RAIRO-Theor. Inf. Appl. 56 (2022) 2
https://doi.org/10.1051/ita/2022002

RAIRO - Theoretical Informatics and Applications
www.rairo-ita.org

WEAKLY PROTECTED NODES IN RANDOM BINARY SEARCH TREES

Ezzat Mohammad Nezhad[1], Mehri Javanian[2,*] and Ramin Imany Nabiyyi[1]

Abstract. Here, we derive the exact mean and variance of the number of weakly protected nodes (the nodes that are not leaves and at least one of their children is not a leaf) in binary search trees grown from random permutations. Furthermore, by using contraction method, we prove normal limit law for a properly normalized version of this tree parameter.

Mathematics Subject Classification. 05C05, 05C80, 60F05.

Received May 23, 2021. Accepted January 17, 2022.

1. Definitions

Let $P = (p_1, p_2, \ldots, p_n)$ be a uniformly random permutation of $\{1, 2, \ldots, n\}$. A *random binary search tree* is generated by P as follows. The elements of P serve as keys. The keys are stored in the internal nodes of the tree. The root of the tree stores the first key p_1. The second key p_2 is compared with p_1. If $p_2 < p_1$, then p_2 becomes root of the left subtree; otherwise, p_2 becomes root of the right subtree. The process repeats on subsequent keys in the same manner. Note that a uniform probability distribution on permutations does not induce a uniform probability distribution on binary search trees [5]. Figure 1 shows an example of a binary search tree.

In a rooted tree, a *protected* node is a node that is not a leaf and none of its children is a leaf. For many types of random trees, protected nodes have been investigated in numerous papers, see for instance [1–4, 6].

By a weakly protected node, we mean a node that is not a leaf and at least one of its children is not a leaf. Figure 1 illustrates the protected nodes and weakly protected nodes in a binary search tree.

In this note, we study the number of *weakly protected* nodes in random binary search trees. Recently, the number of weakly protected nodes have only been studied for ordered trees in [10].

2. The expectation and variance

We denote the number of weakly protected nodes in a random binary search tree of size n by X_n. We denote the sizes of the left subtree and right subtree of the root by U_n and $n-1-U_n$, respectively. In view of the probability distribution on binary search trees, U_n and $n-1-U_n$ have uniform distribution on the set $\{0, 1, \ldots, n-1\}$.

Keywords and phrases: Random binary search tree, Weakly protected nodes, Zolotarev metric, Contraction method, limiting distribution.

[1] Department of Statistics, Faculty of Mathematical Sciences, University of Tabriz, Tabriz Iran.
[2] Department of Statistics, Faculty of Sciences, University of Zanjan, Zanjan, Iran.

* Corresponding author: javanian@znu.ac.ir

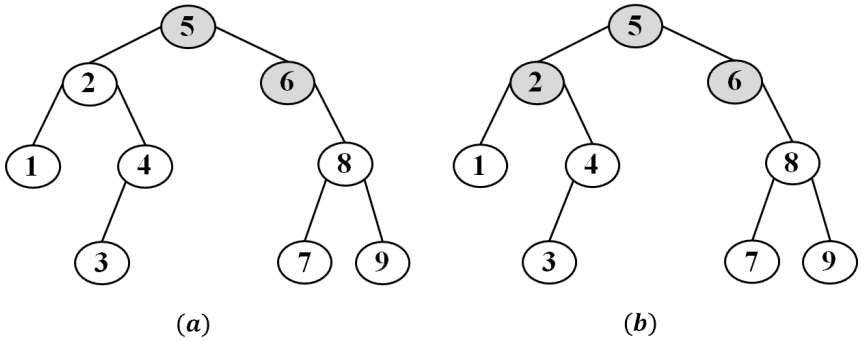

FIGURE 1. A binary search tree built from the keys (5, 2, 6, 1, 8, 7, 4, 9, 3) where gray nodes are: (a) Protected nodes; (b) Weakly protected nodes.

Let the notation $\stackrel{d}{=}$ indicate the equality in distribution. For $n \geq 3$, we have a distributional recurrence for X_n, i.e,

$$X_n \stackrel{d}{=} X_{U_n} + \overline{X}_{n-1-U_n} + 1 - \delta_{n,3}\mathbb{1}_{\{U_n=1\}}, \tag{2.1}$$

where $\mathbb{1}_A$ is the indicator function of A, $\delta_{i,j}$ is Kronecker delta, $X_n \stackrel{d}{=} \overline{X}_n$, and X_n, $\overline{X}_n$ and U_n are independent. Moreover, $X_n = 0$ for $n = 0, 1, 2$.

Theorem 2.1. *Let X_n denote the number of weakly protected nodes in a random binary search tree of size n. Then*

$$\mathbb{E}(X_n) = \frac{7n-8}{15}, \qquad \textit{for } n \geq 4, \tag{2.2}$$

$$\mathrm{Var}(X_n) = \frac{211}{3150}(n+1), \qquad \textit{for } n \geq 10, \tag{2.3}$$

with $\mathbb{E}(X_n) = 0$ for $n = 0, 1, 2$ and $\mathbb{E}(X_3) = \frac{2}{3}$. Moreover,

$$\begin{aligned}
&\mathrm{Var}(X_n) = 0, \quad n = 0, 1, 2, \quad \mathrm{Var}(X_3) = \frac{2}{9},\\
&\mathrm{Var}(X_4) = \frac{2}{9}, \quad \mathrm{Var}(X_5) = \frac{32}{75}, \quad \mathrm{Var}(X_6) = \frac{38}{75},\\
&\mathrm{Var}(X_7) = \frac{286}{525}, \quad \mathrm{Var}(X_8) = \frac{211}{350}, \quad \mathrm{Var}(X_9) = \frac{211}{315}.
\end{aligned}$$

Proof. Taking expectation of (2.1), for $n \geq 3$, we obtain

$$\mathbb{E}(X_n) = \mathbb{E}(X_{U_n}) + \mathbb{E}(\overline{X}_{n-1-U_n}) + 1 - \frac{\delta_{n,3}}{n}. \tag{2.4}$$

By conditioning on U_n, the equation (2.4) gives

$$n\mathbb{E}(X_n) = 2\sum_{j=0}^{n-1} \mathbb{E}(X_j) + n - \delta_{n,3}. \tag{2.5}$$

We subtract from the equation (2.5) a version of itself with n replaced by $n-1$ and unwind the recurrence: for $n \geq 4$,

$$\begin{aligned}
\mathbb{E}(X_n) &= \frac{n+1}{n}\mathbb{E}(X_{n-1}) + \frac{1+\delta_{n-1,3}}{n} \\
&\vdots \\
&= \frac{n+1}{4}\mathbb{E}(X_3) + \sum_{j=3}^{n-1} \frac{(n+1)(1+\delta_{j,3})}{(j+1)(j+2)} \\
&= \frac{4(n+1)}{15} - \frac{1}{n+2} + \sum_{j=5}^{n+1} \frac{n+1}{j} - \sum_{j=5}^{n+1} \frac{n+1}{j+1}.
\end{aligned}$$

By simplifying this we get (2.2). Similarly, by (2.1) and (2.2), we have, for $n \geq 9$,

$$\begin{aligned}
\mathbb{E}(X_n^2) &= \frac{2}{n}\sum_{j=0}^{n-1}\mathbb{E}(X_j^2) + \frac{4}{n}\sum_{j=4}^{n-1}\mathbb{E}(X_j) + \frac{4}{n}\mathbb{E}(X_3) + 1 \\
&\quad + \frac{2}{n}\sum_{j=4}^{n-5}\mathbb{E}(X_j)\mathbb{E}(X_{n-1-j}) + \frac{4}{n}\mathbb{E}(X_3)\mathbb{E}(X_{n-4}) \\
&= \frac{2}{n}\sum_{j=0}^{n-1}\mathbb{E}(X_j^2) + \frac{4}{n}\sum_{j=4}^{n-1}\frac{7j-8}{15} + \frac{8}{3n} + 1 \\
&\quad + \frac{2}{n}\sum_{j=4}^{n-5}\frac{7j-8}{15}\cdot\frac{7(n-1-j)-8}{15} + \frac{8}{3n}\cdot\frac{7(n-4)-8}{15} \\
&= \frac{2}{n}\sum_{j=0}^{n-1}\mathbb{E}(X_j^2) + \frac{49}{675}n^2 + \frac{49}{225}n - \frac{577}{675} + \frac{112}{45n}.
\end{aligned}$$

By the last equation, for $n \geq 10$, it follows that

$$\begin{aligned}
&n\mathbb{E}(X_n^2) - (n-1)\mathbb{E}(X_{n-1}^2) \\
&\qquad = 2\mathbb{E}(X_{n-1}^2) + \frac{49}{225}n^2 + \frac{49}{225}n - 1.
\end{aligned}$$

Therefore, for $n \geq 10$, we obtain

$$\begin{aligned}
\mathbb{E}(X_n^2) &= \frac{n+1}{n}\mathbb{E}(X_{n-1}^2) + \frac{1}{n}\Big(\frac{49}{225}n^2 + \frac{49}{225}n - 1\Big) \\
&\vdots \\
&= \frac{n+1}{10}\mathbb{E}(X_9^2) + \sum_{j=10}^{n}\frac{n+1}{j(j+1)}\Big(\frac{49}{225}j^2 + \frac{49}{225}j - 1\Big) \\
&= \frac{49}{225}n^2 - \frac{1357}{3150}n + \frac{123}{350}
\end{aligned} \tag{2.6}$$

with $\mathbb{E}(X_3^2) = \frac{2}{3}$, $\mathbb{E}(X_4^2) = 2$, $\mathbb{E}(X_5^2) = \frac{11}{3}$, $\mathbb{E}(X_6^2) = \frac{254}{45}$, $\mathbb{E}(X_7^2) = \frac{505}{63}$, $\mathbb{E}(X_8^2) = \frac{759}{70}$ and $\mathbb{E}(X_9^2) = \frac{494}{35}$.

Finally, applying (2.6) and $\mathrm{Var}(X_n) = \mathbb{E}(X_n^2) - \big(\mathbb{E}(X_n)\big)^2$, the assertion in (2.3) follows. □

3. Limiting distribution

In this section, we begin to prove the normality of limiting distribution of X_n. The proof was completed by applying the contraction method, which was first introduced by [9], in studying the Quicksort algorithm.

Here, we prefer the Zolotarev metric ζ_3 (see [8]) as the metric space applied in the contraction method. Let the distribution of a random variable X denoted by $\mathcal{L}(X)$. Then, for any given random variables X and Y, the 3rd order Zolotarev metric between X and Y is defined as

$$\begin{aligned}\zeta_3(X,Y) &= \zeta_3(\mathcal{L}(X),\mathcal{L}(Y))\\ &:= \sup\{|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| : f \in \mathcal{F}\}\end{aligned}$$

where $\mathcal{F} = \{f : f \in C^{(2)}, |f^{(2)}(x) - f^{(2)}(y)| \le |x-y|\}$ denotes the set of all twice differentiable functions, where the second derivative is Lipschitz continuous with Lipschitz constant equal to 1.

The following lemma gives several properties of probability metric $\zeta_3(X,Y)$, which are quite useful in our proof of Theorem 3.5.

Lemma 3.1 (see [8]). *Let $\zeta_3(X,Y)$ be the 3rd order Zolotarev metric between the random variables X and Y. Then,*

(i) For any real number $c > 0$,

$$\zeta_3(cX, cY) = c^3\zeta_3(X,Y); \tag{3.1}$$

(ii) If the random variables Y and (X_1, X_2) are independent mutually, then

$$\zeta_3(X_1 + Y, X_2 + Y) \le \zeta_3(X_1, X_2); \tag{3.2}$$

(iii) For any random variables X and Y,

$$\begin{aligned}&\mathbb{E}(|X|^3) + \mathbb{E}(|Y|^3) < \infty, \quad \mathbb{E}(X^k) = \mathbb{E}(Y^k),\ k = 1,2\\ &\qquad\qquad \Longleftrightarrow \zeta_3(X,Y) < \infty\end{aligned} \tag{3.3}$$

(iv) For the random variables V and $\{V_n\}_{n\ge1}$, as $n \to \infty$

$$\zeta_3(V_n, V) \to 0 \implies V_n \xrightarrow{\mathcal{D}} V, \tag{3.4}$$

where the notation $\xrightarrow{\mathcal{D}}$ denotes the convergence in distribution.

Lemma 3.2 (see [7]). *let X_1, X_2, T_1 and T_2 be random variables such that the pairs $(X_1 + T_1, X_2 + T_2)$ and (X_1, X_2) satisfies (3.3). Then*

$$\begin{aligned}&\zeta_3(X_1 + T_1, X_2 + T_2) \le \zeta_3(X_1, X_2)\\ &+ \sum_{i=1}^{2}\Big\{\frac{\|X_i\|_3^2\|T_i\|_3}{2} + \frac{\|X_i\|_3\|T_i\|_3^2}{2} + \frac{\|T_i\|_3^3}{6}\Big\},\end{aligned}$$

where $\|X\|_3 := \mathbb{E}(|X|^3)^{1/3}$ for a random variable X.

Moreover, the proof of Theorem 3.5 requires the following upper bound for metric ζ_3:

$$\zeta_3(X,Y) \le \frac{1}{2}\big(\|X\|_3^2 + \|X\|_3\|Y\|_3 + \|Y\|_3^2\big)\ell_3(X,Y), \tag{3.5}$$

where the minimal L_3-metric ℓ_3 defined by

$$\begin{aligned}\ell_3(X,Y) &:= \ell_3(\mathcal{L}(X),\mathcal{L}(Y))\\ &:= \inf\{\|X'-Y'\|_3 : X \overset{d}{=} X',\, Y \overset{d}{=} Y'\},\end{aligned}$$

for random variables X and Y with $\|X\|_3 < \infty$, $\|Y\|_3 < \infty$.

We standardize X_n with its mean and variance, *i.e.*,

$$Y_n := \frac{X_n - \mathbb{E}(X_n)}{\sigma(n)}, \qquad \sigma^2(n) := \mathrm{Var}(X_n)$$

Let denote a quantity

$$Y_n \overset{d}{=} \frac{\sigma(U_n)}{\sigma(n)}\, Y_{U_n} + \frac{\sigma(n-1-U_n)}{\sigma(n)}\, \overline{Y}_{n-1-U_n}, \qquad n \ge 4.$$

where $Y_i \overset{d}{=} \overline{Y}_i$, for $0 \le i \le n-1$. The random variables Y_i, $\overline{Y}_i$, U_n, $0 \le i \le n-1$, are independent. To prove the Theorem 3.5, we still require some more arrangements. The following three lemmas are necessary.

Lemma 3.3. *Let W, W_1 and W_2 be independent standard normal random variables. Then we have*

$$W \overset{d}{=} \sqrt{\frac{U_n+1}{n+1}}\, W_1 + \sqrt{\frac{n-U_n}{n+1}}\, W_2, \tag{3.6}$$

where U_n is a random variable with uniform distribution on the set $\{0,1,\ldots,n-1\}$.

Proof. It is sufficient to verify that the characteristic function of the right side of (3.6) is the same as that of a standard normal random variable. From the independence of the random variables W, W_1, W_2, U_n, we have

$$\begin{aligned}&\mathbb{E}\Big[\exp\Big\{it\Big(\sqrt{\frac{U_n+1}{n+1}}\, W_1 + \sqrt{\frac{n-U_n}{n+1}}\, W_2\Big)\Big\}\Big]\\ &= \sum_{j=0}^{n-1} \frac{1}{n}\mathbb{E}\Big[\exp\Big\{it\Big(\sqrt{\frac{j+1}{n+1}}\, W_1 + \sqrt{\frac{n-j}{n+1}}\, W_2\Big)\Big\}\Big]\\ &= \sum_{j=0}^{n-1} \frac{1}{n}\mathbb{E}\Big[\exp\Big\{itW_1\sqrt{\frac{j+1}{n+1}}\Big\}\Big]\mathbb{E}\Big[\exp\Big\{itW_2\sqrt{\frac{n-j}{n+1}}\Big\}\Big]\\ &= \sum_{j=0}^{n-1} \frac{1}{n}\exp\Big(-\frac{j+1}{n+1}\cdot\frac{t^2}{2}\Big)\exp\Big(-\frac{n-j}{n+1}\cdot\frac{t^2}{2}\Big) = e^{-\frac{t^2}{2}}\end{aligned}$$

where, the function $e^{-\frac{t^2}{2}}$ is the characteristic function of a standard normal random variable. Hence we obtain the claim. □

Lemma 3.4. *As $n \to \infty$, $\mathbb{E}[|Y_n^3|] = \mathcal{O}(1)$.*

Proof. By Lyapunov's inequality, $\mathbb{E}[|Y_n|] \le \sqrt{\mathbb{E}[|Y_n|^2]} = 1$. Let $\xi_n := 1 \vee \max_{0\le j\le n} \mathbb{E}[|Y_j^3|]$ and U be a uniform random variable on $(0,1)$. Then from (2.1) we obtain

$$\begin{aligned}\mathbb{E}[|Y_n^3|] &\le 2\sum_{j=0}^{n-1} \frac{1}{n}\Big(\frac{\sigma(j)}{\sigma(n)}\Big)^3 \mathbb{E}\big[\big|Y_j\big|^3\big] + \mathcal{O}(1)\\ &\le \Big(2\mathbb{E}\big(U^{\frac{3}{2}}\big) + o(1)\Big)\xi_{n-1} + \mathcal{O}(1)\\ &\le \big(0.8 + o(1)\big)\xi_{n-1} + \mathcal{O}(1).\end{aligned} \tag{3.7}$$

Hence, there exist an $n_0 \in \mathbb{N}$ and a constant $0 < \alpha < \infty$ such that for $n \ge n_0$

$$\mathbb{E}[|Y_n^3|] \le 0.9\xi_{n-1} + \alpha.$$

By induction, we have $\mathbb{E}[|Y_n^3|] \le \xi_{n_0} \vee (10\alpha)$ for all $n \ge 0$. This implies the claim. □

In the following, we begin to prove the asymptotic normality distribution for X_n.

Theorem 3.5. *Let X_n denote the number of weakly protected nodes in a random binary search tree of size n. Then,*

$$\frac{X_n - \frac{7}{15}n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{211}{3150}\Big), \qquad \text{as } n \to \infty,$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal random variable with mean μ and variance σ^2.

Proof. By (3.4), we just need to show that the Zolotarev metric between the random variables Y_n and N, a standard normal random variable, approaches 0, as $n \to \infty$. For W_1 and W_2 independent standard normal random variables, also independent of U_n, we set

$$\Theta_n := \frac{\sigma(U_n)}{\sigma(n)} W_1 + \frac{\sigma(n-1-U_n)}{\sigma(n)} W_2, \qquad n \ge 4.$$

Note that $\mathrm{Var}(\Theta_n) > 0$ for all $n \ge n_0$, and $\mathrm{Var}(\Theta_n) \to 1$ as $n \to \infty$. Hence there exists a deterministic sequence $(\delta_n)_{n\ge n_0}$ with $\delta_n \to 0$ as $n \to \infty$ such that $\mathrm{Var}\big((1+\delta_n)\Theta_n\big) = 1$ for all $n \ge 4$. So, by Lemma 3.4, each pair from the random variables Y_n, $(1+\delta_n)\Theta_n$ and N satisfies (3.3). Thus we obtain

$$\zeta_3(Y_n, N) \le \zeta_3\big(Y_n, (1+\delta_n)\Theta_n\big) + \zeta_3\big((1+\delta_n)\Theta_n, N\big).$$

Now Lemma 3.2 yields

$$\zeta_3\big(Y_n, (1+\delta_n)\Theta_n\big) \le \zeta_3\big(Y_n, \Theta_n\big) + o(1).$$

Using the bound (3.5) and Lemma 3.3, for some finite constant $M > 0$, we obtain

$$\begin{aligned}\zeta_3\big((1+\delta_n)\Theta_n, N\big) &\le M\ell_3\big((1+\delta_n)\Theta_n, N\big)\\ &\le M\Bigg\|\Bigg((1+\delta_n)\frac{\sigma(U_n)}{\sigma(n)} - \sqrt{\frac{U_n+1}{n+1}}\Bigg)W_1\\ &\quad + \Bigg((1+\delta_n)\frac{\sigma(n-1-U_n)}{\sigma(n)} - \sqrt{\frac{n-U_n}{n}}\Bigg)W_2\Bigg\|_3 \to 0.\end{aligned}$$

From (3.1) and (3.2), we can conclude that

$$\begin{aligned}
\zeta_3(Y_n,N) &\le \big(Y_n,\Theta_n\big)+o(1)\\
&\le \zeta_3\Big(\frac{\sigma(U_n)}{\sigma(n)}\,Y_{U_n}+\frac{\sigma(n-1-U_n)}{\sigma(n)}\,\overline{Y}_{n-1-U_n},\\
&\qquad \frac{\sigma(U_n)}{\sigma(n)}W_1+\frac{\sigma(n-1-U_n)}{\sigma(n)}W_2\Big)+o(1)\\
&\le \sum_{j=0}^{n-1}\frac{1}{n}\zeta_3\Big(\frac{\sigma(j)}{\sigma(n)}\,Y_j+\frac{\sigma(n-1-j)}{\sigma(n)}\,\overline{Y}_{n-1-j},\\
&\qquad \frac{\sigma(j)}{\sigma(n)}W_1+\frac{\sigma(n-1-j)}{\sigma(n)}W_2\Big)+o(1)\\
&= 2\sum_{j=0}^{n-1}\frac{1}{n}\Big(\frac{\sigma(j)}{\sigma(n)}\Big)^3\zeta_3\big(Y_j,N\big)+o(1)\\
&= 2\mathbb{E}\Big[\Big(\frac{\sigma(U_n)}{\sigma(n)}\Big)^3\zeta_3\big(Y_{U_n},N\big)\Big]+o(1) \qquad (3.8)\\
&\le \Big(2\mathbb{E}\big(U^{\frac{3}{2}}\big)+o(1)\Big)\sup_{0\le j\le n-1}\zeta_3(Y_j,N)+o(1).
\end{aligned}$$

This implies, similarly to the inequality (3.7), $\big(\zeta_3(Y_n,N)\big)_{n\ge0}$ that is bounded. We denote $\xi:=\sup_{n\ge0}\zeta_3(Y_n,N)$ and $s:=\limsup_{n\to\infty}\zeta_3(Y_n,N)\ge0$. For any $\varepsilon>0$ there exists an $n_1\ge4$ such that $\zeta_3(Y_n,N)\le s+\varepsilon$ for all $n\ge n_1$. Hence, from (3.8) we obtain

$$\begin{aligned}
\zeta_3(Y_n,N) &\le 2\mathbb{E}\Big[\mathbb{1}_{\{U_n\le n_1\}}\Big(\frac{\sigma(U_n)}{\sigma(n)}\Big)^3\Big]\,\xi\\
&\quad+2\mathbb{E}\Big[\mathbb{1}_{\{U_n>n_1\}}\Big(\frac{\sigma(U_n)}{\sigma(n)}\Big)^3\Big]\,(s+\varepsilon)+o(1)\\
&\sim 2\mathbb{E}\Big[\mathbb{1}_{\{U_n>n_1\}}\Big(\frac{\sigma(U_n)}{\sigma(n)}\Big)^3\Big]\,(s+\varepsilon)+o(1).
\end{aligned}$$

So $0\le s=\limsup_{n\to\infty}\zeta_3(Y_n,N)\le0.8(s+\varepsilon)<s+\varepsilon$. Since $\varepsilon>0$ is arbitrary then we have $s=0$. Therefore, by (3.4), the assertion holds. □

References

[1] M. Bona, *k*-protected nodes in binary search trees. *Adv. Appl. Math.* **53** (2014) 1–11.

[2] L. Devroye and S. Janson, Protected nodes and fringe subtrees in some random trees. *Electr. Commun. Probab.* **19** (2014) 1–10.

[3] C. Holmgren and S. Janson, Asymptotic distribution of two-protected nodes in ternary search trees. *Electr. J. Probab.* **20** (2015) 1–20.

[4] C. Holmgren and S. Janson, Limit laws for functions of fringe trees for binary search trees and recursive trees. *Electr. J. Probab.* **20** (2015) 1–51.

[5] H.M. Mahmoud, Evolution of random search trees. John Wiley & Sons Inc., New York (1992).

[6] H.M. Mahmoud and M.D. Ward, Asymptotic distribution of two-protected nodes in random binary search trees. *Appl. Math. Lett.* **25** (2012) 2218–2222.

[7] R. Neininger, Refined quicksort asymptotics. *Random Struct. Algor.* **46** (2015) 346–361.

[8] S. Rachev, Probability Metrics and the Stability of Stochastic Models. New York (1991).

[9] U. Roesler, A limit theorem for "Quicksort". *RAIRO: ITA* **25** (1991) 85–100.
[10] L. Yang and S.L. Yang, Weakly protected points in ordered trees. *Graphs Combinat.* (2021) https://doi.org/10.1007/s00373-021-02278-w.