Height and Saturation Level of Random Digital Trees

(joint with M. Drmota, H.-K. Hwang and R. Neininger)

Michael Fuchs

Department of Mathematical Sciences
National Chengchi University

August 21st, 2019

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:

011011
010101
101110
010000
101010
001100

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:

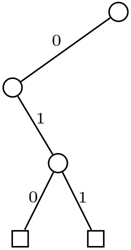
011011
010101
101110
010000
101010
001100

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



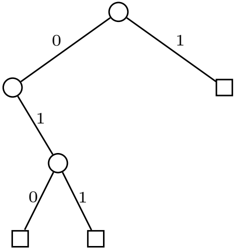
011011
010101
101110
010000
101010
001100

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



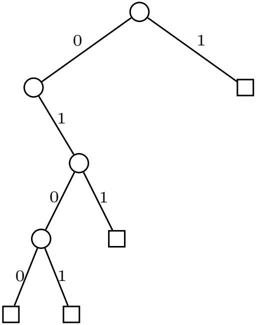
011011
010101
101110
010000
101010
001100

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



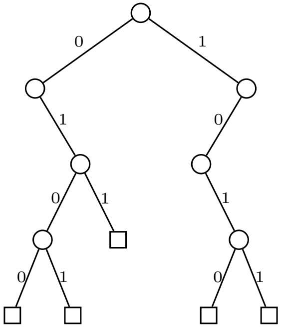
011011
010101
101110
010000
101010
001100

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



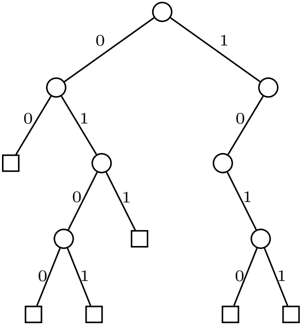
011011
010101
101110
010000
101010
001100

Tries

Proposed by René de la Briandais in 1959.

Name from the word data retrieval (suggested by Fredkin).

Example:



011011
010101
101110
010000
101010
001100

PATRICIA Tries

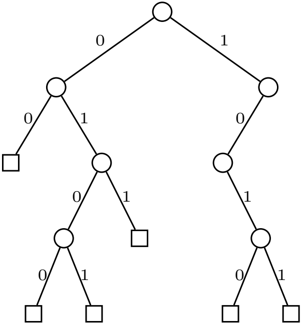
Proposed by Donald R. Morrison in 1968.

PATRICIA=Practical Algorithm To Retrieve Information Coded In Alphanumeric

PATRICIA Tries

Proposed by Donald R. Morrison in 1968.

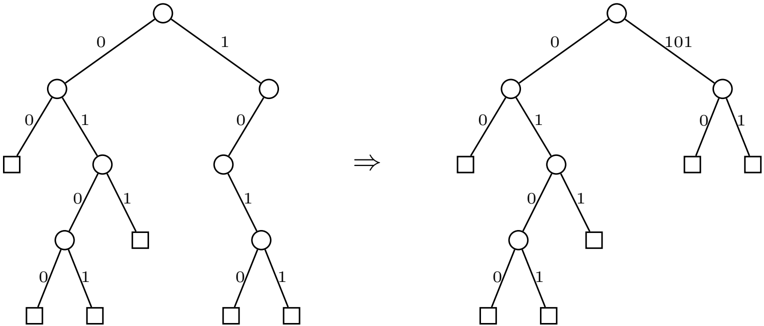
PATRICIA=Practical Algorithm To Retrieve Information Coded In Alphanumeric



PATRICIA Tries

Proposed by Donald R. Morrison in 1968.

PATRICIA=Practical Algorithm To Retrieve Information Coded In Alphanumeric



Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Example:

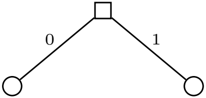
011011
010101
101110
010000
101010
001100

Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Example:



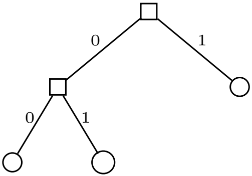
011011
010101
101110
010000
101010
001100

Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Example:



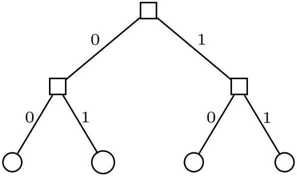
011011
010101
101110
010000
101010
001100

Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Example:



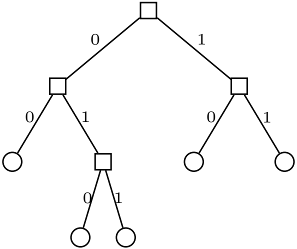
011011
010101
101110
010000
101010
001100

Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Example:



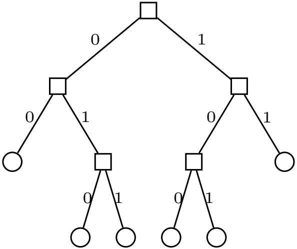
011011
010101
101110
010000
101010
001100

Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Example:



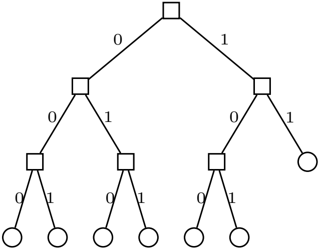
011011
010101
101110
010000
101010
001100

Digital Search Trees (DSTs)

Proposed by Edward G. Coffman and James Eve in 1970.

Closely related to the Lempel-Ziv compression scheme.

Example:



011011
010101
101110
010000
101010
001100

Random Model

Bits are generated by independent Bernoulli random variables with mean p

$\longrightarrow$ *Bernoulli model*

Random Model

Bits are generated by independent Bernoulli random variables with mean p

$\longrightarrow$ *Bernoulli model*

Two types of digital trees:

- $p = 1/2$: *symmetric digital trees;*
- $p \neq 1/2$: *asymmetric digital trees.*

Random Model

Bits are generated by independent Bernoulli random variables with mean p

$\longrightarrow$ *Bernoulli model*

Two types of digital trees:

- $p = 1/2$: *symmetric digital trees;*
- $p \neq 1/2$: *asymmetric digital trees.*

Question: What can be said about the "shape" of the tree?

Random Model

Bits are generated by independent Bernoulli random variables with mean p

$\longrightarrow$ *Bernoulli model*

Two types of digital trees:

- $p = 1/2$: *symmetric digital trees;*
- $p \neq 1/2$: *asymmetric digital trees.*

Question: What can be said about the "shape" of the tree?

This question is important because its answer will shed light on the complexity of algorithms performed on digital trees.

Three Shape Parameters

$H_n =$ longest path to a leaf;

$S_n =$ shortest path to a leaf;

$F_n =$ saturation (or fill-up) level;

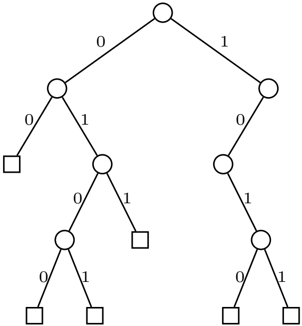
Three Shape Parameters

H_n = longest path to a leaf;

S_n = shortest path to a leaf;

F_n = saturation (or fill-up) level;

Example:



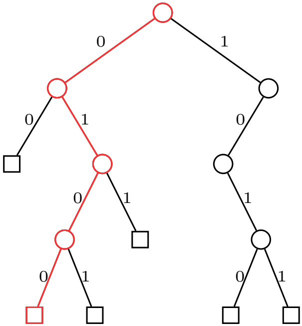
Three Shape Parameters

H_n = longest path to a leaf;

S_n = shortest path to a leaf;

F_n = saturation (or fill-up) level;

Example:



$H_n = 4;$

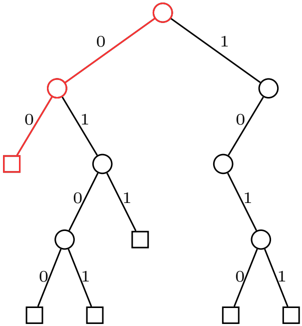
Three Shape Parameters

H_n = longest path to a leaf;

S_n = shortest path to a leaf;

F_n = saturation (or fill-up) level;

Example:



$H_n = 4;$

$S_n = 2;$

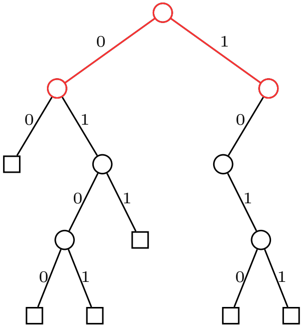
Three Shape Parameters

H_n = longest path to a leaf;

S_n = shortest path to a leaf;

F_n = saturation (or fill-up) level;

Example:



$H_n = 4;$

$S_n = 2;$

$F_n = 1.$

Results for Tries (i)

Flajolet (1983):

Theorem

For symmetric tries,

$$\mathbb{P}(H_n \le k) \to e^{-e^{-t}},$$

where k and n tend to infinity such that $\log(2^{k+1}/n^2) \to t$.

Results for Tries (i)

Flajolet (1983):

Theorem

For symmetric tries,

$$\mathbb{P}(H_n \le k) \to e^{-e^{-t}},$$

where k *and* n *tend to infinity such that* $\log(2^{k+1}/n^2) \to t$.

This shows that the "limit distribution" of the height is a **Gumbel distribution**.

Results for Tries (i)

Flajolet (1983):

Theorem

For symmetric tries,

$$\mathbb{P}(H_n \leq k) \to e^{-e^{-t}},$$

where k *and* n *tend to infinity such that* $\log(2^{k+1}/n^2) \to t$.

This shows that the "limit distribution" of the height is a **Gumbel distribution**.

The above result was generalized to asymmetric tries by Pittel (with a probabilistic approach) and Jacquet & Règnier (with a complex-analytic approach) in 1986.

Results for Tries (ii)

Theorem (Pittel; 1986)

Let $p \geq q$. The distribution of S_n is concentrated on two points:

$$\mathbb{P}(S_n = k_S \text{ or } k_S + 1) \longrightarrow 1, \qquad \text{as } n \longrightarrow \infty.$$

Here, k_S is a sequence of n which satisfies

$$k_S = \begin{cases} \log_2 n - \log_2 \log n + \mathcal{O}(1), & \text{if } p = q; \\ \log_{1/q} n - \log_{1/q} \log\log n + \mathcal{O}(1), & \text{if } p \neq q. \end{cases}$$

Results for Tries (ii)

Theorem (Pittel; 1986)

Let $p \geq q$. The distribution of S_n is concentrated on two points:

$$\mathbb{P}(S_n = k_S \text{ or } k_S + 1) \longrightarrow 1, \qquad \text{as } n \longrightarrow \infty.$$

Here, k_S is a sequence of n which satisfies

$$k_S = \begin{cases} \log_2 n - \log_2 \log n + \mathcal{O}(1), & \text{if } p = q; \\ \log_{1/q} n - \log_{1/q} \log \log n + \mathcal{O}(1), & \text{if } p \neq q. \end{cases}$$

Theorem (Hwang & Nicodème & Park & Szpankowski; 2006)

We have,

$$\mathbb{P}(F_n = S_n - 1) \longrightarrow 1, \qquad \text{as } n \longrightarrow \infty.$$

External and Internal Node Profile

$B_{n,k}$ = number of external nodes at level k;

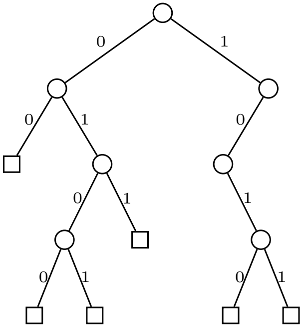
$I_{n,k}$ = number of internal nodes at level k.

External and Internal Node Profile

$B_{n,k}$ = number of external nodes at level k;

$I_{n,k}$ = number of internal nodes at level k.

Example:

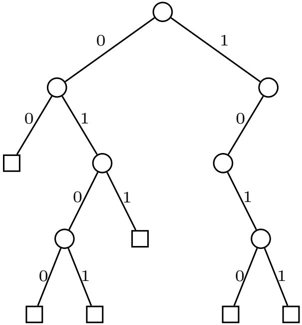


External and Internal Node Profile

$B_{n,k}$ = number of external nodes at level k;

$I_{n,k}$ = number of internal nodes at level k.

Example:



$B_{6,0} = 0;$

$B_{6,1} = 0;$

$B_{6,2} = 1;$

$B_{6,3} = 1;$

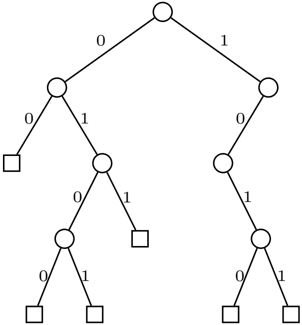
$B_{6,4} = 4;$

External and Internal Node Profile

$B_{n,k}$ = number of external nodes at level k;

$I_{n,k}$ = number of internal nodes at level k.

Example:



$B_{6,0} = 0;$ $I_{6,0} = 1;$

$B_{6,1} = 0;$ $I_{6,1} = 2;$

$B_{6,2} = 1;$ $I_{6,2} = 2;$

$B_{6,3} = 1;$ $I_{6,3} = 2;$

$B_{6,4} = 4;$ $I_{6,4} = 0;$

H_n, S_n, F_n and the Profile of Tries

$$H_n = \max\{k \ : \ B_{n,k} > 0\};$$

$$S_n = \min\{k \ : \ B_{n,k} > 0\};$$

$$F_n = \max\{k \ : \ I_{n,k} = 2^k\}.$$

H_n, S_n, F_n and the Profile of Tries

$H_n = \max\{k \ : \ B_{n,k} > 0\};$

$S_n = \min\{k \ : \ B_{n,k} > 0\};$

$F_n = \max\{k \ : \ I_{n,k} = 2^k\}.$

So, for instance, we have

$$S_n > k \qquad \Longrightarrow \qquad B_{n,k} = 0$$

and

$$S_n < k \qquad \Longrightarrow \qquad B_{n,\ell} > 0 \text{ for some } \ell < k$$

H_n, S_n, F_n and the Profile of Tries

$H_n = \max\{k \ : \ B_{n,k} > 0\};$

$S_n = \min\{k \ : \ B_{n,k} > 0\};$

$F_n = \max\{k \ : \ I_{n,k} = 2^k\}.$

So, for instance, we have

$$S_n > k \qquad \Longrightarrow \qquad B_{n,k} = 0$$

and

$$S_n < k \qquad \Longrightarrow \qquad B_{n,\ell} > 0 \text{ for some } \ell < k$$

and thus

$$\mathbb{P}(S_n > k) \leq \mathbb{P}(B_{n,k} = 0) \qquad \text{and} \qquad \mathbb{P}(S_n < k) \leq \sum_{\ell=0}^{k-1} \mathbb{P}(B_{n,\ell} > 0).$$

First and Second Moment Method

Theorem

Let X be a non-negative, integer-valued random variable. Then,

$$\mathbb{P}(X > 0) \le \mathbb{E}(X).$$

and

$$\mathbb{P}(X = 0) \le \frac{\mathrm{Var}(X)}{(\mathbb{E}(X))^2}.$$

First and Second Moment Method

Theorem

Let X be a non-negative, integer-valued random variable. Then,

$$\mathbb{P}(X > 0) \leq \mathbb{E}(X).$$

and

$$\mathbb{P}(X = 0) \leq \frac{\mathrm{Var}(X)}{(\mathbb{E}(X))^2}.$$

Thus,

$$\mathbb{P}(S_n > k) \leq \frac{\mathrm{Var}(B_{n,k})}{(\mathbb{E}(B_{n,k}))^2}$$

and

$$\mathbb{P}(S_n < k) \leq \sum_{\ell=0}^{k-1} \mathbb{E}(B_{n,\ell}).$$

Profile of Tries (Hwang et al.; 2006)

Let $p \geq q$ and

$$\alpha_1 := \frac{1}{\log(1/q)},\ \alpha_2 := \frac{p^2+q^2}{p^2\log(1/p)+q^2\log(1/q)},\ \alpha_3 := \frac{2}{\log(1/(p^2+q^2))}$$

and

$$\rho := \frac{1}{\log(p/q)}\log\left(\frac{1-\alpha\log(1/p)}{\alpha\log(1/q)-1}\right) \qquad \text{with } \alpha = \lim_n (k/\log n).$$

Profile of Tries (Hwang et al.; 2006)

Let $p \geq q$ and

$$\alpha_1 := \frac{1}{\log(1/q)}, \ \alpha_2 := \frac{p^2 + q^2}{p^2 \log(1/p) + q^2 \log(1/q)}, \ \alpha_3 := \frac{2}{\log(1/(p^2+q^2))}$$

and

$$\rho := \frac{1}{\log(p/q)} \log\left(\frac{1 - \alpha \log(1/p)}{\alpha \log(1/q) - 1}\right) \qquad \text{with } \alpha = \lim_n (k/\log n).$$

Then,

$$\frac{\log \mathbb{E}(B_{n,k})}{\log n} \to \begin{cases} 0, & \text{if } \alpha \leq \alpha_1; \\ -\rho + \alpha \log(p^{-\rho} + q^{-\rho}), & \text{if } \alpha_1 \leq \alpha \leq \alpha_2; \\ 2 + \alpha \log(p^2 + q^2), & \text{if} \alpha_2 \leq \alpha \leq \alpha_3; \\ 0, & \text{if } \alpha \geq \alpha_3 \end{cases}$$

and $\mathrm{Var}(B_{n,k}) = \Theta(\mathbb{E}(B_{n,k}))$.

Concentration of Saturation Level and Height

Saturation Level:

Trees	$p = q$?	Concentration	Reference
Tries	$0 < p < 1$	2 points	HNPS2006
DSTs	$p = \frac{1}{2}$	2 points	DFHN2019+
	$p \neq \frac{1}{2}$	at most 3 points	DF2019+
PATRICIA Tries	$0 < p < 1$	2 points	HNPS2006

Concentration of Saturation Level and Height

Saturation Level:

Trees	$p = q$?	Concentration	Reference
Tries	$0 < p < 1$	2 points	HNPS2006
DSTs	$p = \frac{1}{2}$ $p \neq \frac{1}{2}$	2 points at most 3 points	DFHN2019+ DF2019+
PATRICIA Tries	$0 < p < 1$	2 points	HNPS2006

Height:

Trees	$p = q$?	Concentration	Reference
Tries	$0 < p < 1$	no	F1983; P1986; JR1986
DSTs	$p = \frac{1}{2}$ $p \neq \frac{1}{2}$	2 points ?	DFHN2019+ DF2019+
PATRICIA Tries	$p = \frac{1}{2}$ $p \neq \frac{1}{2}$	3 points ?	Conjectured by KS2002 ?

Profile of Symmetric DSTs: Mean

Let

$$Q(z) = \prod_{\ell=1}^{\infty} \left(1 - z2^{-\ell}\right), \qquad Q_n = \prod_{\ell=1}^{n} \left(1 - 2^{-\ell}\right) = \frac{Q(2^{-n})}{Q(1)}.$$

Profile of Symmetric DSTs: Mean

Let

$$Q(z) = \prod_{\ell=1}^{\infty}\left(1 - z2^{-\ell}\right), \qquad Q_n = \prod_{\ell=1}^{n}\left(1 - 2^{-\ell}\right) = \frac{Q(2^{-n})}{Q(1)}.$$

Theorem (Drmota & F. & Hwang & Neininger; 2019+)

We have,

$$\mathbb{E}(B_{n,k}) = 2^k F(n/2^k) + \mathcal{O}(1),$$

where $F(x)$ is the positive function

$$F(x) = \sum_{j\geq 0} \frac{(-1)^j 2^{-\binom{j}{2}}}{Q_j Q(1)} e^{-2^j x}.$$

Profile of Symmetric DSTs: $F(x)$ (i)

As $x \to \infty$,

$$F(x) = \frac{e^{-x}}{Q(1)} + \mathcal{O}(e^{-2x})$$

Profile of Symmetric DSTs: $F(x)$ (i)

As $x \to \infty$,

$$F(x) = \frac{e^{-x}}{Q(1)} + \mathcal{O}(e^{-2x})$$

and as $x \to 0$,

$$F(x) \sim \frac{X^{1/\log 2}}{\sqrt{2\pi x}} \exp\left(-\frac{(\log(X \log X))^2}{2\log 2} - \sum_{j\in\mathbb{Z}} c_j (X \log X)^{-\chi_j} \right),$$

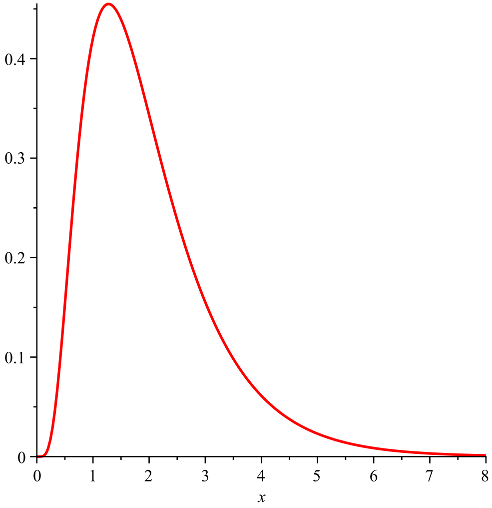
where $X = 1/(x \log 2)$, $\chi_j = 2j\pi i/\log 2$,

$$c_0 = \frac{\log 2}{12} + \frac{\pi^2}{6\log 2}$$

and

$$c_j = \frac{1}{2j\sinh(2j\pi^2/\log 2)}, \qquad (j \neq 0).$$

Profile of Symmetric DSTs: $F(x)$ (ii)



Profile of Symmetric DSTs: Variance

Theorem (Drmota & F. & Hwang & Neininger; 2019+)

We have,

$$\mathrm{Var}(B_{n,k}) = 2^k G(n/2^k) + \mathcal{O}(1),$$

where $G(x)$ *is a function with*

$$G(x) = \frac{e^{-x}}{Q(1)} + \mathcal{O}(xe^{-2x}), \qquad (x \to \infty)$$

and

$$G(x) \sim 2F(x), \qquad (x \to 0).$$

Profile of Symmetric DSTs: $G(x)$ (i)

We have,

$$G(x)=\sum_{j,r=0}^{\infty}\sum_{0\le h,\ell\le j}\frac{2^{-j}(-1)^{r+h+\ell}2^{-\binom{r}{2}-\binom{h}{2}-\binom{\ell}{2}+2h+2\ell}}{Q_rQ(1)Q_hQ_{j-h}Q_\ell Q_{j-\ell}}\varphi(2^{r+j},2^h+2^\ell;x),$$

where

$$\varphi(u,v;x)=\begin{cases}\dfrac{e^{-ux}-((v-u)x+1)e^{-vx}}{(v-u)^2}, & \text{if } u\neq v;\\ x^2e^{-ux}/2, & \text{if } u=v.\end{cases}$$

Profile of Symmetric DSTs: $G(x)$ (i)

We have,

$$G(x) = \sum_{j,r=0}^{\infty} \sum_{0\le h,\ell\le j} \frac{2^{-j}(-1)^{r+h+\ell}2^{-\binom{r}{2}-\binom{h}{2}-\binom{\ell}{2}+2h+2\ell}}{Q_rQ(1)Q_hQ_{j-h}Q_\ell Q_{j-\ell}}\varphi(2^{r+j}, 2^h+2^\ell; x),$$

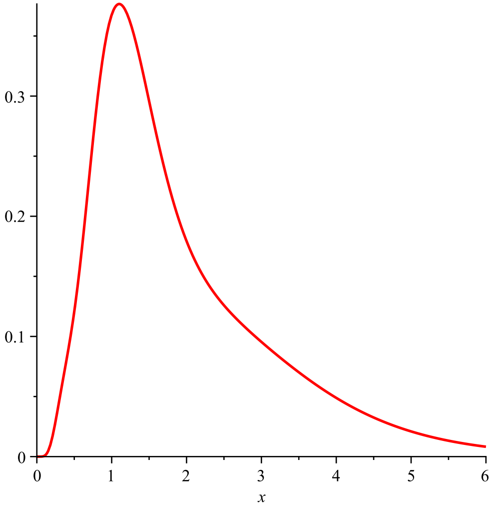
where

$$\varphi(u,v;x) = \begin{cases} \dfrac{e^{-ux}-((v-u)x+1)e^{-vx}}{(v-u)^2}, & \text{if } u \neq v; \\ x^2e^{-ux}/2, & \text{if } u = v. \end{cases}$$

Proposition (Drmota & F. & Hwang & Neininger; 2019+)

$G(x)$ is a positive function on $(0,\infty)$.

Profile of Symmetric DSTs: $G(x)$ (ii)



Major Tools for the Proofs

Major Tools for the Proofs

- Analytic Depoissonization & JS-admissibility

  Developed by Jacquet & Szpankowski (1998).

Major Tools for the Proofs

- Analytic Depoissonization & JS-admissibility

  Developed by Jacquet & Szpankowski (1998).

- Theory of Poisson Variance

  Developed by F., Hwang, Zacharovs (2010,2014).

Major Tools for the Proofs

- Analytic Depoissonization & JS-admissibility

  Developed by Jacquet & Szpankowski (1998).

- Theory of Poisson Variance

  Developed by F., Hwang, Zacharovs (2010,2014).

- Mellin Transform

  Systemized by Flajolet, Gourdon, Dumas (1995).

Major Tools for the Proofs

- Analytic Depoissonization & JS-admissibility

  Developed by Jacquet & Szpankowski (1998).

- Theory of Poisson Variance

  Developed by F., Hwang, Zacharovs (2010,2014).

- Mellin Transform

  Systemized by Flajolet, Gourdon, Dumas (1995).

- Laplace Transform

Major Tools for the Proofs

- Analytic Depoissonization & JS-admissibility

  Developed by Jacquet & Szpankowski (1998).

- Theory of Poisson Variance

  Developed by F., Hwang, Zacharovs (2010,2014).

- Mellin Transform

  Systemized by Flajolet, Gourdon, Dumas (1995).

- Laplace Transform

- Saddle-point Method

Profile of Symmetric DSTs: Limit Laws

$$k_f := \log_2 n - \log_2 \log n + 1 + \frac{\log_2 \log n}{\log n};$$

$$k_h := \log_2 n + \sqrt{2 \log_2 n} - \frac{1}{2} \log_2 \log_2 n + \frac{1}{\log 2} - \frac{3 \log \log n}{4\sqrt{2(\log n)(\log 2)}}.$$

Profile of Symmetric DSTs: Limit Laws

$$k_f := \log_2 n - \log_2 \log n + 1 + \frac{\log_2 \log n}{\log n};$$

$$k_h := \log_2 n + \sqrt{2 \log_2 n} - \frac{1}{2} \log_2 \log_2 n + \frac{1}{\log 2} - \frac{3 \log \log n}{4\sqrt{2(\log n)(\log 2)}}.$$

Theorem (Drmota & F. & Hwang & Neininger; 2019+)

(i) $\mathbb{E}(B_{n,k}), \mathrm{Var}(B_{n,k}) \to \infty$ *iff there exists* $\omega_n \to \infty$ *with*

$$k_f + \frac{\omega_n}{\log n} \le k \le k_h - \frac{\omega_n}{\sqrt{\log n}}.$$

(ii) *If* $\mathbb{E}(B_{n,k}) \to \infty$, *then*

$$\frac{B_{n,k} - \mathbb{E}(B_{n,k})}{\sqrt{\mathrm{Var}(B_{n,k})}} \xrightarrow{d} N(0,1).$$

Saturation Level and Height of Symmetric DSTs (i)

Recall,

$$\mathbb{E}(B_{n,k}) = 2^k F(n/2^k) + \mathcal{O}(1).$$

Saturation Level and Height of Symmetric DSTs (i)

Recall,

$$\mathbb{E}(B_{n,k}) = 2^k F(n/2^k) + \mathcal{O}(1).$$

This result is not precise enough to understand the behavior of the saturation level and height!

Saturation Level and Height of Symmetric DSTs (i)

Recall,

$$\mathbb{E}(B_{n,k}) = 2^k F(n/2^k) + \mathcal{O}(1).$$

This result is not precise enough to understand the behavior of the saturation level and height!

However, it can be refined to

$$\mathbb{E}(B_{n,k}) = 2^k F(n/2^k) + F'(n/2^k) - 2^{-k-1} n F''(n/2^k) + \mathcal{O}(n^{-1} + n/4^k)$$

and for $n/2^k \to \infty$

$$\mathbb{E}(B_{n,k}) \sim \frac{2^k}{Q_k}(1 - 2^{-k})^n.$$

Saturation Level and Height of Symmetric DSTs (i)

Recall,

$$\mathbb{E}(B_{n,k}) = 2^k F(n/2^k) + \mathcal{O}(1).$$

This result is not precise enough to understand the behavior of the saturation level and height!

However, it can be refined to

$$\mathbb{E}(B_{n,k}) = 2^k F(n/2^k) + F'(n/2^k) - 2^{-k-1} n F''(n/2^k) + \mathcal{O}(n^{-1} + n/4^k)$$

and for $n/2^k \to \infty$

$$\mathbb{E}(B_{n,k}) \sim \frac{2^k}{Q_k}(1 - 2^{-k})^n.$$

These results are sufficient!

Saturation Level and Height of Symmetric DSTs (ii)

Theorem (Drmota & F. & Hwang & Neininger; 2019+)

Let

$$k_H := \left\lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{1}{2}\log_2\log_2 n + \frac{1}{\log 2} \right\rfloor.$$

Then, for the height H_n of symmetric DSTs,

$$\mathbb{P}(H_n = k_H \text{ or } k_H + 1) \longrightarrow 1, \qquad \text{as } n \longrightarrow \infty.$$

This was conjectured by Aldous & Shields (1988).

Saturation Level and Height of Symmetric DSTs (ii)

Theorem (Drmota & F. & Hwang & Neininger; 2019+)

Let

$$k_H := \left\lfloor \log_2 n + \sqrt{2\log_2 n} - \frac{1}{2}\log_2\log_2 n + \frac{1}{\log 2} \right\rfloor.$$

Then, for the height H_n of symmetric DSTs,

$$\mathbb{P}(H_n = k_H \textit{ or } k_H + 1) \longrightarrow 1, \qquad \textit{as } n \longrightarrow \infty.$$

This was conjectured by Aldous & Shields (1988).

Theorem (Drmota & F. & Hwang & Neininger; 2019+)

Let $k_F := \lceil \log_2 n - \log_2 \log n \rceil$. Then, for the saturation level F_n of symmetric DSTs,

$$\mathbb{P}(F_n = k_F - 1 \textit{ or } k_F) \longrightarrow 1, \qquad \textit{as } n \longrightarrow \infty.$$

Profile of Asymmetric DSTs: Notation

Assume that $p \geq q$.

Set

$$\alpha_1 = \frac{1}{\log(1/q)}, \qquad \alpha_2 = \frac{1}{\log(1/p)}$$

and

$$\rho = \frac{1}{\log(p/q)} \log\left(\frac{1-\alpha\log(1/p)}{\alpha\log(1/q)-1}\right),$$

where

$$\alpha = \lim_{n\to\infty} \frac{k}{\log n}.$$

Moreover, set

$$v = -\rho + \alpha\log(p^{-\rho}+q^{-\rho}).$$

Profile of Asymmetric DSTs: Mean & Variance

Theorem (Drmota & Szpankowski; 2011)

If $(\alpha_1 + \epsilon)\log n \le k \le (\alpha_2 - \epsilon)\log n$, *then*

$$\mathbb{E}(B_{n,k}) \sim H_1\left(\rho; \log_{p/q} p^k n\right) \frac{p^\rho q^\rho (p^{-\rho} + q^{-\rho})}{\sqrt{2\pi\alpha}\log(p/q)} \cdot \frac{n^v}{\sqrt{\log n}},$$

where $H_1(\rho; x)$ *is a* 1*-periodic function.*

Profile of Asymmetric DSTs: Mean & Variance

Theorem (Drmota & Szpankowski; 2011)

If $(\alpha_1+\epsilon)\log n \le k \le (\alpha_2-\epsilon)\log n$, *then*

$$\mathbb{E}(B_{n,k}) \sim H_1\left(\rho;\log_{p/q} p^k n\right)\frac{p^\rho q^\rho(p^{-\rho}+q^{-\rho})}{\sqrt{2\pi\alpha}\log(p/q)}\cdot\frac{n^v}{\sqrt{\log n}},$$

where $H_1(\rho;x)$ *is a* 1*-periodic function.*

Theorem (Kazemi & Vahidi-Asl; 2011)

If $(\alpha_1+\epsilon)\log n \le k \le (\alpha_2-\epsilon)\log n$, *then*

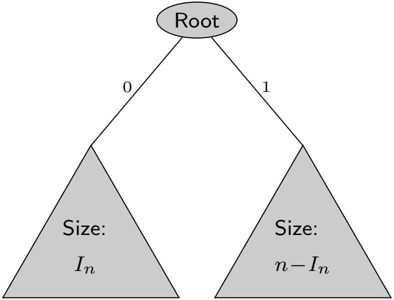
$$\mathrm{Var}(B_{n,k}) \sim H_2\left(\rho;\log_{p/q} p^k n\right)\frac{p^\rho q^\rho(p^{-\rho}+q^{-\rho})}{\sqrt{2\pi\alpha}\log(p/q)}\cdot\frac{n^v}{\sqrt{\log n}},$$

where $H_2(\rho;x)$ *is a* 1*-periodic function.*

Recurrences

$$B_{n+1,k} \stackrel{d}{=} B_{I_n,k-1} + B^*_{n-I_n,k-1}$$

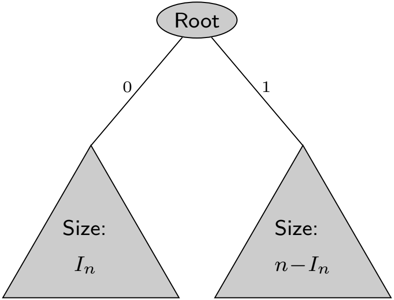
- $I_n \stackrel{d}{=}$ Binomial(n, p);
- $B_{n,k} \stackrel{d}{=} B^*_{n,k}$;
- $B_{n,k}, B^*_{n,k}, I_n$ independent.



Recurrences

$$B_{n+1,k} \stackrel{d}{=} B_{I_n,k-1} + B^*_{n-I_n,k-1}$$

- $I_n \stackrel{d}{=} \text{Binomial}(n,p)$;
- $B_{n,k} \stackrel{d}{=} B^*_{n,k}$;
- $B_{n,k}, B^*_{n,k}, I_n$ independent.



This gives the following recurrence for the mean ($\mu_{n,k} := \mathbb{E}(B_{n,k})$)

$$\mu_{n+1,k} = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} \left(\mu_{j,k-1} + \mu_{n-j,k-1}\right).$$

Solving the Recurrence for the Mean (i)

$$\mu_{n+1,k} = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} \left(\mu_{j,k-1} + \mu_{n-j,k-1}\right).$$

Solving the Recurrence for the Mean (i)

$$\mu_{n+1,k} = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} \left(\mu_{j,k-1} + \mu_{n-j,k-1}\right).$$

- Consider the **Poisson-generating function**:

$$\tilde{f}_k(z) := e^{-z} \sum_n \mu_{n,k} \frac{z^n}{n!}.$$

Then,

$$\tilde{f}_k'(z) + \tilde{f}_k(z) = \tilde{f}_{k-1}(pz) + \tilde{f}_{k-1}(qz).$$

Solving the Recurrence for the Mean (i)

$$\mu_{n+1,k} = \sum_{j=0}^{n} \binom{n}{j} p^j q^{n-j} \left(\mu_{j,k-1} + \mu_{n-j,k-1}\right).$$

- Consider the **Poisson-generating function**:

$$\tilde{f}_k(z) := e^{-z} \sum_{n} \mu_{n,k} \frac{z^n}{n!}.$$

  Then,

$$\tilde{f}'_k(z) + \tilde{f}_k(z) = \tilde{f}_{k-1}(pz) + \tilde{f}_{k-1}(qz).$$

- Consider the (normalized) **Mellin-transform**:

$$F_k(s) := \frac{1}{\Gamma(s)} \int_0^\infty \tilde{f}_k(z) z^{s-1} \mathrm{d}s,$$

  where $\Gamma(s)$ is the Gamma-function.

Solving the Recurrence for the Mean (ii)

Then,

$$F_k(s) - F_k(s-1) = T(s)F_{k-1}(s),$$

where

$$T(s) := p^{-s} + q^{-s}.$$

Solving the Recurrence for the Mean (ii)

Then,

$$F_k(s) - F_k(s-1) = T(s)F_{k-1}(s),$$

where

$$T(s) := p^{-s} + q^{-s}.$$

- Consider the **ordinary generating function**:

$$f(s,\omega) := \sum_k F_k(s)\omega_k.$$

Then,

$$f(s,\omega) = \frac{f(s-1,\omega)}{1-\omega T(s)}$$

Solving the Recurrence for the Mean (ii)

Then,

$$F_k(s) - F_k(s-1) = T(s)F_{k-1}(s),$$

where

$$T(s) := p^{-s} + q^{-s}.$$

- Consider the **ordinary generating function**:

$$f(s,\omega) := \sum_k F_k(s)\omega_k.$$

Then,

$$f(s,\omega) = \frac{f(s-1,\omega)}{1-\omega T(s)}$$

and by iteration

$$f(s,\omega) = \frac{g(s,\omega)}{g(0,\omega)}, \qquad g(s,\omega) := \prod_{j\geq 0} \frac{1}{1-\omega T(s-j)}.$$

Solving the Recurrence for the Mean (iii)

What is left to invert the whole process.

Solving the Recurrence for the Mean (iii)

What is left to invert the whole process.

- From $f(s,\omega)$ to $F_k(s)$:

$$F_k(s) = \frac{1}{2\pi i}\int_{\mathcal{C}_1} \frac{f(s,\omega)}{\omega^{k+1}}\mathrm{d}\omega,$$

where $\mathcal{C}_1$ is a suitable contour.

Solving the Recurrence for the Mean (iii)

What is left to invert the whole process.

- From $f(s,\omega)$ to $F_k(s)$:

$$F_k(s) = \frac{1}{2\pi i}\int_{\mathcal{C}_1} \frac{f(s,\omega)}{\omega^{k+1}}\mathrm{d}\omega,$$

  where $\mathcal{C}_1$ is a suitable contour.

- From $F_k(s)$ to $\tilde{f}_k(z)$:

$$\tilde{f}_k(z) = \frac{1}{2\pi i}\int_{\mathcal{C}_2} \Gamma(s)F_k(s)z^{-s}\mathrm{d}s,$$

  where $\mathcal{C}_2$ is a suitable vertical line.

Solving the Recurrence for the Mean (iii)

What is left to invert the whole process.

- From $f(s,\omega)$ to $F_k(s)$:

$$F_k(s) = \frac{1}{2\pi i}\int_{\mathcal{C}_1} \frac{f(s,\omega)}{\omega^{k+1}}\mathrm{d}\omega,$$

  where $\mathcal{C}_1$ is a suitable contour.

- From $F_k(s)$ to $\tilde{f}_k(z)$:

$$\tilde{f}_k(z) = \frac{1}{2\pi i}\int_{\mathcal{C}_2} \Gamma(s)F_k(s)z^{-s}\mathrm{d}s,$$

  where $\mathcal{C}_2$ is a suitable vertical line.

- From $\tilde{f}_k(z)$ to $\mu_{n,k}$:

$$\mu_{n,k} = \frac{n!}{2\pi i}\int_{\mathcal{C}_3} \frac{e^z\tilde{f}_k(z)}{z^{n+1}}\mathrm{d}z$$

  where $\mathcal{C}_3$ is a suitable contour.

Solving the Recurrence for the Mean (iv)

Drmota & Szpankowski (2011):

$(\alpha_1 + \epsilon) \log n \leq k \leq (\alpha_2 + \epsilon) \log n.$

Solving the Recurrence for the Mean (iv)

Drmota & Szpankowski (2011):

$(\alpha_1 + \epsilon) \log n \leq k \leq (\alpha_2 + \epsilon) \log n.$

- From $f(s, \omega)$ to $F_k(s)$ via residue theorem.
- From $F_k(s)$ to $\tilde{f}_k(z)$ and $\tilde{f}_k(z)$ to $\mu_{n,k}$ via saddle-point method.

Solving the Recurrence for the Mean (iv)

Drmota & Szpankowski (2011):

$(\alpha_1 + \epsilon) \log n \le k \le (\alpha_2 + \epsilon) \log n.$

- From $f(s, \omega)$ to $F_k(s)$ via residue theorem.
- From $F_k(s)$ to $\tilde{f}_k(z)$ and $\tilde{f}_k(z)$ to $\mu_{n,k}$ via saddle-point method.

$\longrightarrow$ "double saddle-point approach" (Hwang et al.; 2006)

Solving the Recurrence for the Mean (iv)

Drmota & Szpankowski (2011):

$(\alpha_1 + \epsilon) \log n \le k \le (\alpha_2 + \epsilon) \log n.$

- From $f(s, \omega)$ to $F_k(s)$ via residue theorem.
- From $F_k(s)$ to $\tilde{f}_k(z)$ and $\tilde{f}_k(z)$ to $\mu_{n,k}$ via saddle-point method.

$\longrightarrow$ "double saddle-point approach" (Hwang et al.; 2006)

Drmota & F. (2019+):

$k \approx \alpha_1 \log n.$

Solving the Recurrence for the Mean (iv)

Drmota & Szpankowski (2011):

$(\alpha_1 + \epsilon) \log n \leq k \leq (\alpha_2 + \epsilon) \log n.$

- From $f(s, \omega)$ to $F_k(s)$ via residue theorem.
- From $F_k(s)$ to $\tilde{f}_k(z)$ and $\tilde{f}_k(z)$ to $\mu_{n,k}$ via saddle-point method.

  $\longrightarrow$ "double saddle-point approach" (Hwang et al.; 2006)

Drmota & F. (2019+):

$k \approx \alpha_1 \log n.$

Saddle point method for the inversion from $\tilde{F}_k(s)$ to $\tilde{f}_k(z)$ has to be replaced by the Poisson summation formula!

Profile of Asymmetric DSTs: Mean

Theorem (Drmota & F.; 2019+)

Let $k = \alpha_1(\log n - \log\log\log n + D)$*, where* $D = \mathcal{O}(1)$*. Then,*

$$\mathbb{E}(B_{n,k}) = \frac{1+o(1)}{\prod_{j\geq 1}(1-q^j)}(\log n)^{\frac{D-\log\log(p/q)-1}{\log(p/q)}}$$

$$\times\left(\frac{(\log(1/q))^{-m_0}}{m_0!}(\log n)^{-\frac{H(m_0\log(p/q)-D+\log\log(p/q))}{\log(p/q)}}\right.$$

$$\left.+\frac{(\log(1/q))^{-m_0-1}}{(m_0+1)!}(\log n)^{-\frac{H((m_0+1)\log(p/q)-D+\log\log(p/q))}{\log(p/q)}}\right)$$

$$+\mathcal{O}\left((\log n)^{\frac{D-\log\log(p/q)-1}{\log(p/q)}-1}\right),$$

where $m_0 := \max(\lfloor(\frac{D-\log\log(p/q)}{\log(p/q)}\rfloor, 0)$ *and* $H(x) := e^x - 1 - x$*.*

Saturation Level of Asymmetric DSTs

Theorem (Drmota & F.; 2019+)

For the saturation level of asymmetric DSTs, we have

$$\mathbb{P}(F_n = k_F - 1 \quad \textit{or} \quad F_n = k_F \quad F_n = k_F + 1) \longrightarrow 1, \qquad \textit{as} \ \longrightarrow \infty,$$

where k_F *is a sequence of* n *which satisfies*

$$k_F = \log_{1/q} n - \log_{1/q} \log\log n + \mathcal{O}(1).$$

Saturation Level of Asymmetric DSTs

Theorem (Drmota & F.; 2019+)

For the saturation level of asymmetric DSTs, we have

$$\mathbb{P}(F_n = k_F - 1 \quad \textit{or} \quad F_n = k_F \quad F_n = k_F + 1) \longrightarrow 1, \qquad \textit{as} \ \longrightarrow \infty,$$

where k_F is a sequence of n which satisfies

$$k_F = \log_{1/q} n - \log_{1/q} \log \log n + \mathcal{O}(1).$$

Remarks:

- Two point concentration holds for almost the whole range of p.

Saturation Level of Asymmetric DSTs

Theorem (Drmota & F.; 2019+)

For the saturation level of asymmetric DSTs, we have

$$\mathbb{P}(F_n = k_F - 1 \quad \textit{or} \quad F_n = k_F \quad F_n = k_F + 1) \longrightarrow 1, \qquad \textit{as} \ \longrightarrow \infty,$$

where k_F is a sequence of n which satisfies

$$k_F = \log_{1/q} n - \log_{1/q} \log\log n + \mathcal{O}(1).$$

Remarks:

- Two point concentration holds for almost the whole range of p.
- We conjecture that two point concentration holds for $1/2 < p < 1$.

Saturation Level of Asymmetric DSTs

Theorem (Drmota & F.; 2019+)

For the saturation level of asymmetric DSTs, we have

$$\mathbb{P}(F_n = k_F - 1 \quad \textit{or} \quad F_n = k_F \quad F_n = k_F + 1) \longrightarrow 1, \qquad \textit{as} \ \longrightarrow \infty,$$

where k_F is a sequence of n which satisfies

$$k_F = \log_{1/q} n - \log_{1/q} \log\log n + \mathcal{O}(1).$$

Remarks:

- Two point concentration holds for almost the whole range of p.
- We conjecture that two point concentration holds for $1/2 < p < 1$.
- We are currently working on a similar concentration result for the height.

Concentration of Saturation Level and Height

Saturation Level:

Trees	$p = q$?	Concentration	Reference
Tries	$0 < p < 1$	2 points	HNPS2006
DSTs	$p = \frac{1}{2}$	2 points	DFHN2019+
	$p \neq \frac{1}{2}$	at most 3 points	DF2019+
PATRICIA Tries	$0 < p < 1$	2 points	HNPS2006

Height:

Trees	$p = q$?	Concentration	Reference
Tries	$0 < p < 1$	no	F1983; P1986; JR1986
DSTs	$p = \frac{1}{2}$	2 points	DFHN2019+
	$p \neq \frac{1}{2}$	?	DF2019+
PATRICIA Tries	$p = \frac{1}{2}$	3 points	Conjectured by KS2002
	$p \neq \frac{1}{2}$	?	?

Profile of Asymmetric PATRICIA Tries

Theorem (Magner & Szpankowski; 2018)

If $(\alpha_1 + \epsilon)\log n \le k \le (\alpha_2 - \epsilon)\log n$, *then*

$$\mu_{n,k} \sim P_1\left(\rho; \log_{p/q} p^k n\right) \frac{p^\rho q^\rho (p^{-\rho} + q^{-\rho})}{\sqrt{2\pi\alpha}\log(p/q)} \cdot \frac{n^v}{\sqrt{\log n}},$$

and

$$\sigma_{n,k}^2 \sim P_2\left(\rho; \log_{p/q} p^k n\right) \frac{p^\rho q^\rho (p^{-\rho} + q^{-\rho})}{\sqrt{2\pi\alpha}\log(p/q)} \cdot \frac{n^v}{\sqrt{\log n}},$$

where $P_1(\rho; x)$ *and* $P_2(\rho; x)$ *are* 1*-periodic functions.*

Moreover,

$$\frac{B_{n,k} - \mu_{n,k}}{\sigma_{n,k}} \xrightarrow{d} N(0,1).$$

Height of PATRICIA tries

By extending the previous study to the boundary.

Height of PATRICIA tries

By extending the previous study to the boundary.

Theorem (Drmota & Magner & Szpankowski; 2019)

With high probability,

$$H_n = \begin{cases} \log_2 n + \sqrt{2\log_2 n} + o(\sqrt{\log n}), & \text{if } p = q; \\ \log_{1/p} n + \frac{1}{2}\log_{p/q}\log n + o(\log\log n), & \text{if } p > q. \end{cases}$$

See the paper:

M. Drmota, A. Magner, W.Szpankowski (2019). Asymmetric Rényi problem, *Combinatorics, Probability and Computing*, **28:4**, 542–573

or the (more detailed) arxiv version of this paper.

Summary and Open Problems

Profile of Random Digital Trees:

Trees	$p = q$?	Mean	Variance	CLT
Tries	$0 < p < 1$	✓	✓	✓
DSTs	$p = \frac{1}{2}$	✓	✓	✓
	$p \neq \frac{1}{2}$	✓	✓	?
PATRICIA Tries	$p = \frac{1}{2}$	?	?	?
	$p \neq \frac{1}{2}$	✓	✓	✓

Summary and Open Problems

Profile of Random Digital Trees:

Trees	$p = q$?	Mean	Variance	CLT
Tries	$0 < p < 1$	✓	✓	✓
DSTs	$p = \frac{1}{2}$	✓	✓	✓
	$p \neq \frac{1}{2}$	✓	✓	?
PATRICIA Tries	$p = \frac{1}{2}$	?	?	?
	$p \neq \frac{1}{2}$	✓	✓	✓

Major Open Tasks:

Summary and Open Problems

Profile of Random Digital Trees:

Trees	$p = q$?	Mean	Variance	CLT
Tries	$0 < p < 1$	✓	✓	✓
DSTs	$p = \frac{1}{2}$	✓	✓	✓
	$p \neq \frac{1}{2}$	✓	✓	?
PATRICIA Tries	$p = \frac{1}{2}$	?	?	?
	$p \neq \frac{1}{2}$	✓	✓	✓

Major Open Tasks:

- profile of symmetric PATRICIA tries;

Summary and Open Problems

Profile of Random Digital Trees:

Trees	$p = q$?	Mean	Variance	CLT
Tries	$0 < p < 1$	✓	✓	✓
DSTs	$p = \frac{1}{2}$	✓	✓	✓
	$p \neq \frac{1}{2}$	✓	✓	?
PATRICIA Tries	$p = \frac{1}{2}$	?	?	?
	$p \neq \frac{1}{2}$	✓	✓	✓

Major Open Tasks:

- profile of symmetric PATRICIA tries;
- refined results for the profile at the boundary of the "central range" for asymmetric PATRICIA tries (very complicated!).