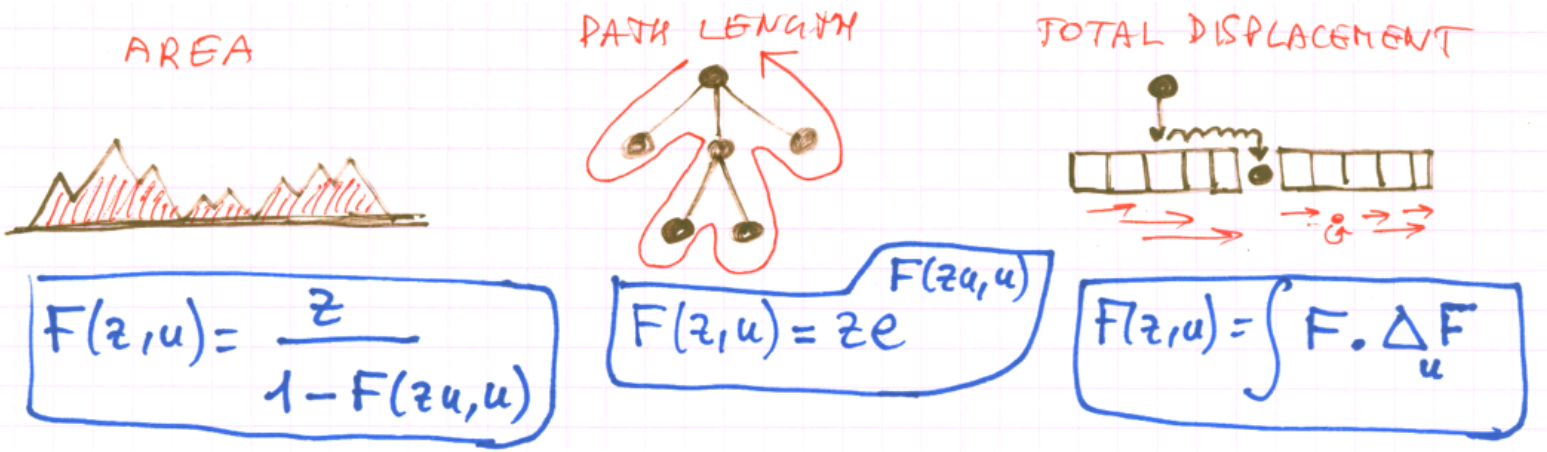


TREES: $Y = z\Phi(Y)$
universality of $\sqrt{\ }$–singularity.
Counting is universally $C \cdot A^n n^{-3/2}$.
Height and width are $\approx \sqrt{n}$.
Path length is $\approx n\sqrt{n}$, &c.

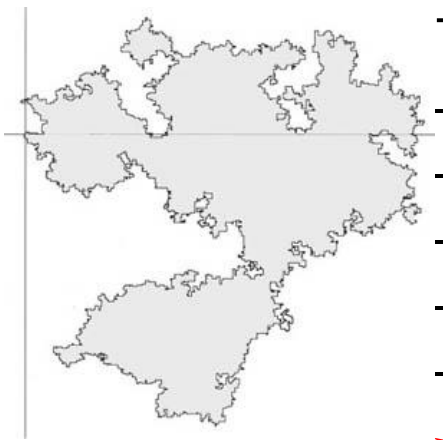
[Tutte$^+$]: universality of $C \cdot A^n n^{-5/2}$ for Rooted maps.

[Bender-Gao-Wormald 2002] $\rightsquigarrow$ Gimenez–Noy [2005$^+$]: Planar graphs $n! C \cdot A^n n^{-7/2}$. Fusy: random generation is $O(n^2)$.

Trees, walks, and hashing: moment pumping $\rightsquigarrow$ Airy distribution.



Louchard, Takacs, F.-Poblete-Viola.



The Guttmann–Richard$^+$ story:

— Analyse simplified models (e.g., 3 choice polygons).

— Observe consistently $C \cdot A^n n^{-5/2}$ and area distribution.

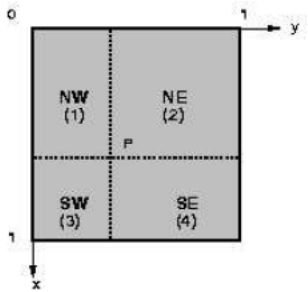
— Postulate this property for SAPs (self-avoiding polygons).

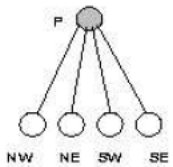
— Compute *exact* values for $n \leq 120$

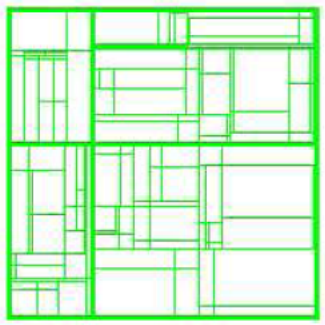
— Verify consistency of lower order asymptotics.

$\star\star$ FACT: $SAP_n \sim CA^n n^{-5/2}$ and area is Airy.!

Quadtrees and the holonomic framework







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

Partial Match Query (1/2) : $PMQ_n^{(1/2)} \approx n^{(\sqrt{17}-3)/2}$.

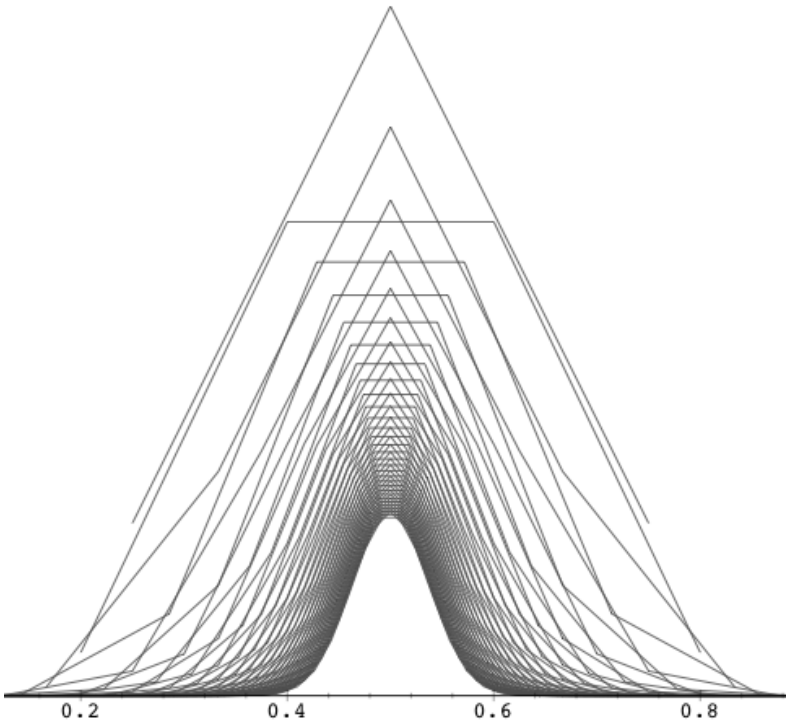
Stanley-Lipshitz-Zeilberger-Gessel: Holonomic framework = linear ODEs with rational coefficients.

A theory of special functions. Equality is decidable; asymptotics are "essentially" decidable.

PART C. DISTRIBUTIONS

Runs in permutations:

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$



For combinatorial class $\mathcal{F}$ with parameter χ, get bivariate GF $F(z, u)$ which is deformation of $F(z, 1) = F(z)$. $[z^n]F(z, u)$ is proportional to the *probability generating function* of χ on $\mathcal{F}_n$.

⋆ For functions $F(z)$ with finite singularities, *usually*,

$$[z^n]F(z) = \rho^{-n} n^{\delta},$$

ρ given by location and n^{δ} by nature of sings.

⋆ For $F(z, u)$, expect to get uniform & analytic

$$[z^n]F(z, u) = \rho(u)^{-n} n^{\delta} \quad \text{or} \quad \rho^{-n} n^{\delta(u)} \equiv \rho^{-n} e^{\delta(u) \log n},$$

via perturbation analysis.

Quasi-Powers approximation:= $PGF_n(u) \approx B(u)^{\text{large}(n)}$.

Theorem (H-K. Hwang's Quasi-Powers Theorem)

In the Quasi-Powers situation, $PGF_n(u) \approx B(u)^{\text{large}(n)}$, *one has:*
(*i*) *convergence to a Gaussian law*

$$\mathbb{P}_n\left[\frac{\chi - \mathbb{E}[\chi]}{\sqrt{\mathbb{V}[\chi]}} \le x\right] \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt;$$

(*ii*) *speed of convergence;* (*iii*) *moment estimates;* (*iv*) *a large deviation principle.*

Works for movable singularity & variable exponent!

[Bender, Richmond, F., Soria, Hwang] Based on: continuity theorem for characteristic functions; Berry-Essen inequalities; differentiability properties of holomorphic functions; basic large deviation theory.

A "conceptual" proof: polynomials over finite fields.

⋆ Polynomials are sequences of coeffs $\Longrightarrow$ $P(z)$ has pole.

⋆ Polynomials are multisets of irreducibles $\Longrightarrow$ $P \approx \exp(I)$, so that $I(z)$ is logarithmic.

The density of irreducibles is $\sim q^n/n$.

⋆ Bivariate relation $P(z,u) \approx e^{uI(z)}$ implies movable exponent implies Gaussian law.

The number of irreducibles is asymptotically normal, with $\log n$ *scaling.*

Cf Prime Number Theorem and Erdös–Kac. Analysis of polynomial factorization [F-Gourdon-Panario]. The **exp-log** schema [F-Soria]. Cf [Arratia-Barbour-Tavare].

Applications to Analysis of Algorithms

— **search trees**: binary, multiway, locally balanced, paged; quicksort and quickslect;

— *__multidimensional search__: k-d-trees, quadtrees; paged, relaxed.

— *__digital structures__: tries, ternary search tree hybrids, multidimensional trees; protocols, leader election; skip lists, . . .

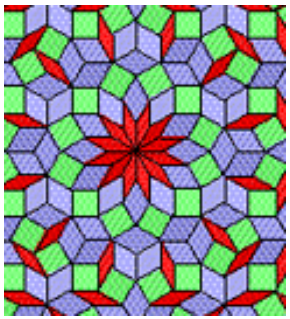
— *__data compression__: LZ algorithms, suffix trees.

— *__hashing__: random/uniform probing; LPH; paged; alternative displacements;

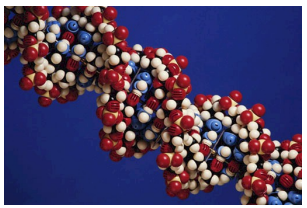
— **priority trees, heaps, mergesort, sorting networks**;

— *__symbolic manipulation__: polynomial GCDs, factorization; symbolic differentiation and term-rewritings;

— **quantitative data mining**: probabilistic & approximate counting; Loglog counting; adaptive sampling.



Patterns in sequences

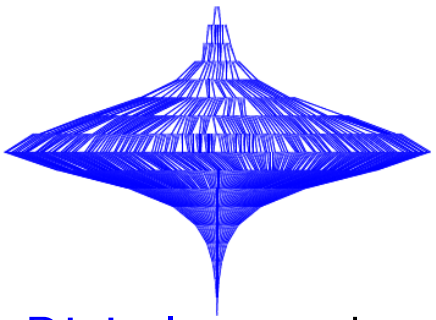


Many kinds of patterns are recognized by finite automata leading to rational functions whose poles move smoothly.

"In random sequence, the number of pattern occurrences is asymptotically normal, for a great variety of patterns and information sources."

[Guibas-Odlyzko; Régnier-Szpankowski; Nicodème-Salvy-F; Vallée]

"Borges' Theorem" for local patterns is known to hold in: words, trees, permutations, search trees, maps, etc. [Devroye, Martinez, F., Bender, Gao, Noy-Elizalde, . . .]



Digital structures and data compression

⋆ Digital trees aka *"tries"* & variants are amenable to analytic combinatorics: GFs, singularity analysis, Mellin transforms, saddle point method ≅ analytic depoissonization.

E.g., Jacquet-Szpankowski DST eqn $\boxed{F(z,u) = \int F(pz,u) \cdot F(qz,u)}$.

Vallée's dynamical sources: *"The cost of radix-sorting of n continued fractions depends on the Riemann hypothesis."*

⋆ Suffix trees too: combine with pattern analyses.

"Redundancy of Lempel-Ziv compression algorithms can be precisely quantified."

The trie saga. [De Bruijn-Knuth, F.-Sedgewick, Devroye, Pittel, Jacquet-Szpankowski-Louchard, Vallée-F.]. [Szpankowski's red book]. Cf [Devroye-Szpankowski@SODA'07] . . .

PART D. FRONTIERS

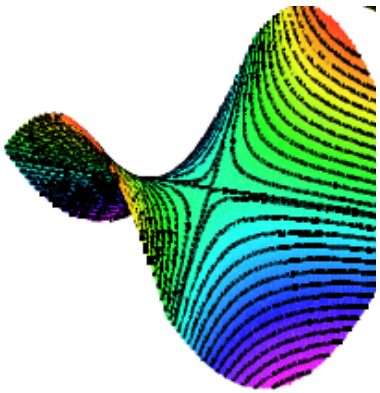
⋆ Organize the field into analytic-combinatorial schemas exhibiting universal properties. Towards a theory of combinatorial processes

⋆ Expand the scope of analytic methods to hard computational problems.

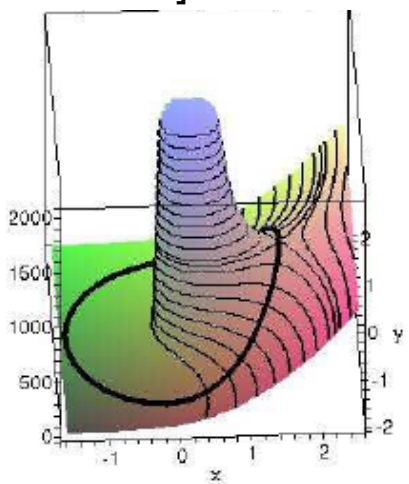
⋆ Determine decidable classes and work out decidability algorithms within symbolic manipulation systems like Maple, Mathematica.

Hard combinatorial problems (NP): *What are the feasible/ unfeasible regions for random problem instances?*

E.g.: answer not known for 3-SAT [2-SAT: BoBoCh+]



Saddle point (SP) method



Represent problem by n–dimensional Cauchy integral; estimate by SP.

E.g. 3–regular graphs from general graphs.

$$RG_n^{(3)} = \frac{1}{(2i\pi)^n} \int \cdots \int \prod_{1 \le i < j \le n} (1 + z_i z_j) \, \frac{dz_1 \cdots dz_n}{z_1^4 \cdots z_n^4}.$$

B. Mc Kay has developed a specific calculus. (Gives access to exponentially sparse families and can "filter" according to many constraints.)

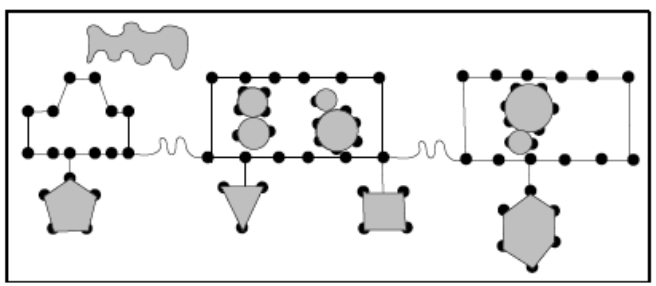
Computability within the calculus of analytic combinatorics.

Algorithms and programs for "automatic combinatorics"?

Theorem (Properties of specifications)

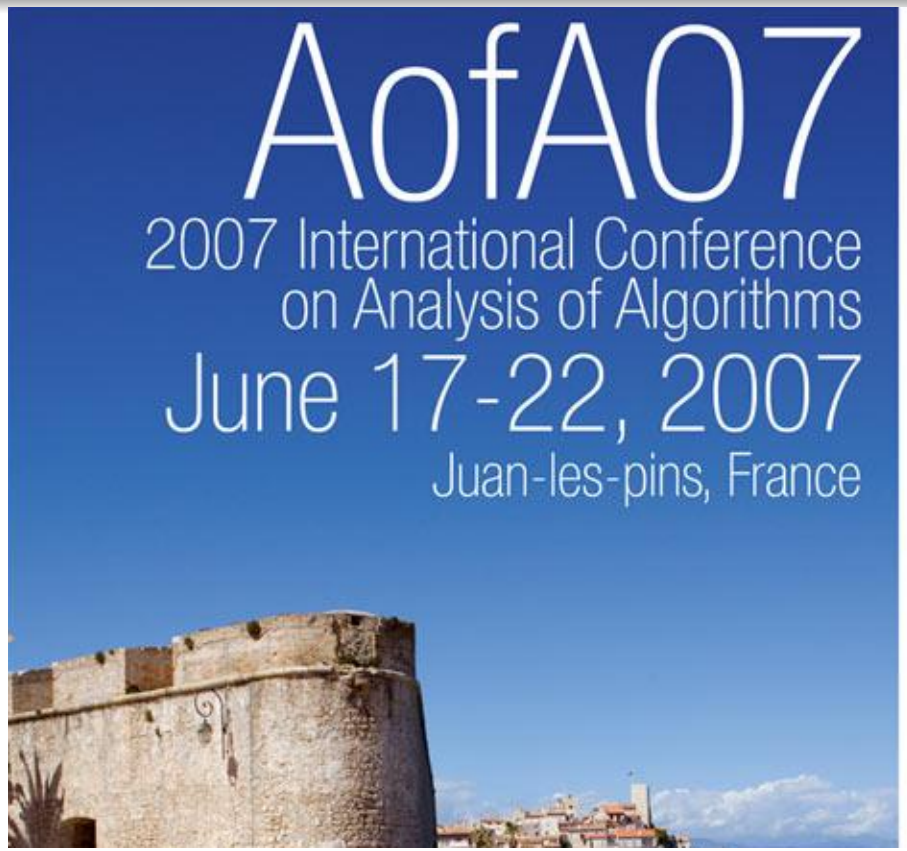
For the core language of constructions:
(*i*) *counting sequences are computable in* $O(n^{1+\epsilon})$;
(*ii*) *GF equations are computable;*
(*iii*) *partially decidable asymptotic properties;*
(*iv*) *random generation by either recursive method or Boltzmann models is achievable in low polynomial time.*

[F-Salvy-Zimmermann] [Duchon-F-Louchard-Schaeffer] [F-Fusy-Pivoteau]



$$Tr(z) = \frac{z^2\left(1+\log((1-z)^{-1})\right)}{\left(1-z^2\left(1+\log((1-z)^{-1})\right)\right)}\left(1-\frac{z^2\left(1+\log((1-z)^{-1})\right)e^{\left(\log((1-z)^{-1})\right)^2}}{1-z^2\left(1+\log((1-z)^{-1})\right)}\right)^{-1}$$

together with the expansion



http://www.aofa2007.org/

That's All, Folks!