# *On the inducibility of small trees*

Audace A. V. Dossou-Olory* Stephan Wagner†

*Department of Mathematical Sciences, Stellenbosch University, South Africa*

*received 15th Apr. 2019, revised 26th Aug. 2019, accepted 10th Sep. 2019.*

The quantity that captures the asymptotic value of the maximum number of appearances of a given topological tree (a rooted tree with no vertices of outdegree 1) $S$ with $k$ leaves in an arbitrary tree with sufficiently large number of leaves is called the inducibility of $S$. Its precise value is known only for some specific families of trees, most of them exhibiting a symmetrical configuration. In an attempt to answer a recent question posed by Czabarka, Székely, and the second author of this article, we provide bounds for the inducibility $J(A_5)$ of the 5-leaf binary tree $A_5$ whose branches are a single leaf and the complete binary tree of height 2. It was indicated before that $J(A_5)$ appears to be 'close' to $1/4$. We can make this precise by showing that $0.24707\ldots \leq J(A_5) \leq 0.24745\ldots$. Furthermore, we also consider the problem of determining the inducibility of the tree $Q_4$, which is the only tree among 4-leaf topological trees for which the inducibility is unknown.

**Keywords:** inducibility, binary tree, ternary tree, leaf-induced subtree, maximum density, $d$-ary tree, topological tree

## 1 Introduction and previous results

The study of graph inducibility was brought forward in 1975 by Pippenger and Golumbic, who investigated the maximum frequency of $k$-vertex simple graphs occurring as subgraphs within a graph whose number of vertices approaches infinity – see Pippenger and Golumbic (1975) for details and first results on the inducibility of graphs. To this day, there is substantial activity regarding this concept. In analogy to (Pippenger and Golumbic, 1975), the inducibility of a rooted tree $S$ with $k$ leaves is defined as the maximum frequency at which $S$ can appear as a subtree induced by $k$ leaves of an arbitrary rooted tree whose number of leaves tends to infinity (Czabarka et al., 2017, 2020; Dossou-Olory and Wagner, 2019). Bubeck and Linial (2016) defined the inducibility of a tree $S$ with $k$ vertices as the maximum proportion of $S$ as a subtree among all $k$-vertex subtrees of a tree whose number of vertices tends to infinity. We also mention that Sperfeld (2011) extended the concept of inducibility to monodirected graphs, and also gave bounds (using Razborov's flag algebra method) for some graphs with at most four vertices.

For any of the aforementioned notions of inducibility, can the exact inducibility of trees (graphs) with a moderate size always be determined explicitly? The answer to this question turns out to be either undecidable or negative in general in the original context of simple graphs (Exoo, 1986; Sperfeld, 2011; Hirst, 2014; Even-Zohar and Linial, 2015; Bubeck and Linial, 2016). The concept of inducibility of a tree with $k$ leaves is still new and the precise value of the inducibility is known only for a few classes of trees, most of them exhibiting a symmetrical configuration. The recent paper (Czabarka et al., 2017) raised some questions on the inducibility of binary trees, one of which is discussed and approximately solved within this note. The present paper also covers a related problem concerning the inducibility of a ternary tree with four leaves.

Since the inducibility of trees is a quantity that was only introduced recently, let us first turn to a preliminary account on the subject.

A rooted tree without vertices of outdegree 1 will be called a *topological* tree as in (Bergeron et al., 1998; Allman and Rhodes, 2004; Dossou-Olory and Wagner, 2019). We are concerned with topological

*Supported in part by Stellenbosch University in association with the African Institute for Mathematical Sciences (South Africa).
†Supported by the National Research Foundation of South Africa, grant 96236.

ISSN 1365–8050 

trees with a given number of leaves. If, in addition, every vertex has $d$ $(\geq 2)$ or fewer children, then the tree will be called a *$d$-ary* tree as in (Czabarka et al., 2020). Instead of 2-ary tree and 3-ary tree, we shall simply say *binary* tree and *ternary* tree, respectively.

A *leaf-induced subtree* of a topological tree $T$ is any subtree produced in the following three steps: consider a subset $L$ of leaves of $T$; take the minimal subtree containing all the leaves in $L$; suppress all vertices whose outdegree is 1.

An illustration of this process of finding a leaf-induced subtree of $T$ is shown in Figure 1. For a topological tree $T$, we shall denote its number of leaves by $|T|$.

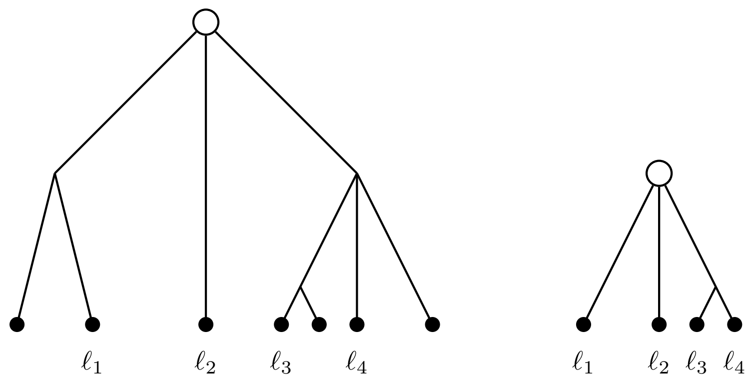



**Fig. 1:** A ternary tree $T$ (left) and the subtree induced by the set of leaves $\{\ell_1, \ell_2, \ell_3, \ell_4\}$ of $T$ (right).

By *density* of a topological tree $S$ in $T$, we mean the proportion of all subsets of $|S|$ leaves of $T$ that induce a leaf-induced subtree isomorphic (in the sense of rooted tree isomorphism) to $S$. We shall denote this density by $\gamma(S,T)$. Thus, it makes sense to set $\gamma(S,T) = 0$ for $|S| > |T|$.

The *inducibility* of $S$ (as defined and studied in (Dossou-Olory and Wagner, 2019)) is its maximum density as a leaf-induced subtree of $T$ as the size of $T$ tends to infinity:

$$J(S) := \limsup_{\substack{|T|\to\infty \\ T \text{ topological tree}}} \gamma(S,T) = \lim_{n\to\infty} \max_{\substack{|T|=n \\ T \text{ topological tree}}} \gamma(S,T)\,.$$

The limit is known to exist; see Dossou-Olory and Wagner (2019, Theorem 3).

Similarly, when the underlying set over which the supremum is taken is restricted to $d$-ary trees, we define

$$I_d(D) := \limsup_{\substack{|T|\to\infty \\ T\ d\text{-ary tree}}} \gamma(D,T) = \lim_{n\to\infty} \max_{\substack{|T|=n \\ T\ d\text{-ary tree}}} \gamma(D,T)$$

to be the inducibility of a $d$-ary tree $D$ in $d$-ary trees (again, the limit is known to exist—Czabarka et al. (2020, Theorem 3)). The subscript $d$ is used to emphasize the fact that we are taking the maximum over the set of all $d$-ary trees.

The initial motivation for studying the quantities $J(S)$ and $I_d(D)$ was twofold: first, they are natural tree analogues of the notion of inducibility in graphs (as outlined above), which is classical in graph theory. The other was a concrete application (Czabarka et al., 2017) to structures called *tanglegrams*, which consist of two binary trees entangled by a perfect matching between the leaves. In order to estimate the crossing number of random tanglegrams, it was necessary to find (asymptotic) bounds on the number $\gamma(S,T)$.

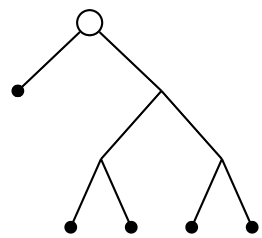
While in the past many results on the inducibility were obtained for graphs, this is not yet the case for trees and many challenging questions remain. The problem of computing the inducibility of a tree appears to be quite difficult even for trees with a small number of leaves—already the inducibilities of some trees with only four or five leaves are not known. The only cases for which an explicit expression is presently known are caterpillars and trees that are highly balanced, thus close to complete $d$-ary trees;

cf. papers Czabarka et al. (2017, 2020); Dossou-Olory and Wagner (2018, 2019). The reason why they are manageable is that the trees $T$ for which the maximum

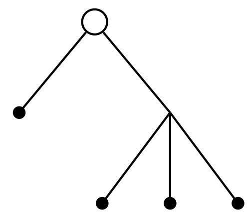
$$\max_{\substack{|T|=n \\ T \text{ } d\text{-ary tree}}} \gamma(D,T)$$

is attained have a simple structure (caterpillars and essentially complete $d$-ary trees) in these cases. In general, this structure appears to be much harder to determine, which is why we have to settle for upper and lower bounds in this paper.

Among 5-leaf binary trees, the tree $A_5$ (see Figure 2(a)) is the only one for which the inducibility has not been determined yet. Also, the inducibility of the 4-leaf ternary tree $Q_4$ shown in Figure 2(b) is unknown. Thus, these are the smallest cases for which we do not have explicit expressions.



(a) The binary tree $A_5$.



(b) The ternary tree $Q_4$.

**Fig. 2:** The topological trees $A_5$ and $Q_4$.

In earlier papers (Czabarka et al., 2020; Dossou-Olory and Wagner, 2018, 2019), various lower bounds were given on the inducibility of topological trees and thus the inducibilities of $Q_4$ and $A_5$. In this note, we shall propose constructions that yield improved lower bounds on the inducibility of the two trees $Q_4$ and $A_5$. Moreover, using a computer search, we shall be able to bound both the inducibility of $A_5$ in topological trees and the inducibility of $Q_4$ in ternary trees from above.

The inducibility of some families of topological trees is known precisely. As such, we have stars, binary caterpillars (Czabarka et al., 2020), complete $d$-ary trees and more generally, the so-called even $d$-ary trees (Dossou-Olory and Wagner, 2018). We already know the inducibility of all topological trees with at most three leaves: each of them has inducibility 1, except for the star with three leaves, which has inducibility $(d-2)/(d+1)$ in $d$-ary trees. There are only five different topological trees with four leaves (see Figure 3), and the precise inducibility of four of them is at least partially known:

$$\begin{aligned}
J\big(CD_2^2\big) &= I_d\big(CD_2^2\big) = \frac{3}{7} \text{ for all } d \text{ (Czabarka et al., 2017; Dossou-Olory and Wagner, 2018)},\\
J\big(F_4^2\big) &= I_d\big(F_4^2\big) = 1 \text{ for all } d \text{ (Czabarka et al., 2017, 2020)},\\
J(S_4) &= 1 \text{ (Czabarka et al., 2020)},\\
I_d(S_4) &= \frac{(d-2)(d-3)}{d^2+d+1} \text{ for all } d \text{ (Czabarka et al., 2020)},\\
I_3\big(E_4^3\big) &= \frac{6}{13} \text{ (Dossou-Olory and Wagner, 2018)},\\
I_d\big(E_4^3\big) &= \text{unknown for } d > 3\,.
\end{aligned}$$

When considering binary trees, we notice that there are only three isomorphism types of 5-leaf trees – see Figure 4 – and the inducibility of two of them has been determined:

$$\begin{aligned}
J\big(E_5^2\big) &= I_d\big(E_5^2\big) = \frac{2}{3} \text{ (Czabarka et al., 2017; Dossou-Olory and Wagner, 2018)},\\
J\big(F_5^2\big) &= I_d\big(F_5^2\big) = 1 \text{ (Czabarka et al., 2017, 2020)}
\end{aligned}$$

for all $d$. The inducibility of the binary tree $A_5$ is of particular interest to us, since it is the smallest binary tree for which the inducibility is not known explicitly. In Czabarka et al. (2017), the authors considered the

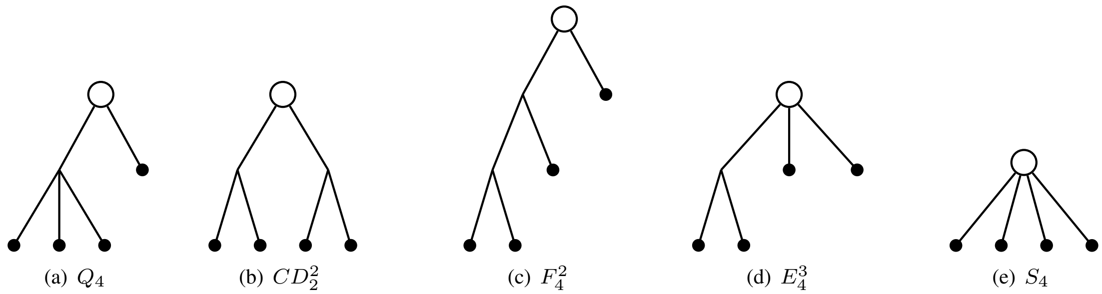



**Fig. 3:** All the topological trees with four leaves.

problem of computing the inducibility of the tree $A_5$ in binary trees, and mentioned that $I_2(A_5)$ appears to be close to $1/4$. This observation came from a computer experiment, but no explicit sequence of binary trees that would yield a value close to $0.25$ in the limit was given. Here we provide a construction which yields the value $0.24707\ldots$ as a lower bound. We also describe how to perform an efficient computer search and obtain $0.24745\ldots$ as an upper bound on $I_2(A_5)$.

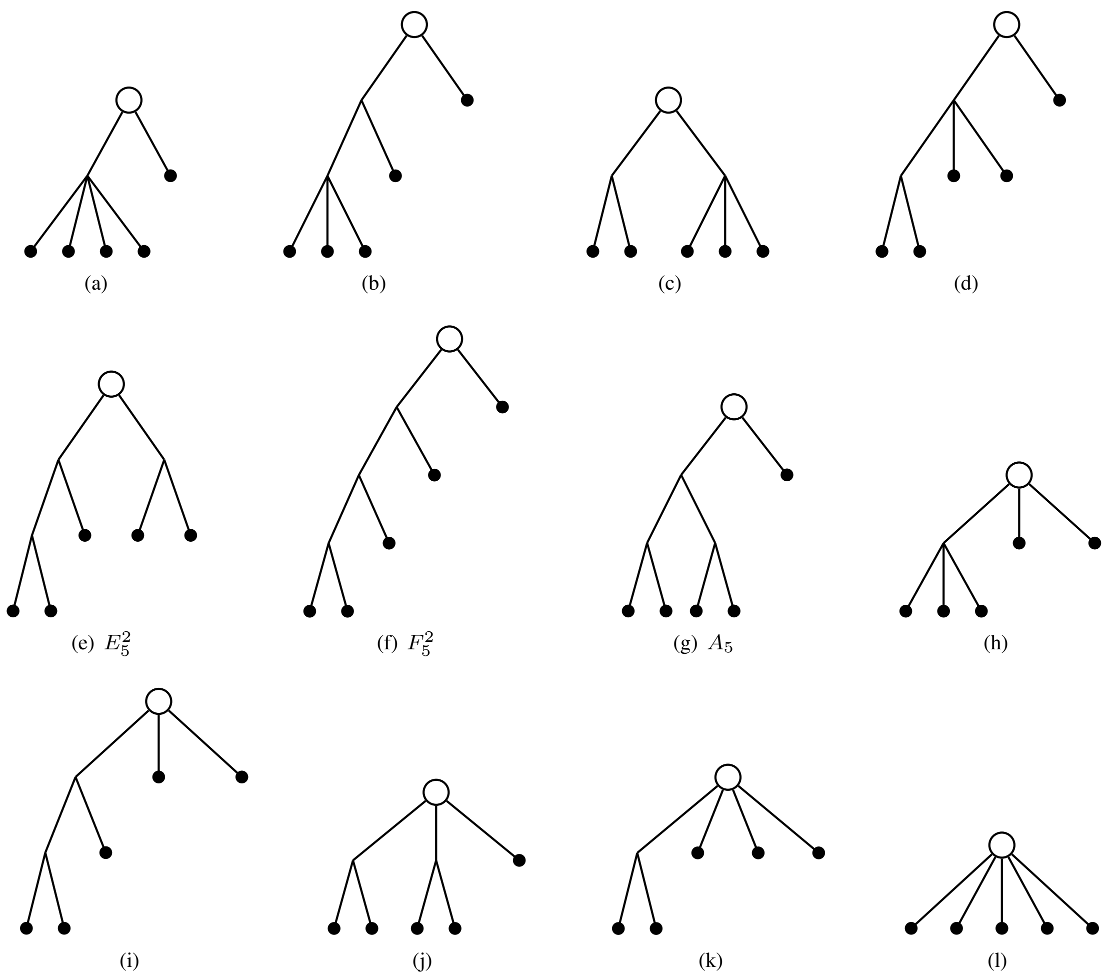



**Fig. 4:** All the 5-leaf topological trees.

In the second part of this work, we consider the problem of finding the inducibility of the ternary tree $Q_4$ in ternary trees. Specifically, we prove that $0.1418\ldots \le I_3(Q_4) \le 0.1435\ldots$. These two trees that we focus on exhibit a non-symmetrical configuration, which makes the computation of their inducibilities harder. For the binary tree $A_5$, we are tempted to conjecture that our candidate is an optimal sequence of binary trees giving the explicit value of $I_2(A_5)$ in the limit, which we obtain as a function of the global maximum of a certain three-variable polynomial over a specific domain.

## 2 Statement of results

Paper (Dossou-Olory and Wagner, 2019) covers, among other things, the relationship between the degree-restricted inducibility $I_d(S)$ in $d$-ary trees and the general inducibility $J(S)$ in topological trees at large. It was proved in (Dossou-Olory and Wagner, 2019) that

$$J(S) = \lim_{d\to\infty} I_d(S)\,.$$

A $d$-ary tree will be called a *strictly* $d$-ary tree if each of its internal vertices has exactly $d$ children. By a result in (Czabarka et al., 2020, Theorem 5), we also know that the underlying set over which the maximum density in $d$-ary trees is taken can be reduced to strictly $d$-ary trees, that is

$$I_d(S) = \lim_{n\to\infty} \max_{\substack{|T|=n \\ T \text{ strictly } d\text{-ary tree}}} \gamma(S,T)\,.$$

In (Czabarka et al., 2017), the authors formulated some questions and conjectures on the inducibility in binary trees, one of which was solved recently in (Czabarka et al., 2020). Among the questions posed, one of them asks for the inducibility of the 5-leaf binary tree $A_5$ (see Figure 2). As mentioned in the introduction, this problem appears to be quite hard and finding a sequence of binary trees that yields $I_2(A_5)$ in the limit also appears to be a difficult task. Czabarka et al. (2017) further mentioned that $I_2(A_5)$ is close to $1/4$, which will be made more precise here with the following result:

**Theorem 1.** *For the binary tree $A_5$, we have*

$$0.247071 \le J(A_5) = I_2(A_5) \le \frac{32828685715097}{132667832500200} \approx 0.247450\,.$$

As part of the ingredients needed to prove this result, let us define a new class of binary trees (which is already considered in recent papers (Czabarka et al., 2017; Dossou-Olory and Wagner, 2018)).

The *even* binary tree $E_n^2$ with $n$ leaves is obtained recursively as follows:

- $E_1^2$ is the tree with only one vertex;
- for $n > 1$, the branches of $E_n^2$ are the even binary trees $E^2_{\lfloor n/2 \rfloor}$ and $E^2_{\lceil n/2 \rceil}$.

An example of an even binary tree can be found in Figure 5.

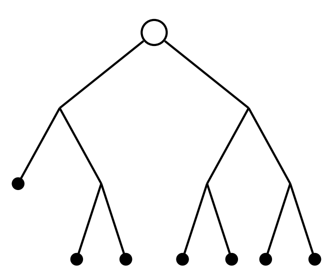


**Fig. 5:** The even binary tree $E_7^2$ with seven leaves.

We shall prove the upper bound in Theorem 1 by means of an algorithmic approach. For the lower bound, we shall make use of the binary tree $S(n_1, n_2, n_3, n_4)$ whose rough picture is shown in Figure 6, where each triangle represents an even binary tree. More specifically, to obtain the tree $S(n_1, n_2, n_3, n_4)$,

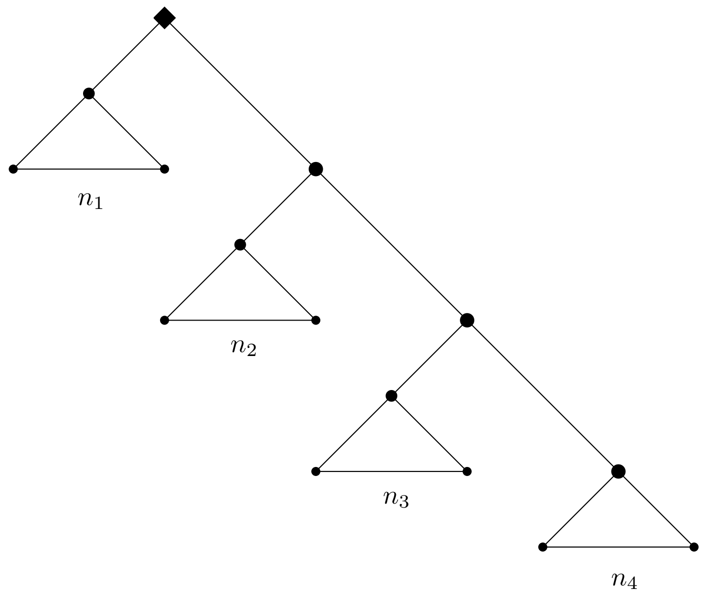



**Fig. 6:** The binary tree $S(n_1, n_2, n_3, n_4)$ described for Theorem 1.

we take the 4-leaf binary tree whose internal vertices form a path beginning at the root (the square vertex on top in Figure 6), and identify the four leaves with the even binary trees whose number of leaves is $n_1, n_2, n_3, n_4$, respectively in this order (starting with the top leaf attached to the root).

As a next step, we set up a formula for the number of copies of $A_5$ in $S(n_1, n_2, n_3, n_4)$; this formula is used together with a result on even binary trees from Dossou-Olory and Wagner (2018) to derive an asymptotic formula for $c(A_5, S(n_1, n_2, n_3, n_4))$ as $n = n_1 + n_2 + n_3 + n_4 \to \infty$. Finally, we compute (at least approximately) the global maximum of the main term in the asymptotic formula of the density $\gamma(A_5, S(n_1, n_2, n_3, n_4))$ in the region defined by $0 < n_1, n_2, n_3, n_4 < n$ and $n_1 + n_2 + n_3 + n_4 = n$.

As a closing comment, when we consider five or more even binary trees instead of four in the tree configuration of Figure 6, we do not seem to get a better lower bound. We therefore expect our construction to be best possible.

Among the topological trees with fewer than five leaves, the 4-leaf ternary tree $Q_4$ (Figure 2) is the only one for which we are yet to determine an exact inducibility. What is the inducibility of $Q_4$ (at least in ternary trees)? In what follows, we shall derive a lower and upper bound on $I_3(Q_4)$. Our second main theorem reads as follows:

**Theorem 2.** *For the ternary tree $Q_4$, we have*

$$0.141827 \approx \frac{59}{416} \leq I_3(Q_4) \leq \frac{73848853}{514606225} \approx 0.143506\,.$$

The proof of the lower bound in Theorem 2 is accomplished by an explicit construction (as in Theorem 1), while the upper bound is obtained by means of a computer search. We defer them to Section 5.

The *star* with $k$ leaves is obtained by joining $k$ distinct vertices to a new vertex (the root of the star). We shall denote it with the symbol $S_k$.

The *complete $d$-ary tree of height $h$* is the strictly $d$-ary tree in which the distance from every leaf to the root is $h$. Such a tree has $d^h$ leaves in total and shall be denoted with the symbol $CD^d_h$.

For a positive integer $k \geq 3$, denote by $Q_k$ the tree whose branches are $S_{k-1}$ and $S_1$ (the single leaf). The following proposition will serve as an intermediary result to proving a new lower bound on the inducibility of the tree $Q_4$. Its proof will be given in Section 5.

**Proposition 3.** *For all positive integers $d, k,$ and $h$ such that $d \geq 2$ and $k \geq 3$, the formula*

$$c\bigl(Q_k, CD_h^d\bigr) = \frac{(d-1)\binom{d}{k-1}}{d^{k-1}-d} \cdot d^h \left( \frac{d^{(k-1)h} - d^{k-1}}{d^{k-1}-1} - \frac{d^h - d}{d-1} \right)$$

*holds. In particular, we have*

$$I_d(Q_k) \geq \frac{k!(d-1)\binom{d}{k-1}}{(d^{k-1}-d)(d^{k-1}-1)}$$

*for every $d \geq 2$ and every $k \geq 3$.*

The next proposition shows that the bounds mentioned in Theorems 1 and 2 are much better than the natural bounds provided by the complete $d$-ary trees, cf. Dossou-Olory and Wagner (2018).

**Proposition 4.** *For the trees $Q_4$ and $A_5$, we have*

$$\lim_{h \to \infty} \gamma\bigl(Q_4, CD_h^3\bigr) = \frac{1}{13}$$

*and*

$$\lim_{h \to \infty} \gamma\bigl(A_5, CD_h^2\bigr) = \frac{1}{7}\,.$$

**Proof:** The specialisation $d = 3$ and $k = 4$ in Proposition 3 yields

$$\lim_{h \to \infty} \gamma\bigl(Q_4, CD_h^3\bigr) = \frac{1}{13}\,.$$

As a special case of a result in (Dossou-Olory and Wagner, 2018, Theorem 1), we know that

$$\lim_{h \to \infty} \gamma\bigl(A_5, CD_h^2\bigr) = \frac{2 \cdot 5}{2^5 - 2} \cdot I_2\bigl(CD_2^2\bigr)\,,$$

while it was proved in the same source (see also Czabarka et al. (2017, Proposition 2)) that

$$I_2\bigl(CD_2^2\bigr) = \frac{3}{7}\,.$$

This completes the proof of the proposition. □

# 3 An algorithm for the maximum

Our next theorem will be used to prove the upper bound on the inducibility of each of the trees $A_5$ and $Q_4$. Here, we shall only discuss the tree $A_5$ (the case of $Q_4$ is analogous, as will become clear from the proof). We know from Czabarka et al. (2020, Theorem 3) that

$$I_d(S) \leq \max_{\substack{|T|=n \\ T \text{ } d\text{-ary tree}}} \gamma(S, T)$$

for all $d$-ary trees $S$ and $n \geq |S|$. Thus it suffices to determine the value on the right (which can be shown to be decreasing in $n$ as $n \geq |S|$) for as large a value of $n$ as possible to obtain an upper bound. This will be the main goal of this section, where an algorithm for this purpose will be presented. We first need a series of lemmas.

If $v$ is a vertex of a topological tree $T$, then the subtree $T[v]$ consisting of $v$ and all its descendants in $T$ is called a *fringe* subtree of $T$. In other words, $T[v]$ is the subtree of $T$ rooted at $v$.

**Lemma 5.** *Let $v$ be a vertex of a binary tree $T$, and let $T[v]$ be the fringe subtree rooted at $v$. The number of copies of $A_5$ in $T$ can be expressed as*

$$c(A_5, T) = c(A_5, T[v]) + (|T| - |T[v]|)c(CD_2^2, T[v]) + R\,,$$

*where $R$ only depends on the size of $T[v]$ (and the rest of $T$), but not its precise structure.*

**Proof:** If a set of leaves contains at most three leaves of $T[v]$, then there is only one possibility for the tree induced by them inside of $T[v]$. Thus the number of copies of $A_5$ in $T$ that contain at most three leaves of $T[v]$ only depends on the size of $T[v]$, but not its shape. This leaves us with

- copies of $A_5$ that are entirely contained in $T[v]$; their number is clearly $c(A_5, T[v])$,
- copies of $A_5$ that contain precisely four leaves of $T[v]$; there are $|T| - |T[v]|$ other leaves, and the four leaves in $T[v]$ have to induce a copy of $CD_2^2$ to obtain a copy of $A_5$. Thus the number of these copies is $(|T| - |T[v]|)c(CD_2^2, T[v])$.

The statement of the lemma follows. □

**Lemma 6.** *Let $v$ be a vertex of a binary tree $T$, and let $T[v]$ be the fringe subtree rooted at $v$. Let $S$ be a binary tree of the same size as $T[v]$ that satisfies*

$$c(CD_2^2, S) \geq c(CD_2^2, T[v]) \quad \text{and} \quad c(A_5, S) \geq c(A_5, T[v])\,,$$

*at least one of them with strict inequality. Let $T'$ be obtained from $T$ by replacing $T[v]$ with $S$; then we have*

$$c(A_5, T') > c(A_5, T).$$

**Proof:** This is immediate from the previous lemma. □

**Lemma 7.** *Let $v$ be a vertex of a binary tree $T$, and let $T[v]$ be the fringe subtree rooted at $v$. Let $S_1$ and $S_2$ be two binary trees of the same size as $T[v]$ that satisfy*

$$c(CD_2^2, S_1) > c(CD_2^2, T[v]) > c(CD_2^2, S_2)$$

*and*

$$c(A_5, S_1) < c(A_5, T[v]) < c(A_5, S_2)\,.$$

*Suppose further that*

$$\frac{c(A_5, S_1) - c(A_5, T[v])}{c(CD_2^2, S_1) - c(CD_2^2, T[v])} \geq \frac{c(A_5, T[v]) - c(A_5, S_2)}{c(CD_2^2, T[v]) - c(CD_2^2, S_2)}. \tag{1}$$

*Let $T_1$ and $T_2$ be obtained from $T$ by replacing $T[v]$ with $S_1$ and $S_2$ respectively; then we have*

$$\max\big(c(A_5, T_1), c(A_5, T_2)\big) \geq c(A_5, T). \tag{2}$$

*If strict inequality holds in* (1)*, then we also have strict inequality in* (2)*.*

**Proof:** Let $k = |T| - |T[v]|$. By Lemma 5, we have

$$\begin{aligned} c(A_5, T_1) - c(A_5, T) &= c(A_5, S_1) - c(A_5, T[v]) + k\big(c(CD_2^2, S_1) - c(CD_2^2, T[v])\big) \\ &= \big(c(CD_2^2, S_1) - c(CD_2^2, T[v])\big)\Big(k + \frac{c(A_5, S_1) - c(A_5, T[v])}{c(CD_2^2, S_1) - c(CD_2^2, T[v])}\Big). \end{aligned}$$

If

$$\frac{c(A_5, S_1) - c(A_5, T[v])}{c(CD_2^2, S_1) - c(CD_2^2, T[v])} \geq -k,$$

then we are done, since $c(A_5, T_1) \geq c(A_5, T)$. Otherwise, (1) implies that

$$\frac{c(A_5, T[v]) - c(A_5, S_2)}{c(CD_2^2, T[v]) - c(CD_2^2, S_2)} < -k.$$

Now it follows that

$$\begin{aligned} c(A_5, T_2) - c(A_5, T) &= c(A_5, S_2) - c(A_5, T[v]) + k\big(c(CD_2^2, S_2) - c(CD_2^2, T[v])\big) \\ &= \big(c(CD_2^2, S_2) - c(CD_2^2, T[v])\big)\left(k + \frac{c(A_5, T[v]) - c(A_5, S_2)}{c(CD_2^2, T[v]) - c(CD_2^2, S_2)}\right) > 0, \end{aligned}$$

so $c(A_5, T_2) \geq c(A_5, T)$. Either way, we have (2). Equality can only hold if both quotients in (1) are equal to $-k$. This completes the proof. □

**Lemma 8.** *Let $v$ be a vertex of a binary tree $T$, and let $T[v]$ be the fringe subtree rooted at $v$. Let $S$ be a binary tree of the same size as $T[v]$ that satisfies*

$$c(CD_2^2, S) > c(CD_2^2, T[v])$$

*and*

$$c(A_5, S) < c(A_5, T[v])\,.$$

*Suppose further that*

$$\frac{c(A_5, S) - c(A_5, T[v])}{c(CD_2^2, S) - c(CD_2^2, T[v])} \geq |T[v]| - |T|. \tag{3}$$

*Let $T'$ be obtained from $T$ by replacing $T[v]$ with $S$; then we have*

$$c(A_5, T') \geq c(A_5, T). \tag{4}$$

*If strict inequality holds in* (3)*, then we also have strict inequality in* (4)*.*

**Proof:** As in the proof of the previous lemma, we have

$$c(A_5, T') - c(A_5, T) = \big(c(CD_2^2, S) - c(CD_2^2, T[v])\big)\left(|T| - |T[v]| + \frac{c(A_5, S) - c(A_5, T[v])}{c(CD_2^2, S) - c(CD_2^2, T[v])}\right).$$

The statement follows immediately. □

Now we are ready to describe the algorithm to determine the maximum number of copies of $A_5$ in a binary tree with $n$ leaves. To this end, we define a sequence of sets of binary trees: intuitively speaking, $\mathcal{L}(n)$ consists of trees with $n$ leaves that can potentially occur as fringe subtrees of "optimal" trees, *i.e.*, binary trees that maximize the number of copies of $A_5$. A formal recursive definition will be provided below. We also associate every tree $T$ with the pair $P(T) = (c(A_5, T), c(CD_2^2, T))$, which can be interpreted as a point in the plane, and we set

$$L(n) = \{P(T) \,:\, T \in \mathcal{L}(n)\}$$

for every $n$. The sets $\mathcal{L}(n)$ are recursively defined as follows:

1. The set $\mathcal{L}(1)$ only consists of one tree, which only has a single vertex.
2. For $n > 1$, we consider all binary trees with $n$ leaves for which each branch lies in one of the sets $\mathcal{L}(m)$ for some $m < n$. Clearly, if one branch lies in $\mathcal{L}(k)$, the other has to lie in $\mathcal{L}(n-k)$. For reasons to become clear later (essentially, we are applying Lemma 8), we will be even more restrictive: we consider all binary trees with $n$ leaves whose branches both lie in

$$\bigcup_{m<n} \Big\{T \in \mathcal{L}(m) \,:\, \text{there is no } S \in \mathcal{L}(m) \text{ such that } c(CD_2^2, S) > c(CD_2^2, T),$$
$$c(A_5, S) < c(A_5, T), \text{ and } \frac{c(A_5, S) - c(A_5, T[v])}{c(CD_2^2, S) - c(CD_2^2, T)} \geq m - n\Big\}.$$

   This gives us a preliminary set $\mathcal{H}_1(n)$.

3. If there are two trees $T$ and $T'$ in $\mathcal{H}_1(n)$ such that

$$c(CD_2^2, T) \geq c(CD_2^2, T') \qquad \text{and} \qquad c(A_5, T) \geq c(A_5, T')\,,$$

remove $T'$ from $\mathcal{H}_1(n)$. If we have equality in both inequalities, we can arbitrarily remove either $T$ or $T'$. In geometric terms, the condition means that the point $P(T')$ lies to the left and below the point $P(T)$ in the plane. We repeat this step until there are no two trees $T$ and $T'$ satisfying the aforementioned condition anymore. At the end, we are left with a set $\mathcal{H}_2(n)$.

4. As a final reduction step, we eliminate all trees $T$ from $\mathcal{H}_2(n)$ for which there exist two trees $S_1$ and $S_2$ in $\mathcal{H}_2(n)$ such that the inequalities of Lemma 7 hold, *i.e.*,

$$c(CD_2^2, S_1) > c(CD_2^2, T) > c(CD_2^2, S_2)$$

and

$$c(A_5, S_1) < c(A_5, T) < c(A_5, S_2)$$

as well as

$$\frac{c(A_5, S_1) - c(A_5, T)}{c(CD_2^2, S_1) - c(CD_2^2, T)} \geq \frac{c(A_5, T) - c(A_5, S_2)}{c(CD_2^2, T) - c(CD_2^2, S_2)}\,.$$

Considering the set of points $\{P(T) \,:\, T \in \mathcal{H}_2(T)\}$ in the plane, this amounts to taking the upper envelope of the points. The resulting set after this reduction is $\mathcal{L}(n)$. At this point, we can arrange the elements of $\mathcal{L}(n)$ as a list of trees $T_1, T_2, \ldots, T_r$ such that

$$c(CD_2^2, T_1) < c(CD_2^2, T_2) < \cdots < c(CD_2^2, T_r)\,,$$
$$c(A_5, T_1) > c(A_5, T_2) > \cdots > c(A_5, T_r)\,,$$

and the sequence of "slopes"

$$\frac{c(A_5, T_{j+1}) - c(A_5, T_j)}{c(CD_2^2, T_{j+1}) - c(CD_2^2, T_j)}$$

is strictly decreasing. This also makes it easier to construct the set in step (2): the trees from $\mathcal{L}(m)$ that are allowed as branches are precisely those starting from the point where the slope is less than $m - n$.

Due to the rules of the two elimination steps, the following holds for all $T \in \mathcal{H}_1(n)$ at the end:

- Either there exists an $S \in \mathcal{L}(n)$ (possibly $T = S$) such that

$$c(CD_2^2, S) \geq c(CD_2^2, T) \qquad \text{and} \qquad c(A_5, S) \geq c(A_5, T)\,,$$

- or there exist two trees $S_1, S_2 \in \mathcal{L}(n)$ such that

$$c(CD_2^2, S_1) > c(CD_2^2, T) > c(CD_2^2, S_2),\ c(A_5, S_1) < c(A_5, T) < c(A_5, S_2)$$

and

$$\frac{c(A_5, S_1) - c(A_5, T)}{c(CD_2^2, S_1) - c(CD_2^2, T)} \geq \frac{c(A_5, T) - c(A_5, S_2)}{c(CD_2^2, T) - c(CD_2^2, S_2)}\,.$$

The following theorem shows that the maximum of $c(A_5, T)$ for binary trees $T$ with a given number of leaves can be determined purely by focusing on the sets $\mathcal{L}(n)$.

**Theorem 9.** *For every positive integer $n$, there exists a binary tree $M_n$ with $n$ leaves such that*

$$c(A_5, M_n) = \max_{\substack{|T|=n \\ T \text{ binary tree}}} c(A_5, T)$$

*and all fringe subtrees of $M_n$ (including $M_n$ itself) lie in $\bigcup_{k \geq 1} \mathcal{L}(k)$. In particular,*

$$\max_{\substack{|T|=n \\ T \text{ binary tree}}} c(A_5, T) = \max_{T \in \mathcal{L}(n)} c(A_5, T) \,.$$

**Proof:** Suppose that the statement does not hold, and let $m$ be minimal with the property that there is a positive integer $n$ such that every "optimal" tree (tree attaining the maximum $\max_{|T|=n} c(A_5, T)$) has a fringe subtree with $m$ or fewer leaves that does not lie in $\bigcup_{1 \leq k \leq m} \mathcal{L}(k)$. Clearly, $m > 1$.

By our choice of $m$, there must be an optimal tree $T$ with $n$ leaves for which all fringe subtrees with less than $m$ leaves lie in $\bigcup_{1 \leq k < m} \mathcal{L}(k)$. Among all possible choices of $T$, we can choose one for which the number of $m$-leaf fringe subtrees that do not lie in $\mathcal{L}(m)$ is minimal. Consider one of these fringe subtrees $T[v]$. Both its branches lie in $\bigcup_{1 \leq k < m} \mathcal{L}(k)$, which leaves us with the following possible reasons why $T[v]$ is not in $\mathcal{L}(m)$:

- The branches of $T[v]$ do not satisfy the condition of step (2) in the construction of $\mathcal{L}(n)$ (*i.e.*, $T[v]$ does not even lie in $\mathcal{H}_1(m)$). Suppose that for one of the branches $B$, there is a tree $S$ in $\mathcal{L}_{|B|}$ such that

  $$c(CD_2^2, S) > c(CD_2^2, B), c(A_5, S) < c(A_5, B) \,,$$

  and

  $$\frac{c(A_5, S) - c(A_5, B)}{c(CD_2^2, S) - c(CD_2^2, B)} \geq |B| - |T[v]| \geq |B| - |T| \,.$$

  We can replace $B$ by $S$, and do likewise with the other branch of $T[v]$ if necessary. We either reach a contradiction to the optimality of $T[v]$ by means of Lemma 8 immediately, or (if equality holds above) a new tree that is still optimal, but where $T[v]$ has been replaced by a tree in $\mathcal{L}(m)$, again contradicting the choice of $T$. So for the remaining cases, we can at least assume that $T[v] \in \mathcal{H}_1(m)$.

- There is a binary tree $S \in \mathcal{L}(m)$ such that

  $$c(CD_2^2, S) \geq c(CD_2^2, T[v]) \qquad \text{and} \qquad c(A_5, S) \geq c(A_5, T[v]) \,.$$

  In this case, we can replace $T[v]$ by $S$ to obtain a new tree with at least as many copies of $A_5$ as $T$ by Lemma 6. This contradicts our choice of $T$ (it is either not optimal, or it does not have the smallest number of $m$-leaf fringe subtrees that do not lie in $\mathcal{L}(m)$).

- There are binary trees $S_1, S_2 \in \mathcal{L}(m)$ such that

  $$c(CD_2^2, S_1) > c(CD_2^2, T) > c(CD_2^2, S_2), \; c(A_5, S_1) < c(A_5, T) < c(A_5, S_2)$$

  and

  $$\frac{c(A_5, S_1) - c(A_5, T)}{c(CD_2^2, S_1) - c(CD_2^2, T)} \geq \frac{c(A_5, T) - c(A_5, S_2)}{c(CD_2^2, T) - c(CD_2^2, S_2)} \,.$$

  In this case, we can replace $T[v]$ by either $S_1$ or $S_2$ to obtain a contradiction in the same way as in the previous case (now by means of Lemma 7).

Since we reach a contradiction in all possible cases, the proof is complete. □

For a practical implementation of this algorithm, it actually suffices to work with the lists

$$L(n) = \{P(T) \,:\, T \in \mathcal{L}(n)\}$$

that contain the values of $P(T) = (c(A_5, T), c(CD_2^2, T))$. These values can be calculated recursively: if the branches of a binary tree $T$ are $B_1$ and $B_2$, we have

$$c(A_5, T) = c(A_5, B_1) + c(A_5, B_2) + |B_1|c(CD_2^2, B_2) + |B_2|c(CD_2^2, B_1) \tag{5}$$

and

$$c(CD_2^2, T) = c(CD_2^2, B_1) + c(CD_2^2, B_2) + \binom{|B_1|}{2}\binom{|B_2|}{2}. \tag{6}$$

These formulas can be explained as follows:

- A subset of five leaves of the leaf-set of $T$ can either be a subset of leaves of $B_1$, or a subset of leaves of $B_2$, or splits into leaves of both $B_1$ and $B_2$. In the latter case, the split must be of the type $1-4$ (or $4-1$) as the branches of $A_5$ are a single vertex and $CD_2^2$. Moreover, the four leaves that lie in one branch have to induce $CD_2^2$ there. This proves the recursion for $A_5$.
- Four leaves of $T$ that induce the tree $CD_2^2$ can either lie entirely in $T_1$ or $T_2$; or precisely two leaves in each of the branches $B_1$ and $B_2$ of $T$ induce the star $S_2$ to obtain a copy of $CD_2^2$. This proves the recursive formula for $CD_2^2$.

Thus, it is never necessary to store full tree structures. At the end, the maximum

$$\max_{\substack{|T|=n \\ T \text{ binary tree}}} c(A_5, T)$$

can be determined easily from $L(n)$.

# 4 Proof of Theorem 1

This section is devoted to proving Theorem 1. Recall that we are going to use the binary tree $S(n_1, n_2, n_3, n_4)$ presented in Figure 6. Moreover, we now need to consider only $I_2(A_5)$ because it is established in (Dossou-Olory and Wagner, 2019, Corollary 8) that $J(B) = I_2(B)$ for every binary tree $B$.

**Proof of Theorem 1:** Let us set $n = n_1 + n_2 + n_3 + n_4$. Recall from equation (5) that a recursion for the number of copies of $A_5$ in any binary tree $T$ with branches $B_1$ and $B_2$ is given by

$$c(A_5, T) = c(A_5, B_1) + c(A_5, B_2) + |B_1| \cdot c\big(CD_2^2, B_2\big) + |B_2| \cdot c\big(CD_2^2, B_1\big)\,.$$

So for the tree $S(n_1, n_2, n_3, n_4)$, we obtain

$$\begin{aligned} c\big(A_5, S(n_1, n_2, n_3, n_4)\big) &= c\big(A_5, E_{n_1}^2\big) + c\big(A_5, E_{n_2}^2\big) + c\big(A_5, E_{n_3}^2\big) + c\big(A_5, E_{n_4}^2\big) \\ &\quad + n_3 \cdot c\big(CD_2^2, E_{n_4}^2\big) + n_4 \cdot c\big(CD_2^2, E_{n_3}^2\big) \\ &\quad + n_2 \cdot c\big(CD_2^2, T_{n_3,n_4}\big) + (n_3 + n_4) \cdot c\big(CD_2^2, E_{n_2}^2\big) \\ &\quad + n_1 \cdot c\big(CD_2^2, T_{n_2,n_3,n_4}\big) + (n_2 + n_3 + n_4) \cdot c\big(CD_2^2, E_{n_1}^2\big)\,, \end{aligned} \tag{7}$$

where $T_{n_3,n_4}$ is the binary tree whose branches are the even binary trees $E_{n_3}^2$ and $E_{n_4}^2$, while $T_{n_2,n_3,n_4}$ is the binary tree whose branches are $E_{n_2}^2$ and $T_{n_3,n_4}$.

Also, recall from equation (6) that a recursion for the number of copies of $CD_2^2$ in any binary tree $T$ with branches $B_1$ and $B_2$ is given by

$$c(CD_2^2, T) = c(CD_2^2, B_1) + c(CD_2^2, B_2) + \binom{|B_1|}{2}\binom{|B_2|}{2}.$$

So for the binary tree $T_{n_2,n_3,n_4}$, we get

$$c\big(CD_2^2, T_{n_2,n_3,n_4}\big) = c\big(CD_2^2, E_{n_2}^2\big) + c\big(CD_2^2, T_{n_3,n_4}\big) + \binom{n_2}{2}\binom{n_3+n_4}{2}.$$

Likewise,

$$c\big(CD_2^2, T_{n_3,n_4}\big) = c\big(CD_2^2, E_{n_3}^2\big) + c\big(CD_2^2, E_{n_4}^2\big) + \binom{n_3}{2}\binom{n_4}{2}.$$

Thus, equation (7) becomes

$$\begin{aligned} c\big(A_5, S(n_1,n_2,n_3,n_4)\big) &= c\big(A_5, E_{n_1}^2\big) + c\big(A_5, E_{n_2}^2\big) + c\big(A_5, E_{n_3}^2\big) + c\big(A_5, E_{n_4}^2\big) \\ &+ n_3 \cdot c\big(CD_2^2, E_{n_4}^2\big) + n_4 \cdot c\big(CD_2^2, E_{n_3}^2\big) + (n_3+n_4)\cdot c\big(CD_2^2, E_{n_2}^2\big) \\ &+ n_2\Big(c\big(CD_2^2, E_{n_3}^2\big) + c\big(CD_2^2, E_{n_4}^2\big) + \binom{n_3}{2}\binom{n_4}{2}\Big) \\ &+ n_1\Big(c\big(CD_2^2, E_{n_2}^2\big) + c\big(CD_2^2, E_{n_3}^2\big) + c\big(CD_2^2, E_{n_4}^2\big) + \binom{n_3}{2}\binom{n_4}{2} \\ &+ \binom{n_2}{2}\binom{n_3+n_4}{2}\Big) + (n_2+n_3+n_4)\cdot c\big(CD_2^2, E_{n_1}^2\big) \end{aligned}$$

after combining everything. As a special case of Dossou-Olory and Wagner (2018, Theorem 12), we have

$$c\big(CD_2^2, E_n^2\big) = \frac{1}{56}\cdot n^4 + \mathcal{O}(n^3)$$

for all $n$. On the other hand, using this asymptotic formula along with the recursion

$$\begin{aligned} c\big(A_5, E_n^2\big) &= c\big(A_5, E_{\lfloor n/2\rfloor}^2\big) + c(A_5, E_{\lceil n/2\rceil}^2) \\ &\quad + \lfloor n/2\rfloor \cdot c\big(CD_2^2, E_{\lceil n/2\rceil}^2\big) + \lceil n/2\rceil \cdot c\big(CD_2^2, E_{\lfloor n/2\rfloor}^2\big), \end{aligned}$$

which follows from (5) by the definition of the even binary tree $E_n^2$, it is not hard to prove that there exist absolute constants $K_1, K_2 \geq 0$ such that the double inequality

$$\frac{1}{840}\cdot n^5 - K_1\cdot n^4 \leq c\big(A_5, E_n^2\big) \leq \frac{1}{840}\cdot n^5 + K_2\cdot n^4$$

holds for all $n$—the details are omitted.

Now, let $x_1, x_2, x_3, x_4$ be positive real numbers with $x_1+x_2+x_3+x_4 = 1$. We set

$$n_i = \lfloor x_i n\rfloor = x_i n + \mathcal{O}(1)$$

for $i = 1,2,3,4$. Combining all the asymptotic formulas, we can now rewrite $c\big(A_5, S(n_1,n_2,n_3,n_4)\big)$ as follows:

$$\begin{aligned} c\big(A_5, S(n_1,n_2,n_3,n_4)\big) &= \frac{n^5}{840}\big(x_1^5+x_2^5+x_3^5+x_4^5\big) + \frac{n^5}{56}\big(x_3\cdot x_4^4 + x_4\cdot x_3^4\big) \\ &+ \frac{n^5}{56}\cdot x_2\big(x_3^4+x_4^4+14\cdot x_3^2\cdot x_4^2\big) + \frac{n^5}{56}(x_3+x_4)x_2^4 \\ &+ \frac{n^5}{56}\cdot x_1\Big(x_2^4+x_3^4+x_4^4+14\cdot x_3^2\cdot x_4^2 + 14\cdot x_2^2(x_3+x_4)^2\Big) \\ &+ \frac{n^5}{56}(x_2+x_3+x_4)x_1^4 + \mathcal{O}(n^4). \end{aligned}$$

Set

$$F(x_1,x_2,x_3) = \frac{1}{840}\big(x_1^5+x_2^5+x_3^5+(1-x_1-x_2-x_3)^5\big) + \frac{1}{56}\Big(x_3(1-x_1-x_2-x_3)^4$$

$$
\begin{aligned}
&+(1-x_1-x_2-x_3)x_3^4 + x_2\big(x_3^4 + (1-x_1-x_2-x_3)^4 \\
&+ 14 \cdot x_3^2(1-x_1-x_2-x_3)^2\big) + (1-x_1-x_2)x_2^4 \\
&+ x_1\Big(x_2^4 + x_3^4 + (1-x_1-x_2-x_3)^4 + 14 \cdot x_3^2(1-x_1-x_2-x_3)^2 \\
&+ 14 \cdot x_2^2(1-x_1-x_2)^2\Big) + (1-x_1)x_1^4\Big)\,.
\end{aligned}
$$

Then we obtain

$$c\big(A_5, S(n_1,n_2,n_3,n_4)\big) = F(x_1,x_2,x_3)\cdot n^5 + \mathcal{O}(n^4)$$

as $x_1+x_2+x_3+x_4=1$. With the help of a computer, we find that the global maximum of the function $F(x_1,x_2,x_3)$ in the region covered by the inequalities $0 < x_1,x_2,x_3 < 1,\ x_1+x_2+x_3 < 1$ is attained at the points whose values are approximately

$$(x_1 = 0.025347732268, x_2 = 0.051425755177, x_3 = 0.788023120078)$$

and

$$(x_1 = 0.025347732268, x_2 = 0.051425755177, x_3 = 0.135203392478).$$

Note here that the two critical points are related by the symmetry between $x_3$ and $x_4$ in the asymptotic formula for $c\big(A_5, S(n_1,n_2,n_3,n_4)\big)$. Thus we have

$$
\begin{aligned}
\max_{\substack{0<x_1,x_2,x_3<1\\ x_1+x_2+x_3<1}} F(x_1,x_2,x_3) &\geq F(0.025347732268, 0.051425755177, 0.788023120078) \\
&\approx 0.002058929182\,.
\end{aligned}
$$

The inequality

$$I_2(A_5) \geq 0.002058929182 \times 5! \approx 0.247071501785$$

follows. This concludes the proof of the first part of the theorem.

*Remark* 1. The precise lower bound is an algebraic number of degree 16; its minimal polynomial is given in the appendix.

For the upper bound, we make use of Theorem 9 which states that the maximum of $c(A_5,T)$ for binary trees $T$ with $n$ leaves can be determined purely by focusing on the sets $\mathcal{L}(n)$ whose algorithmic description is given in Section 3. Recall that by Theorem 3 in (Czabarka et al., 2020), we have

$$I_2(A_5) \leq \max_{\substack{|T|=n\\ T \text{ binary tree}}} \gamma(A_5,T)$$

for every $n \geq 5$. Thus we want to calculate the maximum for different values of $n$. When our algorithm terminates, the maximum number of copies of $A_5$ among all binary trees with $n$ leaves can be read off as the greatest $x$-coordinate (first coordinate) of the elements of $L(n)$, that is the $x$-coordinate of the very first element of $L(n)$—see the discussion before Theorem 9.

We have implemented this algorithm in Mathematica. The notebook can be accessed at `http://math.sun.ac.za/~swagner/TreeA5Final`. The precise values of

$$a_n = \max_{\substack{|T|=n\\ T \text{ binary tree}}} \gamma(A_5,T)$$

have been computed for $n \leq 2000$—see Table 1. It follows that

$$I_2(A_5) \leq a_{2000} = \frac{32828685715097}{132667832500200} \approx 0.24745\,.$$

This completes the proof of the theorem.

□

**Tab. 1:** Maximum density $a_n$ of $A_5$ among $n$-leaf binary trees

| $\boldsymbol{n}$ | 5 | 6 | 7 | 8 | 9 | 10 | 20 | 30 |
|---|---|---|---|---|---|---|---|---|
| $a_n$ | 1 | $\frac{1}{2}$ | $\frac{3}{7}$ | $\frac{11}{28}$ | $\frac{23}{63}$ | $\frac{1}{3}$ | $\frac{553}{1938}$ | $\frac{19219}{71253}$ |
| $\boldsymbol{n}$ | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 150 |
| $a_n$ | $\frac{57793}{219336}$ | $\frac{550621}{2118760}$ | $\frac{351943}{1365378}$ | $\frac{44899}{175406}$ | $\frac{6127045}{24040016}$ | $\frac{930032}{3662439}$ | $\frac{3177631}{12547920}$ | $\frac{24765738}{98600005}$ |
| $\boldsymbol{n}$ | 200 | 250 | 300 | 350 | 400 | 500 | 600 | 700 |
| $a_n$ | 0.250153 | 0.249543 | 0.249142 | 0.248854 | 0.24864 | 0.24834 | 0.248143 | 0.248001 |
| $\boldsymbol{n}$ | 800 | 900 | 1000 | 1200 | 1400 | 1600 | 1800 | 2000 |
| $a_n$ | 0.247894 | 0.247812 | 0.247747 | 0.247648 | 0.247577 | 0.247524 | 0.247483 | 0.24745 |

# 5 Proof of Theorem 2

Let us first provide a proof of Proposition 3, which is an intermediate step in the proof of Theorem 2:

**Proof of Proposition 3:** Let $k \geq 3$ and $d \geq k-1$ be fixed (note that $c\big(Q_k, CD_h^d\big) = 0$ for $d < k-1$). For $h = 1$, we have $c\big(Q_k, CD_h^d\big) = c(Q_k, S_d) = 0$. Since for the case $h = 1$, the statement holds trivially, we can safely assume $h \geq 2$ and proceed by induction on $h$.

We distinguish possible cases that can happen for a subset of $k$ leaves of the tree $CD_h^d$:

- all $k$ leaves belong to the same branch of $CD_h^d$. The total number of these subsets of leaves that induce the tree $Q_k$ is given by $d \cdot c\big(Q_k, CD_{h-1}^d\big)$, as all the branches of $CD_h^d$ are isomorphic to $CD_{h-1}^d$;
- more than two of the branches of $CD_h^d$ contain at least one of the $k$ leaves. In this case the leaf-induced subtree is not isomorphic to $Q_k$ (as the root degree of $Q_k$ is 2);
- exactly two of the branches of $CD_h^d$ contain at least two of the $k$ leaves each. In this case the leaf-induced subtree is not isomorphic to $Q_k$ (as one of the branches of $Q_k$ is the single leaf);
- one branch of $CD_h^d$ contains exactly one of the leaves and another branch of $CD_h^d$ contains $k-1$ leaves. Since $k > 2$, the total number of these subsets of leaves that induce the tree $Q_k$ is given by $2 \cdot d^{h-1} \cdot c\big(S_{k-1}, CD_{h-1}^d\big)$ for every choice of two branches of $CD_h^d$.

Therefore, a recursion for $c\big(Q_k, CD_h^d\big)$ is given by

$$\begin{aligned} c\big(Q_k, CD_h^d\big) &= d \cdot c\big(Q_k, CD_{h-1}^d\big) + 2\binom{d}{2} \cdot d^{h-1} \cdot c\big(S_{k-1}, CD_{h-1}^d\big) \\ &= d \cdot c\big(Q_k, CD_{h-1}^d\big) + (d-1)d^h \binom{d}{k-1} \left( \frac{d^{(k-1)(h-1)} - d^{h-1}}{d^{k-1} - d} \right), \end{aligned}$$

where the last step uses the identity

$$c\big(S_k, CD_h^d\big) = \binom{d}{k} \frac{d^{k \cdot h} - d^h}{d^k - d} \tag{8}$$

valid for every $k \geq 2$ – formula (8) can be found in (Czabarka et al., 2020, proof of Theorem 1). The induction hypothesis gives

$$\begin{aligned} c\big(Q_k, CD_h^d\big) = \; & \frac{(d-1)\binom{d}{k-1}}{d^{k-1} - d} \cdot d^h \left( \frac{d^{(k-1)(h-1)} - d^{k-1}}{d^{k-1} - 1} - \frac{d^{h-1} - d}{d-1} \right) \\ & + (d-1)d^h \binom{d}{k-1} \left( \frac{d^{(k-1) \cdot (h-1)} - d^{h-1}}{d^{k-1} - d} \right) \end{aligned}$$

$$= \frac{(d-1)\binom{d}{k-1}}{d^{k-1}-d} \cdot d^h \left( d^{(k-1)(h-1)} + \frac{d^{(k-1)(h-1)} - d^{k-1}}{d^{k-1}-1} \right. \\ \left. - \frac{d^{h-1}-d}{d-1} - d^{h-1} \right),$$

which, after simplification, yields the desired equality. The statement on the inducibility follows by passing to the limit of the density $\gamma\big(Q_k, CD_h^d\big)$ as $h \to \infty$:

$$I_d(Q_k) \geq \lim_{h\to\infty} \gamma\big(Q_k, CD_h^d\big) = \frac{k!(d-1)\binom{d}{k-1}}{(d^{k-1}-d)(d^{k-1}-1)}\,.$$

□

We can now focus on Theorem 2.

**Proof of Theorem 2:** First off, we construct a new family of ternary trees: given a nonnegative integer $h \geq 0$, attach one copy of each of the complete ternary trees $CD_h^3$ and $CD_{h+1}^3$ to a common vertex (their respective roots are joined to a new vertex) to form a ternary tree which we shall name $W_h^3$. For example, $W_0^3$ is the tree $Q_4$. See also Figure 7 for the ternary tree $W_1^3$.

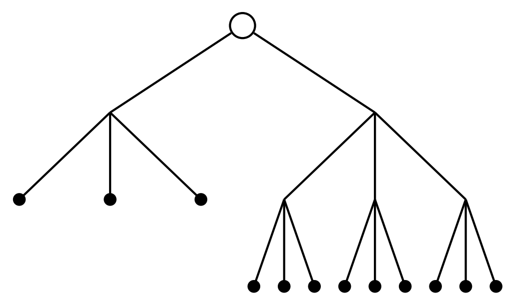


**Fig. 7:** The ternary tree $W_1^3$ defined in the proof of Theorem 2.

We shall prove that

$$\lim_{h\to\infty} \gamma\big(Q_4, W_h^3\big) = \frac{59}{416}\,.$$

To justify the specific choice, let us consider a more general construction. For positive integers $n_1$ and $n_2$, we consider the ternary tree (which we simply denote by $T_{n_1,n_2}$) whose branches are even ternary trees with $n_1$ and $n_2$ leaves, respectively. The even ternary tree $E_n^3$ with $n$ leaves is obtained recursively as follows: $E_1^3$ is the tree with only one vertex; $E_2^3$ is the star with two leaves; for $n > 2$, the branches of $E_n^3$ are the even ternary trees $E_{k_1}^3$, $E_{k_2}^2$ and $E_{k_3}^3$ with $k_1, k_2, k_3$ as equal as possible and $k_1 + k_2 + k_3 = n$.

According to Proposition 3, we have

$$c\big(Q_4, CD_h^d\big) = \frac{(d-2)d^{4\cdot h}}{6(d+1)(d^2+d+1)} + \mathcal{O}(d^{2\cdot h})$$

for every $d$ and all $h \geq 3$. In particular, the asymptotic formula

$$c\big(Q_4, CD_h^3\big) = \frac{1}{312} \cdot 3^{4\cdot h} + \mathcal{O}(3^{2\cdot h})$$

is obtained for all $h \geq 3$. On the other hand, we recall that the specialisation $k = 3$ in equation (8) of the proof of Proposition 3 gives

$$c\big(S_3, CD_h^3\big) = \frac{1}{24} \cdot 3^{3h} + \mathcal{O}(3^h)$$

for all $h \geq 1$, and employing the identity

$$c(S_3, E_n^3) = c(S_3, E_{k_1}^3) + c(S_3, E_{k_2}^3) + c(S_3, E_{k_3}^3) + k_1 \cdot k_2 \cdot k_3\,,$$

it is not difficult to show that

$$c(S_3, E_n^3) = \frac{1}{24} \cdot n^3 + \mathcal{O}(n^2)$$

for all $n$. Using the recursion

$$\begin{aligned} c\big(Q_4, E_n^3\big) = c\big(Q_4, E_{k_1}^3\big) + c\big(Q_4, E_{k_2}^3\big) + c\big(Q_4, E_{k_3}^3\big) + k_1 \cdot c\big(S_3, E_{k_2}^3\big) + k_2 \cdot c\big(S_3, E_{k_1}^3\big) \\ + k_1 \cdot c\big(S_3, E_{k_3}^3\big) + k_3 \cdot c\big(S_3, E_{k_1}^3\big) + k_2 \cdot c\big(S_3, E_{k_3}^3\big) + k_3 \cdot c\big(S_3, E_{k_2}^3\big)\,, \end{aligned}$$

we also find that

$$c\big(Q_4, E_n^3\big) = \frac{1}{312} \cdot n^4 + \mathcal{O}(n^3)$$

for all $n$. Moreover, the number of copies of $Q_4$ in any topological tree $T$ with two branches $T_1, T_2$ is given by

$$c(Q_4, T) = c(Q_4, T_1) + c(Q_4, T_2) + |T_1| \cdot c(S_3, T_2) + |T_2| \cdot c(S_3, T_1)\,.$$

For $x \in (0,1)$, set $n_1 = \lfloor x \cdot n \rfloor$ and $n_2 = \lfloor (1-x)n \rfloor$, and let $n \to \infty$. Combining all the formulas above, we see that an asymptotic formula for $c\big(Q_4, T_{n_1,n_2}\big)$ is given by

$$\begin{aligned} c\big(Q_4, T_{n_1,n_2}\big) &= \frac{1}{312}(x \cdot n)^4 + \frac{1}{312}\big((1-x)n\big)^4 \\ &\quad + x \cdot n \cdot \frac{1}{24}\big((1-x)n\big)^3 + (1-x)n \cdot \frac{1}{24}(x \cdot n)^3 + \mathcal{O}(n^3) \\ &= \frac{1}{312}\big(x^4 + (1-x)^4\big)n^4 + \frac{1}{24}\big(x(1-x)^3 + (1-x)x^3\big)n^4 + \mathcal{O}(n^3) \\ &= \frac{1}{312}\big(1 + 9x - 33x^2 + 48x^3 - 24x^4\big)n^4 + \mathcal{O}(n^3)\,. \end{aligned}$$

Set

$$f(x) = \frac{1}{312}\big(1 + 9x - 33x^2 + 48x^3 - 24x^4\big)\,.$$

The first derivative of this function is given by

$$f'(x) = \frac{-(2x-1)(4x-3)(4x-1)}{104}\,.$$

We see that $f(x)$ attains its maximum at $x = 1/4$ (or $x = 3/4$):

$$f(x) \leq f\Big(\frac{1}{4}\Big) = \frac{59}{9984}$$

for all $x \in (0,1)$. This motivates the choice of the trees $W_h^3$ defined before. We have

$$I_3(Q_4) \geq \lim_{h \to \infty} \gamma\big(Q_4, W_h^3\big) = \frac{4! \cdot 59}{9984} = \frac{59}{416}\,.$$

This completes the proof of the lower bound in the theorem.

The proof of the upper bound is also via an algorithmic approach and is quite similar to the one given for the binary tree $A_5$ in Section 3. Recall again that by Theorem 3 in (Czabarka et al., 2020),

$$I_3(Q_4) \leq \max_{\substack{|T|=n \\ T \text{ ternary tree}}} \gamma(Q_4, T),$$

so the aim is to compute the right hand side for different values of $n$. The algorithm is essentially the same as for $A_5$, with the trees $Q_4$ and $S_3$ assuming the roles of $A_5$ and $CD_2^2$ respectively. The only difference is that trees with two or three branches have to be considered in the construction of the sets $\mathcal{L}(n)$.

For the recursive calculation of $c(Q_4, T)$ and $c(S_3, T)$, we have the formulas

$$\begin{aligned} c(Q_4,T) &= c(Q_4,T_1) + c(Q_4,T_2) + c(Q_4,T_3) + |T_1| \cdot c(S_3,T_2) + |T_2| \cdot c(S_3,T_1) \\ &\quad + |T_1| \cdot c(S_3,T_3) + |T_3| \cdot c(S_3,T_1) + |T_2| \cdot c(S_3,T_3) + |T_3| \cdot c(S_3,T_2) \end{aligned}$$

and

$$c(S_3,T) = c(S_3,T_1) + c(S_3,T_2) + c(S_3,T_3) + |T_1| \cdot |T_2| \cdot |T_3|\,,$$

where $T_1, T_2, T_3$ are the branches of $T$. If there are only two branches, all terms involving $T_3$ can simply be left out.

Again, we have implemented the algorithm in Mathematica—the notebook can be found at `http://math.sun.ac.za/~swagner/TreeQ4Final`. The exact values of

$$b_n = \max_{\substack{|T|=n \\ T \text{ ternary tree}}} \gamma(Q_4, T)$$

have been determined for values of $n$ up to 500; see Table 2.

**Tab. 2:** Maximum density $b_n$ of $Q_4$ among $n$-leaf ternary trees

| $n$ | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| $\boldsymbol{b_n}$ | 1 | $\frac{2}{5}$ | $\frac{2}{5}$ | $\frac{2}{7}$ | $\frac{19}{70}$ | $\frac{5}{21}$ | $\frac{5}{21}$ |
| $n$ | 15 | 20 | 25 | 30 | 35 | 40 | 45 |
| $\boldsymbol{b_n}$ | $\frac{18}{91}$ | $\frac{291}{1615}$ | $\frac{1103}{6325}$ | $\frac{172}{1015}$ | $\frac{1097}{6545}$ | $\frac{7452}{45695}$ | $\frac{7948}{49665}$ |
| $n$ | 50 | 60 | 70 | 80 | 90 | 100 | 150 |
| $\boldsymbol{b_n}$ | 0.158072 | 0.155422 | 0.153588 | 0.152096 | 0.150978 | 0.150264 | 0.147342 |
| $n$ | 200 | 250 | 300 | 350 | 400 | 450 | 500 |
| $\boldsymbol{b_n}$ | 0.145967 | 0.145195 | 0.144651 | 0.144239 | 0.143931 | 0.143691 | 0.143506 |

We conclude that

$$I_3(Q_4) \le b_{500} = \frac{73848853}{514606225} \approx 0.143506\,.$$

This completes the proof of the theorem. □

## Conclusion

Naturally, the main open question left for us is

**Question 1.** *What are the precise values of $J(A_5) = I_2(A_5)$ and $I_3(Q_4)$?*

It is conceivable that the constructions yielding our lower bounds are asymptotically optimal, in which case the lower bounds would in fact be exact. A characterisation (at least an approximate characterisation) of the trees that attain the maxima of $\gamma(A_5, T)$ and $\gamma(Q_4, T)$ would be highly desirable.

Observe that the only known exact values of the inducibility are rational numbers. It was already asked in (Czabarka et al., 2017) whether the inducibility of trees (binary trees in that specific paper) is always rational. A natural generalisation would be the following question:

**Question 2.** *Are $I_d(S)$ and $J(S)$ rational for every rooted tree $S$ (and every integer $d \ge 2$)?*

The tree $A_5$ that we studied in this paper would be a natural candidate for a counterexample. For instance, our bounds show that $J(A_5) = I_2(A_5)$, if rational, would have to have a denominator of at least $89$. Thus an answer to the first question might immediately imply an answer to our second question.

Another natural direction of further research would be to search for other examples of trees whose inducibility can be determined explicitly (especially if this turns out to be too difficult for the trees $A_5$ and $Q_4$).

Finally, we would like to mention that there are other small binary trees $B$ for which the same algorithm described for the tree $A_5$ (see Section 3) can be exploited to determine the maximum number of copies of $B$ in a binary tree with $n$ leaves, thus an upper bound on $I_2(B) = J(B)$. Let $B_1$ and $B_2$ denote the two branches of $B$, and $B_{2,1}, B_{2,2}$ the two branches of $B_2$. If $1 \leq |B_1|, |B_{2,1}|, |B_{2,2}| \leq 3$, then one can apply the same algorithm with only very minor modifications. There are 17 nonisomorphic binary trees satisfying this criterion. The inducibility is presently known explicitly for only 7 of these trees.

# References

E. S. Allman and J. A. Rhodes. *Mathematical models in biology: an introduction*. Cambridge University Press, Cambridge, 2004. ISBN 9780511790911.

F. Bergeron, G. Labelle, and P. Leroux. *Combinatorial species and tree-like structures*. Cambridge University Press, Cambridge, 1st edition, 1998. ISBN 9780521573238.

S. Bubeck and N. Linial. On the local profiles of trees. *J. Graph Theory*, 81(2):109–119, 2016.

É. Czabarka, L. A. Székely, and S. Wagner. Inducibility in binary trees and crossings in random tanglegrams. *SIAM J. Discrete Math.*, 31(3):1732–1750, 2017.

É. Czabarka, A. A. V. Dossou-Olory, L. A. Székely, and S. Wagner. Inducibility of d-ary trees. *Discrete Math., in press. DOI: 10.1016/j.disc.2019.111671*, 2020.

A. A. V. Dossou-Olory and S. Wagner. Further results on the inducibility of d-ary trees. *Preprint arXiv:1811.11235*, 2018.

A. A. V. Dossou-Olory and S. Wagner. Inducibility of topological trees. *Quaest. Math.*, 42(6):749–764, 2019.

C. Even-Zohar and N. Linial. A note on the inducibility of 4-vertex graphs. *Graphs Combin.*, 31(5): 1367–1380, 2015.

G. Exoo. Dense packings of induced subgraphs. *Ars Combin.*, 22:5–10, 1986.

J. Hirst. The inducibility of graphs on four vertices. *J. Graph Theory*, 75(3):231–243, 2014.

N. Pippenger and M. C. Golumbic. The inducibility of graphs. *J. Combin. Theory Ser. B*, 19(3):189–203, 1975.

K. Sperfeld. The inducibility of small oriented graphs. *Preprint arXiv:1111.4813*, 2011.

# Appendix

The minimal polynomial of the lower bound derived in Section 4 (cf. Remark 1) is

$$
\begin{aligned}
&- 219990282547586266429960528777627452703544325176405813669341\\
&+ 1460204372604973227604751998057292514879880570165591870981228 0x\\
&- 4439880648861131188987438585934958372711164527752447209452465 60x^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 8191391786597997025923387108156673725457502710845806254299710400x^3\\
&- 10234975841656619685632205734115514398338672174441604510715235778 0x^4\\
&+ 9149157331040544275493200259078485367578591985556632766829251514 64x^5\\
&- 6021742541574757636997532900244251351617306701953896563191732661600x^6\\
&+ 2955550363332979997835217763567965156206341485426134176713245121 1440x^7\\
&- 1082036419607120379793994900091594730591033619129340814453695335 69710x^8\\
&+ 2918880206226716928188790803748145084433752815802397855148691520 91240x^9\\
&- 5639514531228009102068932876091422253490174867966157045903657251 41664x^{10}\\
&+ 7387718363412123491654794961916027294932665272629313655901961423 74880x^{11}\\
&- 5872133144147083947275071487424416677281361485965916039866930486 73940x^{12}\\
&+ 2119825531604947182889453010481324257697695624994895230703659964 62520x^{13}\\
&+ 2642044260670867601071997587023204415009550503501324422149424398 240x^{14}\\
&+ 22989587770824658009039088651658607134068726582600925337729500 00x^{15}\\
&+ 563916767637963643123242260073274239273437824576171875x^{16}.
\end{aligned}
$$