Methodol Comput Appl Probab (2014) 16:863–884
DOI 10.1007/s11009-013-9333-z

Asymptotic Joint Normality of Counts of Uncorrelated Motifs in Recursive Trees

Mohan Gopaladesikan · Hosam Mahmoud · Mark Daniel Ward

Received: 3 May 2012 / Revised: 15 March 2013 /
Accepted: 18 March 2013 / Published online: 29 March 2013

Abstract We study the fringe of random recursive trees, by analyzing the joint distribution of the counts of uncorrelated motifs. Our approach allows for finite and countably infinite collections. To be able to deal with the collection when it is infinitely countable, we use measure-theoretic themes. Each member of a collection of motifs occurs a certain number of times on the fringe. We show that these numbers, under appropriate normalization, have a limiting joint multivariate normal distribution. We give a complete characterization of the asymptotic covariance matrix. The methods of proof include contraction in a metric space of distribution functions to a fixed-point solution (limit distribution). We discuss two examples: the finite collection of all possible motifs of size four, and the infinite collection of rooted stars. We conclude with remarks to compare fringe-analysis with matching motifs everywhere in the tree.

Keywords Analysis of algorithms · Random trees · Recurrence · Motif · Contraction

AMS 2000 Subject Classifications MSC 05C05 · MSC 60C05 · MSC 68P10 · MSC 68W40 · MSC 11B37

Dedicated to the memory of Philippe Flajolet.

M. Gopaladesikan's & M. D. Ward's research is supported by NSF Science & Technology Center for Science of Information Grant CCF-0939370.

M. Gopaladesikan (✉) · M. D. Ward
Department of Statistics, Purdue University,
150 North University Street, West Lafayette,
IN 47907–2067, USA
e-mail: mgopalad@purdue.edu

M. D. Ward
e-mail: mdw@purdue.edu

H. Mahmoud
Department of Statistics, The George Washington University,
Washington, DC 20052, USA
e-mail: hosam@gwu.edu

1 Introduction

A random recursive tree is a naturally growing structure that underlies several stochastic developments, such as recruiting, the spreading of chain letters, contagion, and the evolution of the Union-Find algorithm. The survey by Smythe and Mahmoud (1995) is a source for numerous facts, references and applications of recursive trees.

The random recursive tree is a rooted nonplanar tree that grows by the successive insertion of nodes labelled 1,2,3,.... The insertions occur at equispaced discrete time points 1, 2, 3, ... At time 1, node 1 is created as the root. The process goes forth in the following manner. For $i = 2, 3, \ldots$, after $i - 1$ insertions, there is a random recursive tree of size $i - 1$. When the ith node appears, a node labeled i joins the tree by randomly choosing any of the existing nodes (with equal probability) as *parent*. This new node becomes the *child* of the selected parent node. After n insertions the tree has n nodes, and we say it is of *size* n, and a tree grown in such a manner is a *random recursive tree* of size n.

It is straightforward to see that the construction of random recursive trees induces a uniform distribution on the space of trees of size n: There are $(n-1)!$ recursive trees of size n and they are all equally likely. We shall use probabilistic methods. However, the uniform distribution of the trees is also amenable to analytic methods (Bergeron et al. 1992).

A *leaf* is a node that has no children. According to the insertion algorithm, the labels of the nodes on any root-to-leaf path are an increasing sequence. Therefore, recursive trees are in the class of increasing trees, which received much attention over the past two decades (see Panholzer and Prodinger 2004, for example).

We use the word *motif* to refer to a specific nonplanar unlabelled rooted tree shape of finite size. For a given motif Γ, of size γ, let $X_{n,\Gamma}$ be a count of the number of occurrences of Γ on the fringe of a random recursive tree of size n. That is, $X_{n,\Gamma}$

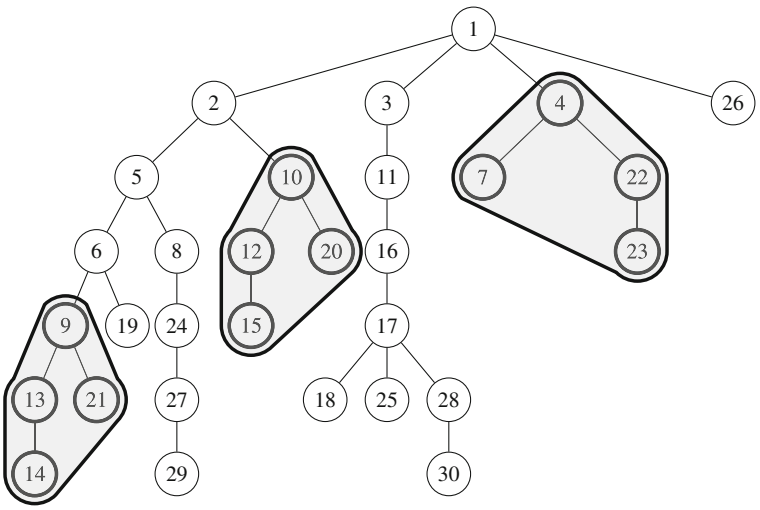

Fig. 1 Example of a recursive tree of size 30 with three occurrences of a motif on the fringe

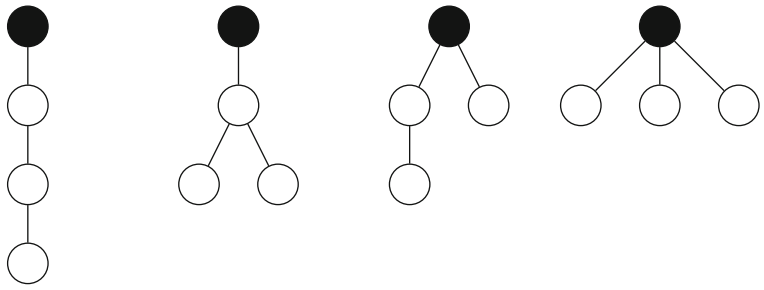

Fig. 2 All motifs of size 4. When generating a recursive tree of size 4, these motifs occur with probabilities $\frac{1}{6}$, $\frac{1}{6}$, $\frac{3}{6}$ and $\frac{1}{6}$, from left to right respectively

counts each rooted subtree[1] with shape isomorphic to Γ, such that the rooted subtree is the motif itself. As an illustration, suppose the realization of the recursive tree after 30 insertions is that in Fig. 1. If Γ is the third motif (of size $\gamma = 4$) from the left in Fig. 2, it occurs $X_{30,\Gamma} = 3$ times on the fringe of the tree of Fig. 1. Each occurrence of Γ on the fringe of the tree in Fig. 1 is shown as a cluster of darkened nodes. There are other occurrences of Γ, such as the nodes $\{5, 6, 8, 24\}$ and $\{17, 25, 28, 30\}$, but these occurrences do not enter our count, as they are not on the fringe. The subgraph $\{5, 6, 8, 24\}$ is not on the fringe as the entire subtree rooted at 5 is larger than the motif; it includes the additional nodes 9, 13, 14, 19, 21, 27, and 29. Also, the subgraph $\{17, 25, 28, 30\}$, is not a motif on the fringe, because of the presence of node 18, so that the entire subtree rooted at 17 is not the same as Γ. Notice that the motifs are nonplanar and the subtrees $\{10, 12, 15, 20\}$ and $\{4, 7, 22, 23\}$ are both counted as matches of Γ, as both are the same (i.e. isomorphic to each other).

2 Applications

Pattern matching in the context of binary search trees is taken up in Flajolet et al. (1997). Similar to the application (Flajolet et al. 1997) gives, we have the following for recursive trees. Knowing the number of occurrences of a particular motif can be of use in data compression. Instead of storing a motif many times in a tree, we can store the content with only one nexus pointing to the motif to realize the shape in the recursive tree. The content itself should be stored in an appropriate canonical order to fit its original position in the recursive tree. In a plain practical implementation not utilizing data compression ideas, each of these nodes would carry a number of pointers (equal to the number of its children). In some applications, like the Union-Find algorithm, the pointers go in the opposite direction (from a child to its parent), as clusters join by adjoining their roots, having the root of one cluster point to the root of the other. In either version, the pointers inside each occurrence of the motif are eliminated in the proposed implementation.

[1] In this manuscript *subtree* refers to a node in the recursive tree and *all* its descendants.

We illustrate this application next. There is more than one isomorphic drawing of a given motif. The different drawings are obtained by permuting the subtrees rooted at the children of a node. Nonetheless, we can consider only one of these drawings as a canonical representation. We can, for instance, require the subtrees rooted at the children of a node to appear in decreasing order of their sizes from left to right, and if the sizes of several subtrees agree, we draw them so that the labels associated with their roots are in increasing order. Take for instance, the motif Γ again to be the the third motif from the left in Fig. 2. As illustrated in Fig. 1, this motif appears three times on the fringe. Each of the nodes 1, 2 and 6 points to an occurrence of Γ. We can let these pointers be directed to the data blocks {9, 13, 14, 21}, {10, 12, 15, 20}, and {4, 22, 23, 7}, (in say array implementation) and each block contains *only one* pointer to the shape of Γ (or a description of it). Note that each of the data blocks is stored to correspond to a root-last and left-to-right traversal of siblings of the canonical form.

3 Uncorrelated Motifs

Let $\mathscr{I}$ be a countable indexing set. Let

$$\mathscr{C} = \{\Gamma_i \,|\, i \in \mathscr{I}\}$$

be a given collection of motifs. We say that two motifs are *uncorrelated*, if neither appears as a subtree on the fringe of the other, and we call a collection of motifs a *collection of uncorrelated motifs*, if its members are pairwise uncorrelated. For instance, two distinct motifs of the same size are always uncorrelated. Let $\mathcal{P}_i$ denote the rooted path of length i. For example, $\mathcal{P}_4$ is the leftmost motif in Fig. 2. The rooted path $\mathcal{P}_i$, of length $i < j$, is correlated with the rooted path $\mathcal{P}_j$, of length j. Knowledge of joint occurrences can lead to a better understanding of the performance of certain algorithms. For instance, if there are various stars in the recursive tree underlying the Union-Find algorithm, it is an indication that several tasks can perform faster in a parallel computing environment.

In many applications the collection of motifs will be finite, but our presentation covers cases of countably infinite collections, too. In the present paper, we consider the joint distribution of X_{n,Γ_i}, for $i \in \mathscr{I}$, in a random recursive tree of size n.

4 Organization

The rest of this paper is organized as follows. In Section 5, we present the main results. In Section 6, we set up some technicalities: We discuss a probability space on which our random variables can be defined for each n, and introduce a univariate linear combination of the number of occurrences of the members of a collection of motifs. In Section 7 we present the proofs. The proofs are structured in sections: Sections 7.2, 7.3, 7.4 are (respectively) for the derivation of the mean, variance, and the Gaussian limit distribution of the univariate linear combination, and a joint multivariate central limit theorem for the number of occurrences of the members of the collection. The rate of convergence is discussed empirically in Section 9, where

we give a supporting simulation study. We give two examples in Sections 8.1 and 8.2. To put the fringe analysis in perspective, we conclude with remarks on how the result compares with matching patterns everywhere in the recursive tree.

5 Results

The genesis of this work is in Feng and Mahmoud (2010). In that reference the authors present a central limit theorem for the number of occurrences of a single motif Γ, of size γ, on the fringe of a recursive tree of size n. The results are presented in terms of $\mathcal{C}(\Gamma)$, the probability that the random construction of recursive tree of size γ realizes the motif Γ. For instance, the four motifs of size 4 in Fig. 2 have $\mathcal{C}(\Gamma)$ equal to $\frac{1}{6}$, $\frac{1}{6}$, $\frac{3}{6}$ and $\frac{1}{6}$, from left to right respectively. A complete characterization of $\mathcal{C}(\Gamma)$ is given in Feng and Mahmoud (2010) in terms of the shape of the motif, and the authors call $\mathcal{C}(\Gamma)$ a shape functional, as its value is derived from the shape of the motif.

The main results of this paper are the following.

Theorem 1 *Let $\mathscr{I}$ be a countable set (finite or infinite). Let $\mathscr{C} = \{\Gamma_i \,|\, i \in \mathscr{I}\}$ be an uncorrelated collection of nonplanar, unlabeled, rooted trees, each of a finite size (motifs). Let $X_{n,\Gamma}$ be the number of occurrences of the motif Γ, of size γ, on the fringe of a random recursive tree of size n. Then, we have*

$$\mathbf{Cov}[\mathbf{X}_{n,\mathscr{C}}] = \mathbf{\Sigma}_{\mathscr{C}}\, n,$$

with

$$(\mathbf{\Sigma}_{\mathscr{C}})_{i,j} = \begin{cases} \left(\dfrac{(\gamma_i+1)(2\gamma_i+1) - (3\gamma_i+2)\,\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)^2(2\gamma_i+1)}\right) \mathcal{C}(\Gamma_i) \\ \qquad \times \mathbf{1}_{\{n>2\gamma_i\}}, & \text{if } i = j; \\ \dfrac{1}{2}\left(\dfrac{2\mathbf{E}[X_{2\gamma^*_{i,j}+1,\Gamma_i} X_{2\gamma^*_{i,j}+1,\Gamma_j}]}{2\gamma^*_{i,j}+1} + \dfrac{\mathscr{W}(2\gamma^*_{i,j}+2, \mathscr{C}, \mathbf{b}_{i,j})}{(2\gamma^*_{i,j}+2)(2\gamma^*_{i,j}+1)}\right. \\ \qquad - \dfrac{\mathcal{C}^2(\Gamma_i)}{\gamma_i^2(\gamma_i+1)^2} - \dfrac{\mathcal{C}^2(\Gamma_j)}{\gamma_j^2(\gamma_j+1)^2} \\ \qquad \left. - \dfrac{2(2\gamma^*_{i,j}+2)\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)}\right) \mathbf{1}_{\{n>2\gamma^*_{i,j}+1\}}, & \text{if } i \neq j; \end{cases}$$

where $\mathbf{X}_{n,\mathscr{C}}$ is the vector with components X_{n,Γ_i}, $\gamma^*_{i,j} = \max\{\gamma_i, \gamma_j\}$, $\mathscr{W}(.,.,.)$ is defined in Eq. 5, and $\mathbf{b}_{i,j}$ is a vector of $|\mathscr{I}|$ dimensions with all entries being zero except positions i and j, where these entries are 1.

Theorem 2 *Let $\mathscr{I}$ be a countable set (finite or infinite). Let $\mathscr{C} = \{\Gamma_i \,|\, i \in \mathscr{I}\}$ be an uncorrelated collection of nonplanar, unlabeled, rooted trees, each of finite size (motifs). Let $X_{n,\Gamma}$ be the number of occurrences of the motif Γ, of size γ, on the fringe of a random recursive tree of size n. Then, we have*

$$\frac{\mathbf{X}_{n,\mathscr{C}} - \boldsymbol{\mu}_{\mathscr{C}} n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}_{|\mathscr{I}|}(\mathbf{0}, \mathbf{\Sigma}_{\mathscr{C}}),$$

where $\mathbf{X}_{n,\mathscr{C}}$*, is the vector with components* X_{n,Γ_i}*, and* $\boldsymbol{\mu}_{\mathscr{C}}$ *is the vector with components*

$$(\boldsymbol{\mu}_{\mathscr{C}})_i = \frac{\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)},$$

for $i \in \mathscr{I}$*, and* $\mathcal{C}(\Gamma_i)$ *is the shape functional of the motif* Γ_i*,* $\mathcal{N}_{|\mathscr{I}|}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathscr{C}})$ *is the jointly multivariate normally distributed random vector in* $|\mathscr{I}|$ *dimensions*[2] *with mean vector* $\mathbf{0}$ *(of* $|\mathscr{I}|$ *components) and* $|\mathscr{I}| \times |\mathscr{I}|$ *covariance matrix* $\boldsymbol{\Sigma}_{\mathscr{C}}$*.*

6 A Probability Space for Recursive Trees

As we intend to discuss a sequence of random variables occurring in growing trees, the matter is made rigorous by considering a probability space on which all the random variables are well defined. Let Ω be the space of all *infinite* recursive trees, which are obtained by perpetuating the insertion ad infinitum. Note that Ω is uncountable. Let $\omega \in \Omega$, thus we can view ω as one stochastic path. On this stochastic path define $T_n = T_n(\omega)$, the corresponding recursive tree of size n. This finite tree with n nodes is obtained by pruning any node labeled greater than n in ω, and destroying any edge that has a child with label $n+1$ or larger. On the other hand, for a given finite recursive tree $\mathcal{T}_n$ there corresponds an uncountable class of recursive trees ω, such that $T_n(\omega) = \mathcal{T}_n$. We can think of this class as the subset of Ω induced by $\mathcal{T}_n$; let us call such a class $C_{\mathcal{T}_n}$, which is a member of the nth cylinder of the space. On this space of trees we impose the measure P that gives the finite cylinder $C_{\mathcal{T}_n}$, the probability $P(C_{\mathcal{T}_n}) = \frac{1}{(n-1)!}$, simply meaning the probability of T_n is $\frac{1}{(n-1)!}$. That is, P is the measure obtained by Kolmogorov's extension to agree with all the finite cylinders. The measure then operates on the sigma field $\mathcal{F}$ generated by the collection of the classes $C_{\mathcal{T}_n}$. Henceforth, $(\Omega, \mathcal{F}, P)$ is the probability space underlying any random variables we deal with. So, $X_{n,\Gamma} = X_{n,\Gamma}(T_n(\omega))$ is a random variable that counts the number of occurrences of a motif Γ in $T_n = T_n(\omega)$.

Toward a multivariate central limit theorem, we work with a univariate linear combination, and prove a univariate central limit theorem for it, to ultimately use the Cramér-Wold device, see Theorem 29.4 on p. 383 of Billingsley (1995). More specifically, we deal with the linear combination

$$Y_{n,\mathscr{C},\boldsymbol{\alpha}} = \boldsymbol{\alpha}\mathbf{X}_{n,\mathscr{C}} = \sum_{i\in\mathscr{I}} \alpha_i X_{n,\Gamma_i},$$

where $\boldsymbol{\alpha}$ is the vector of α_i's, $i \in \mathscr{I}$. The product in the middle is a dot product of the two vectors. This linear combination is well defined on the probability space just described. It will turn out that $Y_{n,\mathscr{C},\boldsymbol{\alpha}}$ is asymptotically normally distributed, and consequently the random variables $\{X_{n,\Gamma_i} \,|\, i \in \mathscr{I}\}$ asymptotically have a joint multivariate normal distribution.

[2] Of course, if $\mathscr{I}$ is finite, the limiting multivariate normal involved is a distribution in $|\mathscr{I}|$ dimensions. If $|\mathscr{I}|$ is infinitely countable, we take the infinite-dimensional limiting multivariate normal to mean that every finite subset of the variables in it has a joint multivariate distribution.

7 Proofs

We discuss the technical proofs in this section, starting with a brief discussion of averages, followed by the necessary computations of covariances.

7.1 A Stochastic Recurrence for the Linear Combination

We shall use a decomposition of a recursive tree introduced in van der Hofstad et al. (2002). Remove the *special edge* joining the nodes labeled 1 and 2. The tree then falls apart into a forest of two trees. One tree is rooted at 2, which we shall recognize as a *special tree* of the original recursive tree (which is a proper subtree of the recursive tree). The other tree is rooted at 1, which is a *nonspecial tree*. Let U_n be the size of the special subtree, and so $n - U_n$ is the size of the nonspecial tree. It is shown in van der Hofstad et al. (2002) that U_n has a uniform distribution on $\{1, 2, \ldots, n-1\}$. Note that the special (respectively, nonspecial) tree is isomorphic to a recursive tree of size U_n (respectively, size $n - U_n$) that has the same uniform probability of a random recursive tree of that size. Also, the two subtrees are conditionally independent (given U_n).

As in Feng and Mahmoud (2010), for $n > \gamma$, we have a stochastic recurrence for $X_{n,\Gamma}$: It can be composed from the number of occurrences of the motif Γ in the special and nonspecial trees, and we need to subtract 1, if the nonspecial subtree is of size γ, and assumes the shape Γ. We shall express the formulation in terms of the *indicator* notation: for any event $\mathcal{E}$, the indicator $\mathbf{1}_{\mathcal{E}} = 1$, if $\mathcal{E}$ occurs, and $\mathbf{1}_{\mathcal{E}} = 0$ otherwise. We shall also refer to a Bernoulli random variable with success probability p as $\mathrm{Ber}(p)$. For $n > \gamma$, we have a stochastic recurrence, which gives rise to an equality in distribution:

$$X_{n,\Gamma} \stackrel{\mathcal{D}}{=} X_{U_n,\Gamma} + \tilde{X}_{n-U_n,\Gamma} - \mathbf{1}_{\{n-U_n=\gamma\}} \mathrm{Ber}\big(\mathcal{C}(\Gamma)\big); \tag{1}$$

the tilded random variable $\tilde{X}_{n-U_n,\Gamma}$ is conditionally independent of $X_{U_n,\Gamma}$ (given U_n). Note also that, for each $j \geq 0$, $\tilde{X}_{j,\Gamma}$ has the same distribution as $X_{j,\Gamma}$.

When $\mathscr{C}$ contains finitely many motifs, we define $\gamma^* := \max_{i \in \mathscr{I}} \gamma_i$. For such a $\mathscr{C}$, a recurrence for $Y_{n,\mathscr{C},\boldsymbol{\alpha}}$ follows naturally from a translation of the stochastic recurrence (Eq. 1) into

$$Y_{n,\mathscr{C},\boldsymbol{\alpha}} \stackrel{\mathcal{D}}{=} Y_{U_n,\mathscr{C},\boldsymbol{\alpha}} + \tilde{Y}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} - \sum_{i \in \mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \mathrm{Ber}\big(\mathcal{C}(\Gamma_i)\big). \tag{2}$$

7.2 The Average of the Linear Combination

Let Γ be a given motif. The average is given in Feng and Mahmoud (2010):

$$\mathbf{E}[X_{n,\Gamma}] = \frac{\mathcal{C}(\Gamma)}{\gamma\,(\gamma\,+1)}\, n, \qquad n > \gamma\,.$$

It then follows that $\mathbf{E}[\mathbf{X}_{n,\mathscr{C}}]$ is a vector with components $\frac{\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)}\, n\, \mathbf{1}_{\{n>\gamma_i\}}$, for each $i \in \mathscr{I}$.

Let us first consider a finite collection of motifs, with a finite indexing set $\mathscr{I}$. If n is not large enough, some or all of these components are 0. Observe that, if the indexing set is finite, we can remove the indicators and simply say that for all $i \in \mathscr{I}$, the ith component in the vector of averages is $\frac{\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)}\, n$, for all $n > \gamma^*$. We have,

$$\mathbf{E}[Y_{n,\mathscr{C},\boldsymbol{\alpha}}] = \sum_{i\in\mathscr{I}} \alpha_i \mathcal{C}(\Gamma_i) \left(\frac{n\mathbf{1}_{\{\gamma_i<n-1\}}}{\gamma_i(\gamma_i+1)} + \frac{\mathbf{1}_{\{\gamma_i=n-1\}}}{\gamma_i} + \mathbf{1}_{\{\gamma_i=n\}} \right), \qquad \text{for any } n.$$

In particular

$$\mathbf{E}[Y_{n,\mathscr{C},\boldsymbol{\alpha}}] = \sum_{i\in\mathscr{I}} \frac{\alpha_i \mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)}\, n, \qquad \text{if } n > \gamma^*.$$

However, if the indexing set is countably infinite, no such γ^* exists, because the collection of motifs must then contain countably many arbitrarily large trees. For any n, however large, there is only a finite number of entries in $\mathbf{E}[\mathbf{X}_{n,\mathscr{C}}]$ that are nonzero; the rest are all zero. On the other hand, any individual motif has a finite size, and at some point in the insertion process, the corresponding entry in the average vector becomes nonzero, and stays nonzero thereafter. That is,

$$\mathbf{E}[\mathbf{X}_{n,\mathscr{C}}]_i = \begin{cases} \dfrac{\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)}\, n, & \text{if } n > \gamma_i; \\ \mathcal{C}(\Gamma_i), & \text{if } n = \gamma_i; \\ 0, & \text{if } n < \gamma_i. \end{cases}$$

Hence, as $n \to \infty$, the vector $\mathbf{E}[\mathbf{X}_{n,\mathscr{C}}]$ "fills out" with nonzero components. A similar argument holds for the variance-covariance matrix.

7.3 The Covariance Structure

Let us again start with a finite collection, with the largest tree among them having size γ^*. For computing the covariance of the linear combination we start with the second moment:

$$\begin{aligned} Y^2_{n,\mathscr{C},\boldsymbol{\alpha}} &= \left(Y_{U_n,\mathscr{C},\boldsymbol{\alpha}} + \tilde{Y}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} - \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \operatorname{Ber}\big(\mathcal{C}(\Gamma_i)\big) \right)^2 \\ &= Y^2_{U_n,\mathscr{C},\boldsymbol{\alpha}} + \tilde{Y}^2_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} + \left(\sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \operatorname{Ber}\big(\mathcal{C}(\Gamma_i)\big) \right)^2 \\ &\quad + 2Y_{U_n,\mathscr{C},\boldsymbol{\alpha}} \tilde{Y}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} - 2Y_{U_n,\mathscr{C},\boldsymbol{\alpha}} \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \operatorname{Ber}\big(\mathcal{C}(\Gamma_i)\big) \\ &\quad - 2\tilde{Y}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \operatorname{Ber}\big(\mathcal{C}(\Gamma_i)\big). \end{aligned}$$

To expand the square of the sum, we utilize $\mathbf{1}_{\{n-U_n=\gamma_i\}}\mathbf{1}_{\{n-U_n=\gamma_j\}} = 0$, for $i \neq j$. After expansion, we take expectations and get

$$\begin{aligned}
\mathbf{E}\left[Y^2_{n,\mathscr{C},\boldsymbol{\alpha}}\right] &= \mathbf{E}\left[Y^2_{U_n,\mathscr{C},\boldsymbol{\alpha}}\right] + \mathbf{E}\left[\tilde{Y}^2_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}\right] + 2\,\mathbf{E}\left[Y_{U_n,\mathscr{C},\boldsymbol{\alpha}}\tilde{Y}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}\right] \\
&\quad + \mathbf{E}\left(\sum_{i\in\mathscr{I}} \alpha_i^2 (\mathbf{1}_{\{n-U_n=\gamma_i\}}\,\mathrm{Ber}(\mathcal{C}(\Gamma_i)))^2\right) \\
&\quad - 2\,\mathbf{E}\left(Y_{U_n,\mathscr{C},\boldsymbol{\alpha}} \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}}\,\mathrm{Ber}(\mathcal{C}(\Gamma_i))\right) \\
&\quad - 2\,\mathbf{E}\left(\tilde{Y}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}}\,\mathrm{Ber}(\mathcal{C}(\Gamma_i))\right) \\
&= \frac{2}{n-1}\left(\sum_{k=1}^{n-1} \left(\mathbf{E}[Y^2_{k,\mathscr{C},\boldsymbol{\alpha}}] + \mathbf{E}[Y_{k,\mathscr{C},\boldsymbol{\alpha}}]\,\mathbf{E}[Y_{n-k,\mathscr{C},\boldsymbol{\alpha}}]\right)\right. \\
&\quad + \sum_{i\in\mathscr{I}} \left(\frac{\alpha_i^2\,\mathcal{C}(\Gamma_i)}{2} - \alpha_i \mathbf{E}[Y_{n-\gamma_i,\mathscr{C},\boldsymbol{\alpha}}]\,\mathcal{C}(\Gamma_i)\right) \\
&\quad \left. - \sum_{i\in\mathscr{I}} \mathbf{E}\left[\alpha_i \tilde{Y}_{\gamma_i,\mathscr{C},\boldsymbol{\alpha}}\,\mathrm{Ber}(\mathcal{C}(\Gamma_i))\right]\right).
\end{aligned}$$

We can decompose the last term of the right hand side:

$$\begin{aligned}
\sum_{i\in\mathscr{I}} \mathbf{E}\left[\alpha_i \tilde{Y}_{\gamma_i,\mathscr{C},\boldsymbol{\alpha}}\,\mathrm{Ber}\left(\mathcal{C}(\Gamma_i)\right)\right] &= \sum_{i\in\mathscr{I}} \mathbf{E}\left[\alpha_i\,\mathrm{Ber}(\mathcal{C}(\Gamma_i)) \sum_{s\in\mathscr{I}} \alpha_s \tilde{X}_{\gamma_i,\Gamma_s}\right] \\
&= \sum_{i\in\mathscr{I}} \alpha_i^2\,\mathcal{C}(\Gamma_i) + \sum_{\substack{s\in\mathscr{I} \\ s\neq i}} \mathbf{E}\left[\alpha_i\,\alpha_s \tilde{X}_{\gamma_i,\Gamma_s}\,\mathrm{Ber}(\mathcal{C}(\Gamma_i))\right].
\end{aligned}$$

Note that the terms in the second summation exist only when $\gamma_i \geq \gamma_s$. In that case, we are looking at the nonspecial tree being Γ_i, and the number of occurrences of the pattern Γ_s in it. For an uncorrelated collection of motifs, this summation vanishes. Hence, for $n > 2\gamma^*$, we have

$$\begin{aligned}
\mathbf{E}\left[Y^2_{n,\mathscr{C},\boldsymbol{\alpha}}\right] = \frac{2}{n-1}\left(\sum_{k=1}^{n-1} \left(\mathbf{E}\left[Y^2_{k,\mathscr{C},\boldsymbol{\alpha}}\right] + \mathbf{E}\left[Y_{k,\mathscr{C},\boldsymbol{\alpha}}\right]\,\mathbf{E}\left[Y_{n-k,\mathscr{C},\boldsymbol{\alpha}}\right]\right)\right. \\
\left. - \sum_{i\in\mathscr{I}} \left(\frac{\alpha_i^2 \mathcal{C}(\Gamma_i)}{2} + \alpha_i \mathbf{E}\left[Y_{n-\gamma_i,\mathscr{C},\boldsymbol{\alpha}}\right]\,\mathcal{C}(\Gamma_i)\right)\right).
\end{aligned}$$

Differencing the recurrence for $(n-2)\mathbf{E}[Y^2_{n-1,\mathscr{C}}]$ from that for $(n-1)\mathbf{E}[Y^2_{n,\mathscr{C}}]$, and using $2\,\mathbf{E}[Y_{n-1,\mathscr{C}}]\,\mathbf{E}[Y_{1,\mathscr{C}}] = 0$ (if the motif is not a single node, see the remark below), we simplify the recurrence to

$$\begin{aligned}(n-1)\mathbf{E}\left[Y^2_{n,\mathscr{C},\boldsymbol{\alpha}}\right] = n\mathbf{E}\left[Y^2_{n-1,\mathscr{C},\alpha}\right] &+ 2\sum_{k=1}^{n-2}\mathbf{E}\left[Y_{k,\mathscr{C},\boldsymbol{\alpha}}\right]\left(\mathbf{E}[Y_{n-k,\mathscr{C},\boldsymbol{\alpha}}] - \mathbf{E}[Y_{n-1-k,\mathscr{C},\boldsymbol{\alpha}}]\right)\\ &+ 2\sum_{i\in\mathscr{I}}\alpha_i\,\mathcal{C}(\Gamma_i)\left(\mathbf{E}[Y_{n-\gamma_i-1,\mathscr{C},\boldsymbol{\alpha}}] - \mathbf{E}[Y_{n-\gamma_i,\mathscr{C},\boldsymbol{\alpha}}]\right).\end{aligned}$$

Remark A motif consisting of a single node is correlated with any other motif. If an uncorrelated $\mathscr{C}$ contains a motif Γ that consists of just one node, it must be the only motif in the collection $\mathscr{C}$. In this case, $X_{n,\mathscr{C}}$ is just the number of leaves in a random recursive tree, which is well studied (see Dondajewski and Szymański 1982; Na and Rapoport 1970; Najock and Heyde 1982). Thus, a collection of two or more uncorrelated motifs cannot contain the single node. Throughout the rest of the paper, we consider collections $\mathscr{C}$ that do not have a motif consisting of a single node.

For $n > 2\gamma^* + 1$, we have a recurrence of the form

$$(n-1)\mathbf{E}\left[Y^2_{n,\mathscr{C},\boldsymbol{\alpha}}\right] = n\mathbf{E}\left[Y^2_{n-1,\mathscr{C},\boldsymbol{\alpha}}\right] + \mathscr{W}(n,\mathscr{C},\boldsymbol{\alpha}), \tag{3}$$

where

$$\begin{aligned}\mathscr{W}(n,\mathscr{C},\boldsymbol{\alpha}) = 2\sum_{k=1}^{n-2}\mathbf{E}\left[Y_{k,\mathscr{C},\boldsymbol{\alpha}}\right]\left(\mathbf{E}[Y_{n-k,\mathscr{C},\boldsymbol{\alpha}}] - \mathbf{E}[Y_{n-1-k,\mathscr{C},\boldsymbol{\alpha}}]\right)\\ + 2\sum_{i\in\mathscr{I}}\alpha_i\mathcal{C}(\Gamma_i)\left(\mathbf{E}[Y_{n-\gamma_i-1,\mathscr{C},\boldsymbol{\alpha}}] - \mathbf{E}[Y_{n-\gamma_i,\mathscr{C},\boldsymbol{\alpha}}]\right).\end{aligned}$$

We shall now evaluate $\mathscr{W}(n,\mathscr{C},\boldsymbol{\alpha})$ term by term:

$$\begin{aligned}&2\sum_{k=1}^{n-2}\mathbf{E}[Y_{k,\mathscr{C},\boldsymbol{\alpha}}]\left(\mathbf{E}[Y_{n-k,\mathscr{C},\boldsymbol{\alpha}}] - \mathbf{E}[Y_{n-1-k,\mathscr{C},\boldsymbol{\alpha}}]\right)\\ &\quad= 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i<k-1}}\frac{\alpha_i\mathcal{C}(\Gamma_i)(k)}{\gamma_i(\gamma_i+1)} + \sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k-1}}\frac{\alpha_i\mathcal{C}(\Gamma_i)}{\gamma_i} + \sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k}}\alpha_i\mathcal{C}(\Gamma_i)\right)\\ &\qquad\times\left(\sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k}}\alpha_j\mathcal{C}(\Gamma_j) + \sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k-1}}\frac{\alpha_j\mathcal{C}(\Gamma_j)(1-\gamma_j)}{\gamma_j} + \sum_{\substack{j\in\mathscr{I}\\ \gamma_j<n-k-1}}\frac{\alpha_j\mathcal{C}(\Gamma_j)}{\gamma_j(\gamma_j+1)}\right).\end{aligned}$$

This cross-product comprises nine terms, namely $a_1(n,\mathscr{C},\boldsymbol{\alpha}), a_2(n,\mathscr{C},\boldsymbol{\alpha}),\ldots,$ $a_9(n,\mathscr{C},\boldsymbol{\alpha})$. Their calculation varies in complexity, with $a_3(n,\mathscr{C},\boldsymbol{\alpha})$ being the most involved. We only show the fine details of how to evaluate $a_3(n,\mathscr{C},\boldsymbol{\alpha})$:

$$\begin{aligned}
a_3(n,\mathscr{C},\boldsymbol{\alpha}) &= 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i<k-1}}\frac{\alpha_i\,k\,\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)}\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j<n-k-1}}\frac{\alpha_j\mathcal{C}(\Gamma_j)}{\gamma_j(\gamma_j+1)}\right)\\
&= 2\sum_{k=1}^{n-2}\sum_{i,\,j\in\mathscr{I}}\frac{\alpha_i\,\alpha_j\,k\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)}\mathbf{1}_{\{\gamma_i<k-1\}}\mathbf{1}_{\{\gamma_j<n-k-1\}}\\
&= 2\sum_{k=\gamma_i+2}^{n-\gamma_j-2}\sum_{i,\,j\in\mathscr{I}}\frac{\alpha_i\,\alpha_j\,k\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)}\\
&= 2\sum_{i,\,j\in\mathscr{I}}\frac{\alpha_i\,\alpha_j\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{2\gamma_i\gamma_j(\gamma_i+1)(\gamma_j+1)}\mathbf{1}_{\{\gamma_i+2\le n-\gamma_j-2\}}\\
&\qquad\times\big((n-\gamma_j-2)(n-\gamma_j-1)-(\gamma_i+2)(\gamma_i+1)\big)\\
&= \sum_{i,\,j\in\mathscr{I}}\frac{\alpha_i\,\alpha_j\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{\gamma_i\gamma_j(\gamma_i+1)(\gamma_j+1)}\mathbf{1}_{\{n>\gamma_i+\gamma_j+2\}}\\
&\qquad\times\big(n^2-n(2\gamma_j+3)+\gamma_j^2-\gamma_i^2+3(\gamma_j-\gamma_i)\big).
\end{aligned}$$

The other eight terms are relatively similar, so we omit the finer details of their evaluation. We compute

$$\begin{aligned}
a_1(n,\mathscr{C},\boldsymbol{\alpha}) &= 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ [-1\,pt]\gamma_i<k-1}}\frac{\alpha_i\,k\,\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)}\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k}}\alpha_j\mathcal{C}(\Gamma_j)\right)\\
&= 2\sum_{i,\,j\in\mathscr{I}}\frac{\alpha_i\,\alpha_j\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)(n-\gamma_j)}{\gamma_i(\gamma_i+1)}\mathbf{1}_{\{n>\gamma_i+\gamma_j+1\}},\\
a_2(n,\mathscr{C},\boldsymbol{\alpha}) &= 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i<k-1}}\frac{\alpha_i\,k\,\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)}\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k-1}}\alpha_j\left(\frac{\mathcal{C}(\Gamma_j)}{\gamma_j}-\mathcal{C}(\Gamma_j)\right)\right)\\
&= 2\sum_{i,\,j\in\mathscr{I}}\frac{\alpha_i\,\alpha_j\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)(1-\gamma_j)(n-\gamma_j-1)}{\gamma_i\gamma_j(\gamma_i+1)}\mathbf{1}_{\{n>\gamma_i+\gamma_j+2\}},\\
a_4(n,\mathscr{C},\boldsymbol{\alpha}) &= 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k-1}}\frac{\alpha_i\,\mathcal{C}(\Gamma_i)}{\gamma_i}\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k}}\alpha_j\mathcal{C}(\Gamma_j)\right)\\
&= 2\sum_{i,\,j\in\mathscr{I}}\frac{\alpha_i\,\alpha_j\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{\gamma_i}\mathbf{1}_{\{n=\gamma_i+\gamma_j+1\}},
\end{aligned}$$

$$a_5(n,\mathscr{C},\boldsymbol{\alpha}) = 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k-1}}\frac{\alpha_i\mathcal{C}(\Gamma_i)}{\gamma_i}\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k-1}}\alpha_j\left(\frac{\mathcal{C}(\Gamma_j)}{\gamma_j}-\mathcal{C}(\Gamma_j)\right)\right)$$

$$= 2\sum_{i,j\in\mathscr{I}}\frac{\alpha_i\alpha_j\mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)(1-\gamma_j)}{\gamma_i\gamma_j}\mathbf{1}_{\{n=\gamma_j+\gamma_i+2\}},$$

$$a_6(n,\mathscr{C},\boldsymbol{\alpha}) = 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k-1}}\frac{\alpha_i\mathcal{C}(\Gamma_i)}{\gamma_i}\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j<n-k-1}}\frac{\alpha_j\mathcal{C}(\Gamma_j)}{\gamma_j(\gamma_j+1)}\right)$$

$$= 2\sum_{i,j\in\mathscr{I}}\frac{\alpha_i\alpha_j\mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)}{\gamma_i\gamma_j(\gamma_j+1)}\mathbf{1}_{\{n>\gamma_j+\gamma_i+2\}},$$

$$a_7(n,\mathscr{C},\boldsymbol{\alpha}) = 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k}}\alpha_i\mathcal{C}(\Gamma_i)\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k}}\alpha_j\mathcal{C}(\Gamma_j)\right)$$

$$= 2\sum_{i,j\in\mathscr{I}}\alpha_i\alpha_j\mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)\mathbf{1}_{\{n=\gamma_j+\gamma_i\}},$$

$$a_8(n,\mathscr{C},\boldsymbol{\alpha}) = 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k}}\alpha_i\mathcal{C}(\Gamma_i)\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j=n-k-1}}\alpha_j\left(\frac{\mathcal{C}(\Gamma_j)}{\gamma_j}-\mathcal{C}(\Gamma_j)\right)\right)$$

$$= 2\sum_{i,j\in\mathscr{I}}\frac{\alpha_i\alpha_j\mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)(1-\gamma_j)}{\gamma_j}\mathbf{1}_{\{n=\gamma_i+\gamma_j+1\}},$$

$$a_9(n,\mathscr{C},\boldsymbol{\alpha}) = 2\sum_{k=1}^{n-2}\left(\sum_{\substack{i\in\mathscr{I}\\ \gamma_i=k}}\alpha_i\mathcal{C}(\Gamma_i)\times\sum_{\substack{j\in\mathscr{I}\\ \gamma_j<n-k-1}}\frac{\alpha_j\mathcal{C}(\Gamma_j)}{\gamma_j(\gamma_j+1)}\right)$$

$$= 2\sum_{i,j\in\mathscr{I}}\frac{\alpha_i\alpha_j\mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)}{\gamma_j(\gamma_j+1)}\mathbf{1}_{\{n>\gamma_i+\gamma_j+1\}}.$$

If we define $a_{10}(n,\mathscr{C},\boldsymbol{\alpha}) = 2\sum_{i\in\mathscr{I}}\alpha_i\mathcal{C}(\Gamma_i)\left(\mathbf{E}[Y_{n-\gamma_i-1,\mathscr{C}}]-\mathbf{E}[Y_{n-\gamma_i,\mathscr{C}}]\right)$, then along the same lines, we have

$$a_{10}(n,\mathscr{C},\boldsymbol{\alpha}) = -2\sum_{i,j\in\mathscr{I}}\alpha_i\alpha_j\mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)\left(\mathbf{1}_{\{n=\gamma_i+\gamma_j\}}+\frac{1-\gamma_i}{\gamma_i}\mathbf{1}_{\{n=\gamma_i+\gamma_j+1\}}\right.$$
$$\left.+\frac{1}{\gamma_i(\gamma_i+1)}\mathbf{1}_{\{\gamma_i+\gamma_j+1<n\}}\right).$$

Assembling the ten terms, we get a complicated expression involving many indicators for $\mathscr{W}(n, \mathscr{C}, \boldsymbol{\alpha})$. However, this expression simplifies greatly for $n > 2\gamma^* + 2$; it becomes

$$\mathscr{W}(n, \mathscr{C}, \boldsymbol{\alpha}) = \left(\sum_{i,j \in \mathscr{I}} \frac{\alpha_i \alpha_j \mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)} \right) n(n-1), \qquad \text{when } n > 2\gamma^* + 2, \quad (4)$$

and at $n = 2\gamma^* + 2$, we have

$$\begin{aligned}
\mathscr{W}(2\gamma^* + 2, \mathscr{C}, \boldsymbol{\alpha}) = {} & 2 \sum_{i,j \in \mathscr{I}} \frac{\alpha_i \alpha_j \mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)(2\gamma^* + 2 - \gamma_j)}{\gamma_i(\gamma_i+1)} \\
& + 2 \sum_{i,j \in \mathscr{I}} \frac{\alpha_i \alpha_j \mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)(1-\gamma_j)}{\gamma_i \gamma_j} \mathbf{1}_{\{\gamma_i = \gamma_j = \gamma^*\}} \\
& + \sum_{i,j \in \mathscr{I}} \frac{\alpha_i \alpha_j \mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)(2\gamma^* + 2)(2\gamma^* + 1)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)} \mathbf{1}_{\{2\gamma^* > \gamma_i + \gamma_j\}} \\
& + \sum_{i,j \in \mathscr{I}} \frac{\alpha_i \alpha_j \mathcal{C}(\Gamma_i)\mathcal{C}(\Gamma_j)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)} \mathbf{1}_{\{2\gamma^* > \gamma_i + \gamma_j\}} \\
& \quad \times \left(2\gamma_j^3 - 4\gamma^* \gamma_j^2 - \gamma_j^2 - 4\gamma^* \gamma_j - \gamma_i^2 - \gamma_i - 3\gamma_j \right). \quad (5)
\end{aligned}$$

Note that both Eqs. 4 and 5 are functions of our collection and can be computed exactly for any given collection $\mathscr{C}$ and a given $\boldsymbol{\alpha}$.

Unwinding recurrence (Eq. 3), and using Eqs. 4 and 5, for $n > 2\gamma^*$ we get

$$\begin{aligned}
\mathbf{E}\left[Y^2_{n,\mathscr{C},\boldsymbol{\alpha}}\right] &= \frac{n}{n-1}\mathbf{E}\left[Y^2_{n-1,\mathscr{C},\boldsymbol{\alpha}}\right] + \frac{1}{n-1}\mathscr{W}(n, \mathscr{C}, \boldsymbol{\alpha}) \\
&= \frac{n}{2\gamma^*+1}\mathbf{E}\left[Y^2_{2\gamma^*+1,\mathscr{C},\boldsymbol{\alpha}}\right] + n \sum_{j=2\gamma^*+2}^{n} \frac{\mathscr{W}(j, \mathscr{C}, \boldsymbol{\alpha})}{j(j-1)} \\
&= \frac{n}{2\gamma^*+1}\mathbf{E}\left[Y^2_{2\gamma^*+1,\mathscr{C},\boldsymbol{\alpha}}\right] + n\frac{\mathscr{W}(2\gamma^*+2, \mathscr{C}, \boldsymbol{\alpha})}{(2\gamma^*+2)(2\gamma^*+1)} \\
&\quad + n \sum_{j=2\gamma^*+3}^{n} \frac{\mathscr{W}(j, \mathscr{C}, \boldsymbol{\alpha})}{j(j-1)} \\
&= \left(\sum_{i \in \mathscr{I}} \frac{\alpha_i \mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)} \right)^2 n^2 + \sigma^2_{\mathscr{C},\boldsymbol{\alpha}} n,
\end{aligned}$$

where

$$\sigma^2_{\mathscr{C},\boldsymbol{\alpha}} = \frac{\mathbf{E}[Y^2_{2\gamma^*+1,\mathscr{C},\boldsymbol{\alpha}}]}{2\gamma^*+1} + \frac{\mathscr{W}(2\gamma^*+2, \mathscr{C}, \boldsymbol{\alpha})}{(2\gamma^*+2)(2\gamma^*+1)} - (2\gamma^*+2)\left(\sum_{i \in \mathscr{I}} \frac{\alpha_i \mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)} \right)^2.$$

We need to evaluate all the terms for $n > 2\gamma^* + 1$. If our collection includes only one motif of size γ, then $\gamma^* = \gamma$, and the variance matches the calculation in Feng and Mahmoud (2010). We thus have the variances (diagonal elements of the covariance matrix) in the form

$$\mathbf{Var}[X_{n,\Gamma_i}] = \left(\frac{(\gamma_i+1)(2\gamma_i+1) - (3\gamma_i+2)\,\mathcal{C}(\Gamma_i)}{\gamma_i(\gamma_i+1)^2(2\gamma_i+1)} \right) \mathcal{C}(\Gamma_i),$$

for $n > 2\gamma^* = 2\gamma_i$.

As our result is valid for all real α_i (not all zero), we can generate the elements of the covariance matrix, by using the variance–covariance relation

$$\mathbf{Var}\left[X_{n,\Gamma_i} + X_{n,\Gamma_j}\right] = \mathbf{Var}\left[X_{n,\Gamma_i}\right] + \mathbf{Var}\left[X_{n,\Gamma_j}\right] + 2\,\mathbf{Cov}\left[X_{n,\Gamma_i}, X_{n,\Gamma_j}\right].$$

We call $\mathbf{b}_{i,j}$ the vector $\boldsymbol{\alpha}$ with all the entries equal to 0, except α_i and α_j, which we set to 1. When $i \neq j$, an off-diagonal entry in the covariance matrix is

$$\begin{aligned}
(\boldsymbol{\Sigma}_{\mathscr{C}})_{i,j} &= \mathbf{Cov}\left[X_{n,\Gamma_i}, X_{n,\Gamma_j}\right] \\
&= \frac{1}{2}\left(\mathbf{Var}\left[X_{n,\Gamma_i} + X_{n,\Gamma_j}\right] - \mathbf{Var}\left[X_{n,\Gamma_i}\right] - \mathbf{Var}\left[X_{n,\Gamma_j}\right]\right) \\
&= \frac{n}{2}\left(\frac{\mathbf{E}[Y^2_{2\gamma^*_{i,j}+1,\mathscr{C},\mathbf{b}_{i,j}}]}{2\gamma^*_{i,j}+1} + \frac{\mathscr{W}(2\gamma^*_{i,j}+2,\mathscr{C},\mathbf{b}_{i,j})}{(2\gamma^*_{i,j}+2)(2\gamma^*_{i,j}+1)} - \frac{\mathcal{C}^2(\Gamma_i)(2\gamma^*_{i,j}+2)}{\gamma_i^2(\gamma_i+1)^2} \right. \\
&\quad - \frac{\mathcal{C}^2(\Gamma_j)(2\gamma^*_{i,j}+2)}{\gamma_j^2(\gamma_j+1)^2} - \frac{2\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)(2\gamma^*_{i,j}+2)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)} - \frac{\mathbf{E}[X^2_{2\gamma^*_{i,j}+1,\Gamma_i}]}{2\gamma^*_{i,j}+1} \\
&\quad \left. + \frac{(2\gamma^*_{i,j}+1)\mathcal{C}(\Gamma_i)^2}{\gamma_i^2(1+\gamma_j)^2} - \frac{\mathbf{E}[X^2_{2\gamma^*_{i,j}+1,\Gamma_j}]}{2\gamma^*_{i,j}+1} + \frac{(2\gamma^*_{i,j}+1)\mathcal{C}(\Gamma_j)^2}{\gamma_i^2(1+\gamma_j)^2} \right),
\end{aligned}$$

for $n > 2\gamma^*_{i,j} + 1$. Expanding $\mathbf{E}[Y^2_{2\gamma^*_{i,j}+1,\mathscr{C},\mathbf{b}_{i,j}}]$, we see that the second moments of $X_{2\gamma^*_{i,j}+1,\Gamma_i}$ and $X_{2\gamma^*_{i,j}+1,\Gamma_j}$ cancel, and thus we get

$$\begin{aligned}
(\boldsymbol{\Sigma}_{\mathscr{C}})_{i,j} = \frac{n}{2}\left(\frac{2\mathbf{E}[X_{2\gamma^*_{i,j}+1,\Gamma_i} X_{2\gamma^*_{i,j}+1,\Gamma_j}]}{2\gamma^*_{i,j}+1} + \frac{\mathscr{W}(2\gamma^*_{i,j}+2,\mathscr{C},\mathbf{b}_{i,j})}{(2\gamma^*_{i,j}+2)(2\gamma^*_{i,j}+1)} \right. \\
\left. - \frac{\mathcal{C}^2(\Gamma_i)}{\gamma_i^2(\gamma_i+1)^2} - \frac{\mathcal{C}^2(\Gamma_j)}{\gamma_j^2(\gamma_j+1)^2} - \frac{2(2\gamma^*_{i,j}+2)\,\mathcal{C}(\Gamma_i)\,\mathcal{C}(\Gamma_j)}{\gamma_i(\gamma_i+1)\gamma_j(\gamma_j+1)} \right),
\end{aligned}$$

for $n > 2\gamma^*_{i,j} + 1$.

This completes the proof of Theorem 1.

Remark The derivation of the variance in Feng and Mahmoud (2010) contains a couple of misprints, where the term $E[R_\gamma]$ inadvertently appears on lines 11 and 17 of Section 5.2, and should be omitted. Nevertheless, their final result is correct and matches the result we derive here.

7.4 Limit Distributions for Varieties of a Fixed Size

In principle, one can continue pumping higher moments of the linear combination by the recurrence methods utilized for the mean and variance, and attempt to determine limit distributions by a method of recursive moments (see Chern and Hwang 2001, for example). The calculation in each higher moment is much more involved than in the previous one. The variance calculation is already complicated enough. The mounting complexity in higher moments would be forbidding. Alternatively, we apply the contraction method, which is transparent for limits.

The contraction method was introduced in Rösler (1999) to analyze the Quick Sort algorithm, and it soon became a popular method, because of the transparency of structure that it provides in the limit. Several useful extensions are added in Rachev and Rüschendorf (1995). General contraction theorems and multivariate extensions appear in Rösler and Rüschendorf (2001) and Neininger (2001). The latter reference deals with the specific context of recursive trees. A general theorem that covers a broad scope of applications is given in Neininger and Rüschendorf (2004). A valuable survey appears in Rösler and Rüschendorf (2001).

According to the definition of a multivariate normal distribution in infinite dimensions, it suffices to consider arbitrary finite collections of size (say) $r \geq 1$ motifs. Also, take $\boldsymbol{\alpha}$ to be an arbitrary vector of r real numbers, not all zero. Let

$$\mathbf{E}[Y_{n,\mathscr{C},\boldsymbol{\alpha}}] =: \mu_{\mathscr{C},\boldsymbol{\alpha}}\, n = n \sum_{i\in\mathscr{I}} \alpha_i (\boldsymbol{\mu}_{\mathscr{C}})_i,$$

$$\mathbf{Var}[Y_{n,\mathscr{C},\boldsymbol{\alpha}}] =: \sigma^2_{\mathscr{C},\boldsymbol{\alpha}}\, n = n \sum_{i\in\mathscr{I}} \sum_{j\in\mathscr{I}} \alpha_i\, \alpha_j\, \mathbf{Cov}[X_{n,\Gamma_i} X_{n,\Gamma_j}];$$

the coefficients $\mu_{\mathscr{C},\boldsymbol{\alpha}}$ and $\sigma^2_{\mathscr{C},\boldsymbol{\alpha}}$, are functions of $\alpha_1, \ldots, \alpha_r$. We start from the recursive representation (Eq. 2), normalized in the centered and scaled form

$$\begin{aligned}\frac{Y_{n,\mathscr{C},\boldsymbol{\alpha}} - \mu_{\mathscr{C},\boldsymbol{\alpha}} n}{\sqrt{n}} &= \frac{Y_{U_n,\mathscr{C},\boldsymbol{\alpha}} - \mu_{\mathscr{C},\boldsymbol{\alpha}} U_n}{\sqrt{U_n}} \sqrt{\frac{U_n}{n}} \\ &\quad + \frac{Y_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} - \mu_{\mathscr{C},\boldsymbol{\alpha}} (n-U_n)}{\sqrt{n-U_n}} \sqrt{\frac{n-U_n}{n}} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \operatorname{Ber}\big(\mathcal{C}(\Gamma_i)\big).\end{aligned}$$

Let

$$Y^*_{n,\mathscr{C},\boldsymbol{\alpha}} := \frac{Y_{n,\mathscr{C},\boldsymbol{\alpha}} - \mu_{\mathscr{C},\boldsymbol{\alpha}}\, n}{\sqrt{n}}.$$

To give an insight in the inner working of the contraction method, we first find the limit heuristically for a finite collection of motifs. Later we prove a Gaussian law. The recurrence equation for the normalized random variables can be written as

$$Y^*_{n,\mathscr{C},\,\boldsymbol{\alpha}} \stackrel{\mathcal{D}}{=} Y^*_{U_n,\mathscr{C},\boldsymbol{\alpha}} \sqrt{\frac{U_n}{n}} + \tilde{Y}^*_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} \sqrt{\frac{n-U_n}{n}} + \frac{\xi_{\mathscr{C},\boldsymbol{\alpha}}(n)}{\sqrt{n}}, \tag{6}$$

where

$$\xi_{\mathscr{C},\boldsymbol{\alpha}}(n) := \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \operatorname{Ber}\big(\mathcal{C}(\Gamma_i)\big)$$

is a bounded toll function, as we have

$$\xi_{\mathscr{C},\boldsymbol{\alpha}}(n) := \sum_{i\in\mathscr{I}} \alpha_i \mathbf{1}_{\{n-U_n=\gamma_i\}} \operatorname{Ber}\big(\mathcal{C}(\Gamma_i)\big) \le \sum_{i\in\mathscr{I}} \alpha_i = |\mathscr{I}| \max_{i\in\mathscr{I}} \alpha_i = O(1).$$

Recall that for the heuristic argument we are considering a finite indexing set, and the maximum in the last expression is merely a number.

If $Y^*_{n,\mathscr{C},\boldsymbol{\alpha}}$ converges to a limit $Y^*_{\mathscr{C},\boldsymbol{\alpha}}$, so would $Y^*_{U_n,\mathscr{C},\boldsymbol{\alpha}}$ and $\tilde{Y}^*_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}$, because both U_n and $n-U_n$ grow to infinity almost surely. The terms on the right-hand side in the representation (Eq. 6) are dependent. However, the correlation between any pair of them gets weaker as n increases, till ultimately their limits become independent. As is well known, we have

$$\frac{U_n}{n} \xrightarrow{\mathcal{P}} U,$$

where U is a standard continuous Uniform(0,1). Subsequently,

$$\sqrt{\frac{U_n}{n}} \xrightarrow{\mathcal{P}} \sqrt{U}, \qquad \sqrt{\frac{n-U_n}{n}} \xrightarrow{\mathcal{P}} \sqrt{1-U}.$$

The limit would satisfy the distributional equation

$$Y^*_{\mathscr{C},\boldsymbol{\alpha}} \overset{\mathcal{D}}{=} Y^*_{\mathscr{C},\boldsymbol{\alpha}}\sqrt{U} + \tilde{Y}^*_{\mathscr{C},\boldsymbol{\alpha}}\sqrt{1-U}.$$

A distributional equation of the latter form has the normal distribution as a solution (see Rösler and Rüschendorf 2001). Such a solution is unique, because it is the fixed-point solution of a contraction operator on distances in a metric space on distribution functions.

We shall next give a formal proof of joint asymptotic normality. Suppose H_1 and H_2 are two random variables, with distribution functions F_{H_1} and F_{H_2}, respectively. Recall the Maejimam-Rachev metric (Maejima and Rachev 1987) of order 3:

$$d_3(F_{H_1}, F_{H_2}) = \sup\big\{\big|\mathbf{E}\big[g(H_1) - g(H_2)\big]\big| : ||g^{(3)}||_\infty \le 1\big\},$$

where the supremum is taken over every three-times-differentiable function $g(\cdot)$, and $||\cdot||_\infty$ is the essential supremum.

Let V_n be a random variable satisfying the recurrence

$$V_n = V_{A_n} + \tilde{V}_{n-A_n} + B_n,$$

where A_n and B_n are sequences of random variables. It is proved in Rachev and Rüschendorf (1995), that $V^*_n = (V_n - \mathbf{E}[V_n])/\sqrt{\mathbf{Var}[V_n]}$ is asymptotically the

standard normal, via the distance calculation $d_s(F_{V_n^*}, F_{\mathcal{N}(0,1)}) \to 0$, if the following conditions hold:

(i) $B_n/\sqrt{n} \to 0$.
(ii) $\mathbf{Var}[V_n^*]$ converges to some $v^2 > 0$.
(iii) $\sup_n \mathbf{E}[|V_n^*|^3] < \infty$.
(iv) $A_n/n \xrightarrow{\mathcal{P}} A$, with $\mathbf{E}[A] > 0$, and

$$\limsup_{n\to\infty} \mathbf{E}\left[\left(\frac{A_n}{n}\right)^{3/2} + \left(\frac{n-A_n}{n}\right)^{3/2}\right] < 1.$$

For a proof see Theorem 3.1 in Rachev and Rüschendorf (1995), and the remarks following the proof, particularly their display (3.25). In our case, V_n is $Y^*_{n,\mathscr{C},\boldsymbol{\alpha}}$, and B_n is $\xi_{\mathscr{C},\boldsymbol{\alpha}}(n) = O(1)$; condition (i) is satisfied. According to Theorem 1, we have $\mathbf{Var}[Y^*_{n,\mathscr{C},\boldsymbol{\alpha}}] \to \sigma^2_{\mathscr{C},\boldsymbol{\alpha}} > 0$, and condition (ii) is satisfied.

For condition (iii), we first formulate a recurrence. Let $M_{n,\mathscr{C},\boldsymbol{\alpha}} = Y_{n,\mathscr{C},\boldsymbol{\alpha}} - \mathbf{E}[Y_{n,\mathscr{C},\boldsymbol{\alpha}}]$; note that $M_{n,\mathscr{C},\boldsymbol{\alpha}}/\sqrt{n}$ is $Y^*_{n,\mathscr{C},\boldsymbol{\alpha}}$. From the stochastic recurrence (2) we can then write a recurrence for absolute third moments:

$$\mathbf{E}\left[|M_{n,\mathscr{C},\boldsymbol{\alpha}}|^3\right] = \mathbf{E}\left[|M_{U_n,\mathscr{C},\boldsymbol{\alpha}} + \tilde{M}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}} + \xi_{\mathscr{C},\boldsymbol{\alpha}}(n)|^3\right].$$

Via the triangle inequality, we first write

$$\mathbf{E}\left[|M_{n,\mathscr{C},\boldsymbol{\alpha}}|^3\right] \le \mathbf{E}\left[\left(|M_{U_n,\mathscr{C},\boldsymbol{\alpha}}| + |\tilde{M}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}| + |\xi_{\mathscr{C},\boldsymbol{\alpha}}(n)|\right)^3\right].$$

Next, we expand the cubic term, which results in ten terms, two of which are recursive, and the rest are $O(n^{3/2})$. The calculation for the O terms are all similar; we argue a couple and omit the rest. For instance, a bound on $\mathbf{E}[|M^2_{U_n,\mathscr{C},\boldsymbol{\alpha}}\tilde{M}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}|]$ follows from the conditional independence, and the bound $\mathbf{E}[|\tilde{M}_{n,\mathscr{C},\boldsymbol{\alpha}}|] \le \sqrt{\mathbf{E}[\tilde{M}^2_{n,\mathscr{C},\boldsymbol{\alpha}}]}$, provided by Jensen's inequality. For this cross-product term we can use the conditional independence of $Y_{U_n,\mathscr{C},\boldsymbol{\alpha}}$ and $\tilde{Y}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}$ (given U_n) to write

$$\begin{aligned}\mathbf{E}\left[|M^2_{U_n,\mathscr{C},\boldsymbol{\alpha}}\,\tilde{M}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}|\right] &= \frac{1}{n-1}\sum_{k=1}^{n-1} \mathbf{E}\left[M^2_{k,\mathscr{C},\boldsymbol{\alpha}}\right]\mathbf{E}\left[|\tilde{M}_{n-k,\mathscr{C},\boldsymbol{\alpha}}|\right] \\ &\le \frac{1}{n-1}\sum_{k=1}^{n-1} \mathbf{Var}[Y^2_{k,\mathscr{C},\boldsymbol{\alpha}}]\sqrt{\mathbf{Var}[Y^2_{n-k,\mathscr{C},\boldsymbol{\alpha}}]}.\end{aligned}$$

Theorem 1 asserts that the variances in the sum are all linear, and we have

$$\mathbf{E}\left[|M^2_{U_n,\mathscr{C},\boldsymbol{\alpha}}\,\tilde{M}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}|\right] \le \frac{\sigma^3_{\mathscr{C},\boldsymbol{\alpha}}}{n-1}\sum_{k=1}^{n-1} k\sqrt{n-k} = O\big(n^{3/2}\big).$$

A cross-product involving $\xi_{\mathscr{C},\boldsymbol{\alpha}}(n)$, when conditioned on U_n, gives only one term in the sum. For instance, we have

$$\begin{aligned}\mathbf{E}\big[|M_{U_n,\mathscr{C},\boldsymbol{\alpha}}\,\tilde{M}_{n-U_n,\mathscr{C},\boldsymbol{\alpha}}\,\xi_{\mathscr{C},\boldsymbol{\alpha}}(n)|\big] &\le \frac{1}{n-1}\sum_{k=1}^{n-1}\mathbf{E}\big[|M_{k,\mathscr{C},\boldsymbol{\alpha}}|\big]\,\mathbf{E}\left[\left|\tilde{M}_{n-k,\mathscr{C},\boldsymbol{\alpha}}\right.\right.\\ &\qquad\left.\left.\times\sum_{i\in\mathscr{I}}\alpha_i\mathbf{1}_{\{n-k=\gamma_i\}}\,\mathrm{Ber}\big(\mathcal{C}(\Gamma_i)\big)\right|\right]\\ &\le \frac{1}{n-1}\sum_{i\in\mathscr{I}}\mathbf{E}\big[|M_{n-\gamma_i,\mathscr{C},\boldsymbol{\alpha}}|\big]\,|\alpha_i|\,\mathbf{E}\big[|\tilde{M}_{\gamma_i,\mathscr{C},\boldsymbol{\alpha}}|\big]\\ &\le \frac{1}{n-1}\sum_{i,\,j\in\mathscr{I}}2\,n\,|\alpha_j|\,|\alpha_i|\,\mathbf{E}\big[|\tilde{M}_{\gamma_i,\mathscr{C},\boldsymbol{\alpha}}|\big]\\ &= O(1).\end{aligned}$$

All the other cross-product terms are $O(n)$. We thus have an asymptotic recurrence:

$$\mathbf{E}\big[|M_{n,\mathscr{C},\boldsymbol{\alpha}}|^3\big] = \frac{2}{n-1}\sum_{k=1}^{n-1}\mathbf{E}\big[|M_{k,\mathscr{C},\boldsymbol{\alpha}}|^3\big] + O(n^{3/2}).$$

A solution for such a recurrence can be obtained by the differencing method we applied to the mean and the variance, and we get

$$\mathbf{E}\left[|Y^*_{n,\mathscr{C},\boldsymbol{\alpha}}|^3\right] = \mathbf{E}\left[\left|\frac{M_{n,\mathscr{C},\boldsymbol{\alpha}}}{\sqrt{n}}\right|^3\right] = O(1);$$

condition (iii) is verified.

In our case A_n is U_n, which is the random variable that is uniformly distributed on the set $\{1,\ldots,n-1\}$. And so, $U_n/n \xrightarrow{\mathcal{D}} U$, where U is the standard continuous Uniform (0, 1) random variable with average $\mathbf{E}[U] = \frac{1}{2}$. Subsequently, we have the computation

$$\limsup_{n\to\infty}\mathbf{E}\left[(U_n/n)^{3/2} + ((n-U_n)/n)^{3/2}\right] = 4/5 < 1,$$

and condition (iv) is satisfied. This completes the proof of Theorem 2.

8 Examples

In this section we discuss two illustrative examples, one with a finite collection of motifs, and one with an infinite collection.

8.1 All the Motifs of Size 4

Consider $\mathscr{C}$ to be all motifs of size 4, as depicted in Fig. 2. There are four such motifs (shown in Fig. 2). Let us call them from left to right $T_1,\ldots,T_4$.

These motifs have shape functionals

$$\mathcal{C}(T_1) = \frac{1}{6}, \qquad \mathcal{C}(T_2) = \frac{1}{6}, \qquad \mathcal{C}(T_3) = \frac{3}{6}, \qquad \mathcal{C}(T_4) = \frac{1}{6}.$$

Hence Theorem 2 has the following realization:

$$\frac{\mathbf{X}_{n,\mathscr{C}} - \begin{pmatrix} 1 \\ 1 \\ 3 \\ 1 \end{pmatrix} \frac{n}{120}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}_4 \left(\mathbf{0}, \frac{1}{16200} \begin{pmatrix} 128 & -7 & -21 & -7 \\ -7 & 128 & -21 & -7 \\ -21 & -21 & 342 & -21 \\ -7 & -7 & -21 & 128 \end{pmatrix} \right).$$

8.2 A Collection of Rooted Stars

Suppose our collection consists of all rooted star trees:

$$\mathscr{C} = \{S_2, S_3, \ldots\},$$

where S_i is the rooted star of size i, consisting of a root and $i-1$ leaves. (The instance S_4 is the rightmost motif in Fig. 2.) Observe that we disallowed S_1, the rooted tree consisting of a single node (root), as it is correlated with any other S_i, for any $i \geq 2$. These stars have shape functionals

$$\mathcal{C}(S_i) = \frac{1}{(i-1)!}.$$

Whence, the vector of counts (indexed starting at 2) has a countably infinite number of components and Theorem 2 realizes the form:

$$\frac{\mathbf{X}_{n,\mathscr{C}} - \begin{pmatrix} 1/3! \\ 1/4! \\ 1/5! \\ \vdots \end{pmatrix} n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}_\infty \left(\mathbf{0}, \begin{pmatrix} 7/90 & -1/36 & -239/5880 & \ldots \\ -1/36 & 15/448 & -11/5760 & \ldots \\ -239/5880 & -11/5760 & 16/2025 & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \right).$$

9 Simulations and Validation

We performed a simulation study on Example 8.1 to validate our claims in Theorem 2 and empirically test the speed of convergence. We generated random recursive

Table 1 Comparison to normal probability($\mu \pm \sigma/2$)

n	Empirical probability − Actual probability
100	0.1669
1000	0.0546
10000	0.0104
100000	0.0018

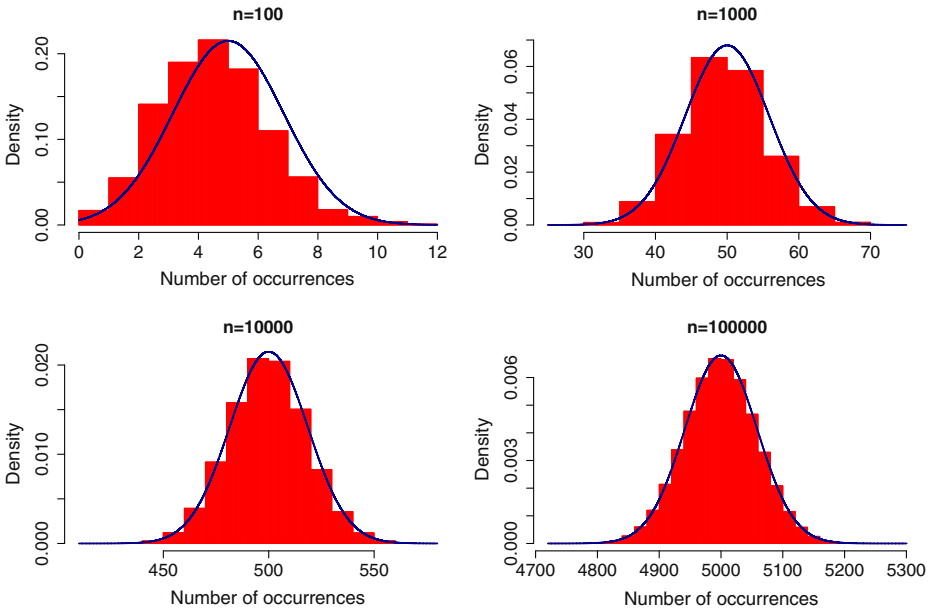

Fig. 3 Plots showing sum of occurrences of the motifs in Fig. 2 converging to normality

trees of sizes $n = 100$, 1000, 10000, and 100000. Generating all the $(n-1)!$ would be computationally expensive, so we sampled $10n$ of them randomly. We counted the sum of the number of occurrences of motifs T_1, T_2, T_3, and T_4 on the fringe. We then compared the analytic asymptotic normal probability of lying close to the mean (within half a standard deviation) with the one estimated by our simulations. These results are tabulated in Table 1, and we find that our analytic results are supported very closely by the simulations; the difference is already as low as about 1 % for trees of the moderate size $n = 10000$, and drops ten fold as n is ten times larger. Figure 3 shows the histograms of the simulated counts superimposed on the asymptotic normal density curve. It is very evident that as $n \to +\infty$ the distribution approaches normality.

10 Concluding Remarks

We presented a multivariate central limit theorem for the number of subtrees on the fringe of a random recursive tree that match a collection of given motifs. A natural question to ask is *How different would the result be, if the matching is made everywhere in the recursive tree, not only on the fringe?*

Some of the results in the fringe analysis will be preserved, though complications may arise because of matches at the root. We can write a recurrence to collect the total number of occurrences of a motif everywhere in a tree, by collecting the contributions from the special and nonspecial trees of the recursive tree, plus an indicator signifying the occurrence of the motif at the root. The latter indicator is in

general complicated. Nevertheless, it can be dealt with explicitly for motifs of small simple structure.

Take for instance S_3 as motif; the rooted tree of size 3 is sometimes called a *cherry* (McKenzie and Steel 2000). The illustration in the example in Section 8.2 tells us that $\mathbf{E}[X_{n,S_3}] = \frac{1}{24}n$. How would this result be different, if we searched for matches everywhere in the recursive tree? Let Q_{n,S_3} be the number of occurrences of a cherry in a random recursive tree. As mentioned, we collect the number of occurrences from the special subtree (Q_{U_n,S_3}) and the nonspecial tree ($\tilde{Q}_{n-U_n,S_3}$) and we need to adjust by additional unaccounted for cherries at the root. All the cherries in the nonspecial tree have been counted in $\tilde{Q}_{n-U_n,S_3}$. We only need to add the number of cherries at the root of the recursive tree that include the special edge (connecting the nodes labelled 1 and 2). If R_n is the degree of the root of the tree, the special edge forms $R_n - 1$ unaccounted for cherries. That is, the stochastic recurrence is

$$Q_{n,S_3} = Q_{U_n,S_3} + \tilde{Q}_{n-U_n,S_3} + R_n - 1.$$

The distribution of R_n is well known (Szymański 1987). For the average, we can plug in the needed average of R_n and construct arguments following the same lines we used in the fringe analysis (differencing then solving recurrences, etc.). These give

$$\mathbf{E}[Q_{n,S_3}] = n - H_{n-1} - 1,$$

where H_n is the harmonic number $\sum_{i=1}^{n} 1/i \sim \ln n$. Note that, on average, the number of matching S_3's everywhere in the tree is considerably larger than the number matching only on the fringe.

Acknowledgements This research was done while the second author was visiting Purdue University. The support H. Mahmoud received from Purdue's Department of Statistics is sincerely appreciated. Special thanks are due to the host of the visit, M. D. Ward, for his great hospitality. The authors thank Anirban DasGupta (of Purdue University) and Robert Smythe (of Oregon State University) for advice on some technical points.

References

Bergeron F, Flajolet P, Salvy B (1992) Varieties of increasing trees. In: Raoult JC (ed) Proceedings of the 17th colloquium on trees in algebra and programming (CAAP '92). Lecture Notes in Computer Science, vol 581. Springer, Berlin/Heidelberg, pp 24–48

Billingsley P (1995) Probability and measure, 3rd edn. Wiley-Interscience

Chern HH, Hwang HK (2001) Phase changes in random m-ary search trees and generalized quicksort. Random Struct Algorithms 19:316–358

Dondajewski M, Szymański J (1982) On the distribution of vertex-degrees in a strata of a random recursive tree. Bulletin de l'Académie Polonaise des Sciences 30:205–209

Feng Q, Mahmoud HM (2010) On the variety of shapes on the fringe of a random recursive tree. J Appl Probab 47:191–200

Flajolet P, Gourdon X, Martínez C (1997) Patterns in random binary search trees. Random Struct Algorithms 11:223–244

van der Hofstad R, Hooghiemstra G, Van Mieghem P (2002) On the covariance of the level sizes in random recursive trees. Random Struct Algorithms 20:519–539

Maejima M, Rachev ST (1987) An ideal metric and the rate of convergence to a self-similar process. Ann Probab 15:708–727

McKenzie A, Steel M (2000) Distributions of cherries for two models of trees. Math Biosci 164:81–92

Na HS, Rapoport A (1970) Distribution of nodes of a tree by degree. Math Biosci 6:313–329

Najock D, Heyde CC (1982) On the number of terminal vertices in certain random trees with an application to stemma construction in philology. J Appl Probab 19:675–680
Neininger R (2001) On a multivariate contraction method for random recursive structures with applications to quicksort. Random Struct Algorithms 19:498–524
Neininger R, Rüschendorf L (2004) A general limit theorem for recursive algorithms and combinatorial structures. Ann Appl Probab 14:378–418
Panholzer A, Prodinger H (2004) Analysis of some statistics for increasing tree families. Discret Math Theor Comput Sci 6:437–460
Rachev ST, Rüschendorf L (1995) Probability metrics and recursive algorithms. Adv Appl Probab 27:770–799
Rösler U (1999) A limit theorem for "Quicksort". Inform Théor Appl 25:85–100
Rösler U, Rüschendorf L (2001) The contraction method for recursive algorithms. Algorithmica 29:3–33
Smythe RT, Mahmoud H (1995) A survey of recursive trees. Theory Probab Math Stat 51:1–27
Szymański J (1987) On a nonuniform random recursive tree. In: Annals of discrete mathematics (33); proceedings of the international conference on finite geometries and combinatorial structures, vol 144. North-Holland, pp 297–306