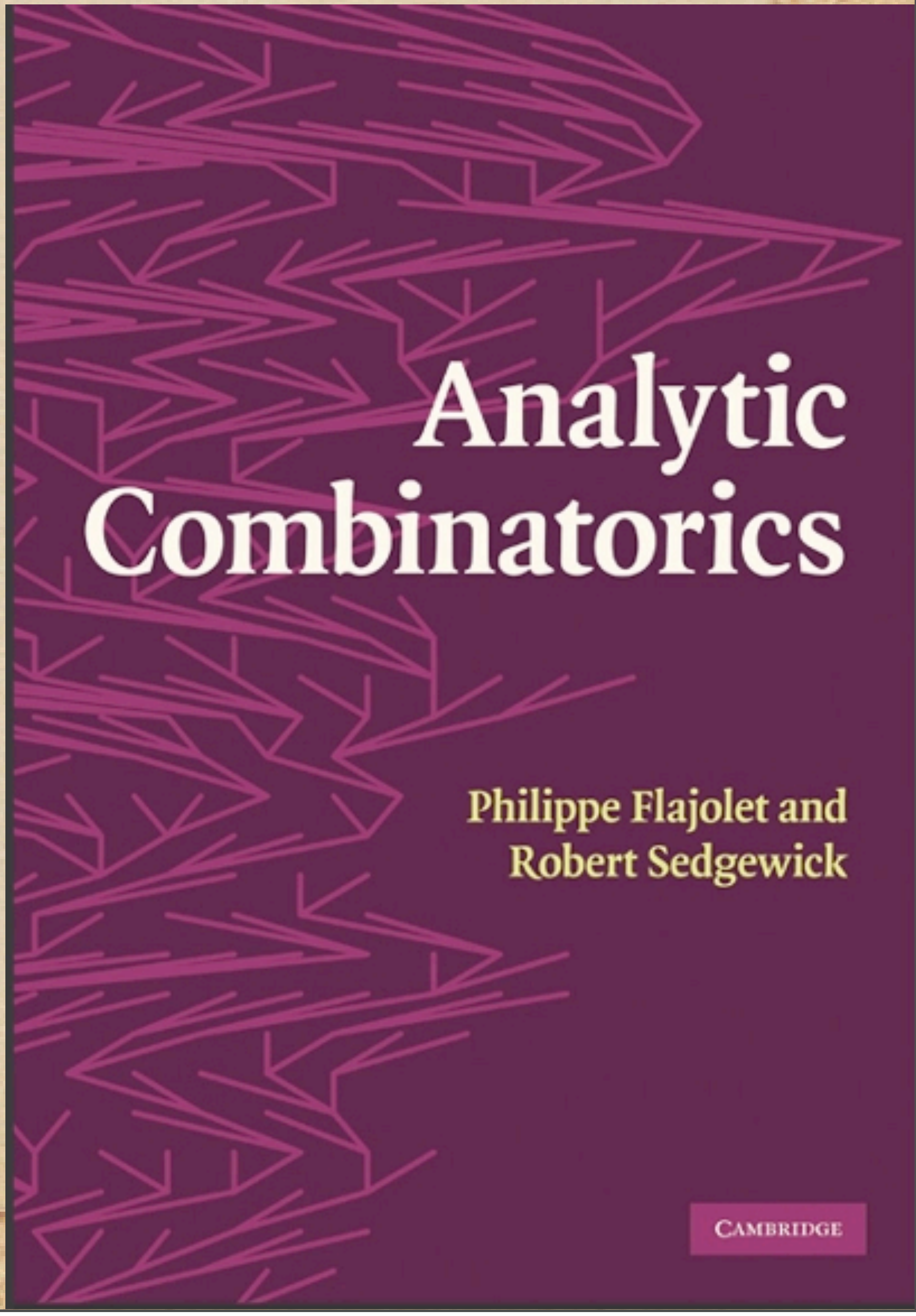
Séminaire de Probabilités, Paris June 2010

The Digital Tree:

Analysis and Applications

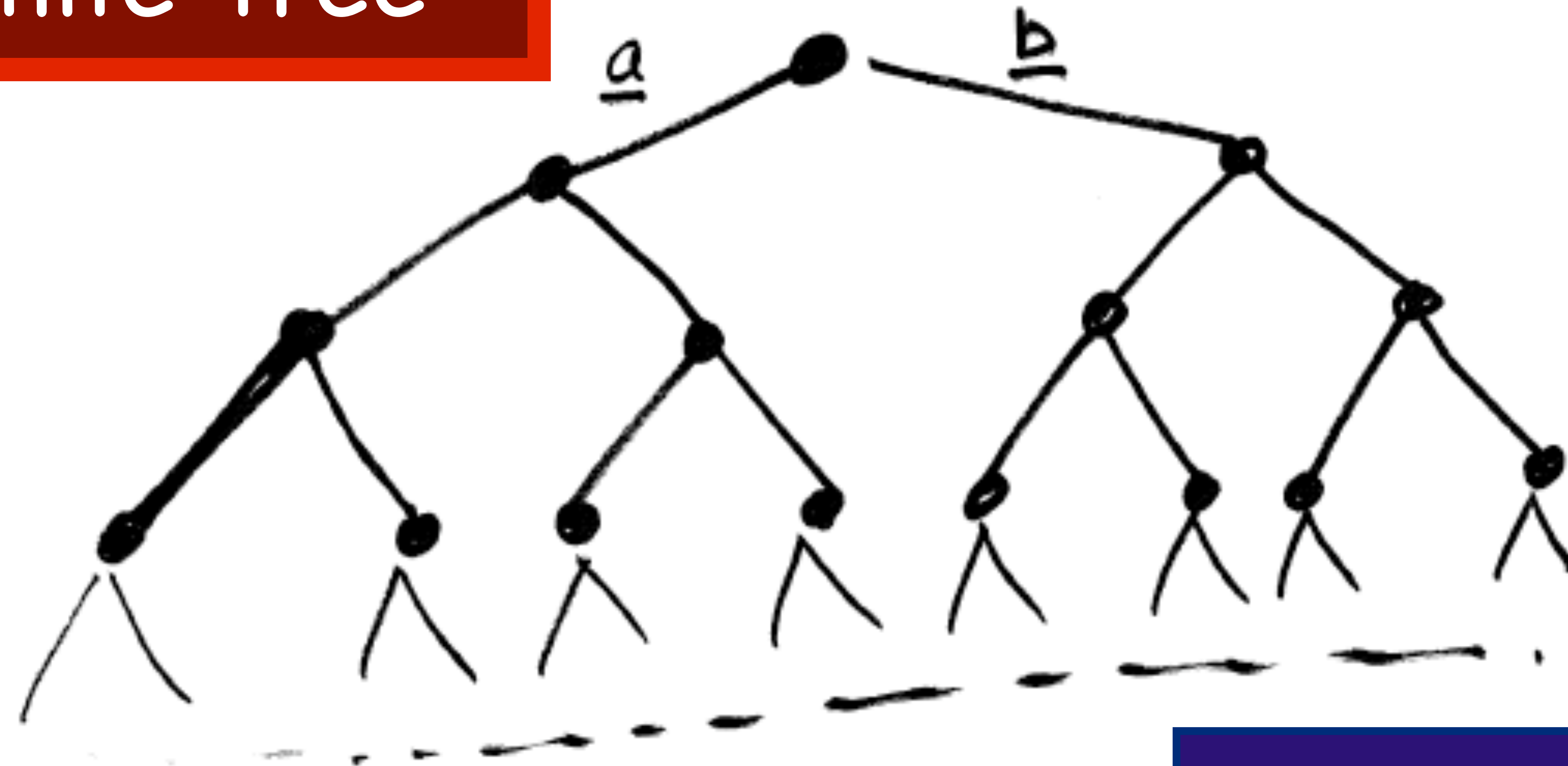
Philippe Flajolet, INRIA Rocquencourt

- A (finite) tree associated with a (finite) set of words over an alphabet A.
- Equipped with a randomness model on words, we get a random tree, indexed by the number n of words.
- Characterize its probabilistic properties, mostly with COMPLEX ANALYSIS.

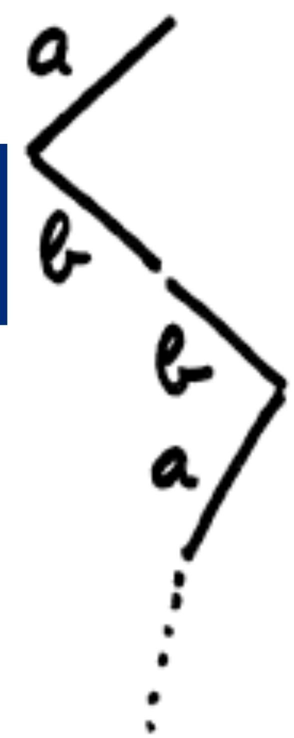


1. Digital Trees & Algorithms

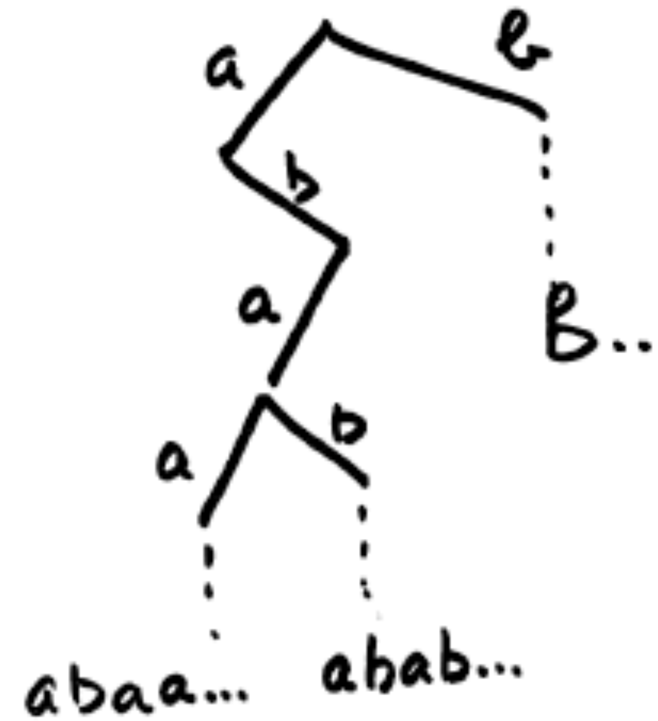
infinite tree



word <--> branch



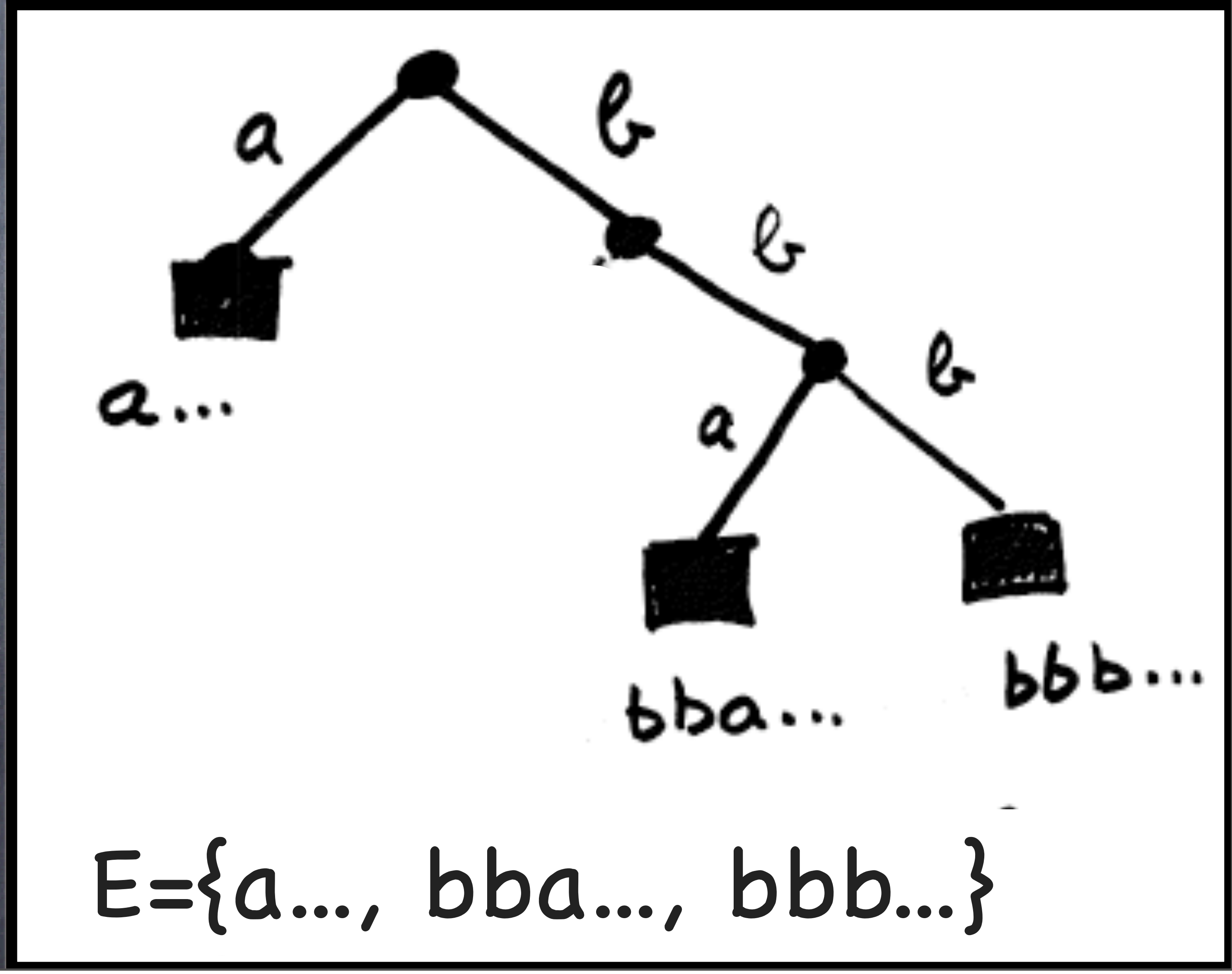
set of words <--> partial tree



- DIGITAL TREE aka "TRIE":= STOP descent by pruning long one-way branches.

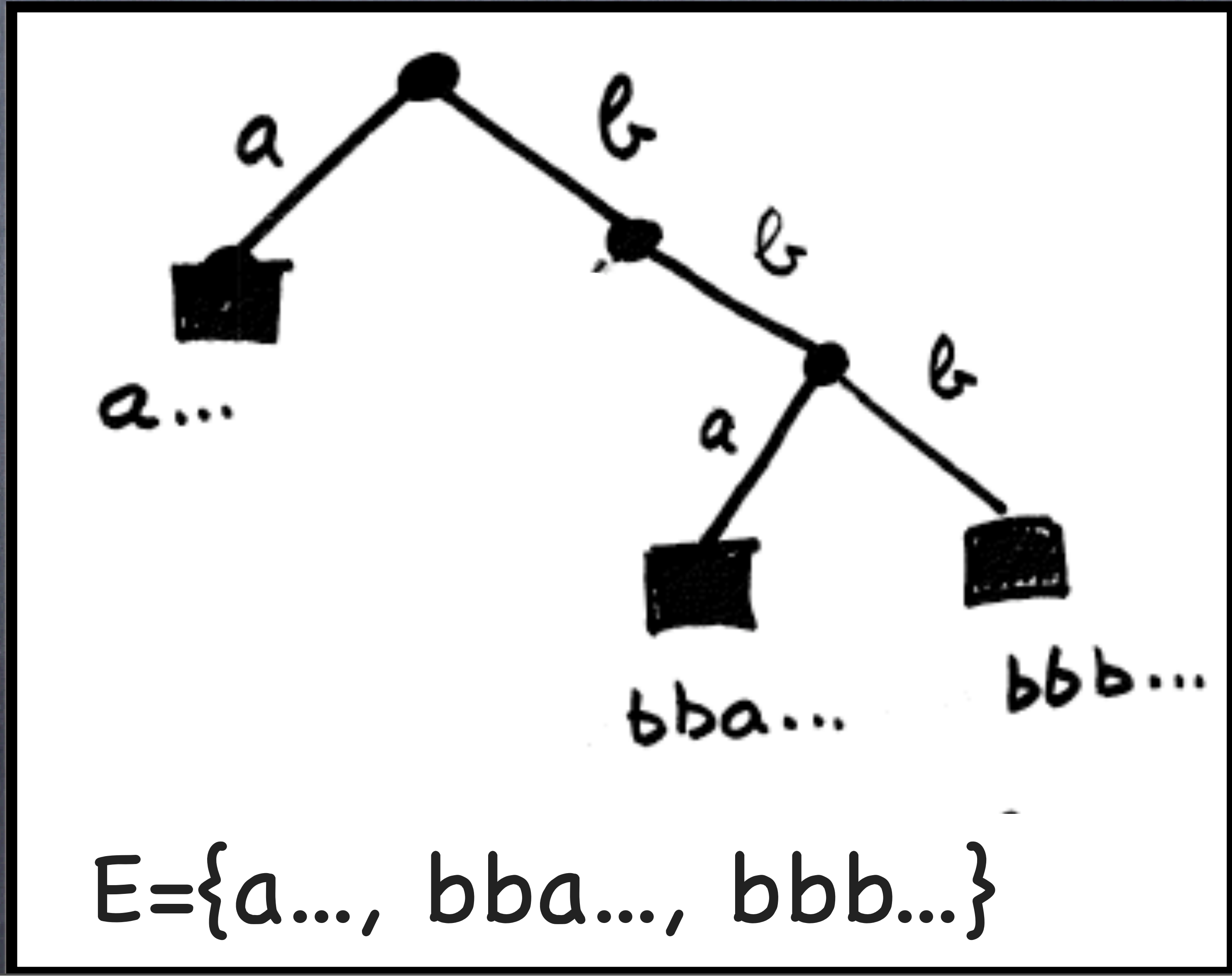
~Only places corresponding to 2+ words (and their immediate descendants) are kept.

~The digital tree is finite as soon as built out of distinct words.

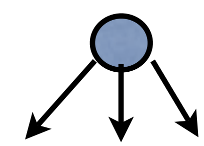


- TOP-DOWN construction: Set E is separated into $E_a, \ldots, E_z$ according to initial letter; continue with next letter...

- INCREMENTAL construction: start with the empty tree and insert elements of E one after the other... (Split leaves as the need arises.)



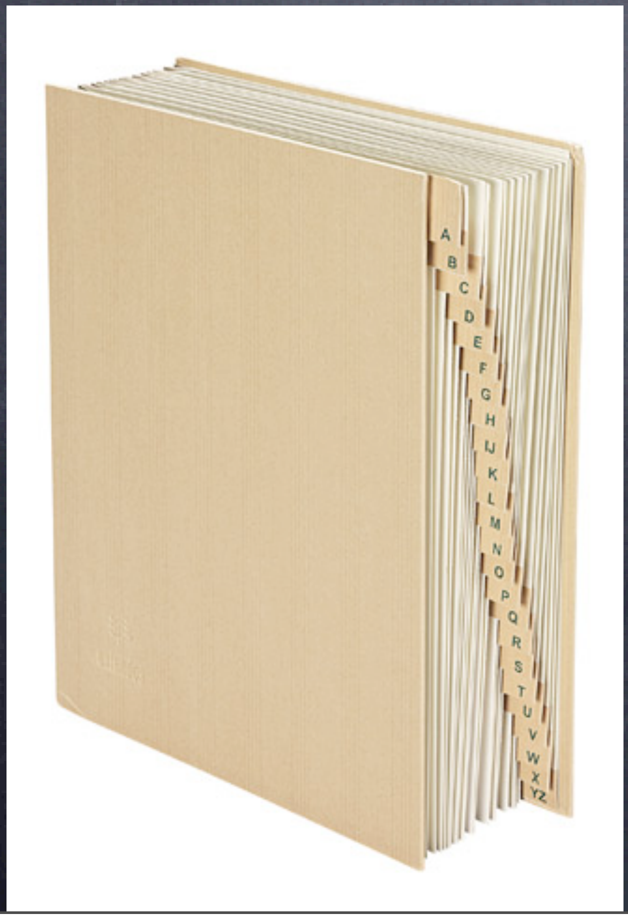
- SUMMARY:

Alphabet $\mathcal{A}$	$\mapsto$	
Words $\mathcal{W} = \mathcal{A}^{\infty}$	$\mapsto$	**Branches**
"Sets" (n–seq.) of words $\mathcal{W}^n$	$\mapsto$	**Digital tree**
Random word model on $\mathcal{W}$	$\mapsto$	**Random Tree$_n$.**

Memoryless (Bernoulli) p,q; Markov, CF

Algorithms: 1 - Dictionaries

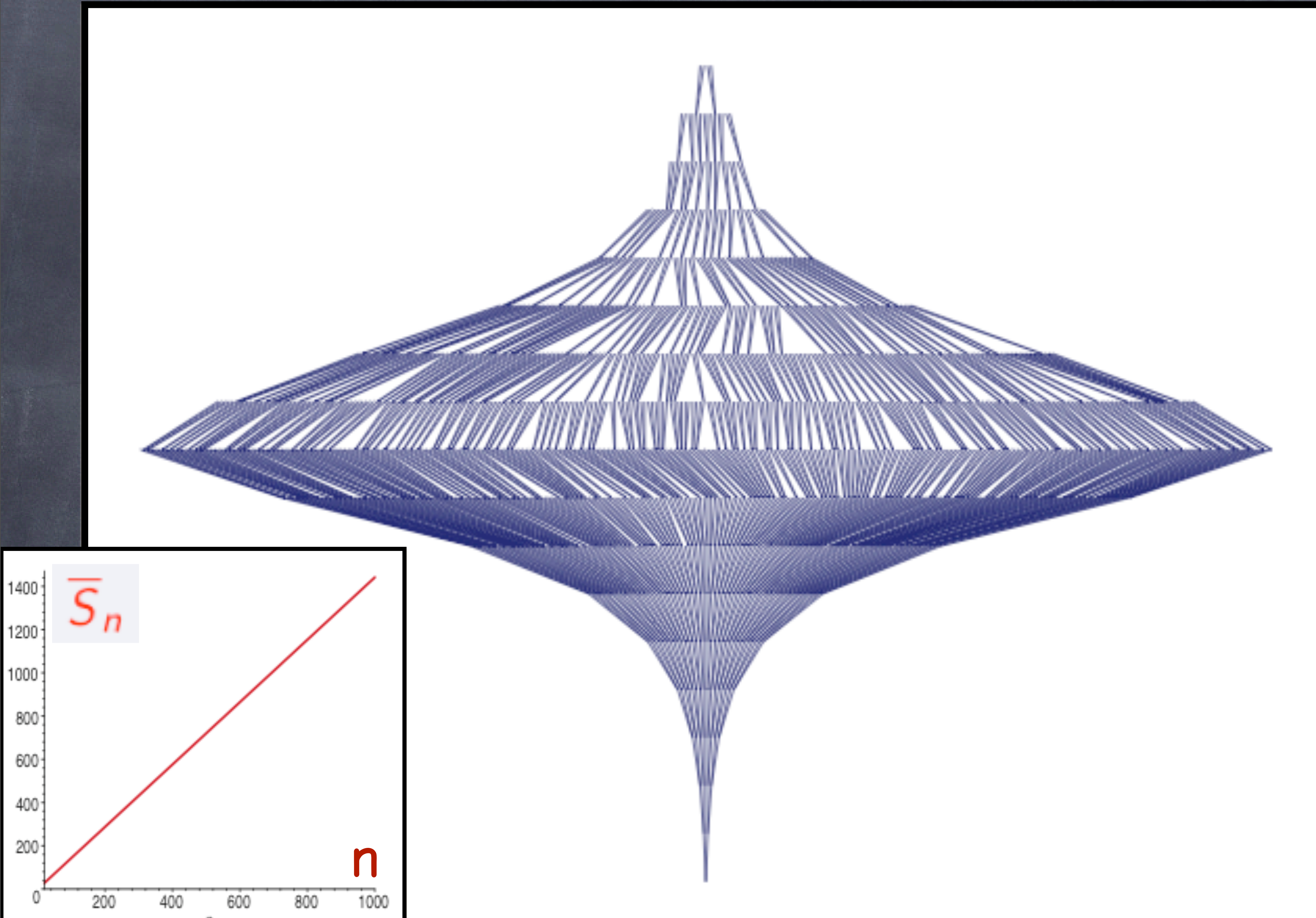
- Manage dynamically dictionaries; hope for O(log n) depth?
- Save space by "factoring" common prefixes; hope for O(n) size?
- However, worst-case is unbounded...



"TRIE"=tree+retrieval

(Fredkin, de la Briandais ~1960)

Analysis?



A random trie on n=500 uniform binary sequences;
size =741 internal nodes; height=18

Algorithms: 2 -Hashing

- Data may be highly structured and share long prefixes. Use a transformation

  h: W -> W'

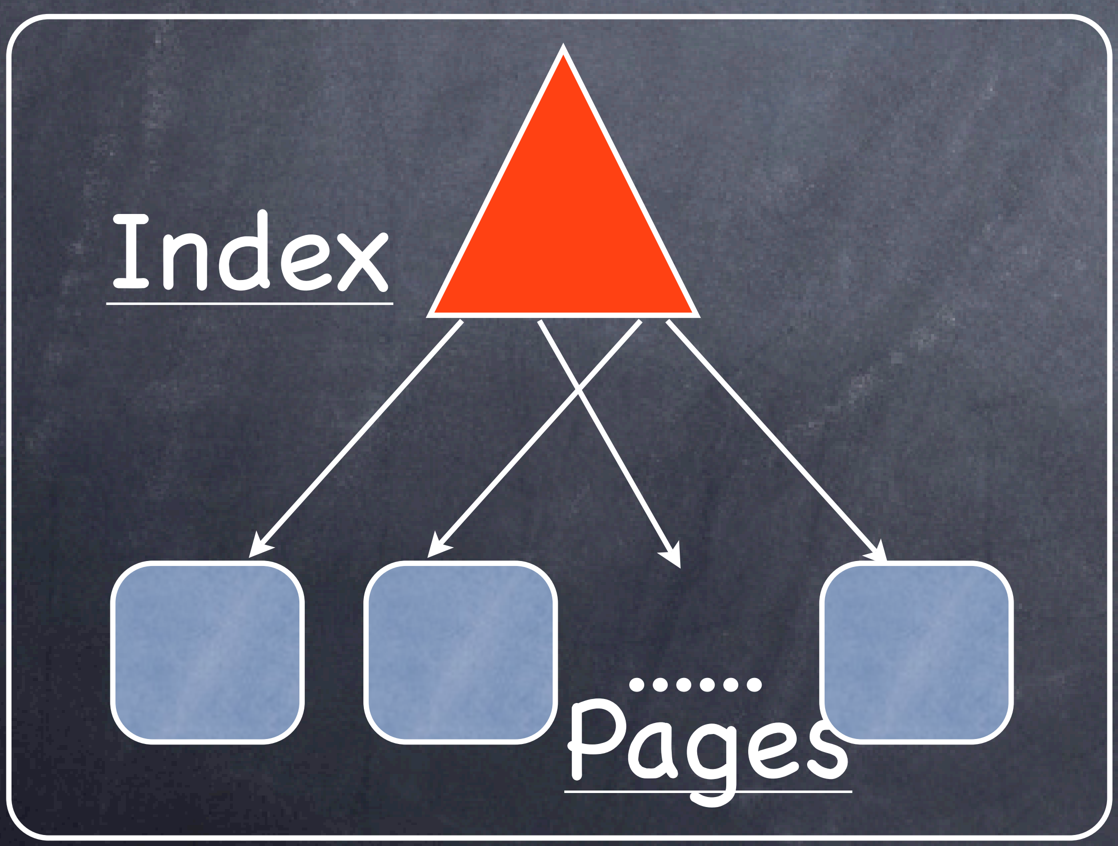
  called "hashing" (akin to random number generators.)

- Uniform binary data are meaningful!

Analysis?

Algorithms: 3 -Paging

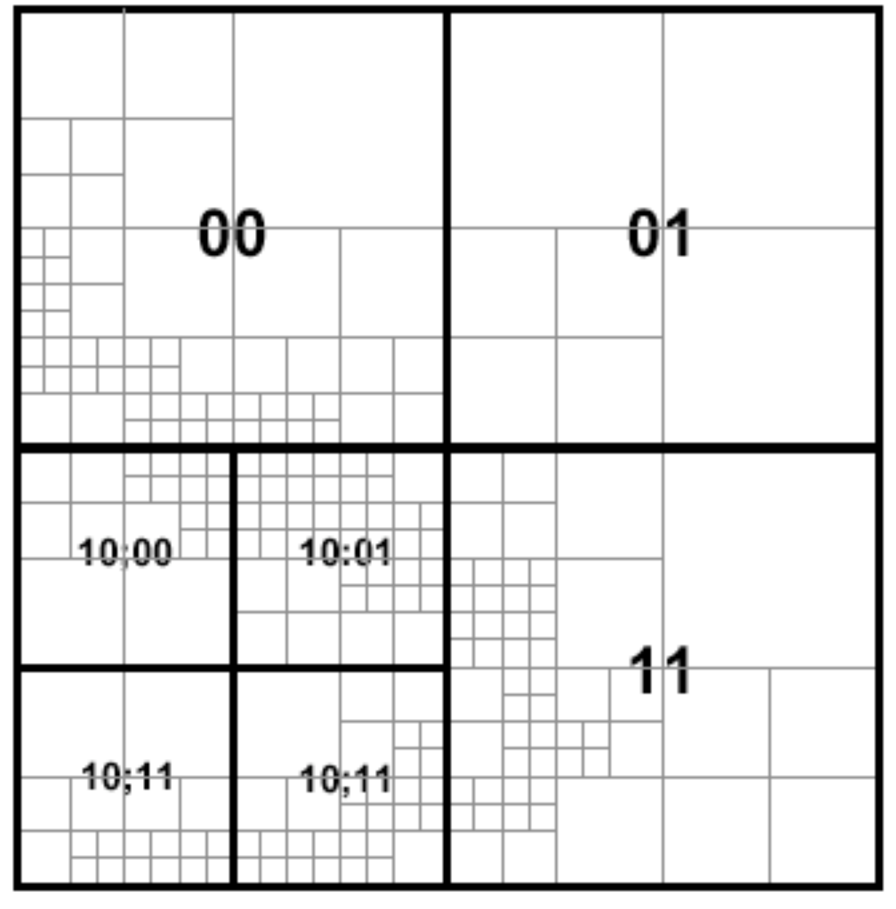
- Data may be accessible by blocks, e.g., pages on disc. Stop recursion as soon as "b" elements are isolated (standard: b=1).
- Combine with hashing = get index structure.



Analysis?

Algorithms: 4-MultiDim

- Data may be multidimensional & numeric/ geometric.

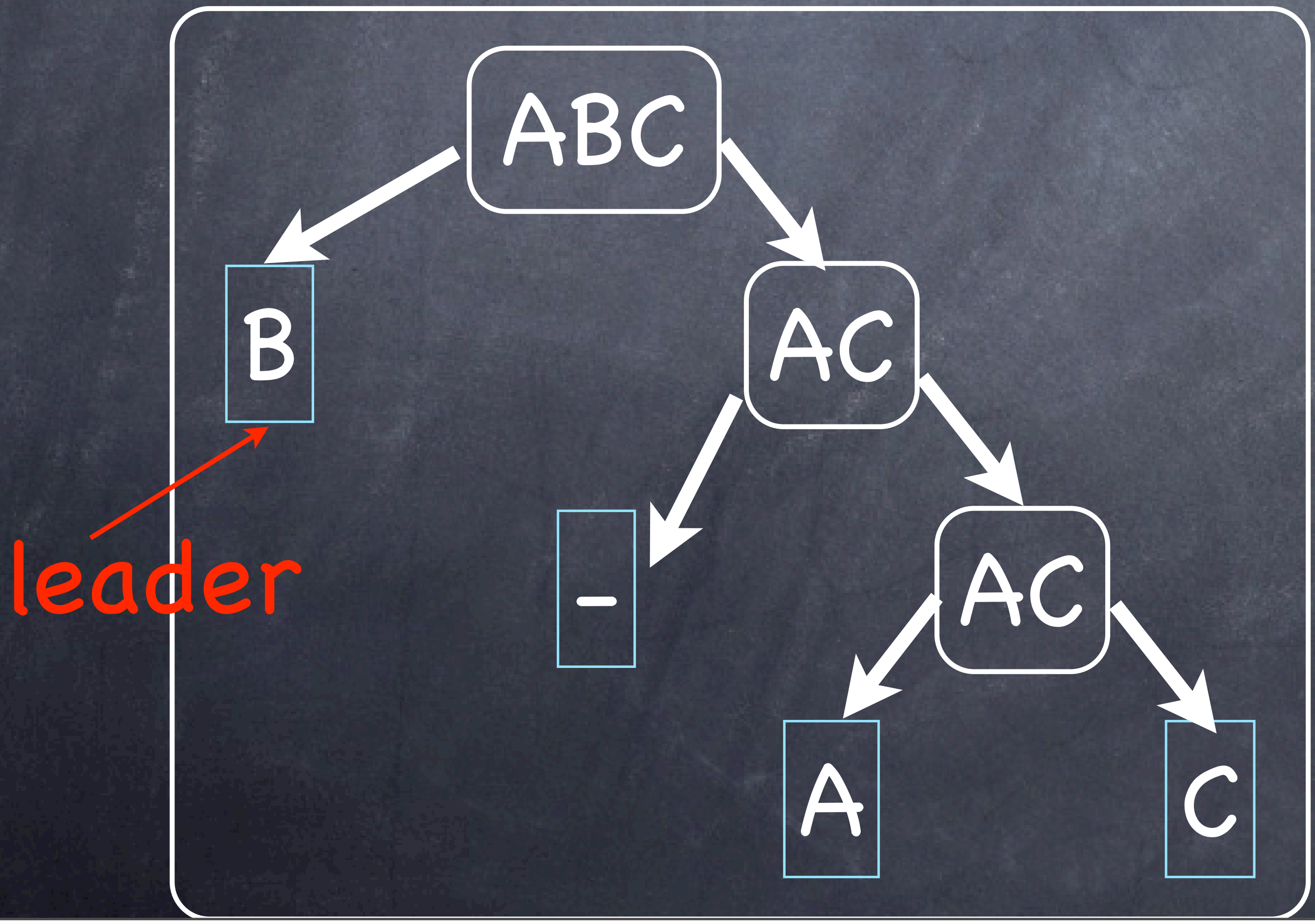


quad-trie

Analysis?

Algorithms: 5-Communication

- Data may be distributed and accessible only via a common channel (network).
- Everybody speaks at the same time; if noise, then SPLIT according to individual coin flips.

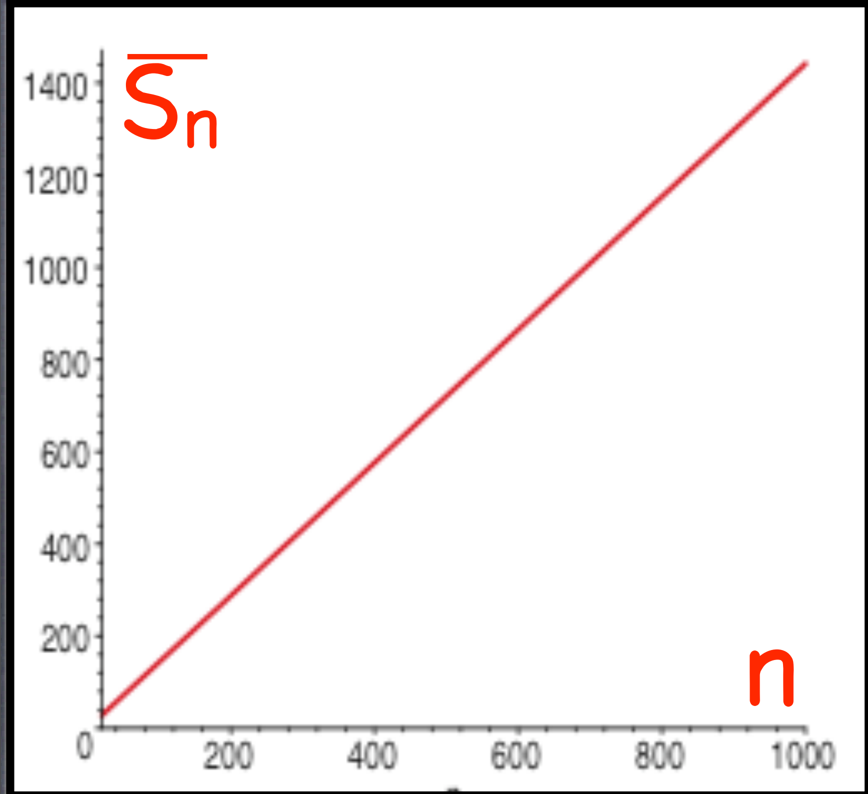


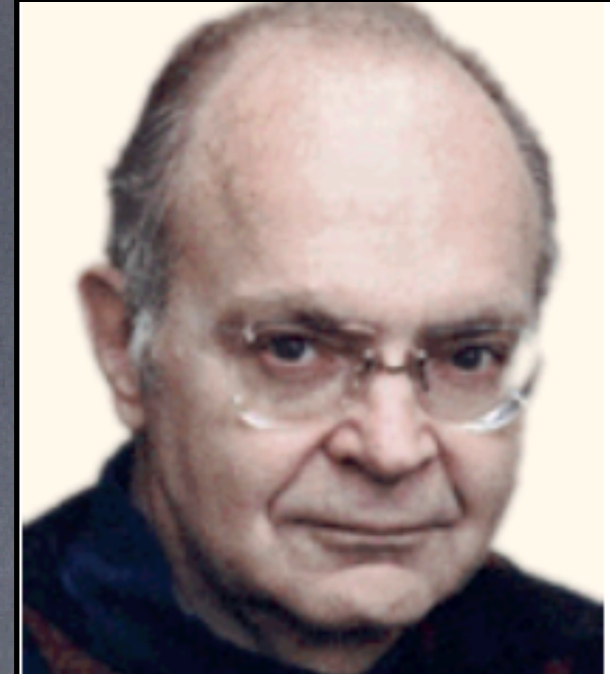
tree protocol

Analysis?

2. Expectations

- Bernoulli vs Poisson models
- Mellin technology
- Fluctuations and error terms





Theorem (Knuth +De Bruijn, 1965+)

For n uniform binary words:

- *Expected number of internal nodes (size) $\overline{S}_n$ is such that $\overline{S}_n/n$ has no limit; it fluctuates with amplitude about 10^{-6}:*

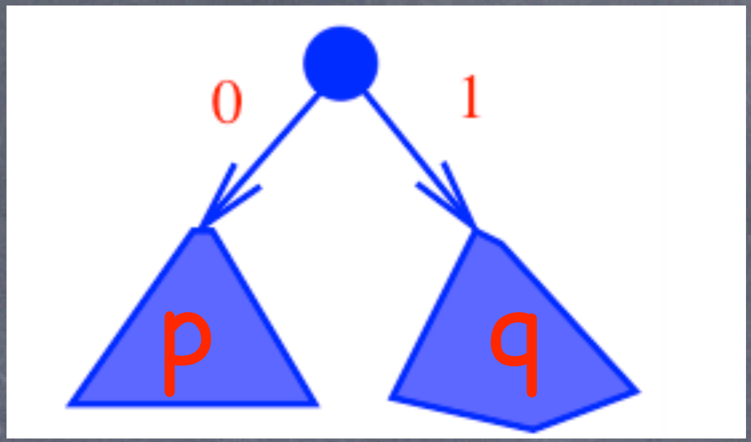
$$\boxed{\frac{\overline{S}_n}{n} \approx \frac{1}{\log 2} \pm 10^{-6}}.$$

- *Expected depth $\overline{D}_n$ of a random leaf satisfies*

$$\boxed{\overline{D}_n = \log_2 n + O(1)}.$$

(Proof in a "modernized" version follows....)

Algebra...



- **Assumption 1.** the number N of elements is $\mathcal{P}$oisson(x).
- **Assumption 2**: a binary alphabet with probabilities p, q.

Let $\sigma(E)$ be the number of internal nodes in the tree:

$$\sigma(E) := \mathbf{1}_{[\#E \geq 2]} + \sigma(E_0) + \sigma(E_1).$$

Let $S(x) := \mathbb{E}_{\mathcal{P}(x)}(\sigma)$. Since thinning of a $\mathcal{P}(x)$ by a Bernoulli RV of parameters p, q gives $\mathcal{P}(px)$, $\mathcal{P}(qx)$:

$$S(x) = \left[1 - (1 + x)e^{-x}\right] + S(px) + S(qx).$$

Algebra...

Solving by iteration

$$S(x) = g(x) \ + \ S(px) \ + \ S(qx).$$

yields, e,g., with $p = q = \frac{1}{2}$ and $g(x) = 1 - (1+x)e^{-x}$, for size:

$$\boxed{S(x) = \sum_{k \geq 0} 2^k g\left(\frac{x}{2^k}\right).}$$

In general, get $S(x) = \sum_{k,\ell} \binom{k+\ell}{k} g(p^k q^\ell x) \equiv \sum_{w \in \{0,1\}^\star} g(p_w x).$

With $\overline{S}_n$ the expected tree size when the tree contains n elements and $S(x)$ the Poisson expectation:

$$S(x) = \sum_{n \geq 0} \overline{S}_n e^{-x} \frac{x^n}{n!}.$$

The Poisson expectation $S(x)$ is like a generating function of $\{\overline{S}_n\}$.
Go back —"depoissonize"— by Taylor expansion. E.g.:

$$\overline{S}_n = \sum_k \left[1 - \left(1 - \frac{1}{2^k}\right)^n - \frac{n}{2^k}\left(1 - \frac{1}{2^k}\right)^{n-1} \right], \qquad p = q = \frac{1}{2}.$$

Many variants are possible and one can justify that

$\overline{S}_n = S(x) + \text{small}$ when $x = n$. **(elementary)**

Analysis...

The Mellin transform

$$f(x) \overset{\mathcal{M}}{\rightsquigarrow} f^{\star}(s) := \int_0^{\infty} f(x) x^{s-1}\, dx$$

(It exists in strips of $\mathbb{C}$ determined by growth of $f(x)$ at $0, +\infty$.)

Property 1. Factors *harmonic sums*:

$$\sum_{(\lambda,\mu)} \lambda f(\mu x) \overset{\mathcal{M}}{\rightsquigarrow} \left(\sum_{(\lambda,\mu)} \lambda \mu^{-s} \right) \cdot f^{\star}(x).$$

Property 2. Maps asymptotics of f on singularities of $f^{\star}$:

$$f^{\star} \approx \frac{1}{(s-s_0)^m} \quad \Longrightarrow \quad f(x) \approx x^{-s_0} (\log x)^{m-1}.$$

Proof of $\mathbf{P_2}$ is from Mellin inversion + residues:

$$f(x) = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} f^{\star}(s) x^{-s}\, ds.$$

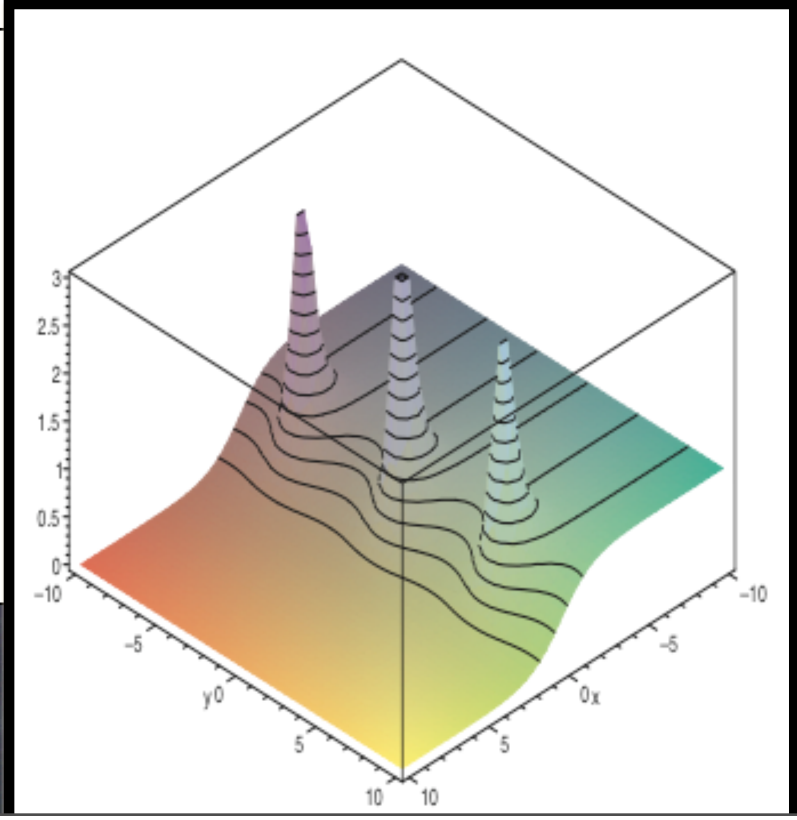
$\boxed{p = q = 1/2}$: $S(x) = \sum_k 2^k g(x/2^k)$, with $g(x) = 1 - (1+x)e^{-x}$.

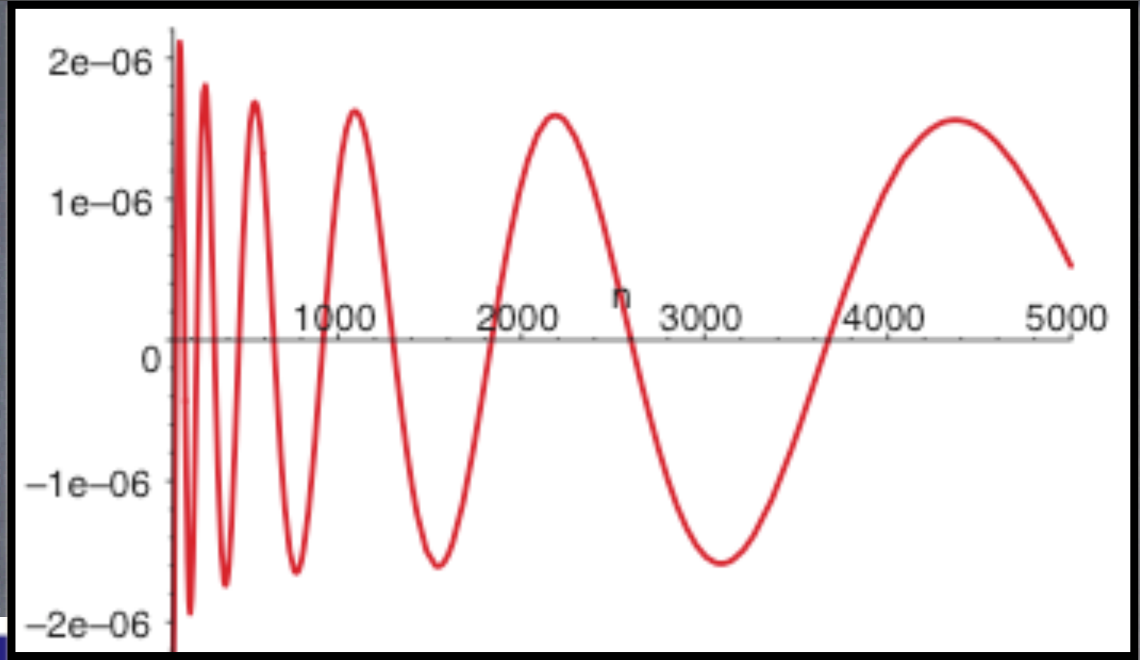
- **Harmonic sum property:**

$$S^\star(s) = \left(\sum 2^k 2^{ks}\right) \cdot (s+1)\Gamma(s) = \frac{\Gamma(s)}{1 - 2^{1+s}}.$$

- **Mapping properties:** $S^\star$ exists in $-2 < \Re(s) < -1$. Poles at $s_k = -1 + 2ik\pi/\log 2$, for $k \in \mathbb{Z}$.

Location of pole (s_0)	$\rightsquigarrow$	Asymptotics of $f(x) \approx x^{-s_0}$
$s_0 = \sigma + i\tau$	$\rightsquigarrow$	$x^{-\sigma} e^{i\tau \log x}$





Theorem (Knuth + De Bruijn, 1965+)

For n uniform binary words, $p = q = \frac{1}{2}$:

- *Expected number of binary nodes (size)* $\overline{S}_n$ *is such that* $\overline{S}_n/n$ *has no limit; it satisfies*

$$\overline{S}_n = \frac{n}{\log 2} + nP(\log_2 n) + O(1),$$

where $P(u)$ *is a Fourier series of amplitude about* 10^{-6}.

Proof above is for Poisson expectation; it transfers to $\overline{S}_n$. Also, things work similarly for depth: $\overline{D}_n = \log_2 n + Q(\log_2 n) + o(1)$.

Memoryless sources (I)

Correspond to $p \neq q$. Dirichlet series is $\dfrac{1}{1 - p^{-s} - q^{-s}}$.

Theorem (Knuth 1973; Fayolle, F., Hofri 1986, ...)

Let $H := p \log p^{-1} + q \log q^{-1}$ be the entropy.

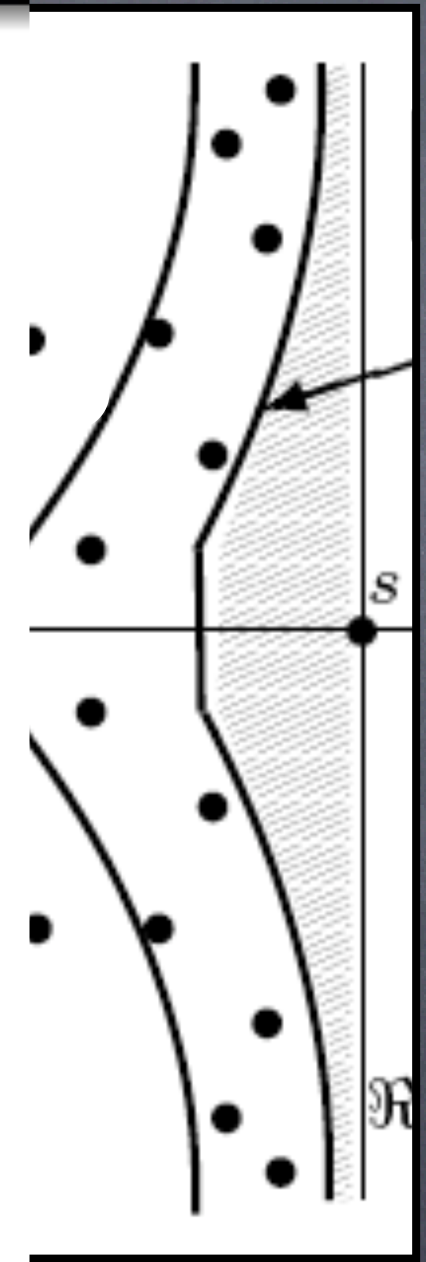
- *In the periodic case, $\frac{\log p}{\log q} \in \mathbb{Q}$, there are fluctuations in $\overline{S}_n$.*
- *In the aperiodic case, $\frac{\log p}{\log q} \notin \mathbb{Q}$:*

$$\overline{S}_n \sim \frac{n}{H} \quad \text{and} \quad \overline{D}_n \sim \frac{1}{H} \log n,$$

Philippe Robert & Hanene Mohamed relate this to the periodic/aperiodic dichotomy of *renewal theory* (2005+).

Memoryless sources (II)

- The geometry of poles of $\dfrac{1}{1 - p^{-s} - q^{-s}}$ intervenes.
- This geometry relates to Diophantine properties of $\alpha := \dfrac{\log p}{\log q}$



Theorem (F., Roux, Vallée, 2010)

If α has a finite irrationality measure, then $\exists\theta$:

$$S_n = \frac{n}{H} + O\left(\exp\left(-(\log n)^{1/\theta}\right)\right), \qquad \theta > 1.$$

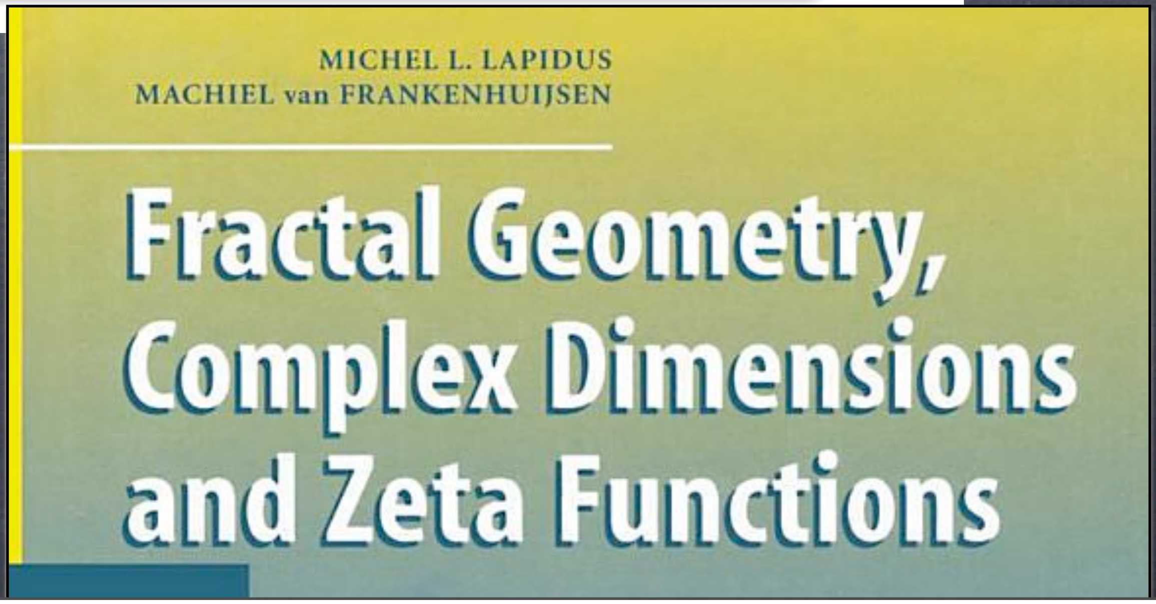
Such is the case for almost all $p \in (0,1)$ and all rational $p \neq \frac{1}{2}$.

Definition (Irrationality measure)

The number $\alpha \notin \mathbb{Q}$ has irrationality measure $\leq m$ iff the number of solutions of $\left|\alpha - \frac{a}{b}\right| < \frac{1}{b^m}$ is finite.

(pi, e, tan(1), log2, z(3), ...)

[Lapidus & van Frankenhuijsen 2006]



3. Distributions

- Analytic depoissonization &
  Saddle-points
- Gaussian laws ...

Saddle-points & analytic depoissonization

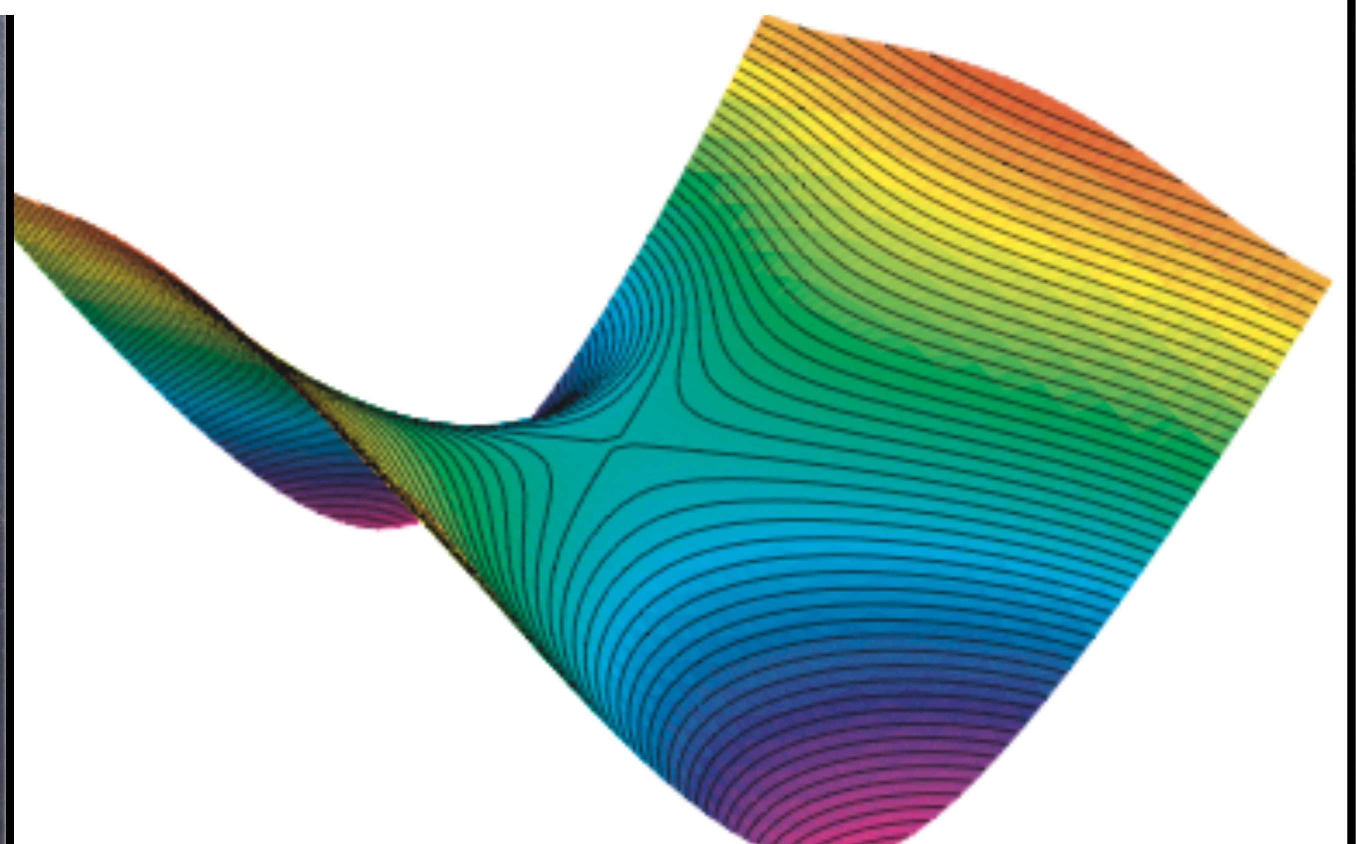
Height H of a b–trie (cf paging) with uniform binary words.

$$\mathbb{P}_n(H \le h) = n! \cdot \text{coeff.}[z^n]\, e_b\left(\frac{z}{2^h}\right)^{2^h}, \quad e_b(z) := 1 + \frac{z}{1!} + \cdots + \frac{z^b}{b!}.$$

- Cauchy: $[z^n]f(z) = \dfrac{1}{2i\pi}\displaystyle\int_\gamma f(z)\,\frac{dz}{z^{n+1}}$.

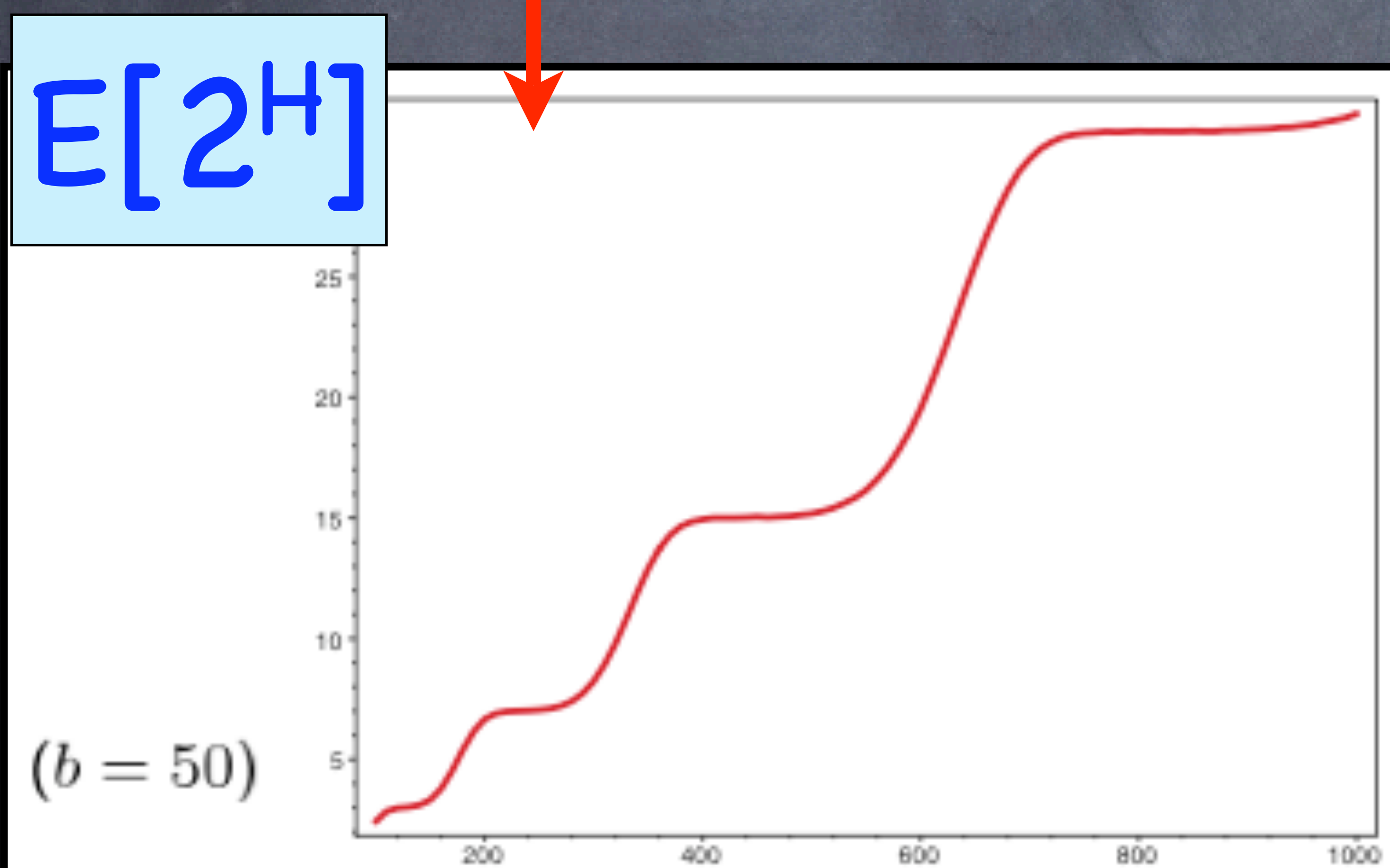
+ Saddle-point contour: concentration + local expansions.

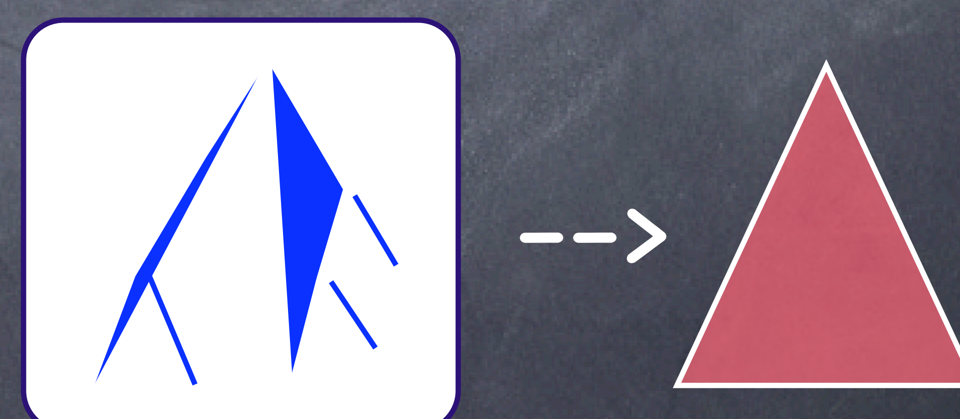
= Throw n balls into 2^h buckets, each of capacity b



Theorem (F. 1983)

The *expected height* *of a b–trie is* $\sim (1 + 1/b)\log_2 n$.
The size of the perfect tree embedding satisfies $\mathbb{E}(2^H) \asymp n^{1+1/b}$.
The *distribution* *is of of double-exponential type* $F(x) = e^{-e^{-x}}$,
with periodicities.





Analytic depoissonization

Theorem (Jacquet–Szpankowski 1995+)

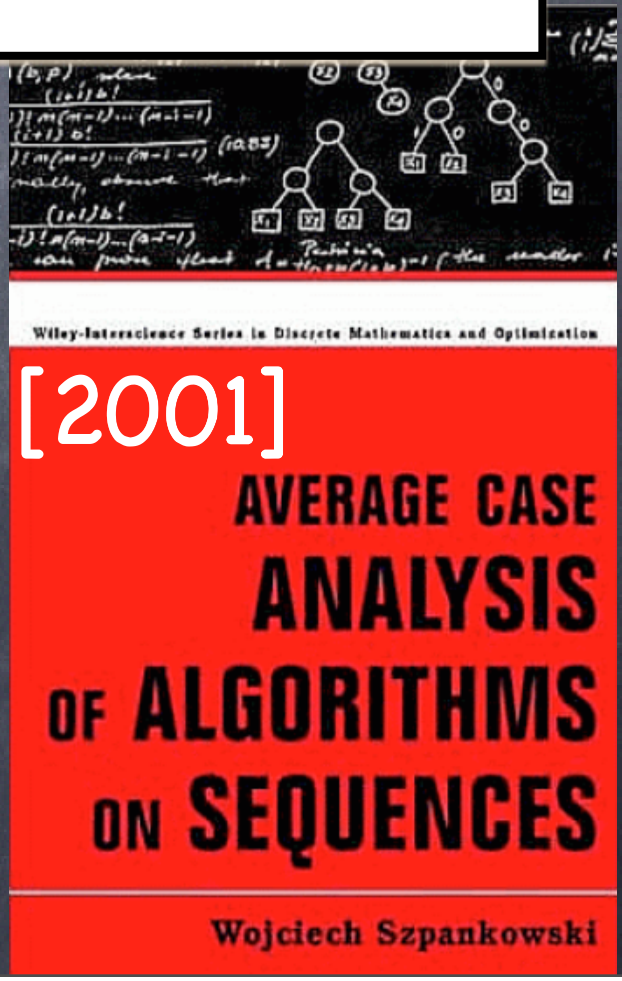
Let $X(\lambda)$ be a Poissonized expectation. Need X_n, which corresponds to conditioning upon Poisson RV $= n$. Assume:

(i) $X(\lambda)$ for complex λ near real axis has standard asymptotics;

(ii) $e^{\lambda}X(\lambda)$ is "small" in complex plane, away from real axis.

Then the Poisson approximation holds: $X_n \sim X(n)$

Proof: use Poisson expectation as a GF, plus Cauchy, plus saddle-point.



DISTRIBUTIONS:
size, depth, and path-length

Theorem (Jacquet-Régnier-Szpankowski, 1990++)

For general (p, q), the distribution of size is asymptotically normal.
The depth of a random leaf is asymptoticaly normal, if $p \neq q$.
The depth of a random leaf is asymptotically $\approx e^{-e^{-x}}$, if $p = q$.
The path-length ($\equiv \sum$depths) is asymptotically normal.

(p=q=1/2)

- Start with bivariate generating function F(z,u).
- Analyse log
- Analyse perturbation near u=1.
- Use analytic depoissonization
- Conclude by continuity theorem for characteristic fns.

$$F(z,u) = uF\left(\tfrac{z}{2},u\right)^2 + (1-u)(1+z)$$

$$\log F(z,u) = 2\log(F(z/2,u) + \cdots$$

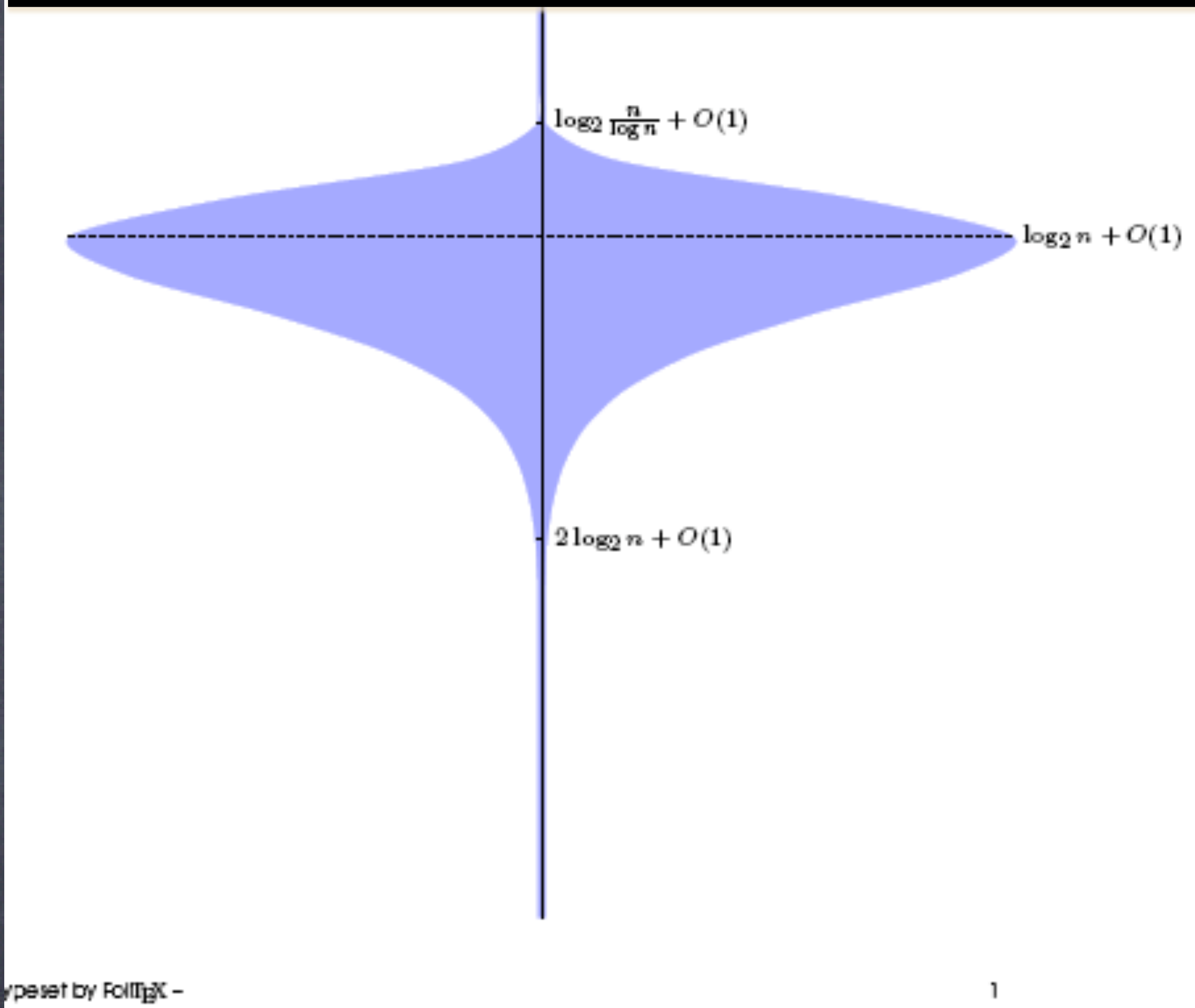
$$\log F(z,e^{it}) \approx z + i\mu_z t - \tfrac{1}{2}\sigma_z^2 t^2 + \cdots$$

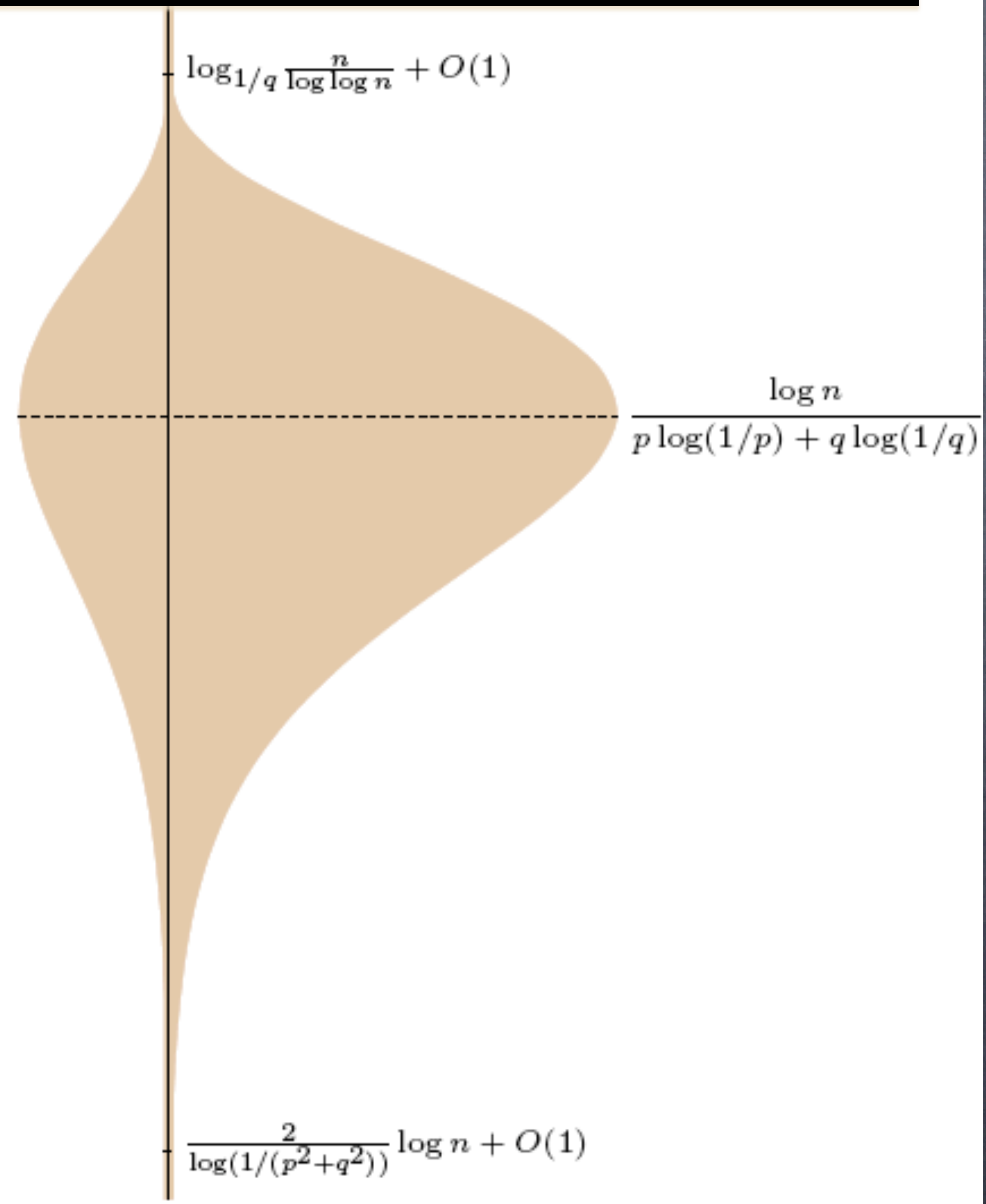
$$\text{get}[z^n]F(z,e^{it}) \approx \cdots$$

$$\mathbb{E}[e^{itS_n}] \rightsquigarrow e^{-t^2/2}$$

(case of size, p=q=1/2)

Profile of tries, after Szpankowski et al.





$(p = 0.5, \alpha_1 = \alpha_2 = 1/\log 2)$ $(p = 0.75)$

+ Cesaratto-Vallée 2010+

4. General sources

- Comparing and sorting real numbers
- Continued fractions
- Fundamental intervals...

Comparing numbers & sorting by continued fractions

$$\text{sign}\left(\frac{a}{b}-\frac{c}{d}\right)=\text{sign}(ad-bc).$$

Requires double precision and/or is unstable with floats.
(Computational geometry, Knuth's Metafont,...)
$\rightsquigarrow$ HAKMEM Algorithm (Gosper, 1972)

$$\frac{36}{113}=\cfrac{1}{3+\cfrac{1}{7+\cfrac{1}{\mathbf{5}}}},\qquad \frac{113}{355}=\cfrac{1}{3+\cfrac{1}{7+\cfrac{1}{\mathbf{16}}}}.$$

Theorem (Clément, F., Vallée 2000+)

Sorting with continued fractions: *mean path length of trie is*

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

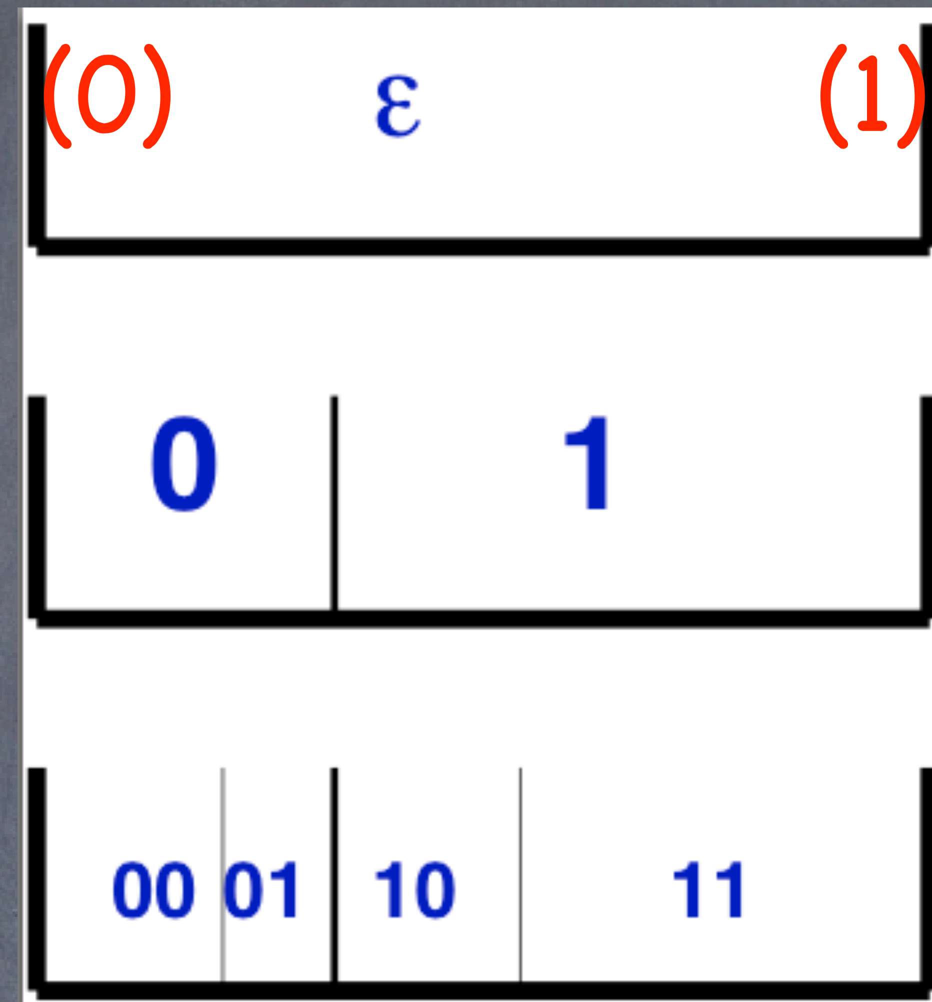
$$K_0=\frac{6\log 2}{\pi^2},\qquad K_1=18\frac{\gamma\log 2}{\pi^2}+9\frac{(\log 2)^2}{\pi^2}-72\frac{\log 2\zeta'(2)}{\pi^4}-\frac{1}{2}.$$

and $\boxed{Q(n)\approx n^{1/4}}$ *is equivalent to* **Riemann Hypothesis**.

[Vallée 1997++]

- View source model in terms of fundamental intervals:

  $w \to p_w$



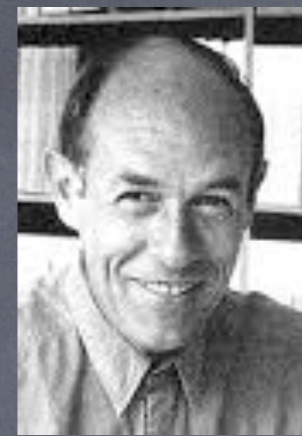
- Revisit the analysis of tries (e.g, size)

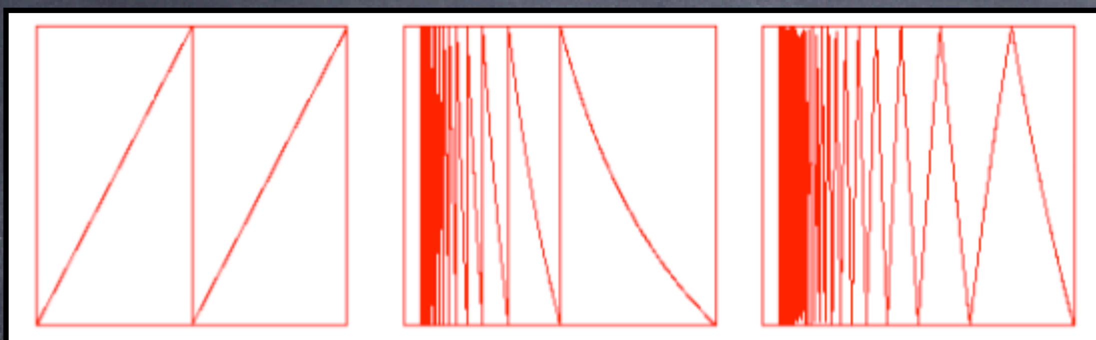
$$\begin{cases} E_{\mathcal{P}(x)}[\text{Size}] = \sum_{w \in \mathcal{A}^\star} g(p_w x) \\ g(x) = 1 - (1+x)e^{-x}. \end{cases}$$

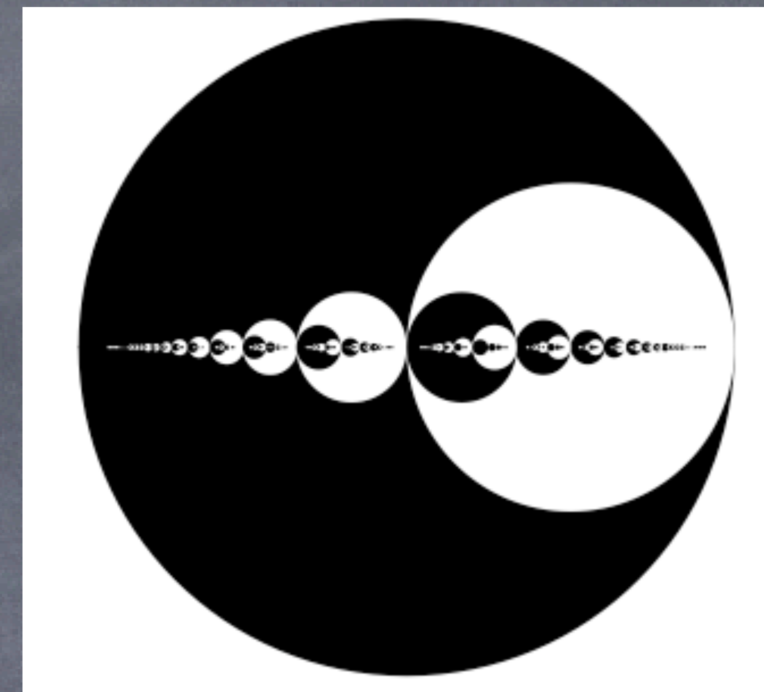
- Mellinize:

$$\begin{cases} S^\star(s) = -(s+1)\Gamma(s)\Lambda(s) \\ \Lambda(s) := \sum_w p_w^{-s} \end{cases}$$

Vallée 1997-2001, Baladi-Vallée 2005+, ...







- For expanding maps T, fundamental intervals are generated by a transfer operator.

$$\mathcal{G}_s[f](x) = \sum_{h \in T^{(-1)}} h'(x)^s f \circ h(x).$$

- For binary system (+Markov) and continued fractions, simplifications occur.

$$\begin{cases} \Lambda(s) = \dfrac{1}{1 - p^{-s} - q^{-s}} \\ \Lambda(s) = \cdots \dfrac{\zeta^{-+}(s,s)}{\zeta(2s)}. \end{cases}$$

- ...and Nörlund integrals complete the job!

- Poisson

- + Mellin = Newton

- -> Nörlund
  = fixed-n model

$$A(x) = \sum_n a_n e^{-x} \frac{x^n}{n!}$$

$$A^\star(s) = \Gamma(s) \sum_n a_n \frac{s(s+1)\cdots(s+n-1)}{n!}$$

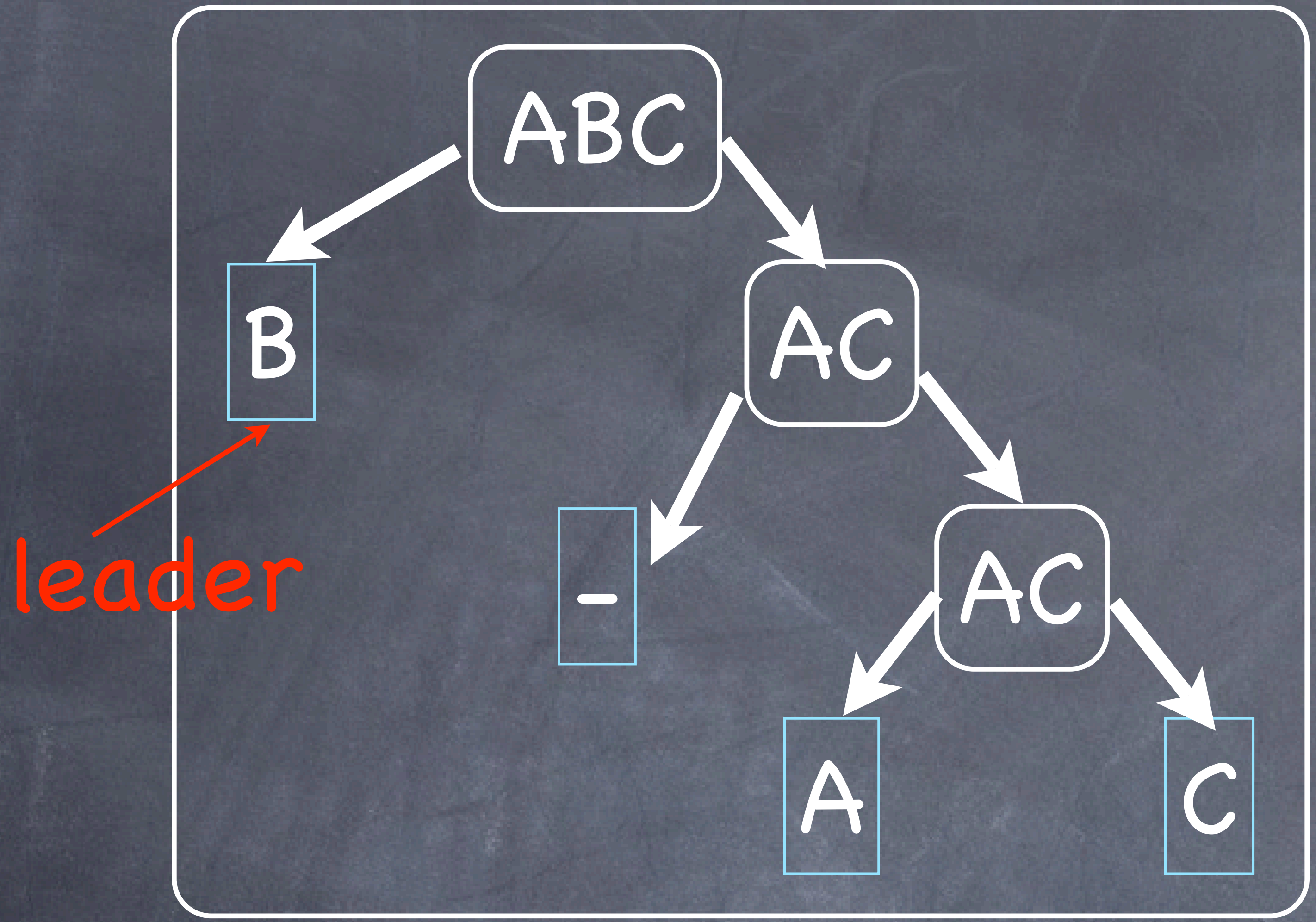
$$a_n = \frac{1}{2i\pi} \int A^\star(-s) \frac{n!\, ds}{s(s+1)\cdots(s+n-1)}$$

cf [F. Sedgewick 1995]

Q.E.D.

5. Other trie algorithms

- Leader election
- The tree communication protocol
- "Patricia" trees
- Data compression: Lempel-Ziv...
- Probabilistic counting
- Quicksort is $O(n\ (\log n)^2)$...



Leader election = leftmost boundary of a random trie (1/2,1/2).

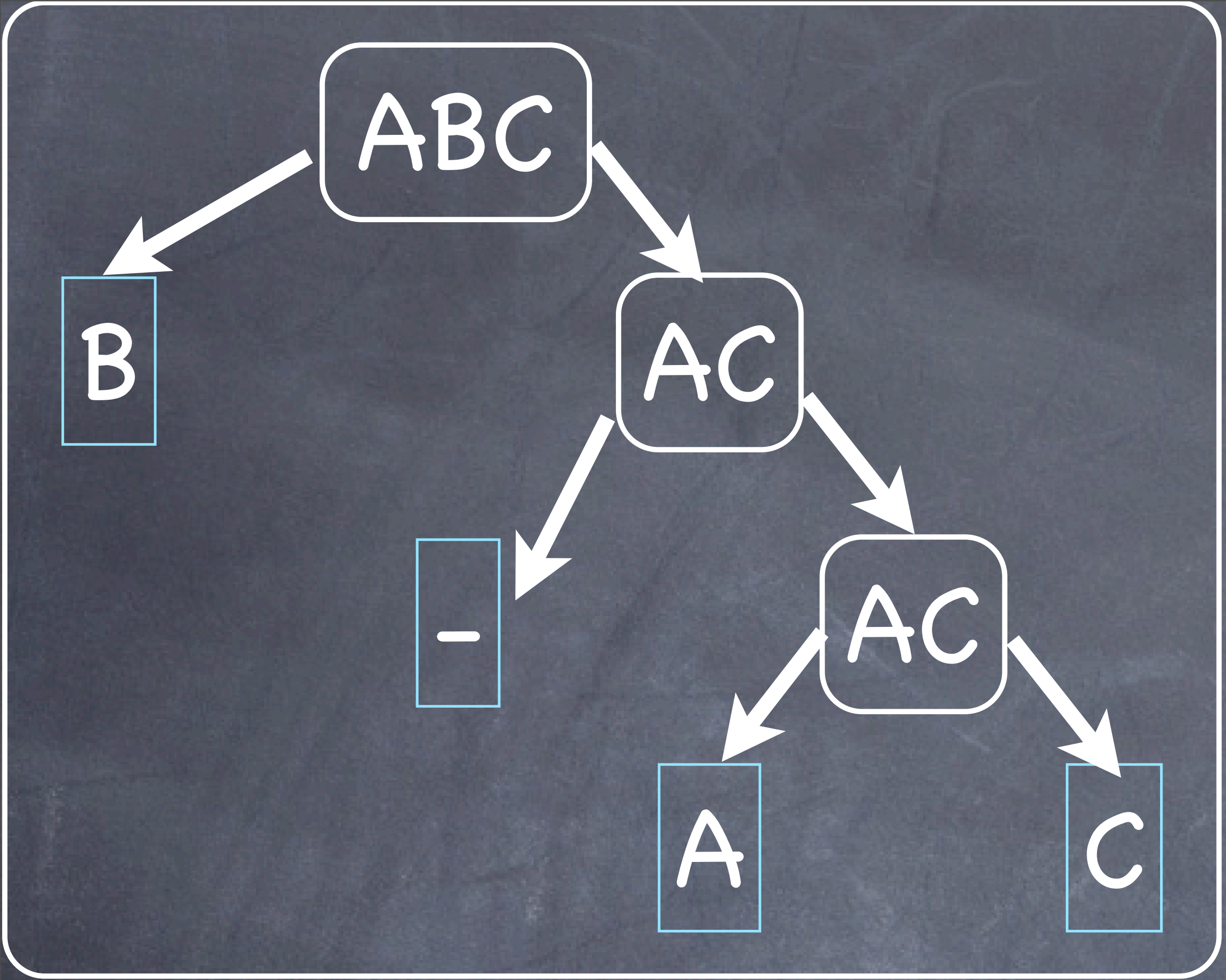
Theorem (Prodinger-Fill-Mamoud-Szpankowski)

The number R_n of rounds satisfies

$$\mathbb{P}(R_n \leq \lfloor \log_2 n \rfloor + k) \sim \frac{\beta(n)2^{-k}}{\exp\left(\beta(n)2^{-k}\right) - 1},$$

where $\beta(n) := n/2^{\lfloor \log_2 n \rfloor}$. There is a family of limit distributions based on $\{\log_2 n\}$, **not** *a single distribution.*

Proof: tree decompositions + Mellin...



tree protocol = trie with arrivals

$$\psi(z) = \tau(z) + \psi(\lambda + pz) + \psi(\lambda + qz).$$

Theorem (Fayolle, Flajolet, Hofri; Robert-Mohamed 2010)

The tree protocol, with $p = q = 1/2$ *is* **stable** *till arrival rate* $\lambda_0 \doteq 0.36017$, *root of*

$$-\frac{1}{2} = \frac{e^{-2y}}{1-2y} \sum_{j \geq 0} 2^j h\left(\frac{y}{2^j}\right), \qquad h(y) \equiv e^{-2y}\left[e^{-y}(1-y) - 1 + 2y + 2y^2\right]$$

A curiosity (cf Mellin):

$$S := \sum_{n=1}^{\infty} (-1)^n \frac{\left(\frac{99}{100}\right)^n}{1 + \left(\frac{99}{100}\right)^{2n}}$$

= !!

−0.2499999999999999999999999999999999999999999999999999
999999999999999999999999999999999999999999999999999999
999999999999999999999999999999999999999999999999999999
999999999999999999999999999999999999999999999999998211

(= $-1/2+10^{-211}$: there are 208 consecutive nines)