Labels distance in bucket recursive trees with variable capacities of buckets

S. Naderi
Department of Statistics,
Science and Research Branch,
Islamic Azad University, Tehran, Iran
email: `snaderi@gmail.com`

R. Kazemi
Department of Statistics,
Imam Khomeini International
University, Qazvin, Iran
email: `r.kazemi@sci.ikiu.ac.ir`

M. H. Behzadi
Department of Statistics,
Science and Research Branch,
Islamic Azad University, Tehran, Iran
email: `behzadi@srbiau.ac.ir`

Abstract. The bucket recursive tree is a natural multivariate structure. In this paper, we apply a trivariate generating function approach for studying of the depth and distance quantities in this tree model with variable bucket capacities and give a closed formula for the probability distribution, the expectation and the variance. We show as $j \to \infty$, limiting distributions are Gaussian. The results are obtained by presenting partial differential equations for moment generating functions and solving them.

1 Introduction

Trees are defined as connected graphs without cycles, and their properties are basics of graph theory. For example, a connected graph is a tree, if and only if the number of edges equals the number of nodes minus 1 [5]. Furthermore,

each pair of nodes is connected by a unique path. A rooted tree is a tree with a countable number of nodes, in which a particular node is distinguished from the others and called the root node. A recursive tree with n nodes is an unordered rooted tree, where the nodes are labelled by distinct integers from $\{1, 2, 3, ..., n\}$ in such a way that the sequence of labels lying on the unique path from the root node to any node in the tree are always forming an increasing sequence. Call a random recursive tree a tree chosen uniformly at random from the $(n-1)!$ possible recursive trees on n nodes. A random recursive tree can also be constructed as follows. The node 1 is distinguished as the root. We imagine the nodes arriving one by one. For $k \geq 2$, node k attaches itself to a node chosen uniformly at random from $1, 2, ..., k-1$ (for more information and applications, see [11]) .

Mahmoud and Smythe introduced bucket recursive trees as a generalization of random recursive trees [10]. In this model the bucket is a node that can hold up to $b \geq 1$ labels. The capacity of a bucket v $(c = c(v))$ is defined by the number of its labels. They applied a probabilistic analysis for studying the height and depth of the largest label in these trees. Kuba and Panholzer analyzed these trees as a special instance of bucket increasing trees which is a family of some combinatorial objects [8]. They obtained exact and limiting distribution results for the parameters depth of a specified label, descendants of a specified label and degree of a specified label. A (probabilistic) description of random bucket recursive trees is given by a generalization of the stochastic growth rule for ordinary random recursive trees (which are the special instance $b = 1$), where a tree grows by progressive attraction of increasing integer labels: when inserting element $n+1$ into an existing bucket recursive tree containing n elements (i.e., containing the labels $\{1, 2, ..., n\}$) all n existing elements in the tree compete to attract the element $n + 1$, where all existing elements have equal chance to recruit the new element. If the element winning this competition is contained in a node with less than b elements (an unsaturated bucket or node), element $n + 1$ is added to this node, otherwise if the winning element is contained in a node with already b elements (a saturated bucket or node), element $n+1$ is attached to this node as a new bucket containing only the element $n+1$. Starting with a single bucket as root node containing only element 1 leads after $n - 1$ insertion steps, where the labels $2, 3, ..., n$ are successively inserted according to this growth rule, to a so called random bucket recursive tree with n elements and maximal bucket-size b. In this paper we consider a model of bucket trees where the nodes are buckets with variable capacities labelled with integers $1, 2, \cdots, n$ (not the same capacities as bucket recursive trees).

Definition 1 [6] *A size-n bucket recursive tree T_n with variable bucket capacities and maximal bucket size b starts with the root labelled by 1. The tree grows by progressive attraction of increasing integer labels: when inserting label $j + 1$ into an existing bucket recursive tree T_j, except the labels in the non-leaf buckets with capacity $< b$ all labels in the tree (containing label 1) compete to attract the label $j+1$. For the root node and buckets with capacity b, we always produce a new bucket $j + 1$. But for a leaf with capacity $c < b$, either the label $j + 1$ is attached to this leaf as a new bucket containing only the label $j + 1$ or is added to that leaf and make a bucket with capacity $c + 1$. This process ends with inserting the label n (i.e., the largest label) in the tree.*

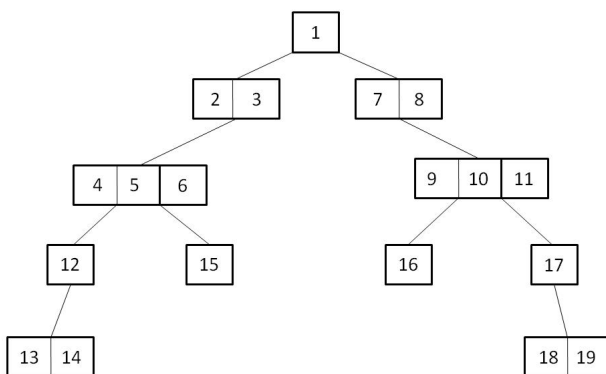Figure 1 illustrates such a tree of size 19 with $b = 3$.

Figure 1: A bucket recursive tree with variable capacities of buckets with 19 elements and $b = 3$.

Bucket recursive trees with variable capacities of buckets are appeared in chemistry, social science, in some computer science applications and furthermore. They are appeared as a model for the spread of epidemics, for pyramid schemes, for the family trees of preserved copies of ancient texts. In the family trees, suppose males with the same ethical traits come together in each generation. Suppose up to 3 people are matched with the same attributes. Then a bucket recursive trees with variable capacities of buckets with maximal bucket size 3 is formed. In this case, and in a genealogy of n people, the distance between two specific individuals is the quantity examined in this article. For another example, if n atoms in a branching molecular structure are stochastically labelled with integers 1, 2, ..., n, then atoms in different functional groups can be considered as the labels of different buckets of a bucket recursive tree

(the size of the largest functional group is b).

In passing, we give the combinatorial description of our model. Let $d(v)$ be the out-degree of node v. It will be convenient to define for trees the size $|T|$ of a tree T via $|T| = \sum_v c(v)$. An increasing labelling of an ordered tree T is then a labelling of T, where the labels $\{1, 2, ..., |T|\}$ are distributed amongst the nodes of T. Then a class $\mathcal{T}$ of a new family of *bucket-increasing trees* can be defined in the following way: A sequence of non-negative numbers $(\alpha_k)_{k \geq 0}$ with $\alpha_0 > 0$ and a sequence of non-negative numbers $\beta_1, \beta_2, ..., \beta_{b-1}$ is used to define the weight $w(T)$ of any ordered tree T by $w(T) := \Pi_v w(v)$, where v ranges over all nodes of T. It is natural that $w(v)$ must be dependent on $c(v)$ and $d(v)$. Thus the weight $w(v)$ of a node v is given as follows:

$$w(v) := \begin{cases} \alpha_{d(v)}, & v \text{ is root or complete } (c(v) = b) \\ \beta_{c(v)}, & v \text{ is incomplete } (c(v) < b). \end{cases} \qquad (1)$$

The above definition is reasonable because the root is the only incomplete node that has outdegree ≥ 1. Thus for complete nodes and root, the weight is dependent on the out-degree and described by the sequence α_k, whereas for incomplete nodes except of root the weights are dependent on the capacities.

Furthermore, $\mathcal{L}(\mathcal{T})$ denotes the set of different increasing labelings of the tree T with distinct integers $\{1, 2, ..., |T|\}$, where $L(T) := |\mathcal{L}(\mathcal{T})|$ denotes its cardinality. Then the family $\mathcal{T}$ consists of all trees T together with their weights $w(T)$ and the set of increasing labelings $\mathcal{L}(T)$. For a given degree-weight sequence $(\alpha_k)_{k \geq 0}$ with a degree-weight generating function $\varphi(t) := \sum_{k \geq 0} \alpha_k t^k$ and a bucket-weight sequence $\beta_1, \beta_2, ..., \beta_{b-1}$, we define the exponential generating function

$$T_{r, k_1, ... k_r}(z) \quad := \quad \sum_{n=1}^{\infty} T_{n, b, r, k_1, ... k_r} \frac{z^n}{n!}, \qquad (2)$$

where $T_{n, b, r, k_1, ... k_r} := \sum_{|T|=n} w(T) \cdot L(T)$ is the total weights. For this model,

$$T_{n, b, r, k_1, ... k_r} = \frac{(n-1)!(b!)^{n(1 - \sum_{i=1}^{r} |\mathcal{P}_{k_i}|)}}{b}, \quad n \geq 1$$

$$\varphi(T_{r, k_1, ... k_r}(z)) = \frac{(b-1)!}{1 - b!^{1 - \sum_{i=1}^{r} |\mathcal{P}_{k_i}|} z}, \qquad (3)$$

where $\mathcal{P}_{k_i}$ is the set of all trees of size k_i and r is the degree of root node [6]. For simplicity, we set $T_{n,b} := T_{n, b, r, k_1, ..., k_r}$ and $T(z) := T_{r, k_1, ..., k_r}(z)$.

Various studies are devoted to a distributional analysis of distances between random nodes in a lot of tree families of interest. For example, Mahmoud and

Neininger [9] for binary search trees, Christophi and Mahmoud [1] for the digital data structure, and Panholzer [13] for simply generated trees. Fewer studies are made to reveal the distribution of distances between specified nodes in labelled tree structures. Dobrow [3] and Dobrow and Smythe [4] have shown a central limit theorem for the distance between the nodes labelled by j and n, respectively, in a random recursive tree of size n and Devroye and Neininger [2] have shown a central limit theorem for the distance between the nodes labelled by j_1 and j_2 in a random binary search tree of size n. Panholzer and Prodinger have studied the level of nodes in increasing trees [14]. Kuba and Panholzer have studied the distribution of distances between specified nodes in increasing trees [7]. Also Moon studied the distance between nodes in recursive trees [12].

If we denote by $D_{n,n}$ the random variable which measures the depth of node containing label n in the our tree model of size n, then it was shown in [6] that $D_{n,n}$ satisfies a central limit law with mean and variance $b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} \log n$. More precisely,

$$P(D_{n,n} = m) = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} \frac{S(n-1, m)}{(n-1)!},$$
$$\mathbb{E}(D_{n,n}) = \mathbb{V}ar(D_{n,n}) = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} \log n + \mathcal{O}(1), \qquad (4)$$

where $S(m_1, m_2)$ are the signless Stirling numbers of first kind. We study the random variable level of label j, i.e., the number of edges from the root node to the bucket containing label j denoted by $D_{n,j}$ in tree T of size $n \geq j$. In this paper we extend the above results for $D_{n,n}$ to $D_{n,j}$. In passing, we study the random variable $H_{n,j}$, which counts the distance, measured by the number of edges lying on the connecting path, between bucket containing label j and bucket containing label n. Finally, we extend our results to the random variable $H_{n,i,j}$ which counts the distance between the bucket containing label i and bucket containing label j in our random tree of size n.

2 The depth of label j

We can to sketch a combinatorial approach to obtain the differential equation on the trivariate generating function related to the level of an arbitrary label j. It is better to think of specifically tricolored trees, where the coloring is as follows: one bucket is colored white (containing label j), all buckets with smaller labels than the all labels in white bucket are colored black, and all buckets with larger labels than the white bucket are colored red. We are interested in

the level of the white bucket. Assume that the out-degree of the root node is $r \geq 1$ and the white bucket of T is not the root node (the case that the white bucket is the root of the tree corresponds to the initial condition, but does not appear explicitly in the differential equation itself). Then the white bucket is located in one of the r subtrees of the root of T; let us assume that it is in the first subtree. After order preserving relabellings, each subtree $\mathsf{T}_1, ..., \mathsf{T}_r$ is a bucket recursive tree with variable capacities of buckets by itself. The first subtree is again a tricolored increasing tree with one white, j_1 black and k_1 red buckets, whereas the remaining $r-1$ subtrees are only bicolored. For a proper description of this combinatorial decomposition we use generating functions which are exponential in both variables z and u, where z marks the black buckets and u marks the red buckets, i.e.,

$$\sum_{j \geq 0} \sum_{k \geq 0} f_{j,k} \frac{z^j}{j!} \frac{u^k}{k!}$$

for sequences $f_{j,k}$ and

$$\sum_{j \geq 0} \sum_{k \geq 0} \sum_{m \geq 0} f_{j,k,m} \frac{z^j}{j!} \frac{u^k}{k!} v^m$$

for sequences $f_{j,k,m}$, where v marks the level of the white bucket. Set $f_{j,k} = T_{k+j,b}$ and $f_{j,k,m} = P(D_{k+j+1,j+1} = m) T_{k+j+1,b}$. Thus the $r-1$ bicolored trees and the tricolored tree lead to

$$\alpha_1^{-\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} T_{n,b}(z+u)^{r-1} L(z,u,v), \tag{5}$$

just similar to [6] where

$$L(z,u,v) = \sum_{k \geq 0} \sum_{j \geq 0} \sum_{m \geq 0} P(D_{k+j+1,j+1} = m) T_{k+j+1,b} \frac{z^j}{j!} \frac{u^k}{k!} v^m.$$

We recall that the total weights of the r subtrees is

$$\alpha_1^{-\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} T_{k_1,b} \cdots T_{k_r,b}.$$

The level of the white bucket in the tree is one more than the level of the white bucket in the subtree. This fact leads to a factor v. We additionally get a factor r, since the white bucket can be in the first, second, ..., r-th subtree. Furthermore, the root has out-degree r that leads to a factor α_r. Thus by summing over $r \geq 1$, (5) leads to

$$\alpha_1^{-\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} v \varphi'(T_{n,b}(z+u)) L(z,u,v).$$

Since the root node labelled by 1 is colored black,

$$\frac{\partial}{\partial z}L(z,u,v) = \alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}v\varphi'(T_{n,b}(z+u))L(z,u,v). \tag{6}$$

Equation (6) has the general solution

$$L(z,u,v) = c(u,v)\exp\left\{\alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}v\int_0^z \varphi'(T_{n,b}(t+u))dt\right\},$$

with a function $c(u,v)$. Evaluating at $z=0$ and adapting to the initial condition gives now $c(u,v) = L(0,u,v) = T'_{n,b}(u) = \alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\varphi(T_{n,b}(u))$. Thus

$$
\begin{aligned}
L(z,u,v) &= \alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\varphi(T_{n,b}(u))\exp\left\{\alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}v\int_0^z \varphi'(T_{n,b}(t+u))dt\right\}\\
&= \alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\varphi(T_{n,b}(u))\\
&\quad \times \exp\left\{\alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}v\int_0^z \frac{\varphi'(T_{n,b}(t+t))T'_{n,b}(t+u)}{\alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\varphi(T_{n,b}(t+u))}dt\right\}\\
&= \alpha_1^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\varphi(T_{n,b}(u))\left(\frac{\varphi(T_{n,b}(z+u))}{\varphi(T_{n,b}(u))}\right)^v\\
&= T'_{n,b}(u)\left(\frac{T'_{n,b}(z+u)}{T'_{n,b}(u)}\right)^v. \tag{7}
\end{aligned}
$$

In the next results we use from the following facts [5]:

$$[z^n]f(qz) = q^n[z^n]f(z), \tag{8}$$

$$\sum_{n\geq0}\sum_{m=0}^{n}S(n,m)\frac{z^n}{n!}v^m = \frac{1}{(1-z)^v}, \tag{9}$$

$$[z^n]\log\left(\frac{1}{1-z}\right)(1-z)^{-1} = H_n, \tag{10}$$

$$[z^n]\log^2\left(\frac{1}{1-z}\right)(1-z)^{-1} = H_n^2 - H_n^{(2)}, \tag{11}$$

where H_n, the n-th harmonic number and $H_n^{(2)}$ is the n-th harmonic number of order 2. In the following lemma we see that distribution of $D_{n,j}$ is independent of n.

Lemma 1 *The probabilities* $P(D_{n,j} = m)$ *are given by the following formula:*

$$P(D_{n,j} = m) = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} \frac{S(j-1, m)}{(j-1)!}, \qquad j \le n \tag{12}$$

and

$$
\begin{aligned}
\mathbb{E}(D_{n,j}) &= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} H_{j-1}, \\
\mathbb{V}ar(D_{n,j}) &= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} H_{j-1}^2 (1 - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}) \\
&+ b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} (H_{j-1} - H_{j-1}^{(2)}).
\end{aligned}
\tag{13}
$$

Proof. By (3), (7) and (10),

$$
\begin{aligned}
\mathbb{E}(D_{n,j}) &= \frac{(j-1)!(n-j-1)!}{T_{n,b}} [z^{j-1} u^{n-j-1}] \frac{\partial L(z, u, v)}{\partial v} \Big|_{v=1} \\
&= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} H_{j-1}
\end{aligned}
$$

and by (11),

$$
\begin{aligned}
\mathbb{E}(D_{n,j}(D_{n,j} - 1)) &= \frac{(j-1)!(n-j-1)!}{T_{n,b}} [z^{j-1} u^{n-j-1}] \frac{\partial^2 L(z, u, v)}{\partial v^2} \Big|_{v=1} \\
&= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} (H_{j-1}^2 - H_{j-1}^{(2)}).
\end{aligned}
$$

Proof of (13) is completed, since $\mathbb{V}ar(D_{n,j}) = \mathbb{E}(D_{n,j}(D_{n,j} - 1)) + \mathbb{E}(D_{n,j}) - \mathbb{E}^2(D_{n,j})$. By (8), the probability generating function

$$
\begin{aligned}
p(v) &= \sum_{m \ge 0} P(D_{k+j,j} = m) v^m = \frac{(j-1)! k!}{T_{k+j,b}} [z^{j-1} u^k] L(z, u, v) \\
&= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} \left\{ \frac{(j-1)! k!}{(k+j-1)!} \binom{v+j-2}{j-1} \binom{k+j-1}{k} \right\} \\
&= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} \binom{v+j-2}{j-1}.
\end{aligned}
$$

Thus one gets (12). Therefore the probability generating function and thus the distribution of $D_{k+j,j}$ is independent of k. $\square$

As an example, in the a tree family with 100 individuals, the probability that the distance from the ancestor to the 45th individual is equal to 10 is calculated from relation (12) under specific conditions. Also, if the ancestor has 5 children, then the average distance to the 45th individual is

$$\mathbb{E}(D_{100,45}) = H_{44} = \frac{5884182435213075787}{1345655451257488800}.$$

Corollary 1 *For* $j = n$,

$$
\begin{aligned}
\mathbb{E}(D_{n,n}) &= b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|} H_{n-1}, \\
\mathbb{V}ar(D_{n,n}) &= b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|} H_{n-1}^2 (1 - b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}) \\
&+ b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|} (H_{n-1} - H_{n-1}^{(2)}).
\end{aligned}
$$

Theorem 1 *As* $j \to \infty$,

$$
\mathbb{E}(D_{n,j}) = \mathbb{V}ar(D_{n,j}) = b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|} \log j + \mathcal{O}(1), \qquad j \le n
$$

and

$$
\sup_{x \in \mathbb{R}} \left| P \left\{ \frac{D_{n,j} - b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|} \log j}{\sqrt{b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|} \log j}} \le x \right\} - \Phi(x) \right| = \mathcal{O}\left(\frac{1}{\sqrt{\log j}} \right).
$$

Proof. This is a direct application of the quasi power theorem for $v = \exp(s)$ in probability generating function $p(v)$ [5]. $\square$

3 Distances

In this section we study the random variable $H_{n,j}$, which counts the distance, measured by the number of edges lying on the connecting path, between bucket containing label j and bucket containing label n in a random bucket recursive tree T with variable capacities of buckets of size n. Let

$$
W(z, u, v) = \sum_{k \ge 1} \sum_{j \ge 1} \sum_{m \ge 0} P(H_{k+j,j} = m) T_{k+j,b} \frac{z^{j-1}}{(j-1)!} \frac{u^{k-1}}{(k-1)!} v^m.
$$

Again we apply a combinatorial description involving the counting of 4-colored bucket recursive tree with variable capacities of buckets. Since the arguments are very similar to [7] we just sketch the derivation. The combinatorial objects considered are all possible 4-colored trees of size ≥ 2 with a coloring as specified next. In each tree T the bucket containing the largest label (i.e., n) is colored green. From the remaining buckets exactly one bucket is colored red (bucket containing label j), all buckets with smaller labels than the red bucket are colored black, and all remaining buckets containing labels larger than the red bucket are colored white. We are interested in the distance between the red bucket and the green bucket. Finally

$$
\frac{\partial W(z, u, v)}{\partial z} = b!^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|} \varphi'(T_{n,b}(z+u)) W(z, u, v)
$$

$$+ \quad b!^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\frac{v^2\varphi''(T_{n,b}(z+u))(T'_{n,b}(z+u))^{2v}}{(T'_{n,b}(u)\alpha_0)^{v-1}} \tag{14}$$

with initial condition

$$
\begin{aligned}
W(0,u,v) &= b!^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\frac{\partial}{\partial u}L(u,0,v) \\
&= b!^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}v T''_{n,b}(u)\left(\frac{T'_{n,b}(u)}{\alpha_0}\right)^{v-1},
\end{aligned}
$$

where z counts the black nodes, u the white nodes, and v the distance between the red and the green label.

Lemma 2 *The probabilities* $P(H_{n,j} = h)$ *are given by the following formula:*

$$
\begin{aligned}
P(H_{n,j} = h) &= \frac{b!^{\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}}{(n-1)\binom{n-2}{j-1}}\left\{ \sum_{\ell=0}^{n-j-1}\binom{n-\ell-2}{j-1}\frac{1}{\ell!}S(\ell, h-1) \right. \\
&+ \left. \sum_{k=0}^{n-j-1}\binom{n-k-2}{j-1}\sum_{\ell=0}^{h-2}\frac{2^\ell}{k!}S(k, h-\ell-2)\right\}, \quad 1 \le j < n \tag{15}
\end{aligned}
$$

Proof. The equation (14) has the following consequence:

$$
\begin{aligned}
W(z,u,v) &= b!^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\frac{v T''_{n,b}(u)T'_{n,b}(z+u)}{T'_{n,b}(u)}\left(\frac{T'_{n,b}(u)}{\alpha_0}\right)^{v-1} \\
&+ b!^{-\sum_{i=1}^{r}|\mathcal{P}_{k_i}|}\frac{v^2 T'_{n,b}(z+u)}{(T'_{n,b}(u)\alpha_0)^{v-1}} \\
&\times \int_0^z \varphi''(T_{n,b}(t+u)(T'_{n,b}(t+u))^{2v-1}dt. \tag{16}
\end{aligned}
$$

By (3),

$$
\begin{aligned}
W(z,u,v) &= \frac{v}{b}\cdot\frac{b!^{1+v+mv}}{(1-b!^{1-m}u)^v(1-b!^{1-m}(z+u))} \\
&+ \frac{v^2}{b^v}\frac{b!^{1-m+v+mv}(b-1)!^{1-v}}{(2v-1)(1-b!^{1-m}u)^{1-v}(1-b!^{1-m}(z+u))^{2v}} \\
&- \frac{v^2}{b^v}\frac{b!^{1-m+v+mv}(b-1)!^{1-v}}{(2v-1)(1-b!^{1-m}u)^v(1-b!^{1-m}(z+u))},
\end{aligned}
$$

where $m = \sum_{i=1}^{r}|\mathcal{P}_{k_i}|$. Thus

$$P(H_{n,j} = h) = \frac{(j-1)!(n-j-1)!}{T_{n,b}}[z^{j-1}u^{n-j-1}v^h]W(z,u,v),$$

and proof is completed (since these computations are essentially straightforward, but quite lengthy computations, they are omitted here. Similar considerations are done in [7] where the somewhat simpler recurrences appearing there are treated analogously). □

Theorem 2 *For* $1 \leq j < n$,

$$\mathbb{E}(H_{n,j}) = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}\left(H_{n-1} + H_j + \frac{1}{j} - 2\right),$$

$$\mathbb{V}ar(H_{n,j}) = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_{n-1}\left(\frac{1}{j} - 1 - \frac{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}}{j} + 2b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}\right)$$

$$- b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_j\left(\frac{3}{j} + 1 + \frac{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}}{j} - 2b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}\right)$$

$$+ \frac{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}}{j}(4b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} - 1) + 4b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(2 - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|})$$

$$- b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_{n-1}^{(2)} - 3b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_j^{(2)} + b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_{n-1}^2(1 - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|})$$

$$+ b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_j^2(1 - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}) + 2b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_jH_{n-1}(1 - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|})$$

$$- \frac{b!^{2\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}}{j^2}.$$

Proof. By (10),

$$\begin{aligned}
\mathbb{E}(H_{n,j}) &= \frac{(j-1)!(n-j-1)!}{T_{n,b}}[z^{j-1}u^{n-j-1}]\frac{\partial W(z,u,v)}{\partial v}\bigg|_{v=1} \\
&= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}\left(H_{n-1} + H_j + \frac{1}{j} - 2\right),
\end{aligned}$$

and by (11),

$$\begin{aligned}
\mathbb{E}(D_{n,j}(D_{n,j}-1)) &= \frac{(j-1)!(n-j-1)!}{T_{n,b}}[z^{j-1}u^{n-j-1}]\frac{\partial^2 W(z,u,v)}{\partial v^2}\bigg|_{v=1} \\
&= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_{n-1}\left(\frac{1}{j} - 2\right) - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_j\left(\frac{3}{j} + 2\right) \\
&\quad - 2\frac{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}}{j} + 10b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|} - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_{n-1}^{(2)} \\
&\quad - 3b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_j^{(2)} + b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_{n-1}^2 + b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_j^2 \\
&\quad + 2b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}H_jH_{n-1}.
\end{aligned}$$

Proof is completed just similar to the Lemma 1. □

Corollary 2 *For* $1 \leq j < n$,

$$\mathbb{E}(H_{n,j}) = \mathbb{V}ar(H_{n,j}) = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log n + \log j) + \mathcal{O}(1).$$

Theorem 3 *As* $n \to \infty$,

$$Z = \frac{H_{n,j} - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log n + \log j)}{\sqrt{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log n + \log j)}} \xrightarrow{d} N(0,1),$$

for arbitrary sequences $(n, j(n))_{n \in \mathbb{N}}$.

Proof. Let $m = \sum_{i=1}^{r} |\mathcal{P}_{k_i}|$ and

$$\psi_{n,j}(v) = \mathbb{E}(v^{H_{n,j}}) = \sum_{h \geq 0} v^h P(H_{n,j} = h),$$

be the probability generating function of $H_{n,j}$. Thus

$$\begin{aligned}
\psi_{n,j}(v) &= \frac{(j-1)!(n-j-1)!}{T_{n,b}}[z^{j-1}u^{n-j-1}]W(z,u,v) \\
&= v\frac{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}\binom{n+v-2}{n-j-1}}{(n-1)\binom{n-2}{j-1}} \\
&\quad + \frac{v^2}{2v-1}\frac{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}\binom{n+v-2}{n-j-1}}{(n-1)\binom{n-2}{j-1}}\binom{2v+j-2}{j-1} \\
&\quad - \frac{v^2}{2v-1}\frac{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}\binom{n+v-2}{n-j-1}}{(n-1)\binom{n-2}{j-1}}.
\end{aligned}$$

Let

$$\begin{aligned}
\mu_{n,j} &:= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log n + \log j), \\
\sigma_{n,j}^2 &= b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log n + \log j),
\end{aligned}$$

and

$$M_{n,j}(t) = \mathbb{E}(e^{tZ}) = \sum_{z \geq 0} e^{tz}P(Z = z),$$

be the moment generating function of

$$Z = \sigma_{n,j}{}^{-1}(H_{n,j} - \mu_{n,j}).$$

Then

$$M_{n,j}(t) = e^{-\frac{\mu_{n,j}}{\sigma_{n,j}}} \psi_{n,j}(e^{\frac{t}{\sigma_{n,j}}}).$$

Now we split the region $1 \le j < n$ into two cases: j big and $j \ge \log n$, and j small and $j \le \log n$. With the same consideration of [7] proof is completed. $\square$

We get also as a corollary similar results for the random variable $H_{n,i,j}$, which counts the distance between the bucket containing label i and bucket containing label j in our random tree of size n.

Corollary 3 *For* $1 \le i < j < n$,

$$\mathbb{E}(H_{n,i,j}) = \mathbb{V}ar(H_{n,i,j}) = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log i + \log j) + \mathcal{O}(1).$$

If $\mu_{n,i,j} := b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log i + \log j)$, $\sigma_{n,i,j}^2 = b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log i + \log j)$, *then*

$$Z = \frac{H_{n,i,j} - b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log i + \log j)}{\sqrt{b!^{\sum_{i=1}^{r} |\mathcal{P}_{k_i}|}(\log i + \log j)}} \xrightarrow{d} N(0, 1)$$

for arbitrary sequences $(n, i(n), j(n))_{n \in \mathbb{N}}$.

4   Conclusion

In this paper, we studied the random variable depth of label j in a bucket recursive tree with variable bucket capacities and maximal bucket size b ($n \ge j$). We obtained a closed formula for the probability distribution, the expectation and the variance. We showed as $j \to \infty$, limiting distributions are Gaussian. In passing, we studied the random variable $H_{n,j}$, which counts the distance, measured by the number of edges lying on the connecting path, between bucket containing label j and bucket containing label n. Finally, we extend our results to the random variable $H_{n,i,j}$ which counts the distance between the bucket containing label i and bucket containing label j in our random tree of size n. We obtained this results by presenting partial differential equations for moment generating functions and solving them.

References

[1] C. Christophi, H. Mahmoud, The oscillatory distribution of distances in random tries, *Ann. Appl. Probab.*, **15** (2005) 1536–1564.

[2] L. Devroye, R. Neininger, Distances and finger search in random binary search trees, *SIAM J. Comput.*, **33** (2004), 647–658.

[3] R. Dobrow, On the distribution of distances in recursive trees, *J. Appl. Probab.*, **33** (1996), 749–757.

[4] R. Dobrow, R. Smythe, Poisson approximations for functionals of random trees, *Random Structures Algorithms*, **9** (1996), 79–92.

[5] P. Flajolet, R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge (2008).

[6] R. Kazemi, Depth in bucket recursive trees with variable capacities of buckets, *Act. Math. Sin, English series*, **30(2)** (2014), 305–310.

[7] M. Kuba, A. Panholzer, On the distribution of distances between specified nodes in increasing trees, *Discrete Appl. Math.*, **158 (5)** (2010), 489–506.

[8] M. Kuba, A. Panholzer, Combinatorial approach to the analysis of bucket recursive trees. *Theoret. Comput. Sci.*, **411** (2010), 34–36, 3253–3273.

[9] H. Mahmoud, R. Neininger, Distribution of distances in random binary search trees, *Ann. Appl. Probab.*, **13** (2003), 253–276.

[10] H. Mahmoud, R. Smythe, Probabilistic analysis of bucket recursive trees. *Theor. Comput. Sci.*, **144** (1995), 221–249.

[11] A. Meir, J.W. Moon, Recursive trees with no nodes of out-degree one, *Congressus Numerantium*, **66** (1988), 49–62.

[12] J.W. Moon, The distance between nodes in recursive trees., *London Mathematical Society Lecture Notes.*, **13** (1974), 125–132.

[13] A. Panholzer, The distribution of the size of the ancestor-tree and of the induced spanning subtree for random trees, *Random Structures Algorithms.*, **25** (2004), 179–207.

[14] A. Panholzer, H. Prodinger, Level of nodes in increasing trees revisited, *Random Structures Algorithms.*, **31** (2007), 203–226.