# VARIANCE OF BINOMIAL SPLITTING PROCESSES

**Hsien-Kuei Hwang, *Academia Sinica, Taiwan***
**(with Michael Fuchs, Vytas Zacharovas)**

**July 16, 2011**


**AofA'11**

22th International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms

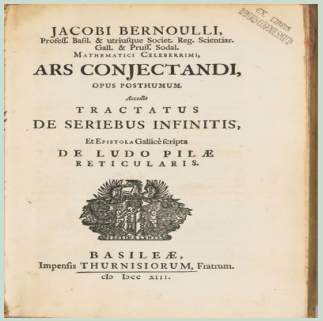
Conference dedicated to the memory of Philippe Flajolet

# BINOMIAL DISTRIBUTION

## Number of successes in $n$ iid Bernoulli trials

### James Bernoulli (1645–1705)

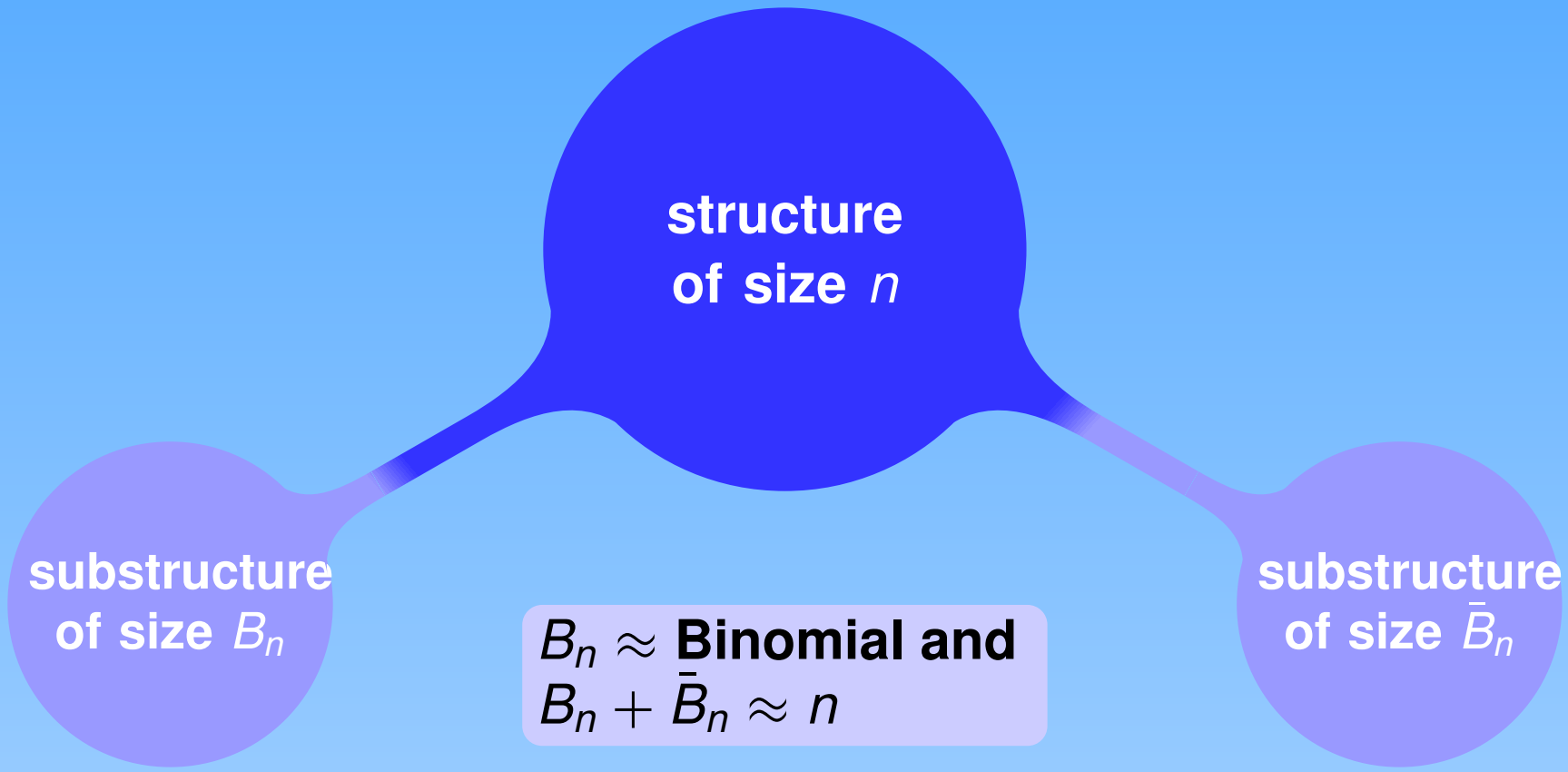
$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$


JACOBI BERNOULLI,
Profess. Basil. & utriusque Societ. Reg. Scientiar. Gall. & Prussic. Sodal.
Mathematici Celeberrimi,
ARS CONJECTANDI,
OPUS POSTHUMUM.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et Epistola Gallicè scripta
DE LUDO PILÆ RETICULARIS.
BASILEÆ,
Impensis THURNISIORUM, Fratrum.

**Ars conjectandi (1713): first appearance of the *binomial distribution* (although binomial theorem was known to Islamic and Chinese mathematicians of the late medieval period)**

***The most widely used probability distribution***

# A BINOMIAL SPLITTING PROCESS




Recursively defined random variables of the form

$$X_{n+b} \stackrel{d}{=} X_{\mathbf{Binom}(n;p)} + X^*_{n-\mathbf{Binom}(n;p)} + T_n$$

# THE MATH PROBLEM

## Q: Asymptotic approximations to mean and variance

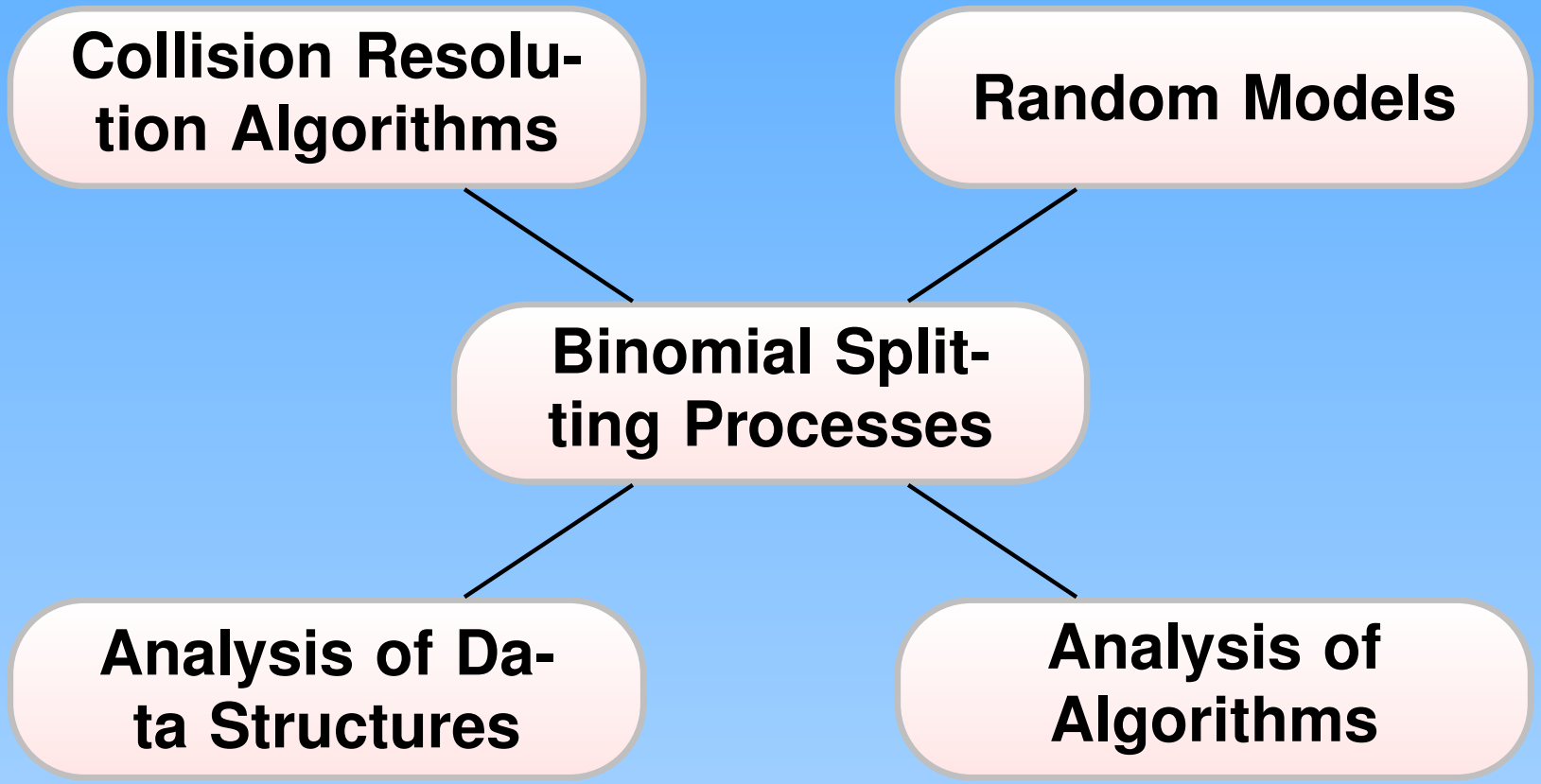
$$\mu_{n+b} = 2 \sum_{0 \le j \le n} \pi_{n,j}(\mu_j + \mu_{n-j}) + t_n \qquad \left( \pi_{n,j} := \binom{n}{j} p^j q^{n-j} \right),$$

$$\sigma^2_{n+b} = 2 \sum_{0 \le j \le n} \pi_{n,j} \left( \sigma_j^2 + \sigma^2_{n-j} \right) + \sum_{0 \le j \le n} \pi_{n,j} \left( \mu_j + \mu_{n-j} - \mu_n + t_n \right)^2 .$$

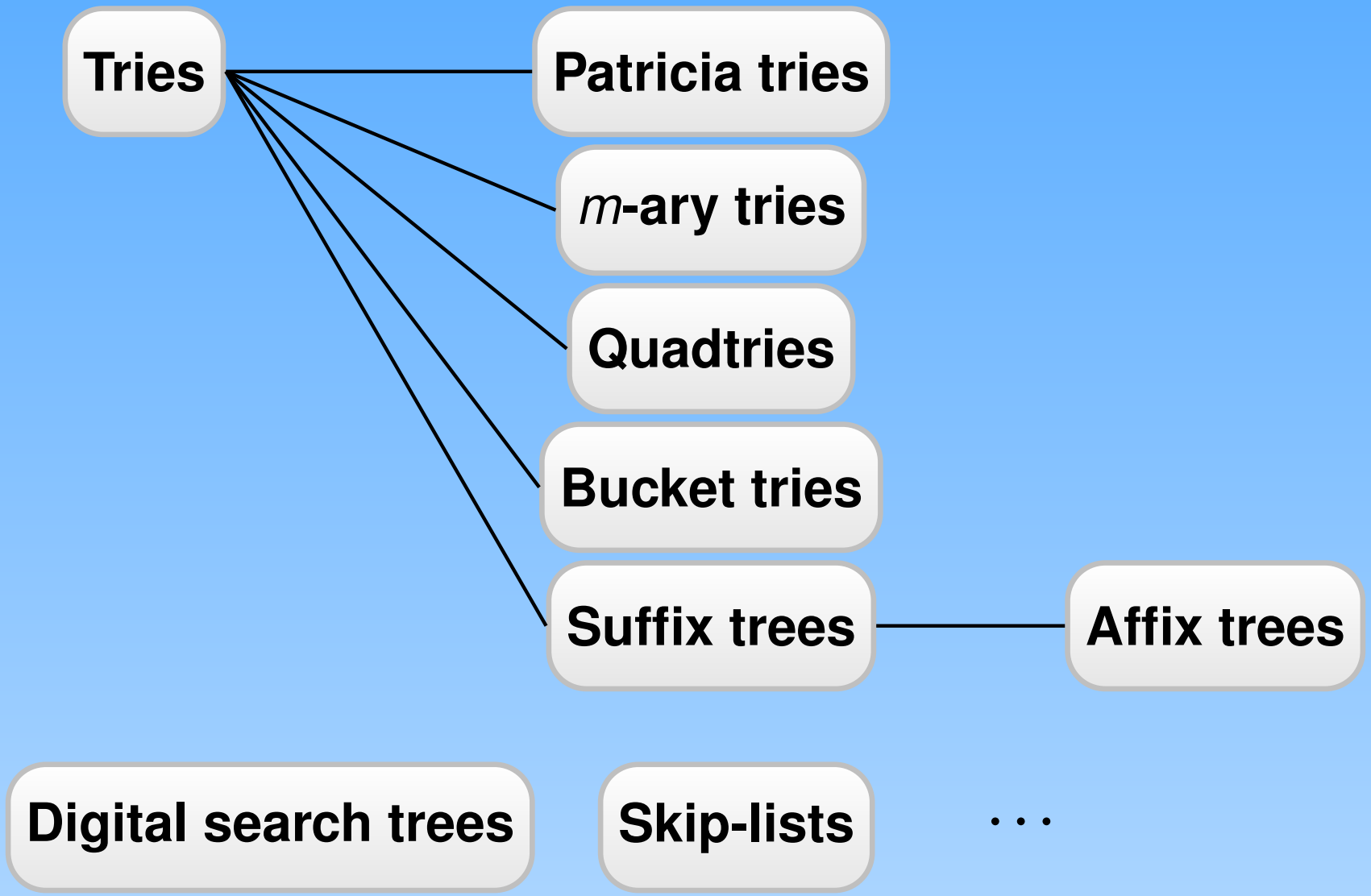
## Features

**(i) $O$-estimate easier, but $\sim$-estimate harder;**
**(ii) mean and limit law easier but variance harder;**
**(iii) periodic fluctuations everywhere;**
**(iv) a Flajolet approach possible**

# GENESIS OF BINOMIAL SPLITTING PROCESSES



# ANALYSIS OF DATA STRUCTURES



# COLLISION RESOLUTION ALGORITHMS

Collision (or contention or conflict) resolution algorithms (or *confusion resolution* ☺)

**Tree protocols in multiaccess channel**

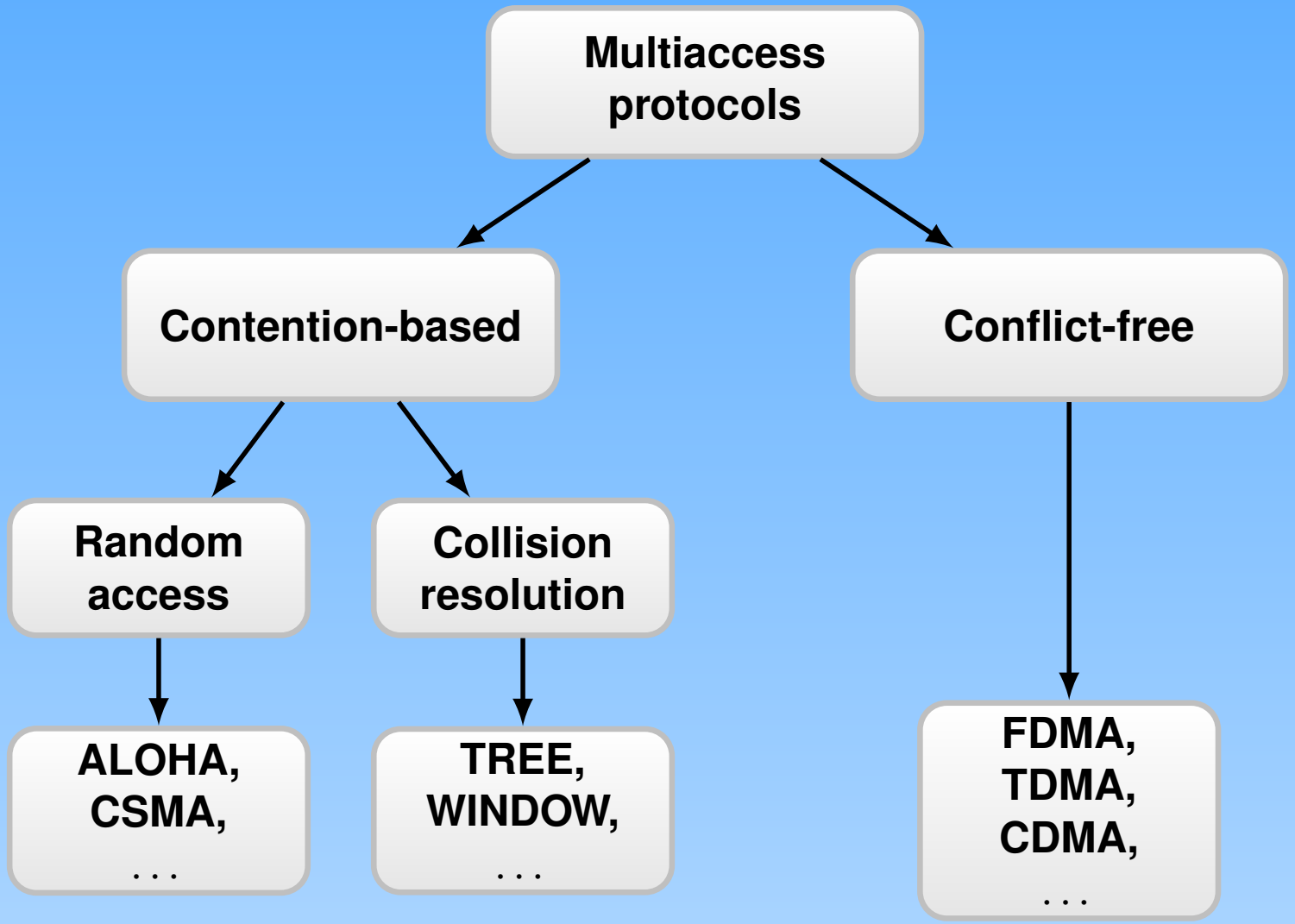
**Initialization problem in radio n/w**

**Mutual exclusion in mobile ad-hoc n/w**

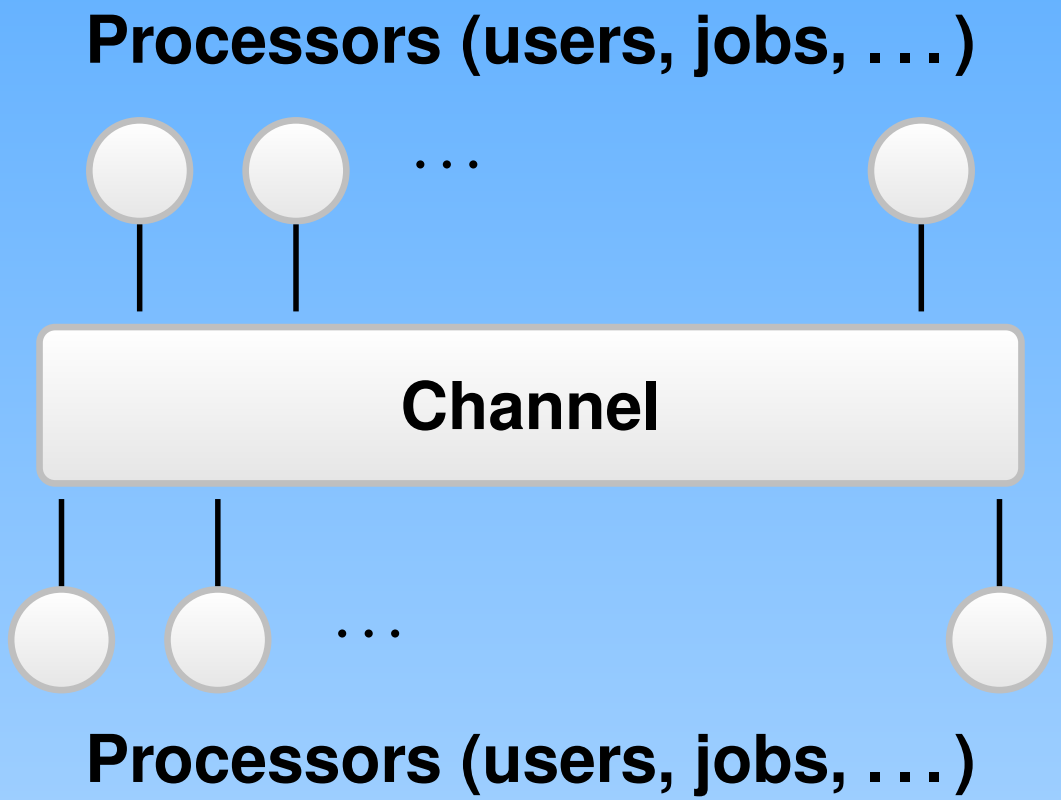
**Tree algorithms in RFID**

**Max-finding in broadcast communication channel**

# CLASSIFICATION OF MULTIACCESS PROTOCOLS



# BROADCAST COMMUNICATION CHANNEL



# ANALYSIS OF ALGORITHMS

**Extendible hashing**

**Probabilistic counting**

**Dynamic hashing**

**Max indep. set in $G_{n,p}$**

**Leader election**

**Transitive closure in random digraphs**

**Approximate counting**

**Polynomial factorization**

# RANDOM MODELS

**Generalized Eden model**

**Rumor-spreading models**

**Evolutionary trees**

**Geometric iid sequences**

**Group testing**

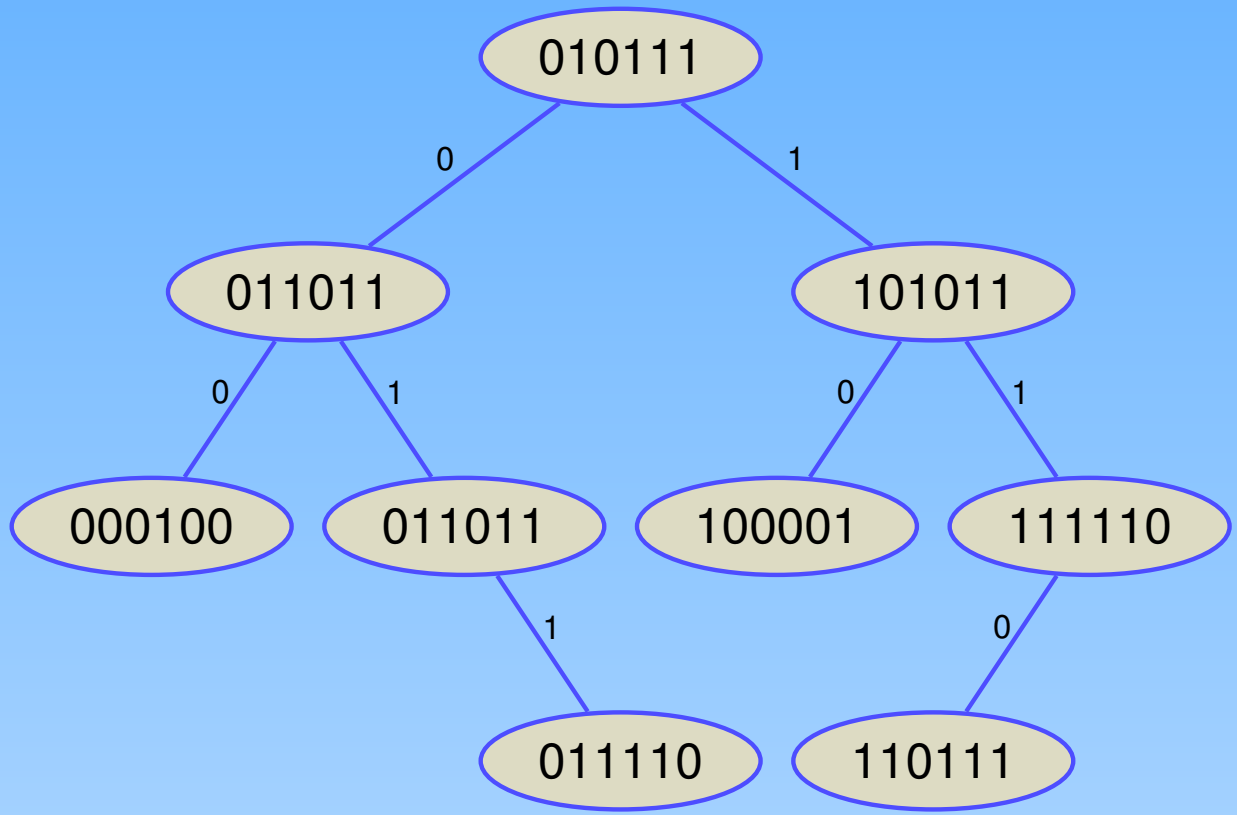
**Random strings under Bernoull**

**Directed diffusion-limited aggregation**

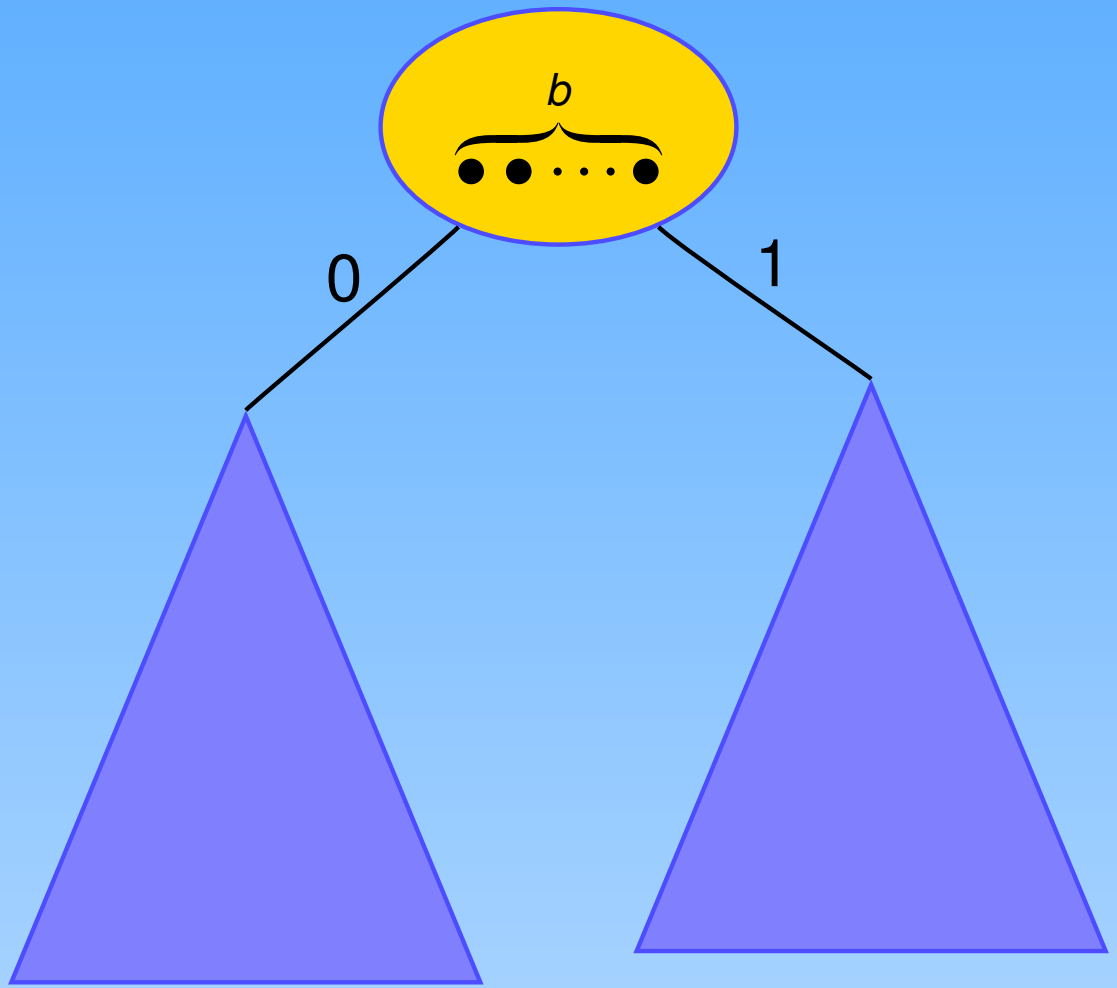
⋮

# CONSTRUCTION OF DIGITAL SEARCH TREES:
## A sequence of binary strings $\Longrightarrow$ a binary tree

010111
101011
100001
011011
111110
110111
010011
011110
000100

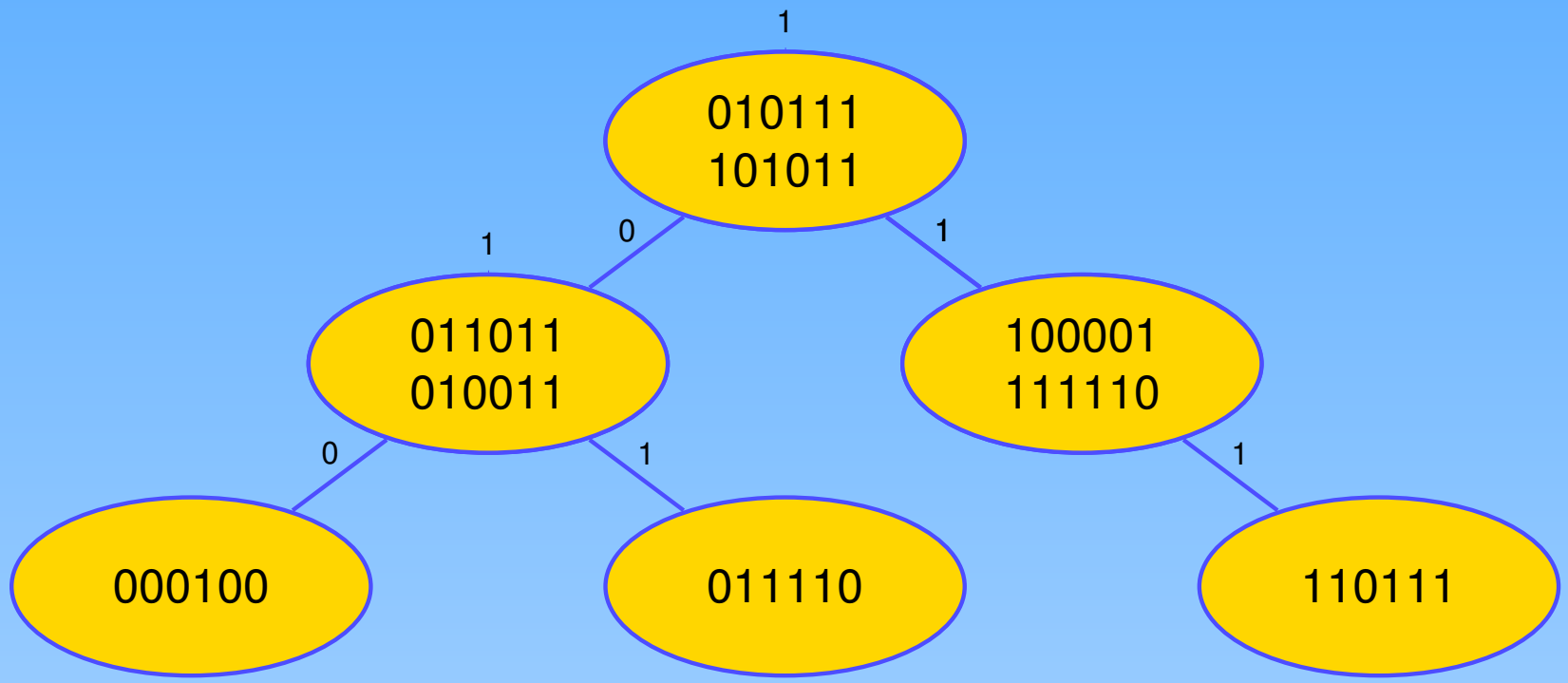


# BUCKET DIGITAL SEARCH TREES WITH $b \geq 2$



# BUCKET DIGITAL SEARCH TREES ($b$-DSTs)
## $b = 2$

010111
101011
100001
011011
111110
110111
010011
011110
000100



# DIGITAL SEARCH TREES (DSTs)

## Introduced by Coffman and Eve (1970) *CACM*

**File Structures Using Hashing Functions**

E. G. Coffman, Jr., and J. Eve
*University of Newcastle upon Tyne, England*

A general method of file structuring is proposed which uses a hashing function to define tree structure. Two types of such trees are examined, and their relation to trees studied in the past is explained. Results for the probability distributions of path lengths are derived and illustrated.

# MOTIVATIONS

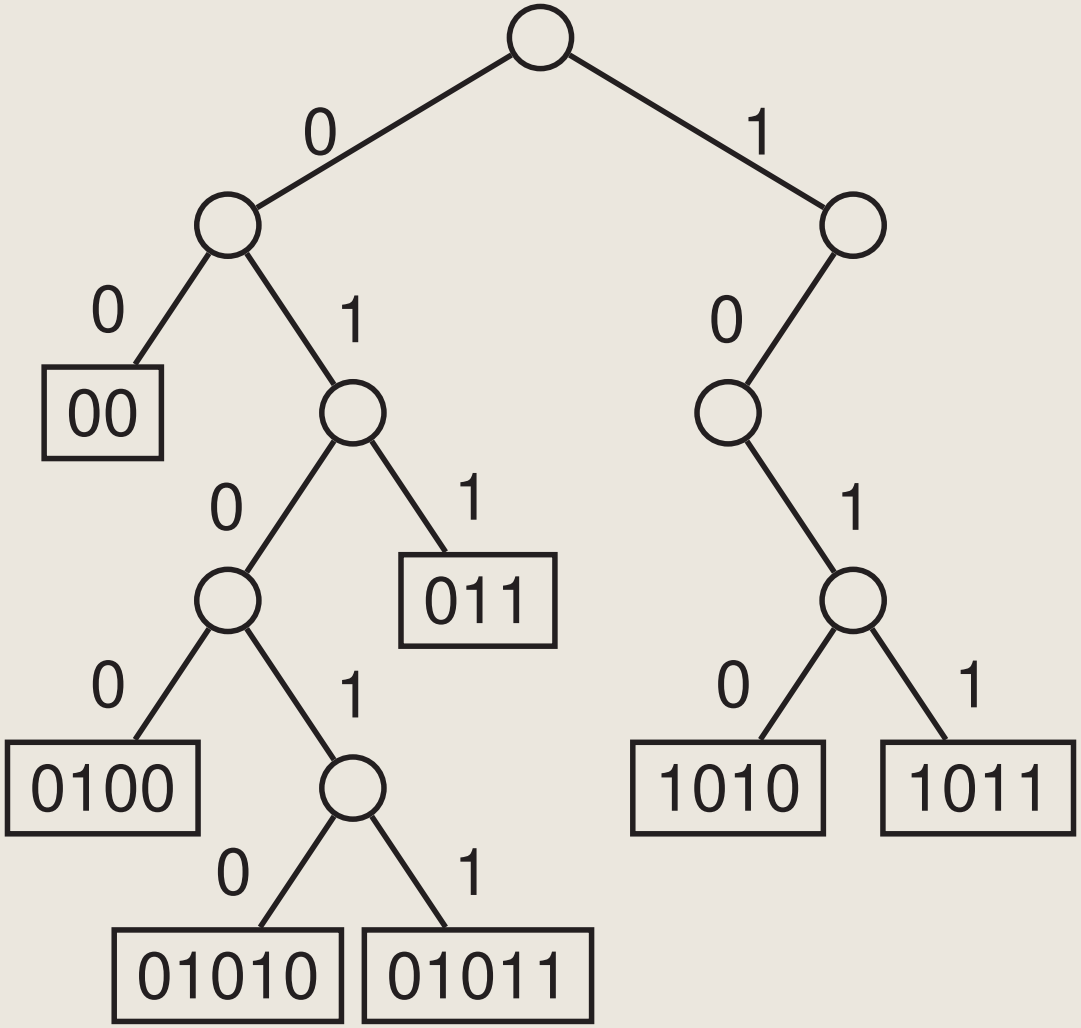
## Useful in diverse applications

**A prototype and simple data structures for indexing and retrieval purposes;**

**useful in many hashing, searching, splitting, sorting schemes,**

**and connected to the analysis of data compression schemes, sequencing algorithms, network protocols, statistical physics, etc.**

# $b = 0$: TRIE (0 $\longrightarrow$ LEFT, 1 $\longrightarrow$ RIGHT)



# FIRST USE OF TRIE IN MATH AND IN CS

## Knuth Vol. III (1997), Sec. 6.3, p. 494

The abstract concept of a trie to represent a family of strings was introduced by Axel Thue in a paper about strings that do not contain adjacent repeated substrings [*Skrifter udgivne af Videnskabs-Selskabet i Christiania*, Mathematisk-Naturvidenskabelig Klasse (1912), No. 1; reprinted in Thue's *Selected Mathematical Papers* (Oslo: Universitetsforlaget, 1977), 413–477].

Trie memory for computer searching was first recommended by René de la Briandais [*Proc. Western Joint Computer Conf.* **15** (1959), 295–298]. He pointed

## Knuth Vol. III (1997), Sec. 6.3, p. 492

The data is represented in Table 1 as a *trie structure*; this name was suggested by E. Fredkin [*CACM* **3** (1960), 490–500] because it is a part of information retrieval. A trie—pronounced "try"—is essentially an $M$-ary tree, whose nodes

# USEFULNESS OF TRIES (A DBLP SEARCH)

## Simple, easily-coded, efficient data structures

**Widely used in diverse applications; for example**

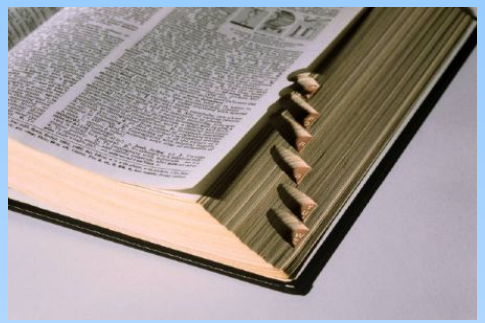
| | |
|---|---|
| indexing | **indexing sorted files, indexing electronic ink, indexing of set-valued attributes, indexing structures in inverted bases** |
| retrieval | **associative searching, orthogonal range search, approximate orthogonal range search, partial-match retrieval, pattern matching, approximate string matching, similarity matching in video databases, on-chip logic minimization** |
| structuring | **VLSI implementation of routing tables, routing chip, IP address or routing lookup, packet classification, classifier management, mining n-most interesting, document classification, hierarchical packet classification, temporal join mechanism, image processing, data compression, natural language dictionaries, peer-to-peer lookup, data mining, cache-conscious sorting, fast frequent itemset discovery, frequent itemset mining algorithms, compacting automata, dictionary-based syntactic pattern recognition, rotamer-pair energy calculations, spatial databases, policy representations for network firewalls, syntactic pattern recognition, Chinese lexical access, logical computing, network motifs, genetic algorithms, text structuring** |

**often used with hashing**

# USEFULNESS OF TRIES

## Fundamental, prototype data structures

- **have a large number of variations, extensions**
- **closely connected to several *adaptive hashing schemes* and *splitting procedures using coin-flipping*: collision resolution in multi-access (or broadcast) communication models, loser selection or leader election, etc.**
- **have direct *combinatorial interpretations* in terms of words, urn models, adaptive hashing, etc.**



# RANDOM $b$-DSTs and RANDOM TRIES

## The simplest Bernoulli model

**Input** $= \{Y_1, Y_2, \ldots, Y_n\}$ **iid, where**

$$Y_i = \{Y_{i,j}\}_{j \geq 1},$$

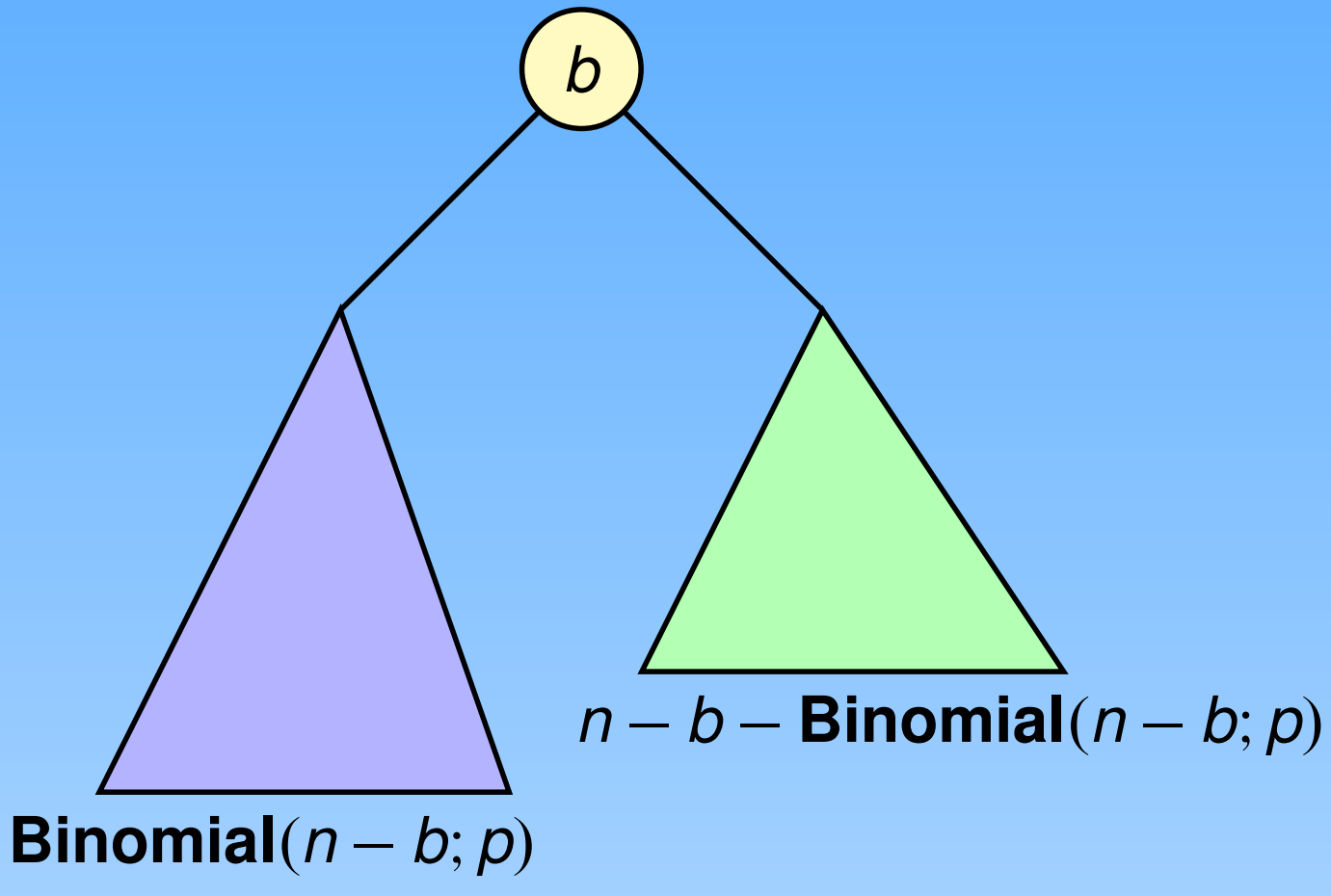
**and**

$$\begin{cases} \mathbb{P}(Y_{i,j} = 1) = p \\ \mathbb{P}(Y_{i,j} = 0) = q = 1 - p. \end{cases}$$

***Random*** *b*-***DST (trie)*****:** *b*-**DST (trie) constructed from this sequence of RVs.**

***Many other general probability models***

# THE TREE STRUCTURE



# b-DST: SHAPE CHARACTERISTICS STUDIED

## related to the cost of algorithms

- ***Depth***, *d*(root, randomly chosen node) : Konheim & Newman (1973), Knuth (1973), Pittel (1986), Louchard (1987), Kirschenhofer & Prodinger (1988), Szpankowski (1988, 1991), Devroye (1992, 1999), Louchard (1994), Jacquet et al. (2001), Janson (2006), Dennert & Grübel (2007).
- ***Height***, size of the longest path from the root : Régnier (1981), Pittel (1985), Devroye (1999), Drmota (2002), Knessl & Szpankowski (2000), Majumdar (2003).
- ***Node sorts*** (leaves, patterns), #(nodes of certain type) : Flajolet & Sedgewick (1986), Kirschenhofer & Prodinger (1988), Flajolet & Richmond (1992), Hubalek et al. (2002).
- ***Internal path length***, Sum(*d*(root,node)) : Konheim & Newman (1973), Knuth (1973), Flajolet & Sedgewick (1986), Flajolet & Richmond (1992), Kirschenhofer et al. (1994), Jacquet & Szpankowski (1995), Hubalek (2000), Hubalek et al. (2002).

# TRIES: SHAPE CHARACTERISTICS STUDIED

## Logarithmic shape parameters

- ***Depth*** : Devroye (1982, 1992, 1999), Mendelson (1982), Pittel (1986), Jacquet & Régnier (1986), Szpankowski (1987, 1988), Kirschenhofer & Prodinger (1988), Jacquet & Szpankowski (1994), Louchard (1994), Schachinger (2001), Fayolle & Ward (2005), Christohpi & Mahmoud (2008).
- ***Height*** : Mendelson (1982), Flajolet & Steyaert (1982), Flajolet (1983), Devroye (1984), Pittel (1985, 1986), Jacquet & Régnier (1986), Szpankowski (1991), Devroye (1992, 1999, 2002), Clément, Flajolet, Vallée (2001), Broutin & Devroye (2008).
- ***Horton-Strahler number and stack-size*** : Devroye & Kruszewski (1996), Nebel (2000, 2002), Bourdon, Nebel & Vallée (2001).
- ***One-sided height (or leader election or loser selection)*** : Prodinger (1995), Fill, Mahmoud, Szpankowski (1996), Janson & Szpankowski (1997), Ward & Szpankowski (2004, 2005), Louchard & Prodinger (2006).

# TRIES: SHAPE CHARACTERISTICS STUDIED

## Linear shape parameters

- ***Sum(internal nodes)*** : Jacquet & Régnier (1988), Régnier & Jacquet (1989), Kirschenhofer & Prodinger (1991), Jacquet & Szpankowski (1994), Rachev & Rüschendorf (1995), Schachinger (1995), Knuth (1998), Clément, Flajolet, Vallée (2001), Schachinger (2001, 2004), Neininger & Rüschendorf (2004), Flajolet, Roux, & Vallée (2010).
- ***Node sorts*** : Tsybakov & Mikhailov (1978), Capetanakis (1979), Mendelson (1982), Flajolet (1983), Mathys & Flajolet (1985), Kaplan & Gulko (1985), Szpankowski (1988), Kirschenhofer and Prodinger (1988, 1991), Janssen & de Jong (2000), Schachinger (2001), Nguyễn-Thế (2003), Mahmoud & Ward (2008), Wagner (2009).
- ***External path length*** : Devroye (1984), Szpankowski (1987), Kirschenhofer, Prodinger, Szpankowski (1989), Schachinger (1995, 2001, 2004), Clément, Flajolet, Vallée (2001), Nguyễn-Thế (2003), Neininger & Rüschendorf (2004),Flajolet, Roux, & Vallée (2010).

# THE MATH PROBLEM

## Recursively defined random variables of the form

$$X_{n+b} \stackrel{d}{=} X_{\mathbf{binomial}(n;p)} + X^{*}_{n-\mathbf{binomial}(n;p)} + t_n$$

## Q: Asymptotic approximations to mean and variance

$$\mu_{n+b} = \sum_{0\le j\le n} \pi_{n,j}(\mu_j + \mu_{n-j}) + t_n \qquad \left(\pi_{n,j} := \binom{n}{j} p^j q^{n-j}\right),$$

$$\sigma^2_{n+b} = \sum_{0\le j\le n} \pi_{n,j}(\sigma^2_j + \sigma^2_{n-j}) + \sum_{0\le j\le n} \pi_{n,j}\left(\mu_j + \mu_{n-j} - \mu_n + t_n\right)^2.$$

## Goal

**A systematic analytic approach (*à la Flajolet*) to precise asymptotics of mean and variance**

# A STANDARD ANALYTIC APPROACH FOR $\mu_n$

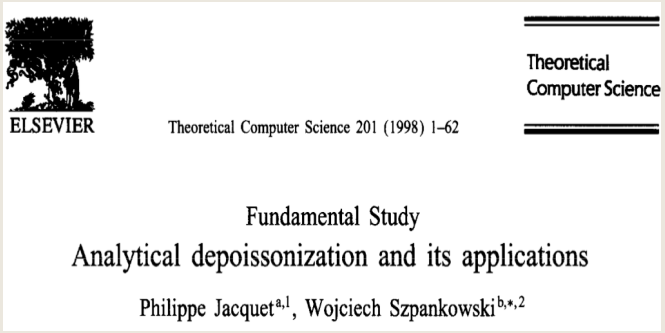
## Poisson GF $\tilde{f}_1(z) := e^{-z} \sum_n \frac{\mu_n}{n!} z^n$

**From** $\mu_n = \sum_{0 \le j \le n} \pi_{n,j}(\mu_j + \mu_{n-j}) + t_n$

$$\tilde{f}_1(z) = \tilde{f}_1(pz) + \tilde{f}_1(qz) + \tilde{g}_1(z).$$

## Poisson heuristic

**If $\mu_n$ is smooth (e.g. regularly varying), then $\mu_n \sim \tilde{f}_1(n)$.**
**Justified by Jacquet-Szpankowski's de-Poissonization tools (saddle-point method).**


ELSEVIER

Theoretical Computer Science 201 (1998) 1–62

Theoretical Computer Science

Fundamental Study

Analytical depoissonization and its applications

Philippe Jacquet[a,1], Wojciech Szpankowski[b,*,2]

## Mellin transform: $\mathscr{M}[\tilde{f}_1; s] := \int_0^\infty \tilde{f}_1(z) z^{s-1} \mathbf{d}z$

$$\tilde{f}_1(z) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} z^{-s} \frac{\mathscr{M}[\tilde{g}_1; s]}{1 - p^{-s} - q^{-s}} \mathbf{d}s.$$

**Need analytic properties of $\mathscr{M}[\tilde{g}_1; s]$: *domain of analyticity, singularities, and behavior at $\sigma \pm i\infty$***

**Poisson heuristic:** $a_n \sim \sum_k a_k \frac{n^k}{k!} e^{-n}$.

## Elementary viewpoint

**By Stirling's formula** $k! \sim \frac{(k/e)^k}{\sqrt{2\pi k}} \left(1 + \frac{1}{12k} + \cdots\right)$ **as** $k \to \infty$**, we obtain**

$$\frac{n^k}{k!} e^{-n} \sim \frac{e^{-x^2/2}}{\sqrt{2\pi n}} \left(1 + \frac{x^3 - 3x}{6\sqrt{n}} + \cdots\right) \quad (k = n + x\sqrt{n}).$$

**Since** $a_n$ **is smooth, we then have**

$$\tilde{f}(n) \sim \sum_{\substack{k=n+x\sqrt{n} \\ x=O(n^{\varepsilon})}} a_k \frac{e^{-x^2/2}}{\sqrt{2\pi n}} \sim a_n \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = a_n.$$

# POISSONIZATION AND DE-POISSONIZATION

## Widely used in diverse problems

**Useful in Borel summability, Tauberian theorems, stochastic processes, statistics, statistical physics, analysis of algorithms, . . .**

***Dated back to at least Ramanujan's Notebooks (P1)***

*Put*

$$\varphi_\infty(x) = e^{-x} \sum_{k=0}^{\infty}{}' \frac{x^k \varphi(k)}{k!},$$

*where the prime on the summation sign indicates that the (finitely many) terms for which $\varphi(k)$ may be undefined are not included in the sum. Then for any fixed positive integer $M$,*

$$\varphi_\infty(x) = \varphi(x) + \sum_{k=2}^{M} \sum_{n=k}^{2k-2} b_{kn} x^{n-k+1} \frac{\varphi^{(n)}(x)}{n!} + O(G(x)x^{-M}), \qquad (10.3)$$

*as $x$ tends to $\infty$, where the numbers $b_{kn}$ are defined by (10.1).*

# POISSONIZATION AND DE-POISSONIZATION

**Poisson heuristic:** $a_n \sim \sum_k a_k \frac{n^k}{k!} e^{-n}$.

## Analytic viewpoint: $\tilde{f}(z) := e^{-z} \sum_n a_n z^n / n!$ is entire

$$a_n = \frac{n!}{2\pi i} \oint_{|z|=n} \underbrace{z^{-n-1} e^z}_{\text{large}} \underbrace{\tilde{f}(z)}_{\text{small}} \, dz$$

$$\sim \tilde{f}(n) \frac{n!}{2\pi i} \oint_{|z|=n} z^{-n-1} e^z \, dz$$

$$= \tilde{f}(n)$$

**Cauchy** $\int$

**Saddle-point**

$\frac{d}{dz}(z^{-n} e^z) = 0$

$\Longrightarrow z = n$

# POISSONIZATION AND DE-POISSONIZATION

**Poisson heuristic:** $a_n \sim \sum_k a_k \frac{n^k}{k!} e^{-n}$.

## A more precise expansion

$$a_n = \frac{n!}{2\pi i} \oint_{|z|=n} z^{-n-1} e^z \tilde{f}(z) dz$$

$$= \sum_{j \geq 0} \frac{\tilde{f}^{(j)}(n)}{j!} \underbrace{\frac{n!}{2\pi i} \oint_{|z|=n} z^{-n-1} e^z (z-n)^j dz}$$

$$= \sum_{j \geq 0} \frac{\tilde{f}^{(j)}(n)}{j!} \underbrace{\tau_j(n)}_{\substack{\text{Charlier} \\ \text{polynomials}}} = \tilde{f}(n) - \frac{n}{2}\tilde{f}''(n) + \cdots .$$

## First few $\tau_j$

$$\begin{aligned}
\tau_0(n) &= 1, \\
\tau_1(n) &= 0, \\
\tau_2(n) &= -n, \\
\tau_3(n) &= 2n, \\
\tau_4(n) &= 3n(n-2), \\
\tau_5(n) &= -4n(5n-6), \\
&\cdots \\
\deg(\tau_j(n)) &= \lfloor j/2 \rfloor
\end{aligned}$$

$$\tau_j(n) = \sum_{0 \leq \ell \leq j} \binom{j}{\ell} (-1)^{j-\ell} \frac{n(n-1)\cdots(n-\ell+1)}{n^\ell}$$

# CHARLIER EXPANSION

## An identity: If $\tilde{f}$ is *entire*, then

$$a_n = \sum_{j \geq 0} \frac{\tilde{f}^{(j)}(n)}{j!} \tau_j(n).$$

**Nothing to do with *growth order* or *variation* or *smoothness* of $\tilde{f}(z)$ at infinity.**

## An example: $\tilde{f}(z) = e^z$

$$2^n = e^n \sum_{j \geq 0} \frac{\tau_j(n)}{j!}.$$

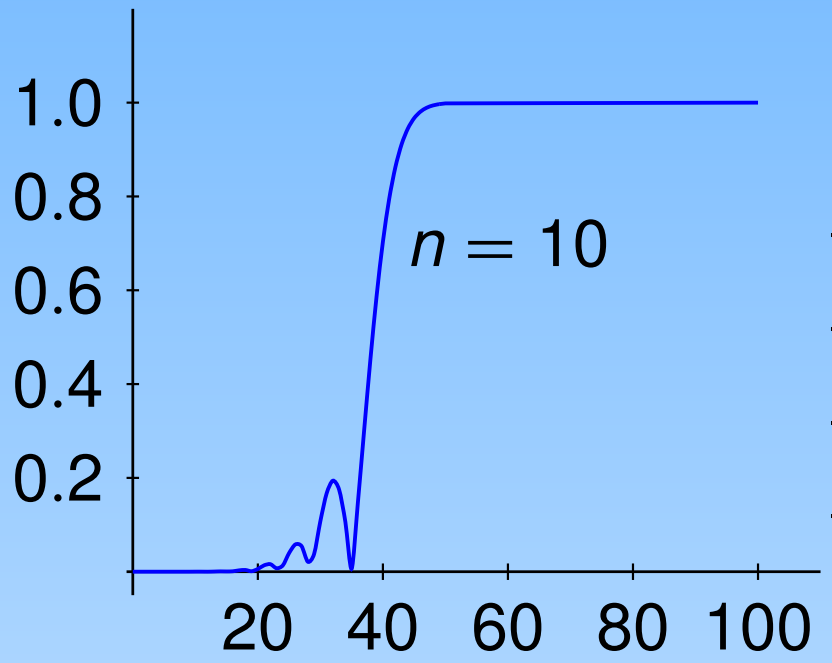
**But $2^n \not\sim e^n$.**

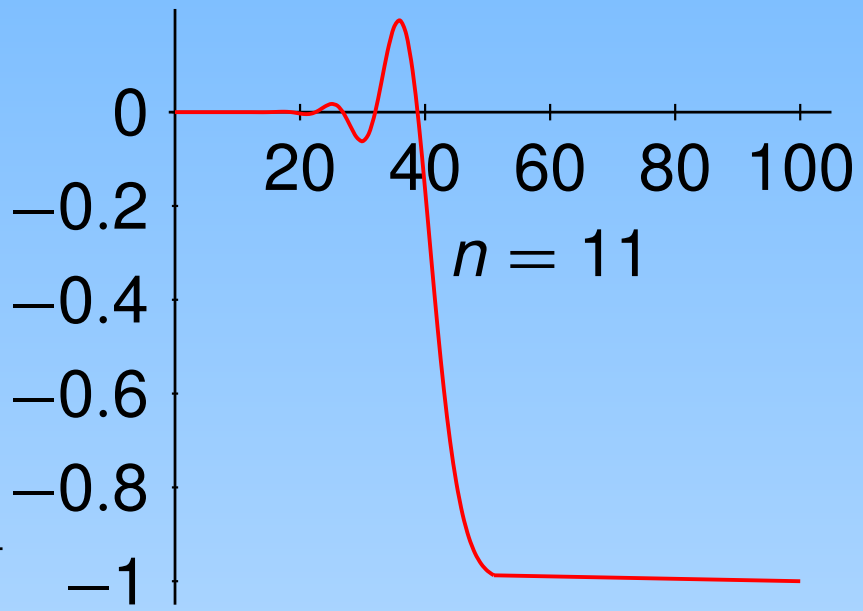
# CHARLIER EXPANSION

## Another example: $\tilde{f}(z) = e^{-2z}$

$$(-1)^n = e^{-2n} \sum_{j \geq 0} \frac{(-2)^j}{j!} \tau_j(n).$$

**But** $(-1)^n \not\sim e^{-2n}$**.**

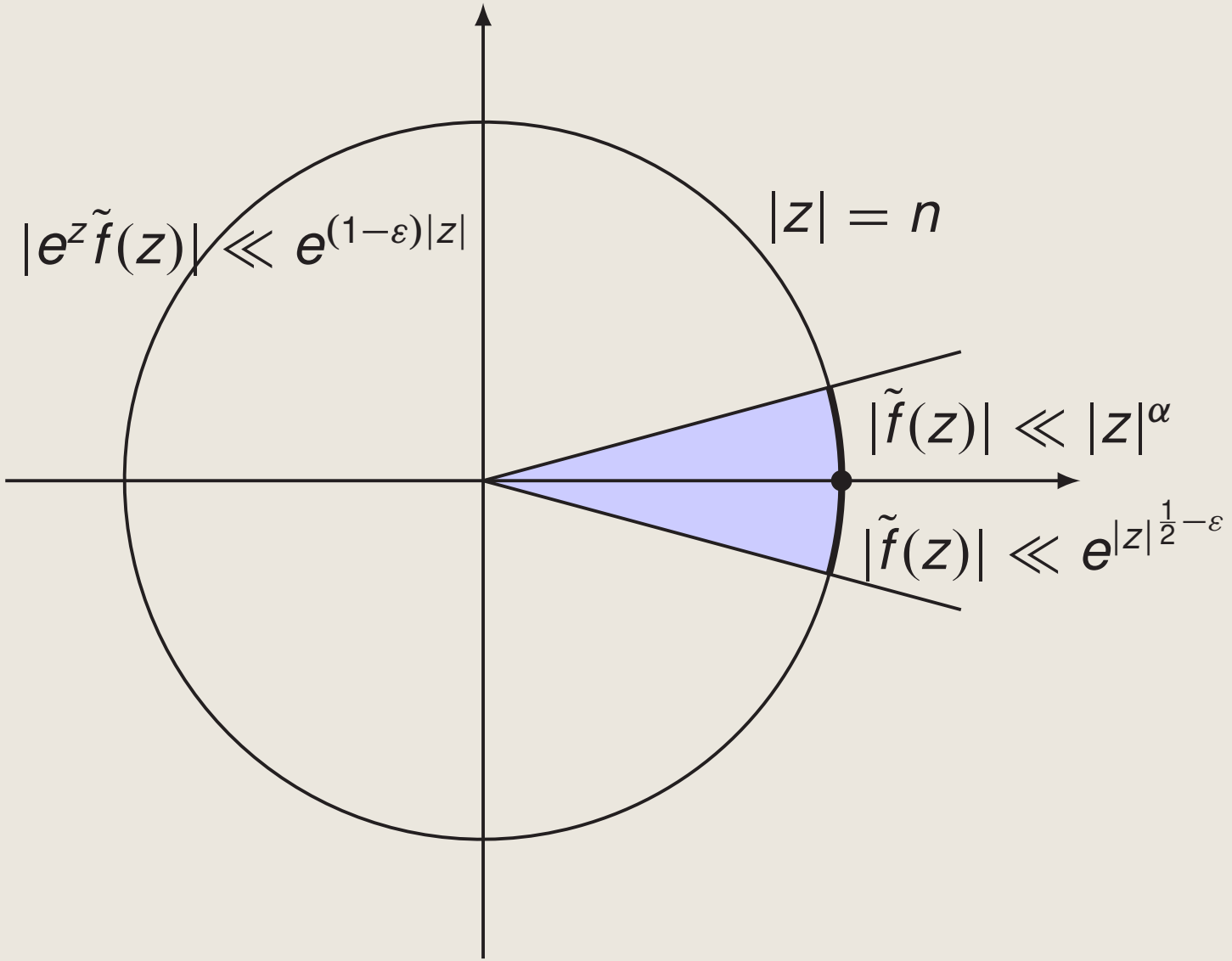



***Major difficulty: prove the asymptotic nature***

# ANALYTIC DE-POISSONIZATION

## Jacquet-Szpankowski (1998): $a_n \sim \tilde{f}(n)$



# JACQUET-SZPANKOWSKI ADMISSIBLE *fs*

## Similar to Hayman (1956) admissibility

$\boxed{\tilde{f} \in \mathscr{JS}}$ **if** $\tilde{f}$ **is entire and there exist** $\varepsilon, \delta > 0$

$$\begin{cases} \tilde{f}(z) = O(|z|^{\alpha}) & \textbf{for } |\arg(z)| \leq \delta < \pi/2; \\ e^{z}\tilde{f}(z) = O(e^{(1-\varepsilon)|z|}) & \textbf{for } \delta < |\arg(z)| \leq \pi, \end{cases}$$

**uniformly in** $z$**. Also use** $\boxed{\tilde{f} \in \mathscr{JS}_{\alpha}}$

## Closure properties: $m \geq 0$ and $\alpha \in (0, 1)$

(i) $z^{m}, e^{-\alpha z} \in \mathscr{JS}$.

(ii) $\tilde{f} \in \mathscr{JS} \Longrightarrow \tilde{f}(\alpha z), \textbf{polynom}(z)\tilde{f} \in \mathscr{JS}$.

(iii) $\tilde{f}, \tilde{g} \in \mathscr{JS} \Longrightarrow \tilde{f} + \tilde{g} \in \mathscr{JS}$.

(iv) $\tilde{f}, \tilde{g} \in \mathscr{JS} \Longrightarrow \tilde{h} \in \mathscr{JS}$, **where** $\tilde{h}(z) := \tilde{f}(\alpha z)\tilde{g}((1-\alpha)z)$.

# CLOSURE PROPERTIES OF JS-ADMISSIBLE *fs*

$$\tilde{f}(z) = \tilde{f}(pz) + \tilde{f}(qz) + \tilde{g}(z)$$

$$\tilde{g} \in \mathscr{JS} \quad \textbf{iff} \quad \tilde{f} \in \mathscr{JS}$$

## Applied to the mean $\tilde{f}_1(z) = \tilde{f}_1(pz) + \tilde{f}_1(qz) + \tilde{g}_1(z)$

$$\tilde{g}_1 \in \mathscr{JS} \implies \mu_n \sim \tilde{f}_1(n)$$

**Then**

$$\tilde{f}_1(n) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} n^{-s} \frac{\mathscr{M}[\tilde{g}_1; s]}{1 - p^{-s} - q^{-s}} \, \mathbf{d}s$$

**(only real parameter for $\tilde{f}_1$ is needed!!)**

# POISSONIZED VARIANCE

**For a random variable $X_n$**

**Let**

$$\tilde{f}_m(z) := e^{-z} \sum_n \frac{\mathbb{E}(X_n^m)}{n!} z^n,$$

**denote the Poisson GF of the $m$-th moment.**

- **Mean:** $\mathbb{E}(X_n) \sim \tilde{f}_1(n)$
- **Second moment:** $\mathbb{E}(X_n^2) \sim \tilde{f}_2(n)$
- **Since** $\mathbb{V}(X_n) = \mathbb{E}(X_n^2) - (\mathbb{E}(X_n))^2$**, does** $\tilde{D}(n) := \tilde{f}_2(n) - \tilde{f}_1(n)^2$ **yield a good approximant to the variance?**

# POISSONIZED VARIANCE

## Now with $\tilde{f}_2(z) = \tilde{D}(z) + \tilde{f}_1(z)^2$

$$\begin{aligned}\sigma_n^2 &= \mathbb{E}(X_n^2) - \mu_n^2 \\ &= \sum_{j\geq 0} \frac{\tilde{f}_2^{(j)}(n)}{j!}\tau_j(n) - \left(\sum_{j\geq 0} \frac{\tilde{f}_1^{(j)}(n)}{j!}\tau_j(n)\right)^2 \\ &= \tilde{D}(n) - n\tilde{f}_1'(n)^2 - n\tilde{f}_1(n)\tilde{f}_1''(n) + \textbf{s.o.t.}\end{aligned}$$

## $f_1 = z^\alpha$

$$n\tilde{f}_1'(n)^2 \asymp n^{2\alpha-1}, \qquad n\tilde{f}_1(n)\tilde{f}_1''(n) \asymp n^{2\alpha-1}.$$

**If we expect $\tilde{D} \asymp n^\alpha$, then $\alpha < 1$.**

## $\tilde{f}_1(z) = z \log z$

$$n\tilde{f}_1'(n)^2 \asymp n(\log n)^2, \qquad n\tilde{f}_1(n)\tilde{f}_1''(n) \asymp n\log n.$$

***In many such cases,* $\sigma_n^2 \asymp n\log n$ *or* $\sigma_n^2 \asymp n$.**

# OUR APPROACH TO $\mathbb{V}(T_n)$

## The crucial step

**Consider**

$$\tilde{V}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2 - z\tilde{f}_1'(z)^2.$$

**Then (with $\tilde{f}_2(n) = \tilde{V}(n) + \tilde{f}_1(n)^2 + n\tilde{f}_1'(n)^2$)**

$$\begin{aligned}\sigma_n^2 &= \sum_{j\geq 0} \frac{\tilde{f}_2^{(j)}(n)}{j!}\tau_j(n) - \left(\sum_{j\geq 0} \frac{\tilde{f}_1^{(j)}(n)}{j!}\tau_j(n)\right)^2 \\ &= \tilde{V}(n) - \underbrace{\frac{n}{2}\tilde{V}''(n) - \frac{n^2}{2}\tilde{f}_1''(n)^2}_{=O(1)} + o(1).\end{aligned}$$

**If $\tilde{V}(z) \asymp z(\log z)^k$, then $\sigma_n^2 \sim \tilde{V}(n)$.**

**So $z\tilde{f}_1'(z)^2$ is the right *correction term* between Poissonized variance and the true variance.**

# ASYMPTOTICS OF $\tilde{V}(n)$

$$X_n \stackrel{d}{=} X_{\text{binomial}(n;p)} + X^*_{n-\text{binomial}(n;p)} + T_n$$

$$\begin{cases} \tilde{f}_1(z) = \tilde{f}_1(pz) + \tilde{f}_1(qz) + \tilde{g}_1(z). \\ \\ \tilde{f}_2(z) = \tilde{f}_2(pz) + \tilde{f}_2(qz) + \tilde{g}_2(z), \end{cases}$$

**where**

$$\tilde{g}_2(z) := 2\tilde{f}_1(pz)\tilde{f}_1(qz) + e^{-z}\sum_n \left(\mathbb{E}(T_n^2) + 2\mathbb{E}(T_n)\left(\mu_n - \mathbb{E}(T_n)\right)\right)\frac{z^n}{n!}$$

## Same type of equation for $\tilde{V}$

$$\tilde{V}(z) := \tilde{f}_2(z) - \tilde{f}_1(z)^2 - z\tilde{f}_1'(z)^2.$$

**Then**

$$\tilde{V}(z) = \tilde{V}(pz) + \tilde{V}(qz) + \tilde{w}(z).$$

# A FRAMEWORK FOR VARIANCE

## $\mathscr{JS}$ closed under Hadamard product

**If** $\tilde{g}_1, \tilde{g}_2(z) \in \mathscr{JS}_\alpha$**,** $\alpha < 1$**,**

- $$\mathbb{V}(X_n) \sim \begin{cases} n\textbf{Periodic}, & \textbf{if } \frac{\log p}{\log q} \in \mathbb{Q}; \\ c_1 n, & \textbf{if } \frac{\log p}{\log q} \notin \mathbb{Q} \end{cases}$$
- **If** $\tilde{g}_1(z) \in \mathscr{JS}$ **and** $\tilde{g}_1(z) = z + O(|z|^\alpha)$**, then**

$$\mathbb{V}(X_n) \sim \frac{pq \log^2(p/q)}{H^3} n \log n + \begin{cases} n\textbf{Periodic}, & \textbf{if } \frac{\log p}{\log q} \in \mathbb{Q}; \\ c_2 n, & \textbf{if } \frac{\log p}{\log q} \notin \mathbb{Q}, \end{cases}$$

**All constants and periodic functions are (easily) explicitly computable.**

# SIZE OF RANDOM TRIES

## $T_n = 1$

$$\mathbb{V}(X_n) \sim \begin{cases} \dfrac{n}{H}\displaystyle\sum_{k\in\mathbb{Z}} G(-1+\chi_k)n^{-\chi_k}, & \textbf{if } \dfrac{\log p}{\log q} \in \mathbb{Q}; \\ \dfrac{G(-1)}{H}n, & \textbf{if } \dfrac{\log p}{\log q} \notin \mathbb{Q}, \end{cases}$$

**where** $\frac{\log p}{\log q} = \frac{\rho}{\ell}$, $(\ell,\rho)=1$, $\chi_k := 2\rho k\pi i/\log p$,

$$G(-1+\chi_k) \qquad (\kappa_l \textbf{ are all zeros of } 1-p^{1+\tau}-q^{1+\tau}=0$$

$$= \chi_k\Gamma(-1+\chi_k)\left(1-\frac{\chi_k^2-2\chi_k+4}{2^{\chi_k+2}}\right) \qquad \textbf{with } \Re(\kappa_l)<0)$$

$$+2\sum_{l\ge 1}\frac{(-1)^l l\Gamma(\chi_k+l)}{(l+1)!(1-p^{l+1}-q^{l+1})}(p^{l+1}+q^{l+1})(l(\chi_k+l)-1)$$

$$+pq\left(\sum_{l\ge 1}\frac{(-1)^{l-1}\Gamma(l+\chi_k+1)\left(p^l-q^l\right)\left(p^{-l}-q^{-l}\right)}{(l-1)!(1-p^{1+l}-q^{1+l})(1-p^{1-l}-q^{1-l})}\right.$$

$$\left.-\sum_{\kappa_l}\frac{\Gamma(\kappa_l+1)\Gamma(-\kappa_l+\chi_k+1)\left(p^{-\kappa_l}-q^{-\kappa_l}\right)\left(p^{\kappa_l}-q^{\kappa_l}\right)}{(1-p^{1-\kappa_l}-q^{1-\kappa_l})(p^{1+\kappa_l}\log p+q^{1+\kappa_l}\log q)}\right).$$

# SIZE OF RANDOM TRIES

## $p = q = \frac{1}{2}$

$$G(-1+\chi_k) = \frac{\chi_k^2}{4}(2-\chi_k)\Gamma(-1+\chi_k) + 2\sum_{l\geq 1}\frac{(-1)^l l\Gamma(\chi_k+l)}{(l+1)!(2^l-1)}(l(\chi_k+l)-1)$$

## Comparing with Régnier & Jacquet (1989)

$$\frac{1}{8} - \frac{1}{2\log 2} - \frac{2\pi^2}{\log^2 2}\sum_{l\geq 1}\frac{l}{\sinh(2l\pi^2/\log 2)} = \sum_{l\geq 1}\frac{(-1)^l l}{2^l-1}.$$

# EXTERNAL PATH LENGTH OF RANDOM TRIES

## $p = q = \frac{1}{2}$

$$\mathbb{V}(X_n) \sim \frac{n}{\log 2} \sum_k G(-1 + \chi_k) n^{-\chi_k}$$

## $\frac{\log p}{\log q} \in \mathbb{Q}$

$$\mathbb{V}(X_n) \sim \frac{pq \log^2(p/q) n \log n}{H^3} + \left( \frac{d}{H} + \frac{p \log^2 p + q \log^2 q}{2h^2} \right) n + \frac{n}{h} \sum_{k \neq 0} G(-1 + \chi_k) n^{-\chi_k}$$

## $\frac{\log p}{\log q} \notin \mathbb{Q}$

$$\mathbb{V}(X_n) \sim \frac{pq \log^2(p/q) n \log n}{H^3} + \left( \frac{d}{H} + \frac{p \log^2 p + q \log^2 q}{2H^2} \right) n$$

# SPECIAL CASE $p = q = \frac{1}{2}$

$$G(s) = s(s+1)\Gamma(s)\left(2^{s+1}(s+3) - \frac{s^3+5s^2+22s+24}{16}\right)$$
$$- 2^{s+2}\sum_{l\geq 1}\frac{(-1)^l\Gamma(s+l+2)}{(l-1)!(2^l-1)}(l(s+l+2)-l-1).$$

**Consistent with Kirschenhofer, Prodinger & Szpankowski (1989)**

**Additional advantage: more transparent Fourier series**

# INTERNAL PATH LENGTH OF RANDOM TRIES

$$P_n \stackrel{d}{=} P_{\textbf{binomial}(n;p)} + P^*_{n-\textbf{binomial}(n;p)} + N_{\textbf{binomial}(n;p)} + N^*_{n-\textbf{binomial}(n;p)},$$

$$N_n \stackrel{d}{=} 1 + N_{\textbf{binomial}(n;p)} + N^*_{n-\textbf{binomial}(n;p)}.$$

## Nguyễn-Thế (2003): Thèse, Ecole polytecnique

**Asymptotically normal with mean $\sim cn \log n$ but variance of order $n(\log n)^2$ in all cases. *Proof incomplete.***

**More precise results can be proved. For example, when $p = q = 0.5$,**

$$\mathbb{V}(I_n) = P_1(\log_2 n) n (\log n)^2 + P_2(\log_2 n) n \log n + P_3(\log_2 n) n + O(1).$$

# GENERALITY OF THE APPROACH

## more examples on random tries

**$w$-parameter (Drmota-Gittenberger-Panholzer-Prodinger-Ward, 2009), pattern occurrences (Nguyễn-Thế, 2003; Wagner, 2009),**

## Also applicable to other splitting processes

**Digital search trees, Patricia tries, collision resolution algorithms, leader elections, etc.**

## Path length of random digital search trees

$$T_{n+1} \stackrel{d}{=} n + T_{\mathbf{binomial}(n;1/2)} + T^*_{n-\mathbf{binomial}(n;1/2)}$$

# THE INVISIBLE PERIODICITY

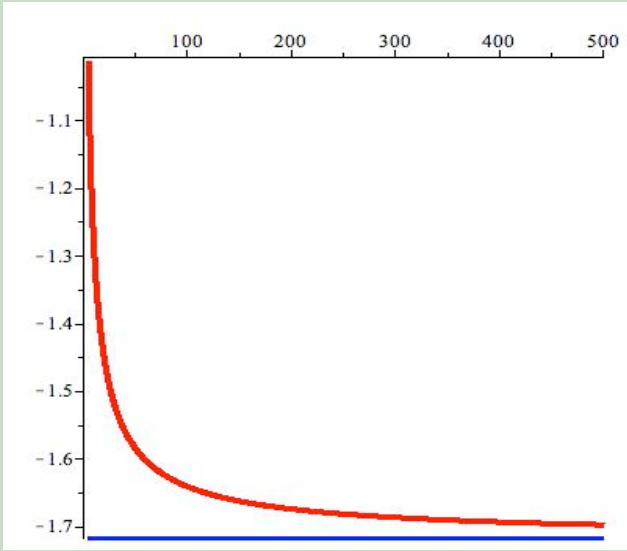
## Expected total path length of random DSTs

**Konheim and Newman (1973), Knuth (1973), Flajolet & Sedgewick (1986)**

$$\mathbb{E}(T_n) = n\log_2 n + n\underbrace{P_1(\log_2 n)}_{\textit{periodic}} + \textbf{smaller order terms}.$$

**Here**

$$P_1(x) = \underbrace{\frac{\gamma-1}{\log 2} + \frac{1}{2} - \sum_{k\geq 1}\frac{1}{2^k-1}}_{\approx -1.7155\ldots} + \underbrace{\frac{1}{\log 2}\sum_{j\neq 0}\Gamma\left(-1-\frac{2j\pi i}{\log 2}\right)e^{2j\pi ix}}_{|\cdot|\leq 10^{-6}}.$$



# VARIANCE OF TPL $T_n$ ($b = 1$)

THEOREM 4. *The variance of the internal path length of digital search trees built from $N$ records becomes*

$$\operatorname{Var} L_N = N \cdot \{C + \delta(\log_2 N)\} + \mathcal{O}(\log^2 N / N)$$

*where $C$ is a constant that can be expressed as*

$$C = -\frac{28}{3L} - \frac{39}{4} - 2\beta_1 + \frac{2\alpha}{L} + \frac{\pi^2}{2L^2} + \frac{2}{L^2} - \frac{2}{L}\sum_{k\geq 3}\frac{(-1)^{k+1}(k-5)}{(k+1)k(k-1)(2^k-1)}$$

(2.13)
$$+\frac{2}{L}\sum_{r\geq 1} b_{r+1}\left(\frac{L(1-2^{-r+1})/2-1}{1-2^{-r}} - \sum_{k\geq 2}\frac{(-1)^{k+1}}{k(k-1)(2^{r+k}-1)}\right)$$

$$+\frac{2}{L}\hat{w}'(3) - 2[\delta_1\delta_2]_0 - [\delta_1^2]_0$$

*with $L = \log 2$, $\alpha = \sum_{n\geq 1} 1/(2^n-1)$, $\beta_1 = \sum_{n\geq 1} n2^n/(2^n-1)^2$, and $b_{r+1} = (-1)^r 2^{-\binom{r+1}{2}}$.*

# VARIANCE OF TPL $T_n$ ($b = 1$)

*The fluctuating function $\delta(x)$ is continuous with period 1, mean zero, and $|\delta(x)| \le 10^{-6}$, and $|[\delta_1^2]_0| \le 10^{-10}$, and $|[\delta_1\delta_2]_0| \le 10^{-10}$. Finally, $\hat{w}(z)$ is a function defined as*

$$(2.14)\quad \frac{\hat{w}(z+1)}{Q_{z-1}} = -2Q_\infty z + \frac{\xi(z+2)}{2^z Q_z} + \frac{\xi(z+3)}{2^{z+1}Q_{z+1}} + \sum_{j\ge 2}\left(\frac{\xi(z+j+2)}{2^{z+j}Q_{z+j}} - \frac{\xi(j+2)}{2^j Q_j}\right)$$

*with $Q_z = Q_\infty/Q(2^{-z})$, where $Q(t) = \prod_{i\ge 1}(1 - t/2^i)$, $Q_\infty = Q(1)$, and*

$$(2.15)\quad \xi(z+1) = \sum_{r\ge 0}\frac{b_{r+1}}{Q_r}\cdot\frac{Q_\infty}{Q(2^{3-z-r})} \cdot\left\{2^z - \frac{2}{1-2^{1-z-r}} - \frac{2z}{1-2^{2-z-r}} + 2\sum_{k\ge 2}\binom{z}{k}\frac{1}{2^{r+k-1}-1}\right\}.$$

*Numerical evaluation of the constant $C$ reveals that $C = 0.26600\ldots$ and all five digits after the decimal point are significant.*

# A BETTER EXPRESSION OF $C$

$$Q_k := \prod_{1\le j\le k}(1-2^{-j})$$

$$C = \frac{Q_\infty}{\log 2}\sum_{k,\ell,m\ge 0}\frac{(-1)^m 2^{-\binom{m+1}{2}}}{Q_k Q_\ell Q_m 2^{k+\ell}}\,\lambda\left(2^{-k-m}+2^{-\ell-m}\right)$$

$$\approx 0.26600\,36454\ldots,$$

**where** $\lambda(x) := \begin{cases} \dfrac{x-\log x-1}{(x-1)^2}, & \textbf{if } x\neq 1; \\ 1, & \textbf{if } x = 1. \end{cases}$

# BEYOND POISSONIZED VARIANCE

**Let $\tilde{V}(z) := \tilde{f}_2(z) - \tilde{S}(z)$**

$$\tilde{S}(z) := \sum_{k\geq 0} \frac{z^k}{k!} \tilde{f}_1^{(k)}(z)^2 = \tilde{f}_1(z)^2 + z\tilde{f}_1'(z)^2 + \frac{z^2}{2}\tilde{f}_1''(z)^2 + \cdots .$$

**Then**

$$\begin{aligned}\mathbb{V}(X_n) &= n![z^n]e^z\tilde{f}_2(z) - \left(n![z^n]e^z\tilde{f}_1(z)\right)^2 \\ &= n![z^n]e^z\tilde{V}(z) = \sum_{j\geq 0}\frac{\tau_j(n)}{j!}\tilde{V}^{(j)}(n).\end{aligned}$$

**An identity**

$$\left(\sum_{j\geq 0}\frac{\tau_j(n)}{j!}\tilde{f}^{(j)}(n)\right)^2 = \sum_{j\geq 0}\frac{\tau_j(n)}{j!}\left(\sum_{k\geq 0}\frac{n^k}{k!}\tilde{f}^{(k)}(n)^2\right)^{(j)}.$$

# EXTERNAL PATH LENGTH OF RANDOM TRIES

## symmetric case

$$\tilde{f}_1(z) = 2\tilde{f}_1(z/2) + z(1 - e^{-z})$$
$$\tilde{f}_2(z) = 2\tilde{f}_2(z/2) + 2\tilde{f}_1(z/2)^2 + 4z\tilde{f}_1(z/2) + 2z\tilde{f}_1'(z/2) + z(1 - e^{-z}) + z^2.$$

## $\tilde{V}(z) := \tilde{f}_2(z) - \sum_{k\geq 0} \frac{z^k}{k!}\tilde{f}_1^{(k)}(z)^2$

**Then**

$$\tilde{V}(z) = 2\tilde{V}(z/2) + \tilde{w}(z)$$
$$\tilde{w}(z) = \sum_{k\geq 2} \frac{2(1 - 2^{1-k})}{k!}(z/2)^k \tilde{f}_1^{(k)}(z/2)^2.$$

# EXTERNAL PATH LENGTH OF RANDOM TRIES

## symmetric case

$$\tilde{V}(z) = \frac{1}{2\pi i} \int_{(c)} \frac{\tilde{w}^*(s) z^{-s}}{1 - 2^{s+1}} \mathbf{d}s$$

## An identity

$$\mathbb{V}(X_n) = \sum_{j \geq 0} \frac{\tilde{V}^{(j)}(n)}{j!} \tau_j(n)$$

$$\tilde{V}(z) = \frac{1}{\log 2} \sum_{k \in \mathbb{Z}} \tilde{w}^*(-1 - \chi_k) z^{1+\chi_k} - \sum_{j \geq 1} 2^{-j} \tilde{w}(2^j z).$$

***leads to many identities***

# MORE CONCRETE EXAMPLES FROM LITERATURE

Random digital trees, radix sort, conflict resolution algorithms, statistical physical models, etc.

- **Total external path length of Patricia tries (Kirschenhofer et al., 1988):**

$$\mathscr{P}(z,y) = \mathscr{P}(pe^y z, y)\mathscr{P}(qe^y z, y) + \mathscr{P}(pz, y) + \mathscr{P}(qz, y) - \mathscr{P}(pe^y z, y) - \mathscr{P}(qe^y z, y).$$

- **Cost of radix sort (Mahmoud et al., 2000)**

$$\mathscr{P}(z,y) = (1-e^y)z + \mathscr{P}\left(\frac{e^y z}{b}, y\right)^b.$$

- **# internal nodes in tries (Jacquet and Régnier)**

$$\mathscr{P}(z,y) = (1-e^y)(1+z) + e^y \mathscr{P}\left(\frac{z}{2}, y\right)^2,$$

# MORE CONCRETE EXAMPLES FROM LITERATURE

## Random digital trees, radix sort, conflict resolution algorithms, statistical physical models, etc.

- **Conflict resolution algorithms (Huang and Berger, 1985)**

$$\mathscr{P}(z,y) = (1-e^y)(1+z) + e^y \mathscr{P}\left(\frac{z}{m}, y\right)^m,$$

**with a large number of variants. Wagner (2009)**

$$\mathscr{P}(z,y) = y\mathscr{P}\left(\frac{z}{m}, y\right)^m + (1-y)\left(ze^{-z} + \left(\mathscr{P}\left(\frac{z}{m}, y\right) - \frac{z}{m}e^{-z/m}\right)^m\right)$$

- **A generalized Eden growth model (Dean and Majumdar, 2006)**

$$\frac{\partial}{\partial z}\mathscr{P}(z,y) + \mathscr{P}(z,y) = e^{-y}\mathscr{P}\left(\frac{z}{b}, y\right)^m$$

- ...

***And examples with*** $p \neq q$

# WHY COMPUTING SO PRECISELY?

- **Mathematically challenging**
- **Developing more systematic methodology (applicable to many other problems)**
- **Discovering new phenomena (in turn requiring more structural interpretation)**
- **Periodicity often not visible (very small amplitude)**

# SCHACHINGER (1995): AN (MOSTLY) ELEMENTARY APPROACH TO THE VARIANCE

**Very general toll sequence**

$$X_n \stackrel{d}{=} X_{\mathbf{Binom}(n;p)} + X^*_{n-\mathbf{Binom}(n;p)} + t_n$$

**$p \neq \frac{1}{2}$**

**If (i)** $t_n = O(n^{1/2-\varepsilon})$ **or** $\Delta t_n = O(n^{-\varepsilon})$ **then** $\mathbb{V}(X_n) = \Theta(n)$

**$p = \frac{1}{2}$**

**If (i)** $t_n = O(n^{1/2-\varepsilon})$ **or** $\Delta t_n = O(n^{-\varepsilon})$ **or** $\Delta^2 t_n = O(n^{-1/2-\varepsilon})$**, then** $\mathbb{V}(X_n) \sim n\mathbf{Periodic}(n)$**.**

# *Dziękuję*