# Combinatorial Aspects of Leaf-Labelled Trees

Peter J. Humphries

---

A thesis
submitted in partial fulfilment
of the requirements for the degree
of
Doctor of Philosophy
in
Mathematics

---

University of Canterbury
Department of Mathematics and Statistics
2008

Πάντα χωρεῖ καὶ οὐδὲν μένει.

Everything changes; nothing remains the same.

(from Plato's *Cratylus*)

# TABLE OF CONTENTS

# ACKNOWLEDGEMENTS

This thesis is the culmination of several years of research that has proven both rewarding and enriching for me on a personal level. Naturally, there have also been many frustrating hours spent gazing blankly at a (sometimes also blank) whiteboard, but this too has been character-building in its own way. I am indebted to a number of people for their input during the last three years.

First and foremost, I am grateful to my supervisor Charles Semple for providing me with the opportunity to work alongside him and for his guidance throughout my graduate studies, and to Mike Steel for his advice and support in the capacity of co-supervisor.

Thank you to Stefan Grünewald and Taoyang Wu, two colleagues with whom I have collaborated on various projects. I have learned much from you both in the course of our discussions and work together. Additionally, thanks to Stefan for inviting me to the PICB in Shanghai for a research visit. Further thanks to Jim Geelen for hosting me at the University of Waterloo as a visiting researcher.

I extend my fullest appreciation to the administrative and technical staff in the Department of Mathematics and Statistics for taking care of those things that I don't understand, and to the University of Canterbury and the New Zealand Marsden Fund for providing me with financial support.

Lastly, thank you to my friends and family for your patience and assistance while I have been preparing this thesis for submission.

*Johnny Humphries*

# ABSTRACT

Leaf-labelled trees are used commonly in computational biology and in other disciplines, to depict the ancestral relationships and present-day similarities between both extant and extinct species. Studying these trees from a mathematical perspective provides a foundation for developing tools and techniques that have practical applications.

We begin by examining some *quartet* problems, namely determining the number of quartets that are required to infer the structure of a particular *supertree*. The *quartet graph* is introduced as a tool for tackling quartet problems, and is subsequently used to give new characterisations of *compatible*, *definitive* and *identifying* quartet sets.

We then turn to investigating some properties of the *subtrees* induced by a collection of trees. This is motivated in part by the problem of reconstructing two or more trees simultaneously from their combined collection of subtrees. We also use some ideas drawn from Ramsey theory to show the existence of arbitrarily large common subtrees.

Finally, we explore some extremal properties of the metric that is induced by the *tree bisection and reconnection* operation. This includes finding new (asymptotically) tight upper and lower bounds on both the size of the neighbourhoods in the metric space and on the diameter of the corresponding *adjacency graph*.

*To my parents, for your unfailing love and support.*

# Chapter 1

# Introduction

The Greek philosopher Heraclitus is famous for his philosophy of $\pi\alpha\nu\tau\alpha$ $\rho\epsilon\iota$ (*panta rhei*), that everything is in a constant state of change. This is certainly true within the context of any living system, and is the central driving force behind all forms of evolution.

By evolution, we are not necessarily limiting ourselves to the biological concept. Rather, we consider all settings in which information reproduces or is transferred to successive generations. Comparative linguistics, or comparative philology, studies the evolution of languages [4]. Stemmatology is another branch of philology that deals primarily with the reconstruction of original texts from surviving copies that may contain transmitted errors [38]. In the areas of anthropology and sociology, the development of material culture and culture in general provide further examples of evolutionary processes [30].

Each of the above examples of evolution exhibits three key features, namely reproduction, mutation and selection. Manuscripts and historical texts that were hand-copied by scribes incurred errors that compounded over successive copyings, with only some of these copies surviving to the present day. Language is passed through generations, and may develop independently in multiple locations depending on the dominant usage, resulting in the divergence of distinct languages. In both cases, there are clearly identifiable reproduction, mutation and selection phases, each helping to guide the overall behaviour of the evolving system.

The traditional approach to reconstructing an evolutionary history assumes that evolution is *tree-like*, and thus attempts to build up the most probable *family tree* to explain the available evidence. Figure 1.1 shows one

hypothesis of how the major modern Germanic languages[1] may have evolved [22]. The branches of the tree represent the lineages of each language since

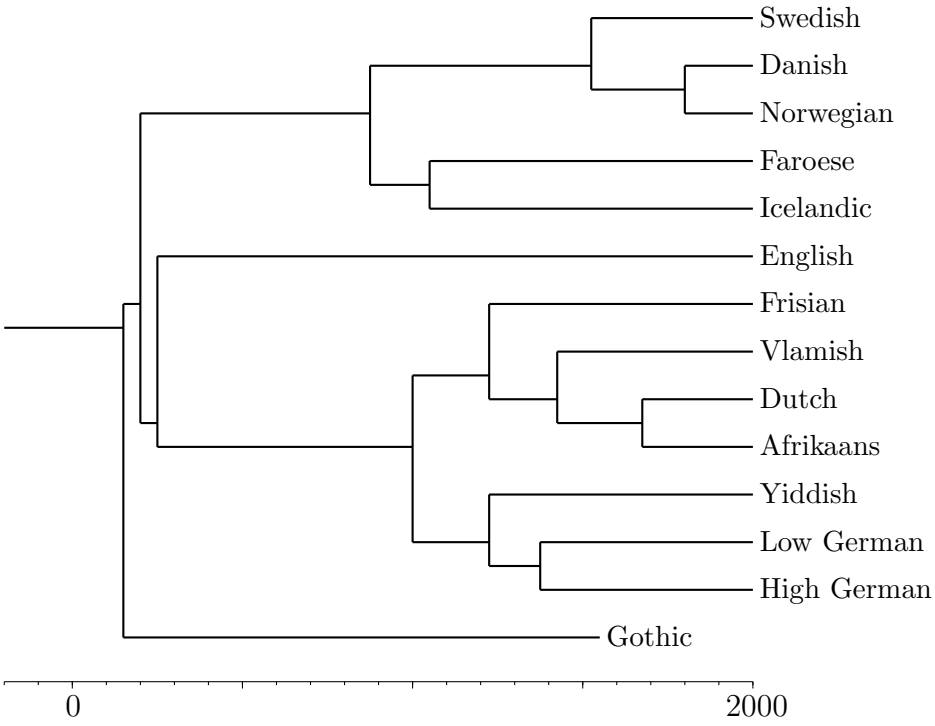


Figure 1.1: One hypothesis of the evolution of the major Germanic languages over the last two thousand years [22].

the emergence of the original Germanic tongue. Such trees allow us not only to trace back through the linguistic history of modern languages, but also to rebuild the vocabulary and grammar of extinct languages with no direct attestation. For example, each of the fourteen languages in Fig. 1.1 is derived from an extinct language known in linguistics as *Proto-Germanic*, the structure of which has been inferred by studying comparative evidence from its derivatives [20, 22].

Trees are not useful solely as a medium for depicting historical and present-day interrelatedness between types, by which we mean the subjects of the evolutionary process, and for piecing together proto-types. The mechanisms behind the mutation and selection phases of evolution can also be

[1]The Germanic languages are more broadly categorised as *Indo-European* [36].

better understood in the context of a reliable ancestral tree. Accurate dating within the tree may help with identifying specific events or patterns that precipitate change, which in turn facilitates projections of the likely future pattern of divergence.

Because there is necessarily a temporal component integral to the unfolding of an evolving population, the lengths of the *branches* within a tree are often indicative of the discrepancy that exists between species. Removing this dimension and ignoring the position of the root results in a tree that illustrates the relative similarities between the extant types in the system without making assumptions about which derive from a common ancestry. Figure 1.2 shows the reduction of the family tree for Germanic languages to this form.

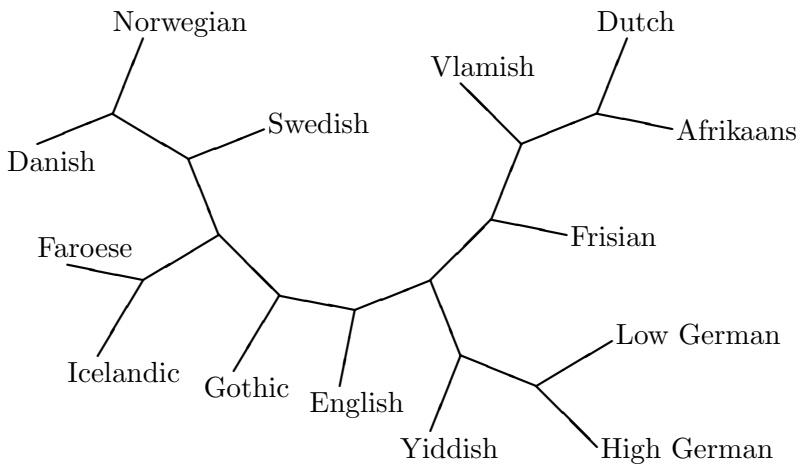


Figure 1.2: The unrooted tree underlying the hypothesised evolution of the Germanic languages as shown in Fig. 1.1.

Both the rooted and the unrooted trees that are used to represent evolution are essentially discrete mathematical objects and may be rigorously defined as such. By exploiting this precise foundation, we are able to better understand the structural properties of these objects. We can also convert vague problems that may arise in a practical setting into well-defined mathematical questions, the solutions to which may be applied back in the original context. Both of these approaches have proven invaluable in the development of analytical tools for studying the phenomenon of evolution across all

disciplines.

Contemporary computational and statistical techniques that are applied in the context of evolutionary systems are to a large extent underpinned by algebraic and combinatorial ideas. With regard to combinatorics, the area of primary importance is undoubtedly *graph theory*. In this thesis, we concentrate on deriving results that centre around unrooted *leaf-labelled trees*[2], with no regard for the length or weight of the edges. Whereas rooted trees demonstrate an ancestral hierarchy that is inherent in a collection of types, unrooted trees are representative of the similarities between the types.

Let us return to our example of the development of the Germanic languages. Starostin and Burlak recently proposed the tree in Fig. 1.3 (reproduced from [7]). The key point to note here is that the underlying unrooted

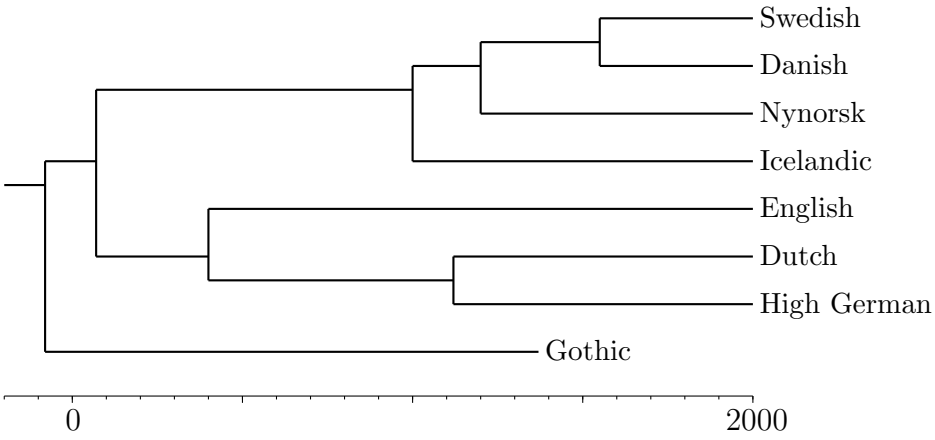


Figure 1.3: A second hypothesis of how the Germanic languages may have evolved (Starostin and Burlak, cited in [7]).

tree is entirely consistent with that displayed in Fig. 1.2. While the exact chronology between the trees differs, the shared ancestral relationships are identical[3]. In fact, the appropriate time scale can be reintroduced in either case by inserting a root on the branch that separates Gothic from the other

[2]In the context of evolutionary biology, these trees are often referred to as *phylogenetic* or *evolutionary trees*. We will use the more inclusive terminology *leaf-labelled trees* to emphasise their wider application in fields other than biology.

[3]*Nynorsk* is based on dialects of Norwegian that predate the Danish rule of Norway, and is currently used by around 10% of the Norwegian population [4].

languages, and by dating the time of each divergence event. This hints at the significance of unrooted trees in an algorithmic role, where they can serve as an intermediate stage between raw empirical evidence and a rooted tree.

To conclude this introduction, we will give a brief synopsis of the main body of the thesis. More detailed overviews of the individual chapters are given prior to each of the three parts. As we have mentioned, we will be concerned primarily with combinatorial problems that arise from unrooted trees. The necessary background for understanding the mathematical content of this thesis is presented in Chapter 2. We define any specific notation and terminology that is used, although the reader should note that a basic knowledge of graph theory is required.

One common and intuitive method of building leaf-labelled trees from data is to begin by constructing smaller trees, and to then piece together these partial trees into one comprehensive tree. In Part I, *Supertrees*, we examine some specific problems related to these methods of reconstruction. In particular, we consider how much information is required to conclusively indicate the entire structure of the original tree. This amounts to distilling a given tree down to its most primitive compositional units, from which the tree can then be inferred with complete accuracy.

The two chapters making up Part II, *Subtrees*, approach a similar idea from two opposing perspectives. Viewing trees as mathematical structures, we can describe their substructures (predictably called *subtrees*) precisely. In Chapter 6, we assume a set of trees is given, and look for a subtree that is as small as possible and is different in each of the given trees. This corresponds to finding a small collection of types that has evolved differently in each tree from a hypothetical collection of trees. In Chapter 7 we turn this around. Again assuming a given set of trees, we look for a subtree that is as large as possible and is the same in each of the trees, or alternatively a large collection of types that has evolved in the same manner in each tree.

The final part of the thesis (Part III, *Tree Rearrangement Operations*) is centred around deformations that can be performed on unrooted trees. Loosely speaking, these deformations (*tree rearrangment operations*) are a way to transform one tree into another by altering the structure in some predefined way. The specific problems that we are interested in are finding the

number of trees that can be obtained from a given tree by a single operation, and also determining the minimum number of operations that separate two trees.

In the appendices that follow the main body of this work are three academic papers that were written during the same period as this thesis. One of these (Appendix C) is related to the work in Part III, while the other two have no connection to the thesis other than the authorship.

The approach taken makes no prior assumption as to the origin of the trees, merely treating them as abstract mathematical objects. Some of the problems considered have been addressed previously by others, whether directly for unrooted trees or in the analogous rooted setting. Other topics of investigation are to the best of our knowledge entirely original, at least in the context of leaf-labelled trees. Unless explicitly noted otherwise, all the results contained in this thesis are new and are the author's own work.

The purpose of this research, then, is firstly to expand on some known results, either by extending a partial solution, by generalising the problem in some way or by providing an alternative proof. Secondly, some of the ideas developed may be directly applicable in the design of efficient algorithms for tackling the large volume of data that is generated on a daily basis in some fields of research. Thirdly, the derivation of some extremal results and their corresponding characterisations that may (tenuously) be of use in complexity analysis. Fourthly, finally, and most importantly, is the satisfaction of intellectual curiosity on a purely combinatorial level.

# Chapter 2

# Mathematical Preliminaries

This chapter is devoted to introducing the notation and terminology required for a complete understanding of the main body of this thesis. It is assumed that the reader is familiar with the fundamentals of graph theory. A comprehensive background to this area of discrete mathematics may be found in *Modern Graph Theory* [10], or a similar introductory text.

Most of the general mathematical nomenclature we use follows accepted convention, but there are some points that we wish to emphasise at this stage to prevent confusion later. For a positive integer $n$, we employ the shorthand $[n]$ to represent the set $\{1, \ldots, n\}$ where convenient. The collection of subsets of a set $X$ is given by $2^X$, while we use $\binom{X}{k}$ to represent the collection of $k$-element subsets of $X$. We also make the distinction between '$\subset$' and '$\subseteq$', with the former denoting strict containment.

A *leaf-labelled tree* is a tree, by which we specifically mean an unrooted tree, that has no vertices of degree two and a unique label assigned to each vertex of degree one. To be more precise, let $T$ be a tree with vertex set $V$ such that the set $\{v \in V : d(v) = 2\}$ is empty, where $d(v)$ denotes the degree of the vertex $v$. Further, let

$$\phi : X \to \{v \in V : d(v) = 1\}$$

be a bijective function for some set $X$. Then $\mathcal{T} = (T; \phi)$ is a leaf-labelled tree. We refer to $X$ as the *leaf set* of $\mathcal{T}$, and write $X = \mathcal{L}(\mathcal{T})$. Further to this, we use $\mathscr{T}_X$ to denote the set of all leaf-labelled trees that have $X$ as the leaf set.

The degree one vertices of $\mathcal{T}$ are called the *leaves* of $\mathcal{T}$, while all other vertices are *interior vertices*. If an interior vertex of a tree $\mathcal{T}$ is adjacent to

exactly two distinct leaves $x$ and $y$, then we call the set $\{x, y\}$ a *cherry* of $\mathcal{T}$. An edge of $\mathcal{T}$ is an *interior edge* if both endpoints are interior vertices, and similarly an *interior path* has two interior vertices as end points. All non-interior edges of $\mathcal{T}$ are called *pendant edges*.

For a tree $\mathcal{T} \in \mathscr{T}_X$, and a subset $Y \subseteq X$ of the leaf set, we define $\mathcal{T}|Y$ to be to be the minimal subgraph of $\mathcal{T}$ that that has the leaf set $Y$ and has all degree two vertices suppressed. We call $\mathcal{T}|Y$ the *restriction* of $\mathcal{T}$ to $Y$, and say that $\mathcal{T}|Y$ is a *subtree* of $\mathcal{T}$ or alternatively, that $\mathcal{T}$ *displays* $\mathcal{T}|Y$.

In the terms of graph minors, forming a subtree of a leaf-labelled tree corresponds to deletion. Let $e$ be an interior edge of $\mathcal{T} \in \mathscr{T}_X$, and let $\mathcal{T}' \in \mathscr{T}_X$ be the tree formed by contracting $e$. In this case, $\mathcal{T}$ is called a *refinement* of $\mathcal{T}'$. However, it should be noted that $\mathcal{T}'$ is not considered a subtree of $\mathcal{T}$.

A leaf-labelled tree is *binary* if every interior vertex has degree three. These are the trees which we are primarily interested in, and to this end we define $\mathscr{T}_n$ to be the set of all binary leaf-labelled trees with the leaf set $\{1, \ldots, n\}$. Both $\mathscr{T}_n$ and $\mathscr{T}_X$ will be referred to as *tree spaces*, with appropriate clarification given if there is any possible ambiguity. It should be noted (see [42] for details) that all binary trees with $n$ leaves have exactly $n-3$ interior edges and $n$ pendant edges, and that the size of $\mathscr{T}_n$ is precisely

$$(2n-5)!! = 1 \times 3 \times \cdots \times (2n-5).$$

One particular tree shape that appears frequently throughout this thesis because of its nice properties is the *caterpillar*. A caterpillar is a binary leaf-labelled tree that has at least four leaves and precisely two cherries. Since any two cherries must by definition be disjoint, it follows that all the trees in $\mathscr{T}_4$ and $\mathscr{T}_5$ are caterpillars. Extending our earlier notation for trees, we use $\mathscr{C}_n \subseteq \mathscr{T}_n$ to denote the set of caterpillars with the leaf set $\{1, \ldots, n\}$.

Figure 2.1 shows a caterpillar $\mathcal{C} \in \mathscr{C}_8$ with cherries $\{3, 4\}$ and $\{5, 7\}$. The *label ordering* of a caterpillar is a permutation of the leaf set in which the leaves occur in the order they appear on the caterpillar. Thus $\mathcal{C}$ has the label ordering $[3, 4, 6, 2, 1, 8, 7, 5]$. We note that as a consequence of graph isomorphisms, a label ordering is not unique. Alternative label orderings may

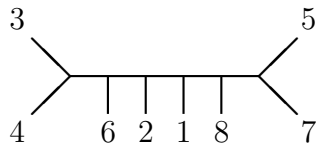


Figure 2.1: A caterpillar $\mathcal{C} \in \mathscr{C}_8$.

be found by either transposing the first pair of elements in the permutation, or by reversing the entire permutation, but we stress that each permutation of a set $X$ is a label ordering for a unique caterpillar.

A *split* of a tree $\mathcal{T}$ is a bipartition of its leaf set induced by deleting an edge. The two blocks of the partition are the leaf sets of each of the two components that result when the edge is deleted. In the example above (see Fig. 2.1), if we let $A = \{3, 4, 6\}$ and $B = \{1, 2, 5, 7, 8\}$, then the bipartiton $A, B$ of $\{1, \ldots, 8\}$ is a split of $\mathcal{C}$. We use the notation $A|B$ to represent this split. The collection of all splits of a tree $\mathcal{T}$ is denoted by $\Sigma(\mathcal{T})$.

Related to the idea of a split is that of a *cluster*. A subset $Y \subset X$ is a cluster of $\mathcal{T} \in \mathscr{T}_X$ if and only if $Y|X - Y$ is a split of $\mathcal{T}$. Clusters give rise to a certain type of subtree. For a cluster $Y$ of $\mathcal{T}$, the subtree $\mathcal{T}|Y$ is known as a *pendant* subtree.

If $A|B$ is a split of a tree $\mathcal{T}$, then for all $A' \subseteq A, B' \subseteq B$ we call $A'|B'$ a *partial split* of $\mathcal{T}$. A (partial) split $A|B$ of a tree is *non-trivial* if both $A$ and $B$ contain at least two elements. If both $A$ and $B$ contain exactly two elements, then $q = A|B$ is a *quartet* of $\mathcal{T}$. Moreover, if $A = \{a_1, a_2\}$ and $B = \{b_1, b_2\}$, then provided no ambiguity arises the notation is simplified to $q = a_1a_2|b_1b_2$. We use $\mathcal{Q}(\mathcal{T})$ to represent the set of all quartets of a tree $\mathcal{T}$.

Paralleling the definition we gave for trees earlier, the leaf set of a partial split $\sigma = A|B$ is written $\mathcal{L}(\sigma)$, and is the union of $A$ and $B$. Similarly, the leaf set of the quartet $q = a_1a_2|b_1b_2$ is $\mathcal{L}(q) = \{a_1, a_2, b_1, b_2\}$. It will frequently be convenient to refer to the leaf set of a collection of partial splits, quartets or trees. To do so, we simply extend the current notation so that the leaf set of a collection is the union over the leaf sets of each member of the collection.

For example, for a set of quartets $\mathcal{Q}$, the leaf set of $\mathcal{Q}$ is

$$\mathcal{L}(\mathcal{Q}) = \bigcup_{q \in \mathcal{Q}} \mathcal{L}(q).$$

We develop the notation for induced quartets in the same way. Thus, for a (partial) split $\sigma = A|B$, the quartet $q = a_1a_2|b_1b_2$ is in $\mathcal{Q}(\sigma)$ if and only if $a_1, a_2 \in A$ and $b_1, b_2 \in B$.

The term quartet is also used to refer a binary tree with four leaves. More specifically, the quartet $q = a_1a_2|b_1b_2$ can be seen as the leaf-labelled tree with four leaves and the single non-trivial split $q$. When a tree displays a quartet, that quartet *distinguishes* a unique interior path of $\mathcal{T}$. That is, if $\mathcal{T}$ displays $q = a_1a_2|b_1b_2$, then $q$ distinguishes the minimal path $v_0, \ldots, v_k$, where $v_0$ lies on the path from $a_1$ to $a_2$ and $v_k$ lies on the path from $b_1$ to $b_2$. Distinguishing edges will be of more importance in this thesis than the more general concept of distinguishing paths.

As the majority of the nomenclature used in this thesis is particular to the individual chapters, we have chosen to exclude a list of commonly used notation and trust that this does not hinder the reader in any way.

# PART I

# SUPERTREES

A fundamental way in which leaf-labelled trees are inferred is by amalgamating a collection $\mathcal{P}$ of smaller trees on overlapping subsets of species into a single parent tree. Collectively, such amalgamation methods are known as *supertree methods* and the resulting parent tree is called a *supertree*. The popularity of supertree methods is highlighted in [5, 6].

If the amalgamating collection $\mathcal{P}$ contains no conflicting information, then $\mathcal{P}$ is said to be *compatible*. Furthermore, $\mathcal{P}$ is *definitive* if $\mathcal{P}$ is compatible and there is exactly one supertree that displays all of the ancestral relationships displayed by the trees in $\mathcal{P}$. Precise definitions of these concepts are given in the ensuing chapters. Within the context of supertree methods, two natural mathematical problems arise:

(i) is $\mathcal{P}$ compatible; and if so,

(ii) is $\mathcal{P}$ definitive?

As computational problems, (i) is known to be NP-complete [9, 45], while the complexity of the second problem continues to remain open. Nevertheless, there are attractive characterisations of these problems in terms of chordal graphs [18, 34, 41, 45]. An overview these characterisations may also be found in Section 4.4.

In practice, while a collection $\mathcal{P}$ of leaf-labelled trees might be compatible, it is unlikely to be definitive. A closely related notion, and one that is essentially as good, is the following: $\mathcal{P}$ *identifies* a supertree $\mathcal{T}$ if $\mathcal{T}$ displays $\mathcal{P}$ and all other supertrees that display $\mathcal{P}$ are refinements of $\mathcal{T}$. This means

that if $\mathcal{P}$ identifies a supertree, then the collection of supertrees that display $\mathcal{P}$ is well understood. This gives rise to a third mathematical problem:

(iii) does $\mathcal{P}$ identify a supertree?

Like problems (i) and (ii), a characterisation of this problem has also been given in terms of chordal graphs [13].

Each of problems (i), (ii), and (iii) are typically stated in terms of collections of quartets—that is, binary leaf-labelled trees with four leaves—rather than an arbitrary collection of trees. The reason for this is that a leaf-labelled tree is completely determined by its collection of induced quartets (see, for example, [42]). Consequently, for the purposes of this thesis, we will view the input to the problems specified above as collections of quartets.

In Chapter 3, we present some results on definitive quartet sets. Chapters 4 and 5 are based on [24], which was written jointly with Stefan Grünewald and Charles Semple. The first of these chapters introduces the *quartet graph*, a new tool for approaching quartet-based problems, while the second applies the quartet graph to finding the minimum size of an identifying quartet set for an arbitrary leaf-labelled tree.

# Chapter 3

# Definitive Quartet Sets

## 3.1 Introduction

The ideal situation when reconstructing a leaf-labelled tree from empirical data is for there to be a unique tree that fits the entire set of data perfectly. In terms of quartet-based reconstruction methods, this means that every quartet in the input is displayed by the output tree, and that there is no other tree with this property.

Suppose that $\mathcal{Q}$ is a set of quartets on the leaf-set $X = \mathcal{L}(\mathcal{Q})$. We say that $\mathcal{Q}$ is *compatible* if and only if some tree $\mathcal{T} \in \mathscr{T}_X$ displays every quartet in $\mathcal{Q}$. Equivalently, $\mathcal{Q}$ is compatible if and only if $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ for some $\mathcal{T} \in \mathscr{T}_X$. For example, the quartet set $\mathcal{Q} = \{12|34, 13|45\}$ is displayed by the tree $\mathcal{T} \in \mathscr{T}_5$ shown in Fig. 3.1, and hence $\mathcal{Q}$ is compatible. Obviously, compatibility is a

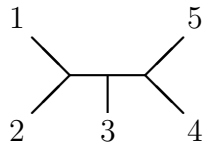


Figure 3.1: A tree $\mathcal{T} \in \mathscr{T}_5$ that displays the quartet set $\mathcal{Q} = \{12|34, 13|45\}$.

desirable property for reconstructing a tree from quartet data. If we attempt to construct a tree from an incompatible set, then some of the original data will necessarily be contradicted.

The compatibility of splits is defined in the same way as for quartets. That is, a collection $\Sigma$ of partial splits of $X$ is *compatible* if there is a leaf-labelled tree that displays each of the splits in $\Sigma$. The following result due

to Buneman [17] shows that every leaf-labelled tree is determined by its collection of non-trivial splits.

**Theorem 3.1.1** (Splits-Equivalence Theorem)**.** *Let $\Sigma$ be a non-trivial collection of splits of a set $X$. Then the following statements are equivalent:*

*(i) there is a leaf-labelled tree $\mathcal{T} \in \mathscr{T}_X$ such that $\Sigma$ is the set of non-trivial splits of $\mathcal{T}$;*

*(ii) $\Sigma$ is pairwise compatible;*

*(iii) for each pair $A_1|B_1$ and $A_2|B_2$ of splits in $\Sigma$, at least one of the sets $A_1 \cap A_2$, $A_1 \cap B_2$, $B_1 \cap A_2$, and $B_1 \cap B_2$ is empty.*

*Moreover, if such a tree exists, then, up to isomorphism, $\mathcal{T}$ is unique.*

A quartet set $\mathcal{Q}$ on the leaf-set $X = \mathcal{L}(\mathcal{Q})$ is *definitive* if and only if the following two conditions hold:

(i) $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ for some $\mathcal{T} \in \mathscr{T}_X$; and

(ii) $\mathcal{Q} \nsubseteq \mathcal{Q}(\mathcal{T}')$ for all $\mathcal{T}' \in \mathscr{T}_X - \mathcal{T}$.

In this case, we say that $\mathcal{Q}$ *defines* $\mathcal{T}$. That is, $\mathcal{T}$ is the unique tree that displays $\mathcal{Q}$ and has no extraneous leaves. It further follows that if $\mathcal{Q}$ defines a tree $\mathcal{T}$, then $\mathcal{T}$ is binary.

Again using the example from earlier, the quartet set $\mathcal{Q} = \{12|34, 13|45\}$ defines the tree $\mathcal{T}$ in Fig. 3.1. To confirm this claim, it suffices to show that no other tree in $\mathscr{T}_5$ displays $\mathcal{Q}$. On the other hand, both $\mathcal{T}_1$ and $\mathcal{T}_2$ shown in Fig. 3.2 display the set $\mathcal{Q} = \{12|34, 12|35\}$, and so this is not a definitive set of quartets.

It is well-known that if $\mathcal{Q}$ defines $\mathcal{T}$, then $\mathcal{Q}$ distinguishes every interior edge of $\mathcal{T}$. Otherwise, suppose that some interior edge $e$ of $\mathcal{T}$ is not distinguished by $\mathcal{Q}$. Then we can contract $e$ to form a tree $\mathcal{T}'$ that is distinct from $\mathcal{T}$ but still displays $\mathcal{Q}$. Moreover, there are definitive quartet sets that contain precisely one quartet for each interior edge of the tree they define. As a binary tree with $n$ leaves has precisely $n-3$ interior edges, we have the following result (see, for example, [42, Corollary 6.3.10]).

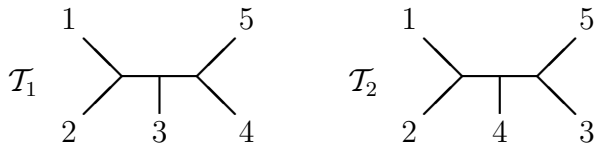


Figure 3.2: Two trees $\mathcal{T}_1, \mathcal{T}_2 \in \mathscr{T}_5$ which both display the quartet set $\mathcal{Q} = \{12|34, 12|35\}$.

**Theorem 3.1.2.** *Let $\mathcal{T} \in \mathscr{T}_n$ be a tree with $n \geq 4$ leaves. Then there is a set of $n-3$ quartets that defines $\mathcal{T}$.*

A definitive quartet set $\mathcal{Q}$ is *minimal* if $\mathcal{Q}$ defines some tree $\mathcal{T}$, but $\mathcal{Q} - q$ does not define $\mathcal{T}$ for all $q \in \mathcal{Q}$. That is, minimal definitive quartet sets contain no redundant information. Using the same example as previously, $\mathcal{Q} = \{12|34, 13|45\}$ is a minimal definitive quartet set for $\mathcal{T}$ shown in Fig. 3.1.

The remainder of the chapter is structured as follows. Section 3.2 introduces the idea of closure, and develops some inference rules for quartets and, more generally, partial splits. In Section 3.3, we use these rules to reprove the result of Mossel and Steel's [35] that a *generous cover* defines a tree, and then show that a slightly weakened version of this theorem does not hold. It has been informally conjectured [14] that the size of a minimal definitive quartet set for $\mathcal{T} \in \mathscr{T}_n$ is bounded by $n + c$ for some fixed constant $c$. We conclude in Section 3.4 by showing that a minimal definitive quartet set may in fact be as large as $\frac{3}{2}n$ for a tree with $n$ leaves.

## 3.2 Closure and Inference Rules

Let us pose the following question. If we are given a set of quartets with the knowledge that they are all displayed by some leaf-labelled tree, what information can we deduce about the tree from the quartet set? Can we infer any further quartets that must also be displayed by this tree? We remark at this point that we have implicitly assumed compatibility of the quartet set, for otherwise the tree we are looking for does not exist.

Consider the example from Fig. 3.1 in the previous section, where a set

of only two quartets defined a five-leafed tree. In terms of the question we asked above, if some tree displays both 12|34 and 13|45, then that same tree is also guaranteed to display the three quartets $\{12|35, 12|45, 23|45\}$.

On the other hand, the following example shows that at least two possible trees display both of 12|34 and 12|35. In fact, a third distinct tree $\mathcal{T}_3 \in \mathscr{T}_5$ (see Fig. 3.3) also displays these two quartets. A quick check shows that $\mathcal{T}_1$,

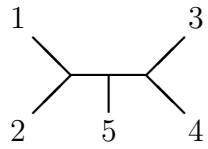


Figure 3.3: A third tree $\mathcal{T}_3 \in \mathscr{T}_5$ that displays the quartet set $\mathcal{Q} = \{12|34, 12|35\}$.

$\mathcal{T}_2$ and $\mathcal{T}_3$ share precisely the quartets 12|34, 12|35 and 12|45.

Using these ideas, we define the *closure* of a compatible quartet set $\mathcal{Q}$ to be

$$\mathrm{cl}(\mathcal{Q}) = \bigcap_{\mathcal{T}} \mathcal{Q}(\mathcal{T}),$$

where the intersection is taken over all trees $\mathcal{T}$ that display $\mathcal{Q}$. From the examples above, we deduce that

$$\begin{aligned} \mathrm{cl}(\{12|34, 12|35\}) &= \mathcal{Q}(\mathcal{T}_1) \cap \mathcal{Q}(\mathcal{T}_2) \cap \mathcal{Q}(\mathcal{T}_3) \\ &= \{12|34, 12|35, 12|45\}, \end{aligned}$$

and

$$\mathrm{cl}(\{12|34, 13|45\}) = \mathcal{Q}(\mathcal{T}_1).$$

For partial splits, we define the closure in much the same way. That is, if $\Sigma$ is a compatible collection of partial splits, then the *closure* of $\Sigma$ is written $\mathrm{cl}(\Sigma)$, and contains all the common partial splits across those trees that display $\Sigma$.

Closure operators appear throughout mathematics and share certain use-

ful properties. We refer the interested reader to [16, Chapter 3] for a more detailed discussion of the closure operator for quartets than has been given here.

We now turn our attention to inference rules for quartets. Suppose that, for a compatible set of quartets $\mathcal{Q}$ and some quartet $q$, we have

$$q \in \mathrm{cl}(\mathcal{Q}).$$

That is, $q \in \mathcal{Q}(\mathcal{T})$ for every tree $\mathcal{T}$ that displays $\mathcal{Q}$. Then we write

$$\mathcal{Q} \vdash q,$$

and refer to this as a *quartet rule*. Returning to our somewhat overused example, we find that

$$\{12|34, 12|35\} \vdash 12|45$$

is a valid quartet inference rule.

For completeness, we generalise the concept of inference rules to deal with partial splits. If $\sigma$ is a partial split in $\mathrm{cl}(\Sigma)$ for some compatible set of partial splits $\Sigma$, then we write

$$\Sigma \vdash \sigma.$$

The statement $\Sigma \vdash \sigma$ is called a *partial split rule*. While it has been demonstrated that no information would be lost by reducing $\Sigma$ to its set of induced quartets, it is frequently more straightforward to deal directly with partial splits than with quartets.

Let us restate the rule that we mentioned above in a more complete form (see [19]):

$$\{ab|cd, ab|ce\} \vdash ab|cde. \tag{3.1}$$

We will refer to (3.1) as the *dyadic closure rule*. A *triadic closure rule* is introduced and used in Chapter 5. Our next rule says that, if $A_1|B_1$ and

$A_2|B_2$ are partial splits with $A_1 \cap A_2 \neq \emptyset$ and $B_1 \cap B_2 \neq \emptyset$, then

$$\{A_1|B_1, A_2|B_2\} \vdash (A_1 \cap A_2)|(B_1 \cup B_2). \tag{3.2}$$

The rule (3.2) is Rule 1 in [34], and is known as the *split closure rule*. We observe also that (3.1) is a special case of (3.2).

The next lemma is obtained by repeated application of (3.1). The proof is routine and thus omitted.

**Lemma 3.2.1.** *Let $\sigma$ be a non-trivial partial split of a set $X$. Then*

$$\mathcal{Q}(\sigma) \vdash \sigma.$$

Suppose that we wish to prove that some partial split rule is valid. That is, we wish to show that $\Sigma \vdash \sigma$ holds for some choice of $\Sigma$ and $\sigma$. Assuming that $\Sigma$ is a compatible set of partial splits, it suffices by Lemma 3.2.1 to show that $\Sigma \vdash q$ for all $q \in \mathcal{Q}(\sigma)$. This fact will come in useful in proving a number of results throughout the first part of this thesis.

We conclude this section with a lemma that will be of more immediate use in Section 3.3. This lemma is essentially a partial split rule, and the proof requires several uses of the split-closure rule (3.2).

**Lemma 3.2.2.** *Let $A_1, A_2, B_1, B_2$ be non-empty disjoint sets with $a_i \in A_i, b_i \in B_i$, and let $A = A_1 \cup A_2, B = B_1 \cup B_2$. If*

$$\Sigma = \{A_1|B, A_2|B, A|B_1, A|B_2, a_1a_2|b_1b_2\},$$

*then $\Sigma \vdash A|B$.*

*Proof.* Applying (3.2), we find that $\Sigma \vdash \{a_1, a_2\}|B_1 \cup b_2$, from which it follows that $\Sigma \vdash \{a_1, a_2\}|B$. Two further applications of (3.2) complete the result. □

## 3.3 Covers of Trees

We earlier stated a theorem (Theorem 3.1.2) which was underpinned by the idea that it is possible to define any binary leaf-labelled tree $\mathcal{T}$ by choos-

ing, for each interior edge $e$ of $\mathcal{T}$, a single quartet that distinguishes that edge. We also justified why distinguishing every interior edge of a tree is necessary if we wish to define that tree. However, this same condition is not sufficient as the example in Fig. 3.4 shows. Both trees $\mathcal{T}_1$ and $\mathcal{T}_2$ in Fig. 3.4

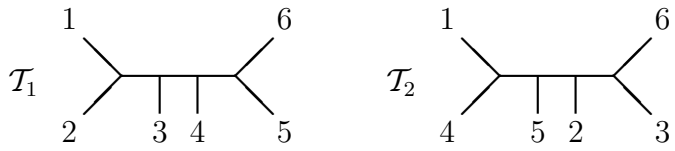


Figure 3.4: Two trees $\mathcal{T}_1, \mathcal{T}_2 \in \mathscr{T}_6$ which both display the quartet set $\mathcal{Q} = \{12|36, 23|45, 14|56\}$.

display the set of quartets $\mathcal{Q} = \{12|36, 23|45, 14|56\}$, and further $\mathcal{Q}$ distinguishes every interior edge of both of these trees. And yet, since they both display $\mathcal{Q}$, neither of them is defined by $\mathcal{Q}$.

Let us instead take this idea to the other extreme. Suppose we choose, for some binary leaf-labelled tree $\mathcal{T}$, an arbitrary set of quartets $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ such that every interior path of $\mathcal{T}$ is distinguished by some element of $\mathcal{Q}$. The question we then ask is whether $\mathcal{Q}$ is guaranteed to define $\mathcal{T}$, or whether in fact we can still find a non-definitive example. To this end, we formalise the concept we have just outlined.

**Definition 3.3.1.** Let $\mathcal{T}$ be a binary leaf-labelled tree. Then a set of quartets $\mathcal{Q}$ is a *generous cover* for $\mathcal{T}$ if and only if

(i) $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$; and

(ii) every interior path of $\mathcal{T}$ is distinguished by some quartet $q \in \mathcal{Q}$.

The notion of a generous cover was introduced by Mossel and Steel in [35] for investigating tree reconstruction under a random cluster model. Let us restate a key theorem from this paper here:

**Theorem 3.3.2** (Theorem 2.4, [35])**.** *Let $\mathcal{T}$ be a binary leaf-labelled tree, and let the set of quartets $\mathcal{Q}$ be a generous cover for $\mathcal{T}$. Then $\mathcal{Q}$ defines $\mathcal{T}$.*

This theorem answers the question we posed above in the affirmative. That is, a single quartet for each interior path of $\mathcal{T}$ suffices to define $\mathcal{T}$. We note that the size of a generous cover for a tree with $n$ leaves is at least $\binom{n-2}{2}$, the number of distinct interior paths in the tree. It can be shown quite easily, however, that a generous cover for a tree is not a minimal defining set. That is, if $\mathcal{Q}$ is a generous cover for $\mathcal{T}$, then there is some strict subset $\mathcal{Q}' \subset \mathcal{Q}$ such that

$$\mathrm{cl}(\mathcal{Q}') = \mathrm{cl}(\mathcal{Q}) = \mathcal{Q}(\mathcal{T}).$$

As a simple example to demonstrate this, consider the tree $\mathcal{T}_1 \in \mathscr{T}_6$ shown in Fig. 3.4. If $\mathcal{Q}$ is a generous cover for $\mathcal{T}_1$, then there is some $x \in \{4, 5, 6\}$ such that $12|3x \in \mathcal{Q}$. If $x = 4$, then there is also some $y \in \{5, 6\}$ such that $12|xy \in \mathcal{Q}$. We may assume without loss of generality that $y = 5$, in which case from (3.1) we have $12|35 \in \mathrm{cl}(\mathcal{Q})$. On the other hand, if $x \neq 4$, then we may assume that $x = 5$. In either case, we have shown that $12|35 \in \mathrm{cl}(\mathcal{Q})$. By following the same logic, we may argue further that $23|56$ is also in the closure of $\mathcal{Q}$. These last two quartets infer $12|56$ by (3.1), but we know that $12|56$ is in $\mathcal{Q}$, since $\mathcal{Q}$ is a generous cover for $\mathcal{T}_1$. That is,

$$\mathrm{cl}(\mathcal{Q} - \{12|56\}) = \mathrm{cl}(\mathcal{Q}).$$

This argument may be extended quite easily to caterpillars of any length, and from there to arbitrary binary trees.

The original proof of Theorem 3.3.2 given in [35] relies on the construction, at least theoretically, of an auxiliary graph. We will reprove the theorem here by first distilling the notion of a generous cover from whole trees to clusters within trees.

**Definition 3.3.3.** Let $\mathcal{T} \in \mathscr{T}_X$ be a binary tree, and let $Y \subset X$ be some cluster of $\mathcal{T}$. For some $y \in Y$, let $\mathcal{Q}'$ be a generous cover for $\mathcal{T}|(X - Y) \cup y$. Then a set of quartets $\mathcal{Q}$ is a *subcover* for $Y$ in $\mathcal{T}$ if and only if $\mathcal{Q} \cup \mathcal{Q}'$ is a generous cover for $\mathcal{T}$.

Essentially, a subcover is a set of quartets that distinguishes each path that has at least one endpoint within the relevant cluster. A straightforward

but useful result is that a subcover for some cluster $Y$ is also a subcover for all clusters $Z$, where $Z \subseteq Y$.

**Lemma 3.3.4.** *Let $\mathcal{T} \in \mathscr{T}_X$ be a binary tree, and let $Z \subseteq Y \subseteq X$. If $\mathcal{Q}$ is a subcover for $Y$ in $\mathcal{T}$, then $\mathcal{Q}$ is a subcover for $Z$ in $\mathcal{T}$.*

*Proof.* Suppose that $\mathcal{Q}'$ is a generous cover for $\mathcal{T}|(X - Y) \cup y$ and that $\mathcal{Q}''$ is a generous cover for $\mathcal{T}|(X - Z) \cup z$, where $y \in Y, z \in Z$. Then the set of interior paths of $\mathcal{T}$ that are distinguished by $\mathcal{Q}'$ is contained in the set distinguished by $\mathcal{Q}''$. Since $\mathcal{Q} \cup \mathcal{Q}'$ distinguishes every interior path of $\mathcal{T}$, the same can be said for $\mathcal{Q} \cup \mathcal{Q}''$, and hence $\mathcal{Q}$ is a subcover for $Z$ in $\mathcal{T}$. □

Recall the closure rules discussed in Section 3.2. Using these, we can obtain some elementary results about the closure of a subcover.

**Lemma 3.3.5.** *Let $\mathcal{T} \in \mathscr{T}_X$ be a binary tree, and let $Y \subset X$ be a cherry on $\mathcal{T}$. If $\mathcal{Q}$ is a subcover for $Y$ in $\mathcal{T}$, then $\mathcal{Q} \vdash Y|X - Y$.*

*Proof.* Let $Z \subseteq X - Y$ be a cherry of $\mathcal{T}$. Then $Y|Z$ is the only quartet that distinguishes the path between the two cherries $Y, Z$, and so $Y|Z \in \mathcal{Q}$. Now, suppose that $Z \subseteq X - Y$ is a minimal cluster of $\mathcal{T}$ such that $\mathcal{Q}$ does not infer the partial split $Y|Z$. Then there is a bipartition $Z_1, Z_2$ of $Z$ such that both $Z_1$ and $Z_2$ are clusters of $\mathcal{T}$.

By the induction assumption, we know that $\mathcal{Q} \vdash Y|Z_i$ for $i \in \{1, 2\}$. From the fact that $\mathcal{Q}$ is a subcover for $Y$ in $\mathcal{T}$, there is some quartet $Y|z_1z_2 \in \mathcal{Q}$, with $z_i \in Z_i$. It now follows from (3.2) that $\mathcal{Q} \vdash Y|Z$, in contradiction to our assumption. □

The purpose of the preceeding lemma is so that we may proof the analagous result for clusters in general.

**Lemma 3.3.6.** *Let $\mathcal{T} \in \mathscr{T}_X$ be a binary tree, and let $Y \subset X$ be a cluster of $\mathcal{T}$. If $\mathcal{Q}$ is a subcover for $Y$ in $\mathcal{T}$, then $\mathcal{Q} \vdash Y|X - Y$.*

*Proof.* By Lemma 3.3.5, this is true when $|Y| = 2$. Let $Y$ be a minimal cluster for which the lemma fails, and let $Z \subseteq X - Y$ be a cherry of $\mathcal{T}$. There is also a bipartition $Y_1, Y_2$ of $Y$ such that both $Y_1$ and $Y_2$ are clusters of $\mathcal{T}$. Note that $\mathcal{Q}$ is a subcover for each of $Y_1, Y_2$ by Lemma 3.3.4, and so

from our induction assumption we have $\mathcal{Q} \vdash Y_i|Z$ for $i \in \{1, 2\}$. Since $\mathcal{Q}$ is a subcover for $Y$ in $\mathcal{T}$, there is some quartet $y_1y_2|Z \in \mathcal{Q}$ where $y_i \in Y_i$. By (3.2), we have $\mathcal{Q} \vdash Y|Z$.

Now let $Z \subseteq X-Y$ be a minimal cluster of $\mathcal{T}$ such that $Y|Z$ is not inferred by $\mathcal{Q}$, and consider the bipartition $Z_1, Z_2$ of $Z$ where $Z_1, Z_2$ are clusters of $\mathcal{T}$. Again, we have some quartet $y_1y_2|z_1z_2 \in \mathcal{Q}$, where $y_i \in Y_i, z_i \in Z_i$. Our induction assumption guarantees that $\mathcal{Q} \vdash \{Y_i|Z, Y|Z_i\}$ for $i \in \{1, 2\}$, and so the result follows from Lemma 3.2.2. □

*Proof of Theorem 3.3.2.* If $\mathcal{Q}$ is a generous cover for $\mathcal{T}$ and $Y$ is a cluster of $\mathcal{T}$, then $\mathcal{Q}$ is a subcover for $Y$ in $\mathcal{T}$. By Lemma 3.3.6 then, every split $Y|X-Y$ of $\mathcal{T}$ is inferred by $\mathcal{Q}$. Thus $\mathcal{Q}$ defines $\mathcal{T}$ by the Splits-Equivalence Theorem (Theorem 3.1.1). □

We demonstrated earlier that a generous cover has an element of redundancy to it. To this end, we wish to briefly explore a weakening of the notion of a generous cover. Instead of distinguishing every interior path within a tree, we are interested in quartet sets that distinguish all interior paths that do not exceed some given length.

**Definition 3.3.7.** Let $\mathcal{T}$ be a binary tree, and $k \geq 1$ be some positive integer. Then a set of quartets $\mathcal{Q}$ is a *$k$-cover* for $\mathcal{T}$ if and only if

(i) $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$; and

(ii) every interior path of $\mathcal{T}$ that is of length $l \leq k$ is distinguished by some quartet $q \in \mathcal{Q}$.

We can rephrase some ideas from earlier in terms of 1-covers. For example, if $\mathcal{Q}$ defines a binary tree $\mathcal{T}$, then $\mathcal{Q}$ contains a 1-cover for $\mathcal{T}$. Further to this, Theorem 3.1.2 follows from the assertion that, for a given binary tree $\mathcal{T}$, there is a definitive set of quartets $\mathcal{Q}$ for $\mathcal{T}$ that is precisely a 1-cover for $\mathcal{T}$. However, the next lemma may be used to show that there is no fixed positive integer $k$ such that every $k$-cover for an arbitrary tree is definitive.

**Lemma 3.3.8.** *For some positive integer $k \geq 1$, let $\mathcal{C} \in \mathscr{C}_n$ be a caterpillar with $n = 2k+4$ leaves. Then there exists a $k$-cover for $\mathcal{C}$ that does not define $\mathcal{C}$.*

*Proof.* Let $\mathcal{C}_1 \in \mathscr{C}_n$ be a caterpillar with $n = 2k+4$ leaves for some $k \geq 1$. We may assume that $\mathcal{C}_1$ has the canonical caterpillar labelling shown in Fig. 3.5. To prove the lemma, we will explicitly construct a $k$-cover $\mathcal{Q}$ for $\mathcal{C}_1$ such that

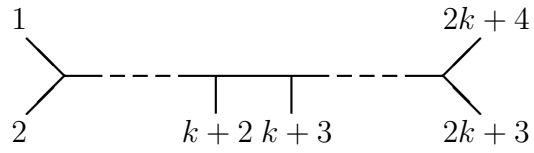


Figure 3.5: The caterpillar $\mathcal{C}_1 \in \mathscr{C}_n$ in the proof of Lemma 3.3.8.

each $q \in \mathcal{Q}$ is also displayed by another tree $\mathcal{T} \in \mathscr{T}_n$.

Let $\mathcal{C}_2 \in \mathscr{C}_n$ be the caterpillar shown in Fig. 3.6, and let $a, b$ be a pair of

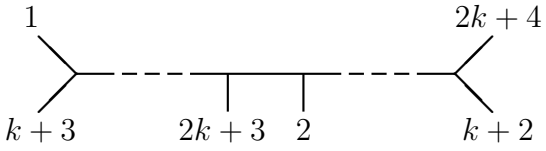


Figure 3.6: The caterpillar $\mathcal{C}_2 \in \mathscr{C}_n$ in the proof of Lemma 3.3.8.

leaves of $\mathcal{C}_1$ such that $2 \leq a < b \leq 2k+3$ and $b - a \leq k$. If either $b \leq k+2$ or $a \geq k+3$, then we let $q = 1a|bn$. Otherwise we have $a > 2$ and $b < n-1$, and we let $q = 2a|b(n-1)$. In either case, the quartet $q$ distinguishes the path from the vertex adjacent to $a$ to the vertex adjacent to $b$. Moreover, $q$ is displayed by $\mathcal{C}_2$, completing the proof. □

To illustrate Lemma 3.3.8, suppose that $k = 2$. Then $n = 8$, and

$$\mathcal{Q} = \{12|38, 13|48, 24|57, 15|68, 16|78, 12|48, 23|57, 24|67, 15|78\}$$

is a 2-cover for the caterpillar $\mathcal{C}_1 \in \mathscr{C}_8$ shown in Fig. 3.7. However, $\mathcal{Q}$ is also displayed by the caterpillar $\mathcal{C}_2 \in \mathscr{C}_8$, and so does not define $\mathcal{C}_1$.

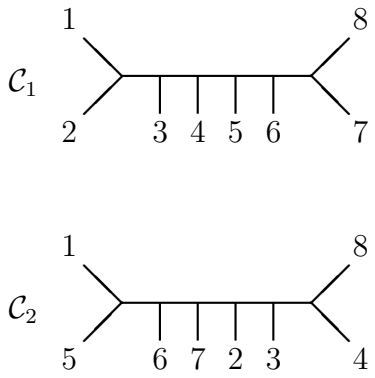


Figure 3.7: Caterpillars $\mathcal{C}_1, \mathcal{C}_2 \in \mathscr{C}_8$ that illustrate Lemma 3.3.8.

Lemma 3.3.8 can be extended quite naturally to trees in general. As a corollary to either this lemma or Theorem 3.3.9, we have our earlier statement that there is a non-definitive $k$-cover for all positive integers $k$.

**Theorem 3.3.9.** *For some positive integer $k \geq 1$, let $\mathcal{T}$ be a binary leaf-labelled tree with some interior path of length $n = 2k + 1$. Then there exists a $k$-cover for $\mathcal{T}$ that does not define $\mathcal{T}$.*

**Corollary 3.3.10.** *There exist non-definitive $k$-covers for all positive integers $k \geq 1$.*

We omit the proofs of both the theorem and the corollary above. The proof of the former follows much the same reasoning as the proof of Lemma 3.3.8, while the proof of Corollary 3.3.10 is a trivial consequence of either Lemma 3.3.8 or Theorem 3.3.9.

## 3.4 The Size of a Minimal Definitive Quartet Set

Up until now, we have made only passing mention of minimal definitive quartet sets. These sets may be thought of as the most basic sets of information that can be used to build up a unique tree. It is still unknown how large a such a set may be as a function of $n$, where $n$ is the size of the leaf-set

under scrutiny. What is known is, that if $\mathcal{Q}$ is a minimal definitive set of quartets, and $|\mathcal{L}(\mathcal{Q})| = n$, then $|\mathcal{Q}| \geq n - 3$.

Let $M(n)$ be the greatest positive integer such that there exists a minimal definitive set of quartets $\mathcal{Q}$ of size $M(n)$, where $n = |\mathcal{L}(\mathcal{Q})|$. It is trivial to show that $M(n)$ is well-defined, since the size of $\mathcal{Q}(\mathcal{T})$ is finite for any finite leaf-labelled tree. Ultimately, we would like to find a tight upper bound on the function $M(n)$, although this has as yet proved beyond reach. We mentioned earlier an anonymous conjecture that $M(n) = n+c$ for some fixed constant $c$. In the remainder of this chapter, we prove Theorem 3.4.1 stated below, immediately invalidating the conjecture.

**Theorem 3.4.1.** *Let $n \geq 4$ be some positive integer. Then*

$$M(n) \geq \frac{3}{2}(n-4).$$

To prove this result, we will use the following two theorems about amalgamating compatible collections of trees that have at least one leaf in common.

**Theorem 3.4.2** (Theorem 6.8.8, [42])**.** *Let $\mathcal{P}$ be a collection of leaf-labelled trees and suppose that $\bigcap_{\mathcal{T}\in\mathcal{P}} \mathcal{L}(\mathcal{T}) \neq \emptyset$. Then $\mathcal{P}$ defines a binary leaf-labelled tree $\mathcal{T}$ if and only if $\mathcal{T}$ displays $\mathcal{P}$ and each interior edge of $\mathcal{T}$ is distinguished by an interior edge of at least one tree in $\mathcal{P}$.*

**Theorem 3.4.3** (Corollary 6.8.9, [42])**.** *Let $\mathcal{T}_1, \mathcal{T}_2$ be binary leaf-labelled trees, and let $Y = \mathcal{L}(\mathcal{T}_1) \cap \mathcal{L}(\mathcal{T}_2)$. Then $\mathcal{T}_1$ and $\mathcal{T}_2$ are compatible if and only if $\mathcal{T}_1|Y = \mathcal{T}_2|Y$.*

As a special case of Theorem 3.4.2, we get the following closure rule. We have already used this rule implicitly in Section 3.1, when stating that the tree $\mathcal{T}$ shown in Fig. 3.1 is defined by $\mathcal{Q} = \{12|34, 13|45\}$.

$$\{ab|cd, ac|de\} \vdash ab|ce \tag{3.3}$$

This is sometimes referred to as the *semi-dyadic closure rule* (see, for example, [42]).

The basic example which we will take as our starting point for proving Theorem 3.4.1 is due to Bordewich and Semple [14], and consists of a set of

six quartets that defines a seven-leafed tree.

**Lemma 3.4.4.** *The set of quartets*

$$\mathcal{Q} = \{12|35, 13|46, 24|57, 35|67, 12|46, 34|67\}$$

*is a minimal definitive quartet set.*

*Proof.* Let us begin by showing a tree that displays $\mathcal{Q}$, namely the caterpillar

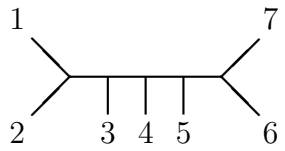


Figure 3.8: A tree $\mathcal{T} \in \mathscr{T}_7$ that displays the quartet set $\mathcal{Q}$ from Lemma 3.4.4.

$\mathcal{T} \in \mathscr{T}_7$ shown in Fig. 3.8. It suffices to show that $\mathcal{Q}$ defines $\mathcal{T}$, and that any strict subset of $\mathcal{Q}$ is displayed by some other tree $\mathcal{T}' \in \mathscr{T}_7$ distinct from $\mathcal{T}$.

Using (3.1), we have $\{12|46, 13|46\} \vdash 23|46$ and $\{34|67, 35|67\} \vdash 45|67$, so both $23|46$ and $45|67$ are in the closure of $\mathcal{Q}$. Further, by (3.3) we can make the inference $\{24|57, 45|67\} \vdash 25|67$. Thus we have

$$\{12|35, 23|46, 24|57, 25|67\} \subseteq \mathrm{cl}(\mathcal{Q}).$$

The quartets in this subset all contain a common leaf, and further they collectively distinguish each interior edge of $\mathcal{T}$. Hence $\mathcal{Q}$ defines $\mathcal{T}$ by Theorem 3.4.2.

Suppose now that there is some $q \in \mathcal{Q}$ such that $\mathrm{cl}(\mathcal{Q} - q) = \mathcal{Q}(\mathcal{T})$. Then in fact $q \in \{12|46, 34|67\}$, since the other four quartets in $\mathcal{Q}$ distinguish distinct interior edges of $\mathcal{T}$. However, the trees $\mathcal{T}_1, \mathcal{T}_2$ in Fig. 3.9 respectively display $\mathcal{Q} - 12|46$ and $\mathcal{Q} - 34|67$. It follows that $\mathcal{Q}$ is indeed a minimal definitive set, completing the lemma. □

The proof of Theorem 3.4.1 now follows relatively easily from Lemma 3.4.4 and Theorem 3.4.2.

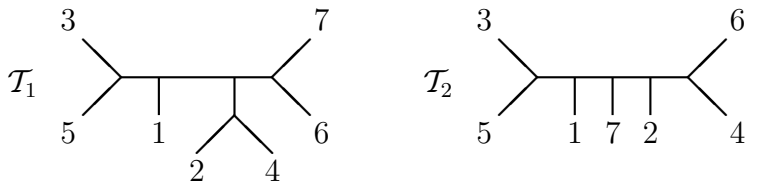


Figure 3.9: Trees $\mathcal{T}_1, \mathcal{T}_2 \in \mathscr{T}_7$ which respectively display $\mathcal{Q} - 12|46$ and $\mathcal{Q} - 34|67$, for $\mathcal{Q}$ as given in Lemma 3.4.4.

*Proof of Theorem 3.4.1.* The result certainly holds for $n \leq 6$ from the observation that $M(n) \geq n - 3$, and for $n = 7$ from Lemma 3.4.4. Suppose instead that $n \geq 8$, and that $\mathcal{T} \in \mathscr{T}_n$ is the caterpillar shown in Fig. 3.10. Let

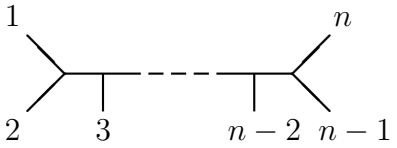


Figure 3.10: The tree $\mathcal{T} \in \mathscr{T}_n$ in the proof of Theorem 3.4.1.

$X_1 = \{1, \ldots, n-4\}$ and $X_2 = \{n-6, \ldots, n\}$. By the induction hypothesis, there is some set $\mathcal{Q}_1$ containing at least $\frac{3}{2}(n-8)$ quartets that minimally defines $\mathcal{T}|X_1$, and by Lemma 3.4.4 there is some set $\mathcal{Q}_2$ containing exactly six quartets that minimally defines $\mathcal{T}|X_2$. Hence, using Theorem 3.4.2, the set $\mathcal{Q} = \mathcal{Q}_1 \cup \mathcal{Q}_2$ defines $\mathcal{T}$.

Suppose now that $q \in \mathcal{Q}_1$. Then there is some tree $\mathcal{T}' \in \mathscr{T}_{X_1}$ that is distinct from $\mathcal{T}|X_1$ but that displays $\mathcal{Q}_1 - q$. Since $|X_1 \cap X_2| = 3$, the trees $\mathcal{T}'$ and $\mathcal{T}|X_2$ are compatible by Theorem 3.4.3, and hence $\mathcal{Q} - q$ does not define $\mathcal{T}$. A similar argument holds for all $q \in \mathcal{Q}_2$, and so $\mathcal{Q}$ in fact minimally defines $\mathcal{T}$. Moreover, since $|X_1 \cap X_2| = 3$, the intersection $\mathcal{Q}_1 \cap \mathcal{Q}_2$ is empty. It now follows that

$$\begin{aligned} |\mathcal{Q}| &\geq \frac{3}{2}(n-8) + 6 \\ &= \frac{3}{2}(n-4), \end{aligned}$$

completing the proof. □

# Chapter 4

# The Quartet Graph

## 4.1 Introduction

We earlier proposed three fundamental questions that arise when dealing with combining the data from a collection of trees $\mathcal{F}$. These are

(i) is $\mathcal{F}$ compatible;

(ii) does $\mathcal{F}$ define some leaf-labelled tree; and

(iii) does $\mathcal{F}$ identify some leaf-labelled tree.

In this chapter, we introduce the quartet graph and show that, in addition to the chordal graph characterisations ([18, 34, 41, 45]), these problems can also be characterised in terms of edge colourings via this graph. One of the main motivations for the quartet graph is that it may provide new insights into not only the complexity of (ii), but also other quartet-based problems that may arise in studying leaf-labelled trees. Indeed, in Chapter 5, we make use of the quartet graph and its associated concepts to determine, for a given tree $\mathcal{T}$, the size of a minimum-sized set of quartets that identifies $\mathcal{T}$. The resulting theorem corrects a previously published result [42].

The remainder of the chapter is organised as follows. The next section consists of preliminaries and formal statements of the main results. For completeness, Section 4.4 contains the chordal graph characterisations of problems (i)-(iii). Section 4.5 contains the proofs of the characterisations of (i)-(iii) in terms of quartet graphs. The proof of the compatibility characterisation is algorithmic and thus provides a leaf-labelled tree that displays the original collection of quartets if this collection is compatible.

## 4.2 The Quartet Graph

For a collection $\mathcal{Q}$ of quartets with leaf set $X$, we define the *quartet graph* of $\mathcal{Q}$, denoted $G_{\mathcal{Q}}$, as follows. The vertex set of $G_{\mathcal{Q}}$ is the set of singletons of $X$ and, for each $q = ab|cd \in \mathcal{Q}$, there is an edge joining $\{a\}$ and $\{b\}$, and an edge joining $\{c\}$ and $\{d\}$ each of which is labelled $q$. Apart from these edges, $G_{\mathcal{Q}}$ has no other edges. Note that if $q_1 = ab|cd, q_2 = ab|ce \in \mathcal{Q}$, then $G_{\mathcal{Q}}$ has edges $\{a, b\}$ and $\{c, d\}$ labelled $q_1$, and separate edges $\{a, b\}$ and $\{c, e\}$ labelled $q_2$. For purposes later in the paper, in reference to $q$, we sometimes use $\{a, b\}_q$ and $\{c, d\}_q$ to denote the two parts of $q$.

As an example, consider the set of quartets $\mathcal{Q} = \{12|45, 23|56, 34|16\}$. The quartet graph of $\mathcal{Q}$ is shown in Fig. 4.1, where, instead of labelling the edges with the appropriate element of $\mathcal{Q}$, we have used solid, dashed, and dotted lines to represent the edges arising from 12|45, 23|56 and 34|16 respectively.

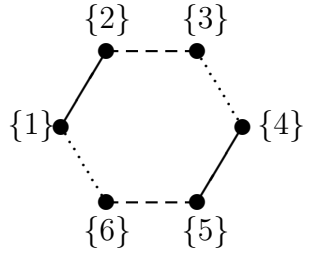


Figure 4.1: The quartet graph for $\mathcal{Q} = \{12|45, 23|56, 34|16\}$.

Each edge of $G_{\mathcal{Q}}$ has a partner, namely, the one which is labelled by the same quartet. Another way we could have indicated this is by assigning a distinct colour to each quartet in $\mathcal{Q}$, and then assigning this colour to each of the two edges corresponding to this quartet. In doing this, we observe that the resulting edge colouring of $G_{\mathcal{Q}}$ is a proper edge colouring. From this viewpoint, we say that an edge is *$q$-coloured* if it is labelled $q$. Recall that an *edge colouring* of a graph $G$ is an assignment of colours to the edges of $G$. An edge colouring is *proper* if no two edges incident with the same vertex have the same colour.

Central to this chapter is a particular graphical operation that *unifies* vertices. Let $X$ be a non-empty finite set, and let $G$ be an arbitrary graph

with no loops and whose vertex set $V$ is a partition of $X$, where no part is the empty set. In other words, $X$ is the disjoint union of the vertices of $G$. Furthermore, suppose that $G$ is properly edge-coloured. Let $U$ be a subset of $V$ with the property that if $e$ and $f$ are distinct edges of $G$ with the same colour, then at most one of these edges is incident with a vertex in $U$. The *unification* of the vertices in $U$ is the graph obtained from $G$ by

(i) replacing the vertices in $U$ together with every edge for which both end-vertices are in $U$ by a single new vertex such that if an edge is incident with exactly one vertex in $U$, then it is incident with the resulting new vertex;

(ii) labelling the new vertex as the union of the elements in $U$; and

(iii) for each edge that joins two vertices in $U$, delete all other edges with the same colour.

Observe that, at the end of (ii), the resulting graph remains properly edge-coloured.

## 4.3 Main Results

Let $\mathcal{Q}$ be a collection of quartets on $X$. The quartet graph $G_{\mathcal{Q}}$ satisfies the properties of being loopless and properly edge-coloured, and so we can apply unification operations to this graph. Let $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$ be a sequence $\mathcal{S}$ of graphs, where $G_i$ is obtained from $G_{i-1}$ by a unification for all $i \in \{1, \ldots, k\}$. We will call such a sequence a *unification sequence* of $G_{\mathcal{Q}}$. If $G_k$ has no edges, then $\mathcal{S}$ is said to be *complete*. As a matter of convenience, for all $i \in \{1, \ldots, k\}$ we denote by $\mathcal{S}_i$ the unification sequence $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_i$.

**Example 4.3.1.** Consider the quartet graph $G_{\mathcal{Q}}$ shown in Fig. 4.1. Figure 4.2 illustrates a unification sequence of $G_{\mathcal{Q}}$ beginning with $G_{\mathcal{Q}}$ on the top left and ending with the graph $G_3$ consisting of three isolated vertices on the bottom right. Initially, we unify the vertices $\{2\}$ and $\{3\}$ to get $G_1$. The third graph, $G_2$, is obtained by unifying $\{1\}$ and $\{6\}$ in $G_1$, while $G_3$ is

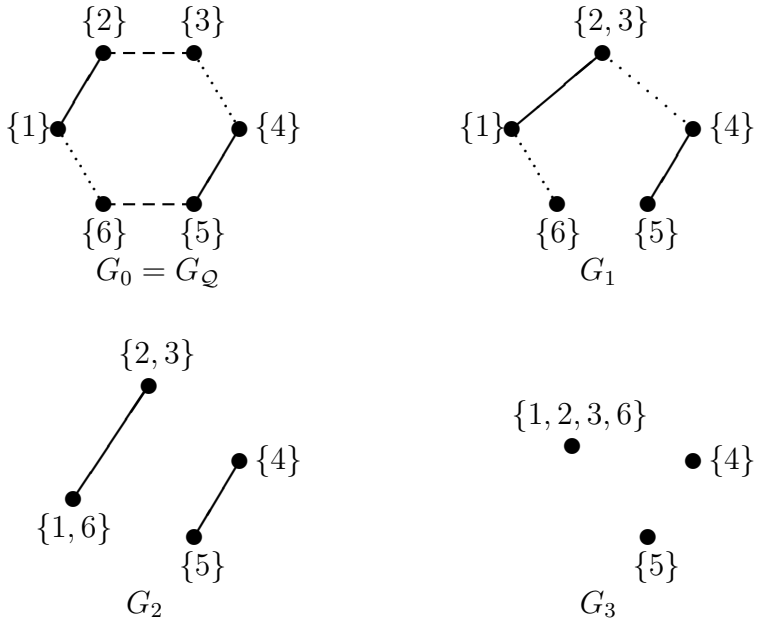


Figure 4.2: A complete unification sequence of the quartet graph in Fig. 4.1.

obtained from $G_2$ by unifying $\{1, 6\}$ and $\{2, 3\}$. Since this last graph has no edges, this unification sequence is complete.

The following theorem characterises the compatibility of a collection of quartets in terms of quartet graphs.

**Theorem 4.3.2.** *Let $\mathcal{Q}$ be a set of quartets. Then $\mathcal{Q}$ is compatible if and only if there is a complete unification sequence of $G_{\mathcal{Q}}$.*

As an illustration of Theorem 4.3.2, the set $\mathcal{Q} = \{12|45, 23|56, 34|16\}$ is compatible since there is a complete unification sequence of $G_{\mathcal{Q}}$ (see Fig. 4.2). Indeed, the tree $\mathcal{T}$ shown in Fig. 4.3 displays $\mathcal{Q}$.

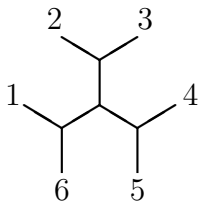


Figure 4.3: A tree $\mathcal{T}$ that displays $\mathcal{Q} = \{12|45, 23|56, 34|16\}$.

We have not yet provided a formal definition of an identifying quartet set. A set of quartets $\mathcal{Q}$ with $X = \mathcal{L}(\mathcal{Q})$ *identifies* a tree $\mathcal{T} \in \mathscr{T}_X$ if and only if

(i) $\mathcal{T}$ displays $\mathcal{Q}$; and

(ii) if $\mathcal{T}' \in \mathscr{T}_X$ displays $\mathcal{Q}$, then $\Sigma(\mathcal{T}) \subseteq \Sigma(\mathcal{T}')$.

The second condition here is equivalent to requiring that all trees $\mathcal{T}' \in \mathscr{T}_X$ displaying $\mathcal{Q}$ are refinements of $\mathcal{T}$.

To describe our characterisations of when a set of quartets identifies and defines a leaf-labelled tree, we require some further definitions. The first of these generalises the notion of distinguishing an edge of a binary tree to an analogous concept for arbitrary leaf-labelled trees.

Let $\mathcal{T} \in \mathscr{T}_X$ be a tree that displays a collection $\mathcal{Q}$ of quartets on $X$, and let $e = uv$ be an interior edge of $\mathcal{T}$. We define $G_{\mathcal{Q}(u,v)}$ to be the graph that has the neighbours of $v$ except $u$ as its vertex set, and where two vertices $w_i, w_j$ are joined by an edge precisely if there is a quartet in $\mathcal{Q}$ that distinguishes $e$ and is of the form $w_i w_j | xy$ for some $x, y \in X$. A set $\mathcal{Q}$ of quartets on $X$ *specially distinguishes* a tree $\mathcal{T} \in \mathscr{T}_X$ if $\mathcal{T}$ displays $\mathcal{Q}$ and, for every interior edge $e = uv$ of $\mathcal{T}$, both $G_{\mathcal{Q}(u,v)}$ and $G_{\mathcal{Q}(v,u)}$ are connected.

Let $\mathcal{Q}$ be a collection of quartets on $X$, and let $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$ be a unification sequence $\mathcal{S}$ of $G_{\mathcal{Q}}$. For all $i$, let $U_i$ denote the subset of vertices of $G_{i-1}$ that are unified to obtain $G_i$ and let $A_i$ denote the union of the elements of $U_i$. We will call $U_1, \ldots, U_k$ the sequence of *unifying sets associated with* $\mathcal{S}$. Observe that, for all $i$ and $j$ with $i < j$, either $A_i \subseteq A_j$ or $A_i \cap A_j = \emptyset$. This observation will be used throughout the paper. Furthermore, we call

$$\Sigma_{\mathcal{S}} = \{A_i | (X - A_i) : i \in \{1, \ldots, k\}\}$$

the set of *splits induced by* $\mathcal{S}$.

Now let $q = ab|cd$ be an element of $\mathcal{Q}$. If, for some $j$, either $\{a, b\}$ or $\{c, d\}$ is a subset of $A_j$, but neither $\{a, b\} \subseteq A_i$ nor $\{c, d\} \subseteq A_i$ for all $i < j$, then we say that $q$ has been *collected* by $U_j$ or, more generally, by $\mathcal{S}$. Moreover, if $\{a, b\} \subseteq A_j$ and, for all $i < j$, neither $\{a, b\} \subseteq A_i$ nor $\{c, d\} \subseteq A_i$, we say that $A_j$ or, again more generally, $\mathcal{S}$ *merged* $\{a, b\}_q$. For a subset $\mathcal{Q}'$ of $\mathcal{Q}$, we

denote the set

$$\{\{a,b\}_q : q = ab|cd \in \mathcal{Q}' \text{ and } \mathcal{S} \text{ merged } \{a,b\}_q\}$$

by $M(\mathcal{Q}')_{\mathcal{S}}$.

Lastly, if $\mathcal{S}$ is complete, then $\mathcal{S}$ is said to be *minimal* if there is no other complete unification sequence $\mathcal{S}'$ with $U_1', \ldots, U_l'$ as its sequence of unifying sets such that $\{A_j' : j \in \{1, \ldots, l\}\}$ is a proper subset of $\{A_i : i \in \{1, \ldots, k\}\}$, where $A_j'$ is the union of the elements in $U_j'$ for all $j$.

**Theorem 4.3.3.** *Let $\mathcal{Q}$ be a set of quartets on $X$. Then $\mathcal{Q}$ identifies a leaf-labelled tree if and only if both of the following conditions hold:*

*(i) There exists a leaf-labelled tree $\mathcal{T} \in \mathscr{T}_X$ that displays $\mathcal{Q}$ and is specially distinguished by $\mathcal{Q}$.*

*(ii) Let $\mathcal{Q}'$ be a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and let $q = A|B \in \mathcal{Q}'$. Let $\mathcal{S}$ and $\mathcal{S}'$ be minimal complete unification sequences of $G_{\mathcal{Q}}$ such that, amongst the quartets in $\mathcal{Q}'$, the quartet $q$ is collected (joint) last and $A$ is merged. Then $M(\mathcal{Q}')_{\mathcal{S}} = M(\mathcal{Q}')_{\mathcal{S}'}$.*

Provided (i) holds in Theorem 4.3.3, we remark here that there is always at least one minimal complete unification sequence that satisfies the assumption conditions in (ii) (see Lemma 4.5.5).

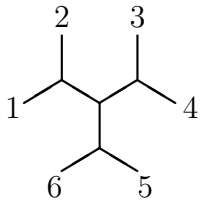


Figure 4.4: A second tree $\mathcal{T}'$ that also displays $\mathcal{Q} = \{12|45, 23|56, 34|16\}$.

**Example 4.3.4.** To illustrate Theorem 4.3.3, again consider the set of quartets $\mathcal{Q} = \{12|45, 23|56, 34|16\}$. As well as the tree $\mathcal{T}$ shown in Fig. 4.3, the tree $\mathcal{T}'$ shown in Fig. 4.4 also displays $\mathcal{Q}$. Since $\mathcal{Q}$ specially distinguishes $\mathcal{T}$,

and the second tree $\mathcal{T}'$ is not a refinement of $\mathcal{T}$, the set $\mathcal{Q}$ does not identify any leaf-labelled tree. This fact is realised by Theorem 4.3.3 as follows.

In addition to the complete unification sequence $\mathcal{S}_1$ shown in Fig. 4.2, Fig. 4.5 shows a second complete unification sequence $\mathcal{S}_2$ of $G_{\mathcal{Q}}$. Now, $\mathcal{Q}$ specially distinguishes $\mathcal{T}$. In both $\mathcal{S}_1$ and $\mathcal{S}_2$, the quartet $12|45$ is the last quartet of $\mathcal{Q}$ that is collected and $\{1, 2\}$ is merged. Consider the quartet $23|56 \in \mathcal{Q}$. In $\mathcal{S}_1$, we have that $\{2, 3\}$ is merged, while, in $\mathcal{S}_2$, we have that $\{5, 6\}$ is merged. Thus $M(\mathcal{Q})_{\mathcal{S}_1} \neq M(\mathcal{Q})_{\mathcal{S}_2}$. It now follows by Theorem 4.3.3 that $\mathcal{Q}$ does not identify a leaf-labelled tree.

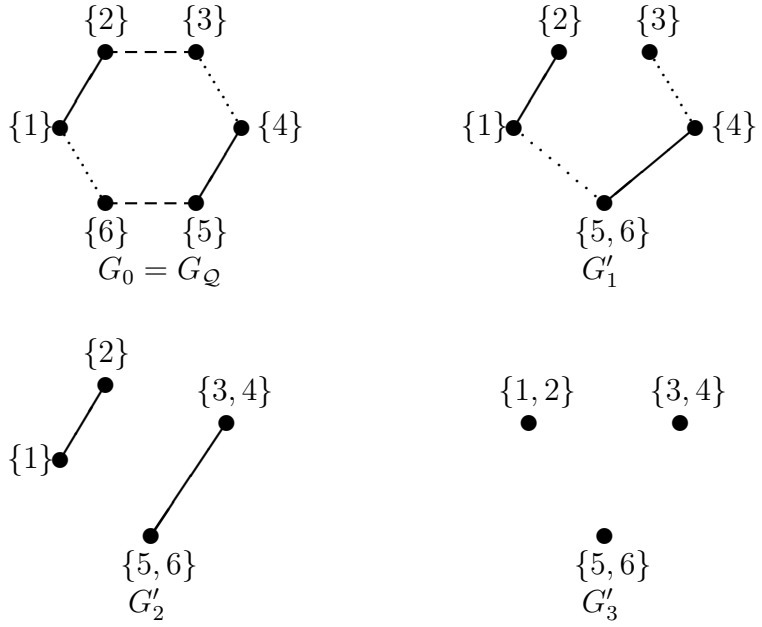


Figure 4.5: Another complete unification sequence of the quartet graph in Fig. 4.1.

We remark here that the quartet set $\mathcal{Q}$ used in Example 4.3.4 shows that condition (i) by itself in Theorem 4.3.3 is not sufficient for a collection of quartets to identify a leaf-labelled tree, as $\mathcal{Q}$ specially distinguishes the tree shown in Fig. 4.3.

As a consequence of Theorem 4.3.3, we have the following corollary.

**Corollary 4.3.5.** *Let $\mathcal{Q}$ be a set of quartets on $X$. Then $\mathcal{Q}$ defines a leaf-labelled tree if and only if both of the following conditions hold:*

(i) *There exists a binary tree* $\mathcal{T} \in \mathscr{T}_X$ *that displays* $\mathcal{Q}$ *and is distinguished by* $\mathcal{Q}$.

(ii) *Let* $\mathcal{Q}'$ *be a minimum-sized subset of* $\mathcal{Q}$ *that distinguishes* $\mathcal{T}$ *and let* $q \in \mathcal{Q}'$. *Let* $\mathcal{S}$ *and* $\mathcal{S}'$ *be minimal complete unification sequences of* $G_{\mathcal{Q}}$ *such that, amongst the quartets in* $\mathcal{Q}'$, *the quartet* $q$ *is collected last. Then* $M(\mathcal{Q}' - q)_{\mathcal{S}} = M(\mathcal{Q}' - q)_{\mathcal{S}'}$.

A *one-split* leaf-labelled tree is a leaf-labelled tree with exactly one interior edge. For example, a quartet is a one-split tree with four leaves. If the single non-trivial split of this tree is $\{a_1, \ldots, a_r\}|\{b_1, \ldots, b_s\}$, then we will denote this tree by $a_1 \cdots a_r|b_1 \cdots b_s$ or $A|B$, where $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$.

## 4.4 Chordal Graph Characterisations

In this section, we state the chordal graph analogues of Theorems 4.3.2 and 4.3.3, and Corollary 4.3.5. This section is independent of the rest of the chapter and so the reader may wish to initially skip it.

The *partition intersection graph* of a collection $\mathcal{Q}$ of quartets, denoted $\text{int}(\mathcal{Q})$, is the vertex-coloured graph that has vertex set

$$\bigcup_{q=A|B \in \mathcal{Q}} \big\{(q, A), (q, B)\big\},$$

and an edge joining $(q', B')$ and $(q'', B'')$ precisely if $B' \cap B''$ is non-empty. Here two vertices are the same colour if they share the same first coordinate.

A graph is *chordal* if none of its vertex-induced subgraphs is isomorphic to a cycle with at least four vertices. A graph $G$ is a *restricted chordal completion* of $\text{int}(\mathcal{Q})$ if $G$ is a chordal graph that can be obtained from $\text{int}(\mathcal{Q})$ by only adding edges between vertices whose first coordinates are distinct. Note that this maintains the property of a *proper vertex colouring*. Theorem 4.4.1, the chordal graph analogue of Theorem 4.3.2, was indicated by Buneman [18] and Meacham [34], and formally proved by Steel [45].

**Theorem 4.4.1.** *Let* $\mathcal{Q}$ *be a set of quartets. Then* $\mathcal{Q}$ *is compatible if and only if there is a restricted chordal completion of* $\text{int}(\mathcal{Q})$.

A restricted chordal completion $G$ of $\text{int}(\mathcal{Q})$ is *minimal* if, for every non-empty subset $F$ of edges of $E(G) - E(\text{int}(\mathcal{Q}))$, the graph $G \backslash F$ is not chordal. The next theorem is due to Semple and Steel [41].

**Theorem 4.4.2.** *Let $\mathcal{Q}$ be a set of quartets on $X$. Then there is a unique leaf-labelled tree $\mathcal{T} \in \mathscr{T}_X$ that displays $\mathcal{Q}$ if and only if the following two conditions hold:*

*(i) there is a binary leaf-labelled tree that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$; and*

*(ii) there is a unique minimal restricted chordal completion of* $\text{int}(\mathcal{Q})$.

To describe the chordal graph analogue of Theorem 4.3.3 requires some further definitions. Let $\mathcal{T} \in \mathscr{T}_X$ be a tree and let $e = u_1u_2$ be an edge of $\mathcal{T}$. Then $e$ is *strongly distinguished* by a one-split tree $A_1|A_2$ if, for each $i \in \{1, 2\}$, the following hold:

(i) $A_i$ is a subset of the vertex set of the component of $\mathcal{T} \backslash e$ containing $u_i$; and

(ii) the vertex set of each component of $T \backslash u_i$, except for the one containing the other end vertex of $e$, contains an element of $A_i$.

For a collection $\mathcal{Q}$ of quartets on $X$, let $\mathcal{G}(\mathcal{Q})$ denote the collection of graphs

$$\{G : \text{there is a leaf-labelled tree } \mathcal{T} \text{ displaying } \mathcal{Q} \text{ with } G = \text{int}(\mathcal{Q}, \mathcal{T})\},$$

where $\text{int}(\mathcal{Q}, \mathcal{T})$ is the graph that has the same vertex set as $\text{int}(\mathcal{Q})$, and an edge joining two vertices $(q, A)$ and $(q', A')$ if the vertex sets of the minimal subtrees of $\mathcal{T}$ connecting the elements in $A$ and $A'$ have a non-empty intersection. Note that if $G$ is a graph in $\mathcal{G}(\mathcal{Q})$, then $G$ is a restricted chordal completion of $\text{int}(\mathcal{Q})$. There is a partial order $\leq$ on $\mathcal{G}(\mathcal{Q})$ which is obtained by setting $G_1 \leq G_2$ for all $G_1, G_2 \in \mathcal{G}(\mathcal{Q})$ if the edge set of $G_1$ is a subset of the edge set of $G_2$. Lastly, a compatible collection $\mathcal{Q}$ of quartets infers a one-split tree if every leaf-labelled tree that displays $\mathcal{Q}$ also displays this one-split tree. Theorem 4.4.3 was established by Bordewich *et al.* [13].

**Theorem 4.4.3.** *Let $\mathcal{Q}$ be a set of quartets on $X$. Then $\mathcal{Q}$ identifies a leaf-labelled tree if and only if the following conditions hold:*

*(i) there is a leaf-labelled tree that displays $\mathcal{Q}$ and, for every edge $e$ of this tree, there is a one-split leaf-labelled tree inferred by $\mathcal{Q}$ that strongly distinguishes $e$; and*

*(ii) there is a unique maximal element in $\mathcal{G}(\mathcal{Q})$.*

Note that if $\mathcal{Q}$ is a collection of quartets, then $\text{int}(\mathcal{Q})$ is the line graph of the quartet graph $G_{\mathcal{Q}}$ where, for a graph $G$, the *line graph* of $G$ has vertex set $E(G)$ and two vertices joined by an edge precisely if they are incident with a common vertex in $G$. The vertex colouring of the partition intersection graph corresponds to the edge colouring of the quartet graph. However, the characterisations of defining and identifying quartet sets described in this section and those derived in this chapter are quite different and we do not use the duality between the partition intersection graph and the quartet graph to prove the new results.

We also point out that the results stated in this section were originally proved for general *characters* (that is, partitions of $X$) rather than for quartets. The concept of the quartet graph can be extended to this more general setup but then hypergraphs must be considered. On the other hand, the information contained in characters can be expressed in terms of quartets thus no generality is lost in restricting our attention to quartets here (see [42, Proposition 6.3.11]).

## 4.5 Proofs of Theorems 4.3.2 and 4.3.3, and Corollary 4.3.5

The proof of Theorem 4.3.2 is an immediate consequence of the next two lemmas.

**Lemma 4.5.1.** *Let $\mathcal{Q}$ be a set of quartets on $X$, and let $\mathcal{S}$ be a unification sequence of $G_{\mathcal{Q}}$. Then the set $\Sigma_{\mathcal{S}}$ of splits induced by $\mathcal{S}$ is compatible. Moreover, if $\mathcal{Q}'$ denotes the subset of $\mathcal{Q}$ collected by $\mathcal{S}$, then the leaf-labelled tree*

*whose set of non-trivial splits is $\Sigma_{\mathcal{S}}$ displays each of the quartets in $\mathcal{Q}'$, but no quartet in $\mathcal{Q} - \mathcal{Q}'$.*

*Proof.* Suppose that $\mathcal{S}$ is the sequence $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$ with unifying sequence $U_1, \ldots, U_k$. For all $i$, let $A_i$ denote the union of the elements of $U_i$. The proof of the proposition is by induction on $k$. If $k = 0$, the result holds trivially. Now suppose that the result holds for all unification sequences of $G_{\mathcal{Q}}$ of smaller length, in particular, the result holds for the unification sequence $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_{k-1}$. Denote this last sequence by $\mathcal{S}'$.

Consider the split $A_k|(X - A_k)$, and note that, by the induction assumption, $\Sigma_{\mathcal{S}'}$ is compatible. Let $A_i|(X-A_i) \in \Sigma_{\mathcal{S}'}$. Since $A_i$ is a subset of a vertex of $G_{k-1}$, either $A_i \subseteq A_k$, in which case $A_i \cap (X - A_k) = \emptyset$, or $A_i \cap A_k = \emptyset$. In either case, by the Splits-Equivalence Theorem (Theorem 3.1.1), $A_i|(X - A_i)$ and $A_k|(X - A_k)$ are compatible. It follows by the induction assumption and the Splits-Equivalence Theorem (Theorem 3.1.1) that $\Sigma_{\mathcal{S}}$ is compatible.

Let $\mathcal{T}$ denote the leaf-labelled tree whose set of non-trivial splits is $\Sigma_{\mathcal{S}}$, and let $\mathcal{T}'$ denote the leaf-labelled tree whose set of non-trivial splits is $\Sigma_{\mathcal{S}'}$. By the induction assumption, $\mathcal{T}'$ displays each of the quartets collected by $\mathcal{S}'$, but no other quartet in $\mathcal{Q}$. Assume that $ab|cd$ is a quartet collected by $U_k$. Then either $a, b \in A_k$ and $c, d \in X - A_k$, or $c, d \in A_k$ and $a, b \in X - A_k$, and so $\mathcal{T}$ displays $ab|cd$. Since $\mathcal{T}$ is a refinement of $\mathcal{T}'$, it follows that $\mathcal{T}$ displays each of the quartets collected by $\mathcal{S}$. Moreover, if $wx|yz$ is a quartet of $\mathcal{Q}$ not collected by $\mathcal{S}$, then, for all $i \in \{1, \ldots, k\}$,

$$\{w, x, y, z\} \cap A_i \notin \{\{w, x\}, \{y, z\}\},$$

and so $wx|yz$ is not displayed by $\mathcal{T}$. □

Given Lemma 4.5.1, we call the tree $\mathcal{T}$ whose set of non-trivial splits is equal to the set of splits induced by a unification sequence $\mathcal{S}$ the *leaf-labelled tree induced by* $\mathcal{S}$.

Lemma 4.5.1 provides one direction of the proof of Theorem 4.3.2. The next lemma gives the other direction.

Let $\mathcal{Q}$ be a set of quartets on $X$ and let $\mathcal{T} \in \mathscr{T}_X$ be a tree that displays $\mathcal{Q}$. Let $v$ be an interior vertex of $\mathcal{T}$. Order the elements $A_1|(X-A_1), \ldots, A_k|(X-$

$A_k$) of $\Sigma(\mathcal{T})$ as follows:

(i) If $e_i$ is the edge of $\mathcal{T}$ that induces $A_i|(X - A_i)$, then $A_i$ is the subset of the vertex set of the component that does not contain $v$ in $\mathcal{T}\backslash e_i$.

(ii) If $i < j$, then either $A_i \subseteq A_j$ or $A_i \cap A_j = \emptyset$.

It is easily checked that such an ordering is possible. Now let $\mathcal{S}_v$ denote the sequence of graphs $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$, where, for all $i$, the graph $G_i$ is obtained from $G_{i-1}$ by unifying the vertices whose disjoint union is $A_i$. It is easily seen that $\mathcal{S}_v$ is well-defined. The next lemma shows that $\mathcal{S}_v$ is a complete unification sequence of $G_{\mathcal{Q}}$.

**Lemma 4.5.2.** *Let $\mathcal{Q}$ be a set of quartets on $X$ and let $\mathcal{T} \in \mathscr{T}_X$ be a tree that displays $\mathcal{Q}$. Let $v$ be an interior vertex of $\mathcal{T}$. Then $\mathcal{S}_v$ (as described above) is a complete unification sequence of $G_{\mathcal{Q}}$.*

*Proof.* Suppose that $\mathcal{S}_v$ is not such a sequence and let $j$ denote the smallest index for which $G_j$ is not a unification of $G_{j-1}$. Since $G_j$ is not a unification of $G_{j-1}$, there is a quartet, $ab|cd$ say, in $\mathcal{Q}$ not yet collected by $\mathcal{S}_v$ such that $|\{a, b, c, d\} \cap A_j| \geq 2$, where, in the case $|\{a, b, c, d\} \cap A_j| = 2$, we have $\{a, b, c, d\} \cap A_j \notin \{\{a, b\}, \{c, d\}\}$. If $|\{a, b, c, d\} \cap A_j| = 2$, then, by the construction of $\mathcal{S}_v$, the tree $\mathcal{T}$ does not display $ab|cd$; a contradiction. So we may assume that $|\{a, b, c, d\} \cap A_j| \geq 3$. But then by our choice of $q$, $U_j$ contains three distinct vertices each having a non-empty intersection with $\{a, b, c, d\}$. This implies that no split of $\mathcal{T}$ displays $q$; a contradiction. Hence $\mathcal{S}_v$ is a unification sequence of $G_{\mathcal{Q}}$. To see that $\mathcal{S}_v$ is complete, note that $\mathcal{T}$ displays $\mathcal{Q}$ and so, for each quartet, $ab|cd$ in $\mathcal{Q}$, there exists some $i$ with the property that either $a, b \in A_i$ or $c, d \in A_i$. This establishes the lemma. □

*Proof of Theorem 4.3.2.* This is now an immediate consequence of Lemmas 4.5.1 and 4.5.2. □

We begin the proof of Theorem 4.3.3 with three lemmas.

**Lemma 4.5.3.** *Let $\mathcal{Q}$ be a collection of quartets on $X$. If $\mathcal{Q}$ identifies a leaf-labelled tree $\mathcal{T}$, then $\mathcal{Q}$ specially distinguishes $\mathcal{T}$.*

*Proof.* Suppose that $\mathcal{Q}$ identifies $\mathcal{T}$, but does not specially distinguish $\mathcal{T}$. Then there exists an interior edge, $uv$ say, of $\mathcal{T}$ such that $G_{\mathcal{Q}(u,v)}$ contains $k > 1$ components $C_1, \ldots, C_k$. We next construct a leaf-labelled tree $\mathcal{T}'$ from $\mathcal{T}$ that displays $\mathcal{Q}$ but is not a refinement of $\mathcal{T}$.

Recalling the definition of $G_{\mathcal{Q}(u,v)}$, delete $v$ and all its incident edges from $\mathcal{T}$. For each $i \in \{1, \ldots, k\}$, either add a new edge joining $u$ and the vertex of $C_i$ if $C_i$ contains exactly one vertex, or adjoin a new vertex $v_i$ to $u$ via a new edge and, for each vertex $w$ of $C_i$, add a new edge joining $v_i$ and $w$. It is now easily seen that the resulting tree $\mathcal{T}'$ displays $\mathcal{Q}$. But $\mathcal{T}'$ is not a refinement of $\mathcal{T}$. It now follows that $\mathcal{Q}$ specially distinguishes $\mathcal{T}$. □

A leaf-labelled tree is *minimally refined* with respect to displaying a set $\mathcal{Q}$ of quartets if it is not a proper refinement of another tree that displays $\mathcal{Q}$.

**Lemma 4.5.4.** *Let $\mathcal{Q}$ be a compatible set of quartets on $X$. If $\mathcal{S}$ is a minimal complete unification sequence of $G_{\mathcal{Q}}$, then the leaf-labelled tree whose set of non-trivial splits is $\Sigma_{\mathcal{S}}$ is minimally refined with respect to displaying $\mathcal{Q}$.*

*Proof.* Suppose that $\mathcal{S}$ is the sequence $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_k$ with unifying sequence $U_1, \ldots, U_k$, and let $\mathcal{T}$ be the leaf-labelled tree whose set of non-trivial splits is $\Sigma_{\mathcal{S}}$. If $\mathcal{T}$ is not minimally refined with respect to displaying $\mathcal{Q}$, then there is an edge $e$ of $\mathcal{T}$ whose contraction results in another tree, $\mathcal{T}'$ say, that displays $\mathcal{Q}$. Let $A_e|(X - A_e)$ denote the split of $\mathcal{T}$ induced by $e$, where, for some $i$, $A_e$ is the union of the elements of $U_i$.

Let $\mathcal{S}'$ be the sequence that is obtained from $\mathcal{S}$ by replacing the sequence of unifying sets associated with $\mathcal{S}$ with $U_1, \ldots, U_{i-1}, U'_{i+1}, \ldots, U'_k$, where, for all $j \in \{i+1, \ldots, k\}$,

$$U'_j = \begin{cases} (U_j - A_e) \cup U_i, & \text{if } A_e \text{ is an element of } U_j; \\ U_j, & \text{otherwise.} \end{cases}$$

Note that if, for some $j$, $U'_j \neq U_j$, then there is exactly one such $j$. To prove the lemma, it suffices to show that $\mathcal{S}'$ is a complete unification sequence of $G_{\mathcal{Q}}$.

Clearly, $\mathcal{S}_{i-1}$ is a unification sequence of $G_{\mathcal{Q}}$. Consider $G'_{i+1}$. If $U'_{i+1} = U_{i+1}$, then it is easily seen that $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_{i-1}, G'_{i+1}$ is

a unification sequence of $G_{\mathcal{Q}}$. Therefore assume that $U'_{i+1} \neq U_{i+1}$. If $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_{i-1}, G'_{i+1}$ is not a unification sequence, then there is a quartet, $q$ say, in $\mathcal{Q}$ such that the two $q$-coloured edges are both incident with vertices in $U'_{i+1}$. Since $\mathcal{S}_{i+1}$ is a unification sequence of $G_{\mathcal{Q}}$, this implies that one of these $q$-coloured edges, $ab$ say, is incident with two vertices in $U_i$, while the other $q$-coloured edge, $cd$ say, is incident with at least one vertex in $U_{i+1} - A_e$. It now follows that $A_e|(X - A_e)$ is the unique split in $\Sigma_{\mathcal{S}}$ that displays $q$. In turn, this implies that $\mathcal{T}'$ does not display $\mathcal{Q}$; a contradiction. Thus $G_0 = G_{\mathcal{Q}}, G_1, \ldots, G_{i-1}, G'_{i+1}$ is a unification sequence of $G_{\mathcal{Q}}$. Moreover, $G'_{i+1} = G_{i+1}$ and, for all $j \in \{i+2, \ldots, k\}$, we have $U'_j = U_j$. It now follows that in this case $\mathcal{S}'$ is a complete unification sequence of $G_{\mathcal{Q}}$.

Considering, in turn, each of the graphs $G'_{i+2}, \ldots, G'_k$ and repeatedly using the same argument as that in the previous paragraph, we eventually deduce that either $\mathcal{S}'$ is a complete unification sequence of $G_{\mathcal{Q}}$ or $\mathcal{S}'$ is a unification sequence but not complete. In the latter case, there is a $q' \in \mathcal{Q}$ such that $G'_k$ contains two $q'$-coloured edges. By Lemma 4.5.1, the leaf-labelled tree whose set of non-trivial splits is $\Sigma_{\mathcal{S}'}$ does not display $q'$. But, as $\mathcal{S}'$ is not complete, $U'_j = U_j$ for all $j$ and so $\Sigma_{\mathcal{S}'} = \Sigma_{\mathcal{S}} - A_e|(X - A_e)$. But $\Sigma_{\mathcal{S}'}$ is the set of non-trivial splits of $\mathcal{T}'$ and so $\mathcal{T}'$ does not display $q'$; a contradiction. This completes the proof of the lemma. □

**Lemma 4.5.5.** *Let $\mathcal{Q}$ be a set of quartets on $X$ and let $\mathcal{T} \in \mathscr{T}_X$ be a tree that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$. Let $q = A|B$ be a quartet in $\mathcal{Q}$ that distinguishes an edge $e = uv$ of $\mathcal{T}$. Then there is a minimal complete unification sequence of $G_{\mathcal{Q}}$ such that, amongst the quartets in $\mathcal{Q}$, the quartet $q$ is collected (joint) last and $A$ is merged. In particular, by choosing $v$ to be the vertex of $\mathcal{T}$ such that the elements in $A$ are in a different component of $\mathcal{T}\backslash e$ from $v$, the sequence $\mathcal{S}_v$ described prior to Lemma 4.5.2 is such a sequence.*

*Proof.* Suppose that $q$ distinguishes the edge $e = uv$ of $\mathcal{T}$, and let $A_e|(X - A_e)$ denote the split of $\mathcal{T}$ induced by $e$. Without loss of generality, we may assume that the elements in $A$ are in the same component of $\mathcal{T}\backslash e$ as $u$. Let $\mathcal{S}_v$ be the complete unification sequence of $G_{\mathcal{Q}}$ as described prior to Lemma 4.5.2 with the additional proviso that $A_e|(X - A_e)$ is last in the associated ordering

of the non-trivial splits induced by the edges of $\mathcal{T}$. It is easily seen using Lemma 4.5.2 that such an ordering and sequence is possible.

To complete the proof of the lemma, we show that $\mathcal{S}_v$ is minimal. If not, then there is a complete unification sequence $\mathcal{S}$ of $G_{\mathcal{Q}}$ such that $\Sigma_{\mathcal{S}}$ is a proper subset of $\mathcal{S}_v$. But then $\mathcal{T}$ is a proper refinement of the tree whose set of non-trivial splits is $\Sigma_{\mathcal{S}}$. Since this last tree also displays $\mathcal{Q}$, we contradict the fact that $\mathcal{Q}$ distinguishes $\mathcal{T}$. Thus $\mathcal{S}_v$ is minimal. □

*Proof of Theorem 4.3.3.* First suppose that $\mathcal{Q}$ identifies a leaf-labelled tree $\mathcal{T}$. Then, by Lemma 4.5.3, (i) holds for $\mathcal{T}$. We next show that (ii) holds for $\mathcal{T}$. Let $\mathcal{Q}'$ be a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$ and let $q = A|B \in \mathcal{Q}'$. Let $\mathcal{S}$ and $\mathcal{S}'$ be two minimal complete unification sequences of $G_{\mathcal{Q}}$ such that amongst the quartets in $\mathcal{Q}'$, the quartet $q$ is collected (joint) last and $A$ is merged. Let $q' = A'|B' \in \mathcal{Q}'$ and suppose that, in $\mathcal{S}$, the set $A'$ is merged, while, in $\mathcal{S}'$, the set $B'$ is merged. Furthermore, suppose that $A_i$ merged $A'$ and $A_j$ merged $A$ in $\mathcal{S}$, and that $A_{i'}$ merged $B'$ and $A_{j'}$ merged $A$ in $\mathcal{S}'$.

Since $\mathcal{Q}$ identifies $\mathcal{T}$, it follows by Lemma 4.5.4 that the leaf-labelled trees whose sets of non-trivial splits are $\Sigma_{\mathcal{S}}$ and $\Sigma_{\mathcal{S}'}$ are both isomorphic to $\mathcal{T}$, in particular, $\Sigma_{\mathcal{S}} = \Sigma_{\mathcal{S}'}$. Since $\mathcal{Q}'$ is a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$, both $q$ and $q'$ distinguish edges of $\mathcal{T}$, and so exactly one split of $\Sigma_{\mathcal{S}}$ displays $q$ and exactly one split of $\Sigma_{\mathcal{S}}$ displays $q'$. This implies that $A_i = (X - A_{i'})$ (so $A_{i'} = (X - A_i)$) and $A_j = A_{j'}$. Up to symmetry, there are two cases to consider:

(I) $A_i \subseteq A_j$ and $A_{i'} \subseteq A_{j'}$; and

(II) $A_i \subseteq A_j$ and $A_{i'} \cap A_{j'} = \emptyset$.

If (I) holds, then $A_j$ contains $X - A_i$. But $A_j$ contains $A_i$, and so $A_j$ contains $X$; a contradiction. Consider (II). Since $A_{i'} \cap A_{j'} = \emptyset$, we have $(X - A_i) \cap A_j = \emptyset$. But $A_i \subseteq A_j$, so $A_i = A_j$. Therefore, as $q$ is collected (joint) last amongst the quartets in $\mathcal{Q}'$ in $\mathcal{S}$, $i = j$. Thus, as $A_i = A_j = A_{j'}$, we have $A_{i'} = (X - A_{j'})$. But $i' \leq j'$, and so $\mathcal{S}'$ merges $B$; a contradiction. Hence (II) does not hold. It now follows that (ii) does indeed hold.

To prove the converse, suppose that, in the size of its leaf set, $\mathcal{Q}$ is a minimal collection of quartets that satisfies (i) and (ii), but does not identify a tree. Since $\mathcal{T}$ is specially distinguished by $\mathcal{Q}$, it follows that $\mathcal{T}$ is minimally refined with respect to displaying $\mathcal{Q}$. Let $\mathcal{T}' \in \mathscr{T}_X$ be another tree that is minimally refined with respect to displaying $\mathcal{Q}$.

We will show that every split of $\mathcal{T}$ is also a split of $\mathcal{T}'$, contradicting the assumption that $\mathcal{T}'$ is minimally refined and different from $\mathcal{T}$. Assume not. Let $\mathcal{Q}'$ be a minimal subset of $\mathcal{Q}$ that specially distinguishes $\mathcal{T}$, and let $q = ab|cd$ be a quartet in $\mathcal{Q}'$ such that the subset of splits in $\Sigma(\mathcal{T}')$ that display $q$ is minimal and does not contain any split of $\mathcal{T}$. Such a quartet exists, since every quartet in $\mathcal{Q}'$ distinguishes an edge of $\mathcal{T}$ and thus is displayed by exactly one split of $\mathcal{T}$. Therefore, a quartet that is displayed by a split in $\Sigma(\mathcal{T}) - \Sigma(\mathcal{T}')$ is not displayed by any split in $\Sigma(\mathcal{T}) \cap \Sigma(\mathcal{T}')$. Let $A|B$ be the split of $\mathcal{T}$ that displays $q$. Without loss of generality, we may assume that $a, b \in A$. Let $H$ be the graph that has vertex set $X$ and an edge joining two vertices $g$ and $h$ precisely if $\{g, h\} \in M(\mathcal{Q}')_{\mathcal{S}}$, where $\mathcal{S}$ is a minimal complete unification sequence of $G_{\mathcal{Q}}$ that collects $q$ (joint) last amongst the quartets in $\mathcal{Q}'$ and merges $\{a, b\}$.

We claim that the vertex set of the connected component of $H$ that contains $a$ and $b$ also contains $A$. Assume the claim is wrong and choose $A'|B' \in \Sigma(\mathcal{T})$ such that $A'$ is minimal with the property that $A' \subseteq A$ and that there is no component of $H$ whose vertex set contains $A'$. Let $L_1, \ldots, L_k$ be the (pairwise different) maximal proper subsets of $A'$ such that, for all $i \in \{1, \ldots, k\}$, the bipartition $L_i|(X - L_i)$ is a split of $\mathcal{T}$. For all $i$, it follows from the minimality of $A'$ that there is a component of $H$ that contains $L_i$. Let $H'$ be the graph that has vertex set $L_1, \ldots, L_k$ and an edge joining to vertices $L_i$ and $L_j$ precisely if there is a quartet $gg'|hh' \in \mathcal{Q}'$ with $g \in L_i$, $g' \in L_j$, and $h, h' \in B'$. Since $\mathcal{Q}'$ specially distinguishes $\mathcal{T}$ the graph $H'$ is connected. It now follows by Lemma 4.5.5 and the fact that (ii) holds for $\mathcal{T}$ that, for all such $gg'|hh'$, we have $\{g, g'\} \in M(\mathcal{Q}')_{\mathcal{S}}$. Hence there is a connected component of $H$ whose vertex set contains $A'$; a contradiction. This establishes the claim.

By Lemma 4.5.5, there is a minimal complete unification sequence $\mathcal{S}'$ of $G_{\mathcal{Q}}$ that collects $q$ (joint) last amongst the quartets in $\mathcal{Q}'$ and merges $\{a, b\}$

such that $\mathcal{T}'$ is the leaf-labelled tree induced by $\mathcal{S}'$. Noting that $M(\mathcal{Q}')_{\mathcal{S}'} = M(\mathcal{Q}')_{\mathcal{S}}$, it is easily seen that, as there is a connected component of $H$ whose vertex set contains $A$, the graph obtained from $\mathcal{T}'$ by deleting all edges corresponding to the splits that display $q$ has a connected component whose vertex set contains $A$. By repeating the above argument using $\{c, d\}$ instead of $\{a, b\}$, the same graph also has a connected component whose vertex set contains $B$. Hence $A|B \in \Sigma(\mathcal{T}')$. This completes the proof of the converse and thus the theorem. □

*Proof of Corollary 4.3.5.* Suppose that $\mathcal{Q}$ defines a leaf-labelled tree $\mathcal{T}$. Then it is clear that (i) holds for $\mathcal{T}$. To show that (ii) holds for $\mathcal{T}$, let $\mathcal{Q}'$ be a minimum-sized subset of $\mathcal{Q}$ that distinguishes $\mathcal{T}$. First note that, for distinct $q, q' \in \mathcal{Q}'$, the quartets $q$ and $q'$ distinguish different edges of $\mathcal{T}$. Let $q = A|B \in \mathcal{Q}'$. Let $\mathcal{S}$ and $\mathcal{S}'$ be two minimal complete unification sequences of $G_{\mathcal{Q}}$ so that amongst the quartets in $\mathcal{Q}'$, the quartet $q$ is collected last. If both $\mathcal{S}$ and $\mathcal{S}'$ merge $A$, or both $\mathcal{S}$ and $\mathcal{S}'$ merge $B$, then, by Theorem 4.3.3, (ii) holds. Furthermore, making use of the note, the argument for the case that one of the sequences, $\mathcal{S}$ say, merges $A$ and the other sequence, $\mathcal{S}'$ say, merges $B$ is similar to that used in the analogous part in the proof of Theorem 4.3.3. We omit the straightforward details.

Now suppose that (i) and (ii) hold. Then, by Theorem 4.3.3, $\mathcal{Q}$ identifies a leaf-labelled tree. Since $\mathcal{T}$ is a binary tree that displays $\mathcal{Q}$ and is distinguished by $\mathcal{Q}$, we deduce that $\mathcal{Q}$ defines $\mathcal{T}$. This completes the proof of the corollary. □

# Chapter 5

# Minimum Identifying Sets of Quartets

## 5.1 Introduction

In Chapter 3, we were interested in sets of quartets that defined a given binary leaf-labelled tree. We turn our attention now instead to the analogous notion for non-binary trees. That is, identifying a tree. Recall that a set of quartets $\mathcal{Q}$ on the set $X$ identifies a leaf-labelled tree if and only if all trees in $\mathscr{T}_X$ that display $\mathcal{Q}$ are refinements of $\mathcal{T}$.

As an example, consider the tree $\mathcal{T}$ in Fig. 5.1, with the single split $12|3456$. Using (3.2), we can easily verify that the quartet set $\{12|34, 12|45, 12|56\}$ identifies $\mathcal{T}$. It is not possible, however, to identify $\mathcal{T}$

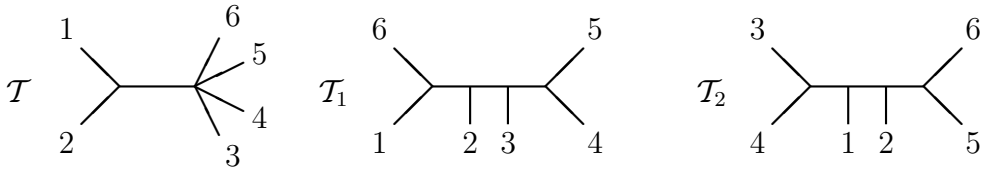


Figure 5.1: The tree $\mathcal{T}$ that is identified by $\{12|34, 12|45, 12|56\}$, and trees $\mathcal{T}_1$ and $\mathcal{T}_2$ that display $\{12|34, 12|45\}$ and $\{12|34, 12|56\}$ respectively.

with any fewer than three quartets. To see this, assume that such a quartet set $\mathcal{Q}$ contains exactly two quartets. Without loss of generality, we either have $\mathcal{Q} = \{12|34, 12|45\}$ or $\mathcal{Q} = \{12|34, 12|56\}$. Examples of trees $\mathcal{T}_1$ and $\mathcal{T}_2$

that display the aforementioned quartet sets but that are not refinements of $\mathcal{T}$ are shown in Fig. 5.1.

We are interested in determining the minimum size of quartet set $\mathcal{Q}$ required to identify an arbitrary tree. If $\mathcal{Q}$ identifies a binary leaf-labelled tree $\mathcal{T}$, then $\mathcal{Q}$ in fact defines $\mathcal{T}$. We already know that any binary tree may be defined by a set of only $n-3$ quartets, but it turns out that in general, $n-3$ quartets are insufficient to identify a tree.

The main result of this chapter is Theorem 5.1.1. This corrects [42, Theorem 6.3.9] which incorrectly states that for any tree $\mathcal{T}$ with $n$ leaves, there is a set of at most $n-3$ quartets that identifies $\mathcal{T}$. As a counterexample to this, consider the tree $\mathcal{T}$ shown in Fig. 5.2. Suppose that $\mathcal{Q}$ is a set of quartets

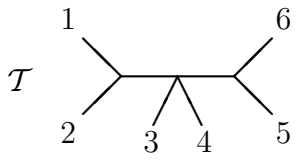


Figure 5.2: A tree $\mathcal{T}$ with $n = 6$ leaves that cannot be identified by a set of $n - 3$ quartets.

that identifies this tree. Then $\mathcal{Q}$ must contain the quartet $12|34$. To see this, the tree $\mathcal{T}_1$ in Fig. 5.3 displays all the quartets in $\mathcal{Q}(\mathcal{T}) - \{12|34\}$. Further to this, $\mathcal{Q}$ must contain one of $\{12|35, 12|36, 12|45, 12|46\}$, for otherwise the tree $\mathcal{T}_2$ in Fig. 5.3 displays $\mathcal{Q}$. By symmetry then, $\mathcal{Q}$ also contains $34|56$ and one of $\{13|56, 14|56, 23|56, 24|56\}$. Thus an identifying quartet set for $\mathcal{T}$ contains at least four distinct quartets.

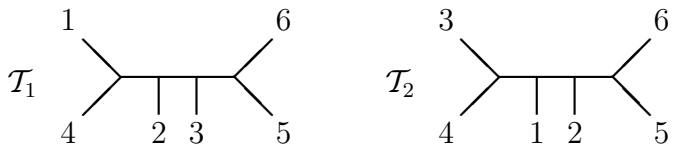


Figure 5.3: Trees $\mathcal{T}_1$ and $\mathcal{T}_2$ which demonstrate that $\mathcal{T}$ from Fig. 5.2 cannot be identified by fewer than four quartets.

We will use $\mathring{E}(\mathcal{T})$ to denote the set of interior edges of $\mathcal{T}$, and remind the reader that $d(v)$ denotes the degree of a vertex $v$ of $\mathcal{T}$. Let $q(\mathcal{T})$ denote the size of a minimum-sized set of quartets that identifies $\mathcal{T}$.

The main theorem of this chapter is the following:

**Theorem 5.1.1.** *Let $\mathcal{T}$ be a leaf-labelled tree and let $\mathcal{Q}$ be a collection of quartets that identifies $\mathcal{T}$. Then, for each interior edge $e = uv$ of $\mathcal{T}$ with $d(u) \leq d(v)$, the collection $\mathcal{Q}$ contains at least $q(d(u) - 1, d(v) - 1)$ quartets that distinguish $e$, where*

$$q(r, s) = \left\lceil \frac{r(s-1)}{2} \right\rceil$$

*for all $r, s \geq 2$. In particular,*

$$|\mathcal{Q}| \geq \sum_{uv \in \mathring{E}(\mathcal{T})} q(d(u) - 1, d(v) - 1).$$

*Moreover, there exists a collection of quartets that identifies $\mathcal{T}$ and has size*

$$q(\mathcal{T}) = \sum_{uv \in \mathring{E}(\mathcal{T})} q(d(u) - 1, d(v) - 1).$$

Restricting Theorem 5.1.1 to binary trees, we obtain the well-known result that $n - 3$ quartets are necessary to define a binary tree with $n$ leaves. See, for example, [42, Corollary 6.3.10].

Section 5.2 contains the proof of Theorem 5.1.1. The proof of this theorem requires extensive use of closure rules to show the special case where the tree of interest in a one-split tree. For this reason, we begin the following section by developing some inference rules for partial splits. The final section of this chapter characterises the trees that respectively maximise and minimise the size of $q(\mathcal{T})$.

## 5.2 Proof of Theorem 5.1.1

Closure rules for quartet sets and more generally, splits, were discussed in Section 3.2. Let us introduce another rule that we will use repeatedly in

proving Theorem 5.1.1. This *triadic closure rule* can be found in [19].

$$\{ab|de, ac|df, bc|ef\} \vdash abc|def \tag{5.1}$$

We remarked earlier that the dyadic closure rule (3.1) is a special case of the split closure rule (3.2). Lemma 5.2.1 generalises (5.1) in a similar manner.

**Lemma 5.2.1.** *Let $\Sigma = \{A_1|B_1, A_2|B_2, A_3|B_3\}$ be a set of partial splits of $X$ such that $A_i \cap A_j \neq \emptyset, B_i \cap B_j \neq \emptyset$ for all $i \neq j$. Then*

$$\Sigma \vdash \bigcup_{i\neq j}(A_i \cap A_j)|\bigcup_{i\neq j}(B_i \cap B_j).$$

*Proof.* By Lemma 3.2.1, it suffices to show that every $q = xy|wz$, where $x, y \in \bigcup_{i\neq j}(A_i \cap A_j)$ and $w, z \in \bigcup_{i\neq j}(B_i \cap B_j)$, is inferred by $\Sigma$. Clearly, this holds if $x, y \in A_i$ and $w, z \in B_i$ for some $i$. Therefore assume that this does not happen. Then, without loss of generality, we may assume that $x \in A_1 \cap A_2$, $y \in A_1 \cap A_3$, and $z \in B_2 \cap B_3$. By symmetry, there are two cases to consider depending on whether $w \in B_1 \cap B_2$ or $w \in B_2 \cap B_3$.

Let $a \in A_2 \cap A_3$ and $b \in B_1 \cap B_3$. If $w \in B_1 \cap B_2$, then, as $xy|wb \in \mathcal{Q}(A_1|B_1)$, $xa|wz \in \mathcal{Q}(A_2|B_2)$, and $ya|zb \in \mathcal{Q}(A_3|B_3)$, it follows by (5.1) that

$$\{xy|wb, xa|wz, ya|zb\} \vdash xya|wzb.$$

Hence, in this case, $q$ is inferred by $\Sigma$.

If $w \in B_2 \cap B_3$, then $xa|wz \in \mathcal{Q}(A_2|B_2)$ and $ya|wz \in \mathcal{Q}(A_3|B_3)$. Therefore, by (3.1), $\Sigma$ infers $xya|wz$ which in turn infers $q$. This completes the proof of the lemma. □

Analogously to a collection of trees, a collection $\Sigma$ of partial splits on $X$ *identifies* a tree $\mathcal{T} \in \mathscr{T}_X$ if $\mathcal{T}$ displays $\Sigma$ and all other trees in $\mathscr{T}_X$ that display $\Sigma$ are refinements of $\mathcal{T}$.

**Lemma 5.2.2.** *Let $\mathcal{T}$ be a one-split leaf-labelled tree in which the unique non-trivial split is $A|B$ with $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$. Then, for positive integers $m$ and $n$ with $r \leq 2m-1$ and $s \leq 2n-1$, the 2-element*

*collection*

$$\Sigma = \left\{a_1 \cdots a_m | b_1 \cdots b_n, a_{r-m+1} \cdots a_r | b_{s-n+1} \cdots b_s\right\}$$

*of partial splits together with the collection*

$$\mathcal{Q} = \left\{a_i a_{m+i} | b_j b_{n+j} : 1 \leq i \leq r-m, 1 \leq j \leq s-n\right\}$$

*of quartets identifies $\mathcal{T}$.*

*Proof.* Let

$$A' = \{a_1, \ldots, a_m\} \cap \{a_{r-m+1}, \ldots, a_r\}$$

and

$$B' = \{b_1, \ldots, b_n\} \cap \{b_{s-n+1}, \ldots, b_s\}.$$

Since $r \leq 2m-1$ and $s \leq 2n-1$, it follows that both $A'$ and $B'$ are non-empty. Therefore, by Lemma 5.2.1, the two partial splits in $\Sigma$ together with the quartet $a_i a_{m+i} | b_j b_{n+j}$ infer the partial split

$$(A' \cup \{a_i, a_{m+i}\}) | (B' \cup \{b_j, b_{n+j}\}) \tag{5.2}$$

for all $i$ and $j$. Furthermore, by repeated applications of (3.2), the partial splits of the form (5.2) infer $(A' \cup \{a_i, a_{m+i}\})|B$ for all $i$. Repeatedly using (3.2) again, these last partial splits infer $A|B$. It now follows that the partial splits in $\Sigma$ together with the quartets in $\mathcal{Q}$ identify $\mathcal{T}$. □

For a one-split tree $\mathcal{T}$ whose non-trivial split is $A|B$ with $|A| \leq |B|$, the size of a minimum-sized set of quartets that identifies $\mathcal{T}$ is given by $q(|A|, |B|)$. Much of the work in proving Theorem 5.1.1 goes into proving the next lemma, a special case of that theorem.

**Lemma 5.2.3.** *Let $\mathcal{T}$ be a one-split leaf-labelled tree in which the only non-trivial split is $A|B$ with $|A| = r$ and $|B| = s$, where $2 \leq r \leq s$. Then*

$$q(\mathcal{T}) = q(r, s) = \left\lceil \frac{r(s-1)}{2} \right\rceil.$$

*Proof.* Throughout the proof, we will assume that $A = \{a_1, \ldots, a_r\}$ and $B = \{b_1, \ldots, b_s\}$. We first show that $q(r,s) \geq \lceil \frac{r(s-1)}{2} \rceil$.

Suppose that $\mathcal{Q}$ is a set of quartets that identifies $\mathcal{T}$ with $|\mathcal{Q}| < \frac{r(s-1)}{2}$, and consider the quartet graph $G_{\mathcal{Q}}$. Since $\mathcal{Q}$ identifies $\mathcal{T}$, no edge in $G_{\mathcal{Q}}$ joins a singleton of $A$ to a singleton of $B$, and, in view of Lemma 4.5.3, $G_{\mathcal{Q}}$ consists of two components whose vertex sets are the set of singletons of $A$ and the set of singletons of $B$. Furthermore, if $q \in \mathcal{Q}$, then there is a $q$-coloured edge joining a pair of singletons of $A$ and a $q$-coloured edge joining a pair of singletons of $B$. Since $|\mathcal{Q}| < \frac{r(s-1)}{2}$ and $r \leq s$, there is a vertex $\{a\} \subset A$ that is incident with at most $s-2$ differently coloured edges.

Let $G_a$ be the subgraph of $G_Q$ that is obtained by deleting all of the singletons of $A$ and deleting all edges whose colour is not that of any coloured edge incident with $\{a\}$ in $G_{\mathcal{Q}}$. Hence, $G_a$ has $s$ vertices and at most $s-2$ edges and is therefore disconnected. Let $C_1, \ldots, C_k$ be the connected components of $G_a$ containing at least two vertices. As $\mathcal{Q}$ specially distinguishes $\mathcal{T}$, we have $k \geq 1$. Now consider the unification sequence $\mathcal{S}$ of $G_{\mathcal{Q}} = G_0, G_1, \ldots, G_{k+1}$ in which we make the following unifications:

(i) For $1 \leq i \leq k$, unify the vertices in $C_i$ of $G_{i-1}$ to obtain $G_i$;

(ii) unify $\{a\}$ together with the set of vertices whose union is $B$ to obtain $G_{k+1}$.

It is easily checked that $\mathcal{S}$ is a complete-unification sequence of $G_{\mathcal{Q}}$. By Lemma 4.5.1, the tree $\mathcal{T}'$ whose set of non-trivial splits is $\Sigma_{\mathcal{S}}$ displays $\mathcal{Q}$. But $A|B$ is not a split of $\mathcal{T}'$, and so $\mathcal{T}'$ is not a refinement of $\mathcal{T}$, contradicting that $\mathcal{Q}$ identifies $\mathcal{T}$. We conclude that $q(r,s) \geq \lceil \frac{r(s-1)}{2} \rceil$.

We next show that $q(r,s) \leq \lceil \frac{r(s-1)}{2} \rceil$ for all $r$ and $s$. We begin with the case $r = 2$.

**5.2.3.1.** *For all $s$, we have $q(2,s) \leq \lceil \frac{2(s-1)}{2} \rceil = s-1$.*

*Proof.* Here $A|B = \{a_1, a_2\}|\{b_1, \ldots, b_s\}$ and it follows by repeated applications of (3.2) that the collection

$$\mathcal{Q} = \left\{a_1a_2|b_1b_i : i \in \{2, \ldots, s\}\right\}$$

of quartets identifies $\mathcal{T}$. As $|\mathcal{Q}| = s - 1$, the inequality holds for $r = 2$. □

**5.2.3.2.** *For all $r$, we have $q(r, r) \leq \frac{r(r-1)}{2}$.*

*Proof.* Let $\mathcal{Q}_r$ be the collection $\{a_i a_j | b_i b_j : 1 \leq i < j \leq r\}$ of quartets. Then $|\mathcal{Q}_r| = \binom{r}{2} = \frac{r(r-1)}{2}$. The proof is by induction on $r$. Clearly, the result holds for $r = 2$. Now suppose that $r \geq 3$ and that the result holds for all smaller values of $r$. Then the partial split $a_1 \cdots a_{r-1} | b_1 \cdots b_{r-1}$ can be identified by $\mathcal{Q}_{r-1}$. By (5.1), the quartets in $\mathcal{Q}_{r-1}$ and $\mathcal{Q}_r - \mathcal{Q}_{r-1}$ infer each of the partial splits in

$$\{a_i a_j a_r | b_i b_j b_r : 1 \leq i < j < r\}.$$

Moreover, by repeatedly applying (3.2), we deduce that the elements in this set infer $a_1 \cdots a_r | b_1 \cdots b_r$. □

**5.2.3.3.** *For all $r$ and all $s$ with $r \leq s \leq 2r - 2$, we have $q(r, s) \leq \lceil \frac{r(s-1)}{2} \rceil$.*

*Proof.* The proof is by induction on $r$. If $r = 2$, then the result holds by (5.2.3.1). Now suppose that $r \geq 3$, and that the result holds for all smaller values of $r$. There are five cases to consider.

**Case 1.** $s = 2l - 1$ for some integer $l \geq 2$.

By Lemma 5.2.2, the 2-element collection

$$\Sigma_1 = \left\{ a_1 \cdots a_l | b_1 \cdots b_l, a_{r-l+1} \cdots a_r | b_l \cdots b_s \right\}$$

of partial splits together with the collection

$$\mathcal{Q}_1 = \left\{ a_i a_{l+i} | b_j b_{l+j} : 1 \leq i \leq r - l, 1 \leq j \leq l - 1 \right\}$$

of quartets identify $\mathcal{T}$. By the induction assumption, each partial split in $\Sigma_1$ can be identified by a collection of $\frac{l(l-1)}{2}$ quartets. Furthermore, $\mathcal{Q}_1$ contains $(r - l)(l - 1)$ quartets. Thus

$$\begin{aligned} q(r, s) &\leq l(l - 1) + (r - l)(l - 1) \\ &= \frac{r(s - 1)}{2}. \end{aligned}$$

**Case 2.** $r = 2k$ and $s = 2l$ for some integers $k \geq 2$ and $l \geq 3$, where either $k$ is odd or $l$ is even.

By Lemma 5.2.2, the 2-element collection

$$\Sigma_2 = \left\{a_1 \cdots a_{k+1}|b_1 \cdots b_{l+1}, a_k \cdots a_r|b_l \cdots b_s\right\}$$

of partial splits together with the collection

$$\mathcal{Q}_2 = \left\{a_i a_{k+i+1}|b_j b_{l+j+1} : 1 \leq i \leq k-1, 1 \leq j \leq l-1\right\}$$

of quartets identify $\mathcal{T}$. By the induction assumption, each partial split in $\Sigma_2$ can be identified by a collection of $\frac{(k+1)l}{2}$ quartets. Without loss of generality, we may assume that these last collections share the quartet $a_k a_{k+1}|b_l b_{l+1}$. Furthermore, $\mathcal{Q}_2$ contains $(k-1)(l-1)$ quartets. Thus

$$\begin{aligned} q(r,s) &\leq (k+1)l - 1 + (k-1)(l-1) \\ &= \frac{r(s-1)}{2}. \end{aligned}$$

**Case 3.** $r = 2k-1$ and $s = 2l$ for some integers $k \geq 2$ and $l \geq 2$, where either $k$ is odd or $l$ is even.

By Lemma 5.2.2, the 2-element collection

$$\Sigma_3 = \left\{a_1 \cdots a_{k+1}|b_1 \cdots b_{l+1}, a_{k-1} \cdots a_r|b_l \cdots b_s\right\}$$

of partial splits together with the collection

$$\mathcal{Q}_3 = \left\{a_i a_{k+i+1}|b_j b_{l+j+1} : 1 \leq i \leq k-2, 1 \leq j \leq l-1\right\}$$

of quartets identify $\mathcal{T}$. By the induction assumption, each partial split in $\Sigma_3$ can be identified by a collection of $\frac{(k+1)l}{2}$ quartets. Without loss of generality, we may assume that these last collections share the quartet $a_k a_{k+1}|b_l b_{l+1}$.

Furthermore, $\mathcal{Q}_3$ contains $(k-2)(l-1)$ quartets. Thus

$$\begin{aligned} q(r,s) &\leq (k+1)l - 1 + (k-2)(l-1) \\ &= \left\lceil \frac{r(s-1)}{2} \right\rceil. \end{aligned}$$

**Case 4.** $r = 4k$ and $s = 4l-2$ for integers $k \geq 1$ and $l \geq 2$.

This case includes an anomaly, in particular when $k = 1$ and $l = 2$; that is, $(r,s) = (4,6)$. We will prove this subcase first before proving Case 4 in general.

Let

$$\mathcal{Q}_1' = \big\{a_1a_2|b_1b_2, a_1a_3|b_1b_3, a_2a_3|b_2b_3\big\},$$

$$\mathcal{Q}_2' = \big\{a_2a_3|b_4b_5, a_2a_4|b_4b_6, a_3a_4|b_5b_6\big\},$$

and

$$\mathcal{Q}_3' = \big\{a_1a_2|b_3b_4, a_3a_4|b_3b_4, a_1a_4|b_1b_5, a_1a_4|b_2b_6\big\}.$$

By (5.1), $\mathcal{Q}_1'$ and $\mathcal{Q}_2'$ infer the partial splits $a_1a_2a_3|b_1b_2b_3$ and $a_2a_3a_4|b_4b_5b_6$, respectively. Furthermore, together with $\mathcal{Q}_3'$, these partial splits infer $a_1a_2|b_1b_2b_3b_4$ and $a_3a_4|b_3b_4b_5b_6$ by (3.2). By (5.1), the partial splits $a_1a_2|b_1b_4$, $a_2a_4|b_4b_5$ $a_1a_4|b_1b_5$ infer $a_1a_2a_4|b_1b_4b_5$. Similarly, by (5.1), we infer

$$a_1a_2a_4|b_2b_4b_6, a_1a_3a_4|b_1b_3b_5, a_1a_3a_4|b_2b_3b_6.$$

In turn, again using (5.1), we infer

$$a_1a_2a_3|b_3b_4b_5, a_1a_2a_3|b_3b_4b_6, a_2a_3a_4|b_1b_3b_4, a_2a_3a_4|b_2b_3b_4.$$

The last eight partial splits now infer $a_1a_2|B$, $a_2a_3|B$, and $a_3a_4|B$ which, by (3.2), infers $A|B$. Thus $q(4,6) \leq 10 = \frac{4(6-1)}{2}$.

Now assume that $k \geq 2$ and $l \geq 3$. By Lemma 5.2.2, the 2-element collection

$$\Sigma_4 = \big\{a_1 \cdots a_{2k+2}|b_1 \cdots b_{2l+1}, a_{2k-1} \cdots a_r|b_{2l-2} \cdots b_s\big\}$$

of partial splits together with the collection

$$\mathcal{Q}_4 = \{a_i a_{2k+i+2} | b_j b_{2l+j+1} : 1 \leq i \leq 2k-2, 1 \leq j \leq 2l-3\}$$

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in $\Sigma_4$ can be identified by a collection of $(2k+2)l$ quartets. Consider one of these partial splits, say $a_1 \cdots a_{2k+2} | b_1 \cdots b_{2l+1}$. Since the size of the larger side is $2l+1 \geq 7$ and odd, we may make up the set of $(2k+2)l$ quartets that identify this partial split as in Case 1, where, by (5.2.3.2), we may assume that this set contains

$$\{a_{2k-1}a_{2k} | b_{2l-2}b_{2l-1}, a_{2k-1}a_{2k+1} | b_{2l-2}b_{2l}, a_{2k}a_{2k+1} | b_{2l-1}b_{2l},$$
$$a_{2k-1}a_{2k+2} | b_{2l-2}b_{2l+1}, a_{2k}a_{2k+2} | b_{2l-1}b_{2l+1}, a_{2k+1}a_{2k+2} | b_{2l}b_{2l+1}\}.$$

Similarly, we may assume the set of $(2k+2)l$ quartets that identifies the other partial split in $\Sigma_4$ also contains the six quartets in this set. Since $\mathcal{Q}_4$ contains $(2k-2)(2l-3)$ quartets, it now follows that

$$\begin{aligned} q(r,s) &\leq 2(2k+2)l - 6 + (2k-2)(2l-3) \\ &= \frac{r(s-1)}{2}. \end{aligned}$$

**Case 5.** $r = 4k-1$ and $s = 4l-2$ for some integers $k \geq 1$ and $l \geq 2$.

By Lemma 5.2.2, the 2-element collection

$$\Sigma_5 = \{a_1 \cdots a_{2k} | b_1 \cdots b_{2l}, a_{2k} \cdots a_r | b_{2l-1} \cdots b_s\}$$

of partial splits together with the collection

$$\mathcal{Q}_5 = \{a_i a_{2k+i} | b_j b_{2l+j} : 1 \leq i \leq 2k-1, 1 \leq j \leq 2l-2\}$$

of quartets identifies $\mathcal{T}$. By the induction assumption, each partial split in $\Sigma_5$ can be identified by a collection of $k(2l-1)$ quartets. Furthermore, $\mathcal{Q}_5$

contains $(2k-1)(2l-2)$ quartets. Thus

$$\begin{aligned} q(r,s) &\leq 2k(2l-1)+(2k-1)(2l-2) \\ &= \left\lceil \frac{r(s-1)}{2} \right\rceil. \end{aligned}$$

Combining Cases 1-5, we conclude that $q(r,s) \leq \lceil \frac{r(s-1)}{2} \rceil$ whenever $r \leq s \leq 2r-2$. □

We complete the proof of Lemma 5.2.3 by showing that, for any fixed $r$, the result holds for all $s$ with $r \leq s$. By (5.2.3.3), the result holds whenever $s \leq 2r-2$. Now assume that $s > 2r-2$ and that the result holds for all smaller values of $s$.

Consider the 2-element collection

$$\Sigma = \left\{ a_1 \cdots a_r | b_1 \cdots b_r, a_1 \cdots a_r | b_r \cdots b_s \right\}$$

of partial splits. Observe that, as $s > 2r-2$, we have $|\{a_1, \ldots, a_r\}| \leq |\{b_r, \ldots, b_s\}|$. By a single application of (3.2), $\Sigma$ infers the full split $A|B$. Furthermore, by (5.2.3.2), the first partial split in $\Sigma$ can be identified by a collection of $\frac{r(r-1)}{2}$ quartets and, by the induction assumption, the second partial split in $\Sigma$ can be identified by a collection of $\lceil \frac{r(s-r)}{2} \rceil$ quartets. Hence

$$\begin{aligned} q(r,s) &\leq \frac{r(r-1)}{2} + \left\lceil \frac{r(s-r)}{2} \right\rceil \\ &= \left\lceil \frac{r(s-1)}{2} \right\rceil. \end{aligned}$$

Running over all values of $r$, we deduce that

$$q(r,s) \leq \left\lceil \frac{r(s-1)}{2} \right\rceil$$

for all $r$ and all $s$ with $2 \leq r \leq s$. This completes the proof of the lemma. □

The next lemma is an immediate consequence of the definition of identifying quartet sets.

**Lemma 5.2.4.** *Let $\mathcal{T} \in \mathscr{T}_X$ be a one-split tree in which the only non-trivial split is $A|B$, and suppose that $\mathcal{T}$ displays a collection $\mathcal{Q}$ of quartets. If $\mathcal{Q}$ does not identify $\mathcal{T}$, then there is another tree $\mathcal{T}' \in \mathscr{T}_X$ that displays $\mathcal{Q}$, but for which $A|B \notin \Sigma(\mathcal{T}')$.*

Before proving Theorem 5.1.1, we require one further definition. An interior vertex of a tree that is adjacent to exactly $k$ leaves is called a *$k$-bud*, or more generally a *bud*.

*Proof of Theorem 5.1.1.* First suppose that for some interior edge $e = uv$ of $\mathcal{T}$, the subset $\mathcal{Q}_e$ of $\mathcal{Q}$ containing exactly the quartets that distinguish $e$ has the property that

$$|\mathcal{Q}_e| < q(d(u) - 1, d(v) - 1).$$

Suppose the neighbours of $u$ that are not $v$ are $U = \{u_1, \ldots, u_r\}$ and the neighbours of $v$ that are not $u$ are $V = \{v_1, \ldots, v_s\}$. Let $\mathcal{T}_e$ denote the leaf-labelled tree that is the minimal subtree of $\mathcal{T}$ containing the vertices in $U \cup V$. Furthermore, let $\mathcal{P}_e$ be the collection of quartets obtained from $\mathcal{Q}_e$ by replacing each quartet, $aa'|bb'$ say, with $u_iu_j|v_kv_l$, where $u_i$ is on the path from $u$ to $a$, $u_j$ is on the path from $u$ to $a'$, $v_k$ is on the path from $v$ to $b$, and $v_l$ is on the path from $v$ to $b'$. Since $\mathcal{T}$ displays $\mathcal{Q}_e$, it follows that $\mathcal{T}_e$ displays $\mathcal{P}_e$. However, because of the cardinality of $\mathcal{Q}_e$, it follows by Lemma 5.2.3 that $\mathcal{P}_e$ does not identify $\mathcal{T}_e$.

By Lemma 5.2.4, there is a leaf-labelled tree $\mathcal{T}'_e$ with leaf set $U \cup V$ that displays $\mathcal{P}_e$ but does not contain the split $U|V$. Let $\mathcal{T}' \in \mathscr{T}_X$ be the tree that is obtained by adjoining, for all $w \in U \cup V$, the maximal subtree of $\mathcal{T}$ that contains $w$ and neither $u$ nor $v$ to $\mathcal{T}'_e$ by identifying the common vertices denoted by $w$. Clearly, $\mathcal{T}'$ displays $\mathcal{Q}_e$. Moreover, it is easily seen by the construction of $\mathcal{T}'$ that every quartet in $\mathcal{Q} - \mathcal{Q}_e$ is also displayed by $\mathcal{T}'$. Since $\mathcal{T}'$ does not contain the split of $\mathcal{T}$ induced by $e$, we deduce that $\mathcal{Q}$ does not identify $\mathcal{T}$. This contradiction means that, for every interior edge $e = uv$, the collection $\mathcal{Q}$ contains $q(r, s)$ quartets that distinguish $e$. Thus

$$|\mathcal{Q}| \geq \sum_{e \in \mathring{E}} q(d(u) - 1, d(v) - 1).$$

We prove the second part of the theorem by induction on the number $m$ of interior edges of $\mathcal{T}$. If $m = 1$ and the unique interior edge is $uv$, then, by Lemma 5.2.3, there exists a collection of quartets of size $q(d(u)-1, d(v)-1)$ that identifies $\mathcal{T}$. Now assume that $m \geq 2$ and that the result holds for every tree with $m-1$ interior edges.

Let $e = uv$ be an interior edge of $\mathcal{T}$ such that $u$ is a bud of $\mathcal{T}$. First assume that $d(u) \leq d(v)$. Let $r = d(u) - 1$ and $s = d(v) - 1$. Furthermore, let $a_1, \cdots, a_r$ be the leaves of $\mathcal{T}$ adjacent to $u$, and let $b_1, \cdots, b_s$ be leaves of $\mathcal{T}$ such that, for all distinct $i$ and $j$, the path from $b_i$ to $b_j$ contains $v$, but not $u$. Let $\mathcal{T}' = \mathcal{T}|(X - \{a_2, \cdots, a_r\})$. Now $\mathcal{T}'$ is a leaf-labelled tree with precisely $m-1$ interior edges, and so by our induction assumption $\mathcal{T}'$ can be identified by a collection $\mathcal{Q}'$ of quartets of size $q(\mathcal{T}')$.

Let $\mathcal{Q}_e$ be a minimum-sized set of quartets that identifies the one-split leaf-labelled tree whose non-trivial split is $a_1 \cdots a_r | b_1 \cdots b_s$. By Lemma 5.2.3, $|\mathcal{Q}_e| = q(r, s)$. Consider $\mathcal{Q}_e \cup \mathcal{Q}'$. Clearly, $\mathcal{T}$ displays $\mathcal{Q}_e \cup \mathcal{Q}'$. Let $\mathcal{T}''$ be a leaf-labelled tree that displays $\mathcal{Q}_e \cup \mathcal{Q}'$. Since $\mathcal{Q}'$ identifies $\mathcal{T}'$, we have that $\mathcal{T}''|(X - \{a_2, \ldots, a_r\})$ is a refinement of $\mathcal{T}'$. Using this fact and the fact that $\mathcal{T}''$ displays $\mathcal{Q}_e$, it is easily seen that $\mathcal{T}''$ displays the partial split $a_1 \cdots a_r | b_1 \cdots b_s$. It now follows that $\mathcal{Q}_e \cup \mathcal{Q}'$ identifies $\mathcal{T}$. Moreover,

$$|\mathcal{Q}_e \cup \mathcal{Q}'| = q(d(u)-1, d(v)-1) + q(\mathcal{T}') = q(\mathcal{T}).$$

The same argument holds if $d(v) < d(u)$. This completes the proof of the theorem. □

## 5.3 Characterisations of the Extremal Cases

Recall that $q(\mathcal{T})$ denotes the size of a minimum-sized set of quartets that identifies a leaf-labelled tree $\mathcal{T}$. We end this chapter with two results that determine, for all $n$, those trees $\mathcal{T}$ with $n$ leaves for which $q(\mathcal{T})$ is minimised and maximised.

**Theorem 5.3.1.** *Let $\mathcal{T}$ be a leaf-labelled tree with $n$ leaves and at least one interior edge. Then $q(\mathcal{T}) \geq n - 3$. Moreover, $q(\mathcal{T}) = n - 3$ if and only if*

*(i) $\mathcal{T}$ has exactly one interior edge and contains a 2-bud or two 3-buds; or*

*(ii) $\mathcal{T}$ has at least two interior edges and every vertex with degree at least four is a bud.*

As the proof will show, part (i) of the above theorem follows as a reasonably simple consequence of Lemma 5.2.3. We can justify part (ii) intuitively by noting that non-binary interior vertices increase the number of quartets required to specially distinguish any incident interior edges. That is, to minimise $q(\mathcal{T})$, we require any non-binary interior vertex of $\mathcal{T}$ to be incident with at most one interior edge. To illustrate this, both six-leafed trees in Fig. 5.4 have two interior vertices of degree three and one of degree four. However, whereas $\mathcal{T}_1$ can be identified by a set of three quartets, by Theorem 5.1.1 (and as demonstrated earlier in Section 5.1), we require at least four quartets to identify $\mathcal{T}_2$.

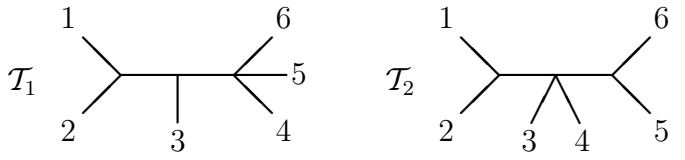


Figure 5.4: Two six-leafed trees which illustrate part (ii) of Theorem 5.3.1.

*Proof of Theorem 5.3.1.* First suppose that $\mathcal{T}$ has exactly one interior edge $uv$. Let $r = d(u) - 1 \geq 2$ and $s = d(v) - 1 \geq 2$. Without loss of generality, we may assume that $r \leq s$. Then, by Theorem 5.1.1,

$$q(\mathcal{T}) = q(r, s) = \left\lceil \frac{r(s-1)}{2} \right\rceil.$$

It is easily checked that $q(\mathcal{T}) \geq r + s - 3$. Furthermore, a routine check also shows that $q(\mathcal{T}) = r + s - 3$ if and only if $r = 2$ or $s = 3$. As $r + s - 3 = n - 3$, the proposition holds over all leaf-labelled trees with exactly one interior edge.

Next we show that the proposition holds in general. The proof is by induction on $n$. Clearly, the result holds if $n = 4$. Let $\mathcal{T}$ be a leaf-labelled tree with $n$ leaves, where $n \geq 5$, and suppose that $q(\mathcal{T})$ is of minimum size. Suppose that the proposition holds for all trees $\mathcal{T}'$ with fewer leaves for

which $q(\mathcal{T}')$ is of minimum size. Since we already know that the result holds if $\mathcal{T}$ has exactly one interior edge, we may assume that $\mathcal{T}$ has at least two interior edges. Since every binary leaf-labelled tree with $n$ leaves is defined by $n-3$ quartets (see, for example, [42]), $q(\mathcal{T}) \leq n-3$. Let $w$ be a bud of $\mathcal{T}$ of maximum size. Let $j$ be the size of this bud, let $x_1, \ldots, x_j$ denote the leaves adjacent to $w$, let $v$ be the non-leaf vertex adjacent to $w$, and let $\mathcal{T}'$ be the restriction of $\mathcal{T}$ to $X - \{x_j\}$. By the induction assumption, $q(\mathcal{T}') \geq (n-1) - 3 = n-4$. We consider the two cases $j \geq 3$ and $j = 2$ separately.

Suppose firstly that $j \geq 3$. If $d(w) \leq d(v)$, then, by Theorem 5.1.1,

$$\begin{aligned} q(\mathcal{T}) - q(\mathcal{T}') &= q(j, d(v)-1) - q(j-1, d(v)-1) \\ &= \left\lceil \frac{j(d(v)-2)}{2} \right\rceil - \left\lceil \frac{(j-1)(d(v)-2)}{2} \right\rceil \\ &\geq 1. \end{aligned}$$

Therefore

$$q(\mathcal{T}) \geq q(\mathcal{T}') + 1 \geq n - 4 + 1 = n - 3. \tag{5.3}$$

Since $q(\mathcal{T}) \leq n-3$, it follows that equality holds throughout (5.3) and so $q(\mathcal{T}) = n-3$ and $q(\mathcal{T}') = n-4$. Since $\mathcal{T}$ has at least two interior edges and $k \geq 3$, the tree $\mathcal{T}'$ has at least two interior edges and so, by the induction assumption, (ii) holds for $\mathcal{T}'$. Hence (ii) holds for $\mathcal{T}$. A similar argument also shows that (ii) holds for $\mathcal{T}$ if $d(w) > d(v)$.

Now suppose that $j = 2$. Here every bud of $\mathcal{T}$ has size two. Note that, in this case, $d(w) \leq d(v)$. By Theorem 5.1.1,

$$q(\mathcal{T}) - q(\mathcal{T}') = q(2, d(v)-1) = d(v) - 2 \geq 1.$$

Arguing as in (i), we now deduce that $q(\mathcal{T}) = n-3$ and $q(\mathcal{T}') = n-4$. This implies that $d(v) - 2 = 1$ and so $d(v) = 3$. If $\mathcal{T}'$ has at least two interior edges, then (ii) holds for $\mathcal{T}'$ and so (ii) holds for $\mathcal{T}$. Furthermore, if $\mathcal{T}'$ has exactly one interior edge, then $\mathcal{T}'$ is a quartet and again it follows that (ii) holds for $\mathcal{T}$. This completes the proof of the theorem. □

For two non-negative integers $k$ and $l$ with $k+l \geq 3$, we will denote by $\mathcal{T}_k^{2l}$ the leaf-labelled tree with $k+2l$ leaves that has an interior vertex adjacent to $k$ leaves while all other $l$ neighbours are 2-buds. As an example, Fig. 5.5 shows the shape of the tree $\mathcal{T}_2^6$.

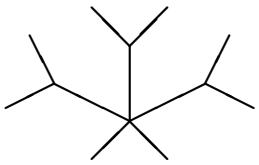

Figure 5.5: The tree shape for $\mathcal{T}_2^6$.

**Theorem 5.3.2.** *Let $\mathcal{T}$ be a leaf-labelled tree with $n$ leaves. Then $q(\mathcal{T}) \leq \left\lfloor \left(\frac{n}{2}-1\right)^2 \right\rfloor$. Moreover, $q(\mathcal{T}) = \left\lfloor \left(\frac{n}{2}-1\right)^2 \right\rfloor$ if and only if $\mathcal{T}$ is isomorphic to*

*(i) $\mathcal{T}_2^{n-2}$ if $n$ is even; or*

*(ii) $\mathcal{T}_1^{n-1}$ or $\mathcal{T}_3^{n-3}$ if $n$ is odd.*

*Proof.* First note that, for $1 \leq k \leq 3$, a routine check using Theorem 5.1.1 shows that $q(\mathcal{T}_k^{n-k}) = \left\lfloor \left(\frac{n}{2}-1\right)^2 \right\rfloor$. In other words, $q(\mathcal{T}_2^{n-2}) = \left(\frac{n}{2}-1\right)^2$ if $n$ is even and $q(\mathcal{T}_1^{n-1}) = q(\mathcal{T}_3^{n-3}) = \frac{(n-1)(n-3)}{4}$ if $n$ is odd. The proof is by induction on $n$. A simple check shows that the result holds if $n \in \{4,5\}$. Let $\mathcal{T}$ be a leaf-labelled tree with $n$ leaves, where $n \geq 6$, and suppose that $q(\mathcal{T})$ is of maximum size. Note that

$$q(\mathcal{T}) \geq \left\lfloor \left(\frac{n}{2}-1\right)^2 \right\rfloor. \tag{5.4}$$

Suppose that the theorem holds for all trees $\mathcal{T}'$ with fewer leaves for which $q(\mathcal{T}')$ is of minimum size. Say $\mathcal{T}$ has exactly one interior edge. Then one of the interior vertices is a $j$-bud with $j \leq \frac{n}{2}$ and the other interior vertex is an $(n-j)$-bud. Consequently, by Theorem 5.1.1,

$$q(\mathcal{T}) = \frac{1}{2}j(n-j-1) \leq \frac{1}{2}\left(\frac{n-1}{2}\right)^2 < \left\lfloor \left(\frac{n}{2}-1\right)^2 \right\rfloor$$

as $n \geq 6$. It now follows that $\mathcal{T}$ has at least two interior edges, which also means that $\mathcal{T}$ has no adjacent buds.

Let $w$ be a bud of $\mathcal{T}$ of maximum size and let $k$ be the size of this bud. Let $x_1, \ldots, x_k$ denote the leaves adjacent to $w$, let $v$ be the non-leaf vertex adjacent to $w$, and let $\mathcal{T}'$ be the restriction of $\mathcal{T}$ to $X - \{x_k\}$. By the induction assumption, $q(\mathcal{T}') \leq \left\lfloor \left(\frac{n-1}{2} - 1\right)^2 \right\rfloor$. Combining this with (5.4), we deduce that

$$q(\mathcal{T}) - q(\mathcal{T}') \geq \left\lceil \frac{n-3}{2} \right\rceil. \tag{5.5}$$

First suppose $k \geq 3$. Then, by Theorem 5.1.1, $q(\mathcal{T}) - q(\mathcal{T}') = q(k, d(v) - 1) - q(k-1, d(v)-1)$ and a routine check shows that $q(\mathcal{T}) - q(\mathcal{T}') \leq \frac{d(v)}{2}$. Together with (5.5), this implies that $d(v) \geq n-2$ if $n$ is even and $d(v) \geq n-3$ if $n$ is odd. Since $\mathcal{T}$ has at least two interior edges and $w$ is adjacent to $k \geq 3$ leaves, this is only possible if $n$ is odd, $k = 3$, and $v$ is adjacent to $n-5$ leaves and a 2-bud. Assuming $n$ is odd, $n \geq 7$ and so, by Theorem 5.1.1,

$$q(\mathcal{T}) = q(2, n-4) + q(3, n-4) = \frac{5}{2}(n-5) < \frac{(n-1)(n-3)}{4};$$

a contradiction.

Now suppose that $k = 2$. By Theorem 5.1.1, $q(\mathcal{T}) - q(\mathcal{T}') = q(2, d(v) - 1) = d(v) - 2$. Therefore, by (5.5), $d(v) \geq \frac{n+1}{2}$. Assume that $\mathcal{T}$ has an interior vertex $v' \neq v$ such that $v'$ is adjacent to a bud. Then, as $v$ is adjacent to a bud, there are at least $d(v) \geq \frac{n+1}{2}$ leaves $\ell$ of $\mathcal{T}$ for which $v'$ is not contained in the path from $\ell$ to $v$. Interchanging $v$ and $v'$ in this argument, we also deduce that there are at least $d(v) \geq \frac{n+1}{2}$ leaves $\ell$ of $\mathcal{T}$ for which $v$ is not contained in the path from $\ell$ to $v'$. Hence $\mathcal{T}$ has at least $n+1$ leaves; a contradiction.

It follows from the above arguments that $\mathcal{T}$ has exactly one interior vertex that is not a bud and all buds are 2-buds. Thus, for some $k$, we have that $\mathcal{T}$

is isomorphic to $\mathcal{T}_k^{n-k}$. Now

$$\begin{aligned} q(\mathcal{T}_k^{n-k}) &= \frac{n-k}{2} \cdot q\left(2, \frac{n+k}{2} - 1\right) \\ &= \frac{n-k}{2}\left(\frac{n+k}{2} - 2\right) \\ &= \frac{1}{4}(n-2+(k-2))(n-2-(k-2)) \end{aligned}$$

and, since $k$ and $n$ must have the same parity, $q(\mathcal{T}_k^{n-k})$ is maximum for $k = 2$ if $n$ is even and for $k \in \{1, 3\}$ if $n$ is odd. This completes the proof of the theorem. □

# PART II

# SUBTREES

Up until now, the problems we have addressed have all involved the reconstruction or recognition of a single tree. There are various combinatorial and practical reasons why we might also be interested in comparing the relative structures of two or more trees. These include, but are not limited to, the study of mixture models (see, for example, [32, 33]) and tree rearrangement operations, which are covered in more depth in Part III. The remaining four chapters of this thesis deal, in some way or another, with the differences and similarities inherent in a collection of trees that have the same leaf set.

Given a collection of trees $\mathcal{P} \subseteq \mathscr{T}_X$, the two general questions that we ask are

(i) how do the trees in $\mathcal{P}$ agree; and

(ii) how do the trees in $\mathcal{P}$ disagree.

The first of these questions may be rephrased in terms of finding a subset of $X$ that has evolved identically on every element of $\mathcal{P}$. The obvious extension of this is to find as large as possible a subset of $X$ for which this holds. The converse of this is to find a minimum-sized subset of $X$ that has evolved distinctly on every tree in $\mathcal{P}$. It should be immediately clear that $\mathcal{P}$ exhibits both internal agreement and disagreement, albeit perhaps only on a trivial level. Since we are dealing with unrooted trees, any three-element subset of $X$ can be displayed in only one way by any tree in $\mathscr{T}_X$. Moreover, all trees in $\mathcal{P}$ are distinct, and so the entire set $X$ is resolved differently by each tree.

Finding a solution to (ii) is not an end in itself. The origin of this lies in determining whether a set of induced subtrees of $\mathcal{P}$ is displayed by another collection of trees $\mathcal{P}'$, where $\mathcal{P}$ and $\mathcal{P}'$ have the same size. This concept, termed *disentangling* by Matsen *et al.* [33], is explained more precisely in Chapter 6. The relevance of (ii) to disentangling is that if there is a relatively small subset of $X$ that gives a distinct subtree on each tree in $\mathcal{P}$, then this effectively *separates* the induced subtrees of $\mathcal{P}$ into equivalence classes from which each member of $\mathcal{P}$ can be reconstructed uniquely. The main result of Chapter 6 gives a logarithmic lower bound and a linear upper bound on the *disentangling number*.

The problem of finding a large common subtree[1], as indicated by (i), has a close connection to classical Ramsey Theory. In Chapter 7, we begin by examining a simplification of the problem where all members of $\mathcal{P}$ are caterpillars, and then extend this to a more general result for binary trees.

[1]This is also known as the *maximum agreement subtree* problem.

# Chapter 6

# Disentangling Sets Of Trees

## 6.1 Introduction

In Chapters 3 and 5, we considered problems centred around reconstructing a tree from incomplete information. That is, determining the structure of a tree from some subset of its induced quartets or, more generally, partial splits. It is well-known, for example, that the collection of all induced quartets for a binary leaf-labelled tree defines that tree.

However, it is shown in [33] that a pair of trees cannot necessarily be uniquely reconstructed from the union of their induced subtrees with five leaves. We reproduce the key example from the aforementioned paper in Fig. 6.1. It can easily be seen that, for any five-element subset $Y \subset \{1, \ldots, 6\}$, the two sets $\{\mathcal{T}_1|Y, \mathcal{T}_2|Y\}$ and $\{\mathcal{T}_1'|Y, \mathcal{T}_2'|Y\}$ are the same. We remark further

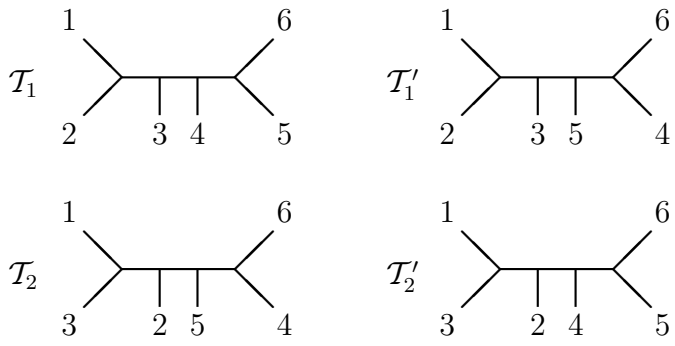


Figure 6.1: Four trees in $\mathscr{T}_6$ for which $\{\mathcal{T}_1|Y, \mathcal{T}_2|Y\} = \{\mathcal{T}_1'|Y, \mathcal{T}_2'|Y\}$ for all subsets $Y$ of size five.

that

$$\Sigma(\mathcal{T}_1) \cup \Sigma(\mathcal{T}_2) = \Sigma(\mathcal{T}_1') \cup \Sigma(\mathcal{T}_2'),$$

and so an arbitrary pair of trees from $\mathscr{T}_n$ cannot in general be reconstructed from their combined splits.

Let us define more strictly what we mean by simultaneously defining collections of $k$ trees on the same leaf-set.

**Definition 6.1.1.** For a collection of trees $\mathcal{P} \subseteq \mathscr{T}_X$ and $\mathcal{Y} \subseteq 2^X$, we write the *restriction* of $\mathcal{P}$ to $\mathcal{Y}$ as

$$\mathcal{P}|\mathcal{Y} = \{\mathcal{T}|Y : \mathcal{T} \in \mathcal{P}, Y \in \mathcal{Y}\}.$$

**Definition 6.1.2.** Let $\mathscr{P}$ be a subset of $2^{\mathscr{T}_X}$ for some $X$, and let $\mathcal{Y}$ be a collection of subsets of $X$. We say that $\mathcal{Y}$ *disentangles* $\mathscr{P}$ if and only if

$$\mathcal{P}|\mathcal{Y} \neq \mathcal{P}'|\mathcal{Y}$$

for all distinct $\mathcal{P}, \mathcal{P}' \in \mathscr{P}$.

We emphasise that while disentangling and defining are somewhat related, they are certainly not interchangeable. If $\mathcal{T}|\mathcal{Y}$ defines $\mathcal{T}$ for some $\mathcal{Y}$, then $\mathcal{T}$ has a unique restriction to $\mathcal{Y}$. On the other hand, for $\mathcal{Y}$ to disentangle $\mathscr{P}$ we require every element of $\mathscr{P}$ to have a unique restriction to $\mathcal{Y}$. We illustrate this with an example.

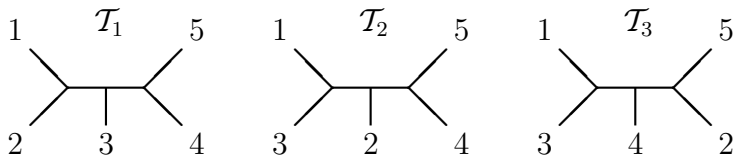


Figure 6.2: Three distinct trees in $\mathscr{T}_5$ that are not disentangled by $\mathcal{Y} = \{\{1,2,3,4\},\{1,3,4,5\}\}$.

If we take $\mathcal{Y} = \{\{1,2,3,4\},\{1,3,4,5\}\}$, then $\mathcal{T}_1$ shown in Fig. 6.2 is

defined by $\mathcal{T}|\mathcal{Y}$. However, the set

$$\mathscr{P} = \{\{\mathcal{T}_1\}, \{\mathcal{T}_2\}, \{\mathcal{T}_3\}\}$$

is not disentangled by $\mathcal{Y}$, as $\mathcal{T}_2|\mathcal{Y} = \mathcal{T}_3|\mathcal{Y}$. In general, the property of disentangling is a much stronger property than defining.

The term disentangle was introduced by Matsen *et al.* in [33], although we remark that the above definition generalises their original concept to disentangling arbitrary sets of trees. The original motivation for these ideas has its roots in the study of mixture models. That is, combining data from more than one data set according to some weighting scheme. It was shown in [32] that a mixture model on a specified tree can, under the right conditions, imitate an unmixed model on a different tree. We are interested purely in the combinatorial aspects of the problem.

## 6.2 The Disentangling Number

In this section, we will consider the problem of disentangling the set $\mathscr{P} = \binom{\mathscr{T}_n}{k}$ for some $k$. That is, all possible $k$-element subsets of $\mathscr{T}_n$. The relevance of this is that if $\mathcal{Y}$ disentangles $\binom{\mathscr{T}_n}{k}$, then the restriction of any element $\mathcal{P} \in \binom{\mathscr{T}_n}{k}$ to $\mathcal{Y}$ acts as a fingerprint from which we can in some sense recognise $\mathcal{P}$.

It is clear that if $\mathcal{Y}$ contains all four-element subsets of $[n]$, then $\mathcal{Y}$ disentangles $\binom{\mathscr{T}_n}{1}$, for otherwise there are two binary trees with precisely the same set of quartets. On the other hand, all three-element subsets alone clearly does not suffice, since there is a single leaf-labelled tree that displays a given three-element leaf-set. This next lemma, however, shows that if $\mathcal{Y}$ does disentangle $\binom{\mathscr{T}_n}{1}$, then each three-element subset of $[n]$ is contained in some member of $\mathcal{Y}$.

**Lemma 6.2.1.** *For some $n \geq 4$, let $\mathcal{Y} \subseteq 2^{[n]}$. If $\mathcal{Y}$ disentangles $\binom{\mathscr{T}_n}{1}$, then for all $\{a, b, c\} \subset [n]$, there is some $Y \in \mathcal{Y}$ such that $\{a, b, c\} \subset Y$.*

*Proof.* Suppose that for some $\{a, b, c\} \subset [n]$, there is no $Y \in \mathcal{Y}$ that strictly contains $\{a, b, c\}$. Choose some $\mathcal{T}_a \in \mathscr{T}_n$ such that $\{b, c\}$ is a cherry of $\mathcal{T}_a$, and $\{a, b, c\}|[n] - \{a, b, c\}$ is a split of $\mathcal{T}_a$. Then let $\mathcal{T}_b$ be identical to $\mathcal{T}_a$ but

with the leaves $a$ and $b$ swapped, and let $\mathcal{T}_c$ also be identical to $\mathcal{T}_a$ but with the leaves $a$ and $c$ swapped. Figure 6.3 gives a general depiction of the three trees.

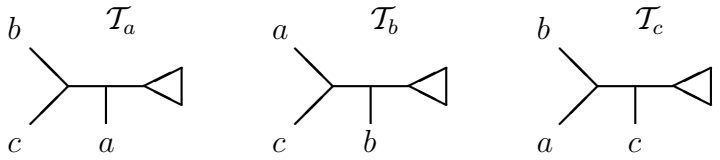


Figure 6.3: The trees $\mathcal{T}_a, \mathcal{T}_b, \mathcal{T}_c$ from the proof of Lemma 6.2.1.

Now, since there is no $Y \in \mathcal{Y}$ containing $\{a, b, c\}$, the restrictions of each of the three trees to $\mathcal{Y}$ are identical. Hence $\mathcal{Y}$ does not disentangle $\binom{\mathscr{T}_n}{1}$. $\square$

The condition stated in Lemma 6.2.1 is necessary but not sufficient for a set to disentangle $\binom{\mathscr{T}_n}{1}$. A counter-example is shown in Fig. 6.4. If we take $\mathcal{Y}$ to be all four-element subsets of $\{1, \ldots, 8\}$ that contain one of the pairs $\{1, 2\}$, $\{3, 4\}$, $\{5, 6\}$ and $\{7, 8\}$, then we can see that both trees have the same restriction to $\mathcal{Y}$, and yet all three-element subsets of $\{1, \ldots, 8\}$ are present in $\mathcal{Y}$.

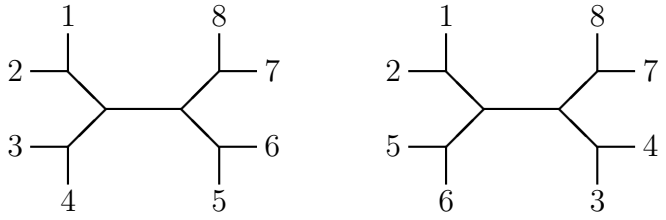


Figure 6.4: Two trees that provide a counter-example to the converse of Lemma 6.2.1.

From earlier, we know that $\binom{[n]}{5}$ does not disentangle $\binom{\mathscr{T}_n}{2}$, so the natural question to ask is whether all six-element subsets will, and more generally, for what $j$ does $\mathcal{Y} = \binom{[n]}{j}$ disentangle $\binom{\mathscr{T}_n}{k}$ for all $n \geq j$. Indeed, it is possibly not obvious that such a $j$ even exists. Theorem 18 in [33] answers the first of these questions in the affirmative. We concentrate now on the second question.

**Definition 6.2.2.** For $k > 0$, let $j \geq 4$ be the least positive integer for which $\binom{[n]}{j}$ disentangles $\binom{\mathscr{T}_n}{k}$ for all $n \geq j$. We call $j = D(k)$ the $k$-th *disentangling number*.

**Lemma 6.2.3.** *The function $D(k)$ is monotonic in $k$.*

*Proof.* For some $k > 0$, $n \geq 4$, let $\mathcal{Y} \subset 2^{[n]}$ be a set that does not disentangle $\binom{\mathscr{T}_n}{k}$. That is, there are distinct $k$-element subsets $\mathcal{P}, \mathcal{P}' \subset \mathscr{T}_n$ such that

$$\mathcal{P}|\mathcal{Y} = \mathcal{P}'|\mathcal{Y}.$$

Let $\mathcal{T}$ be some tree in $\mathscr{T}_n - (\mathcal{P} \cup \mathcal{P}')$. We may assume that such a tree exists by choosing $n$ to be large enough. Then

$$\begin{aligned}(\mathcal{P} \cup \mathcal{T})|\mathcal{Y} &= \mathcal{P}|\mathcal{Y} \cup \mathcal{T}|\mathcal{Y} \\ &= (\mathcal{P}' \cup \mathcal{T})|\mathcal{Y}.\end{aligned}$$

That is, $\mathcal{Y}$ does not disentangle $\binom{\mathscr{T}_n}{k+1}$ and hence $D(k) \leq D(k+1)$, completing the proof. □

The first and second disentangling numbers are four and six respectively, as set out previously. Since $D(k)$ is monotonic, as shown in Lemma 6.2.3, and $D(1) = 4$, the definition of the disentangling number remains consistent for all $k > 0$.

Though it may not be immediately obvious, we shall see that the function $D(k)$ is well-defined. That is, for all $k > 0$, some $j_0 > 0$ exists such that, for all $j_0 < j \leq n$, the set $\binom{[n]}{j}$ disentangles $\binom{\mathscr{T}_n}{k}$. Let us first consider the problem of, for some collection of trees, finding a reasonably small subset of the leaf-set on which none of the trees agree.

**Definition 6.2.4.** Let $\mathcal{P} \subseteq \mathscr{T}_n$ be a set of trees. We say that a non-empty set $Y \subseteq [n]$ *separates* $\mathcal{P}$ if

$$\mathcal{T}_i|Y \neq \mathcal{T}_j|Y$$

for all distinct $\mathcal{T}_i, \mathcal{T}_j \in \mathcal{P}$.

The specification that a separating set is non-empty ensures that the definition remains consistent when $\mathcal{P}$ contains only one tree.

**Definition 6.2.5.** For $k > 0$, let $j > 0$ be the smallest integer such that, for all $n \geq 4$, all $x \in [n]$, and all $\mathcal{P} \in \binom{\mathscr{T}_n}{k}$, there exists some $Y \in \binom{[n]}{j}$ containing $x$ that separates $\mathcal{P}$. We write $S(k) = j$.

It is clear from Definition 6.2.5 that $S(2) = 4$. This follows from two basic facts. Firstly, for all $k \geq 2$, we must have $S(k) > 3$, and secondly any two distinct trees must differ in at least one quartet.

**Lemma 6.2.6.** *The function $S(k)$ is monotonic in $k$.*

*Proof.* For some $k > 0$, $n \geq 4$, let $Y \subseteq [n]$ be a set that does not separate some $\mathcal{P} \subset \mathscr{T}_n$ of size $k$. Clearly, $Y$ will not separate any set containing $\mathcal{P}$ either, and so $S(k) \leq S(k+1)$. □

The above lemma is included purely for completeness. We prove next that $S(k)$ exists for all $k$, and use this fact to obtain the parallel result for the $k$-th disentangling number.

**Lemma 6.2.7.** *The function $S(k)$ is well-defined. Moreover,*

$$S(k) \leq 3k - 2.$$

*Proof.* The lemma is trivially true for $k = 1, 2$. Suppose now that $k > 2$, and let $\mathcal{T}_1, \ldots, \mathcal{T}_k$ be distinct trees from $\mathscr{T}_n$ for some $n$. For any $x \in [n]$, there is some $Y \subseteq [n]$ containing $x$ of size at most $3k-5$ that separates $\mathcal{T}_1, \ldots, \mathcal{T}_{k-1}$. If $\mathcal{T}_i|Y \neq \mathcal{T}_k$ for all $1 \leq i < k$, then $Y$ separates $\mathcal{T}_1, \ldots, \mathcal{T}_k$, and we are done.

Otherwise, $\mathcal{T}_i|Y = \mathcal{T}_k|Y$ for at most one $i < k$. Then there is a quartet $q \in \mathcal{Q}(\mathcal{T}_i) - \mathcal{Q}(T_k)$ that contains $x$. Hence $Y \cup \mathcal{L}(q)$ separates $\mathcal{T}_1, \ldots, \mathcal{T}_k$. Since $|Y| \leq 3k - 2$ the proof is complete. □

**Theorem 6.2.8.** *The function $D(k)$ is well-defined. Moreover,*

$$D(k) \leq S(k) + 2$$

*for all $k \geq 2$.*

*Proof.* For $k \geq 2$, let $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_k\}$ be $k$ distinct trees in $\mathscr{T}_n$ for some $n \geq S(k)+2$, and let $\mathcal{Y} = \binom{[n]}{S(k)+2}$. We wish to show that $\mathcal{P}$ can be uniquely reconstructed from $\mathcal{P}|\mathcal{Y}$.

If $n = S(k)+2$, then $\mathcal{P}|\mathcal{Y} = \mathcal{P}$, and so there is nothing to prove. We may therefore assume that $n > S(k)+2$. Suppose that $Z$ is a subset of $[n]$ of size $S(k)$ that separates $\mathcal{P}$, and let $\mathcal{Y}_Z$ be the set

$$\mathcal{Y}_Z = \{Y \in \mathcal{Y} : Z \subset Y\}.$$

We can partition $\mathcal{P}|\mathcal{Y}_Z$ into $k$ disjoint sets $\mathcal{S}_1, \ldots, \mathcal{S}_k$ so that, for all $i \in \{1, \ldots, k\}$ and for each pair $\mathcal{T}, \mathcal{T}'$ of distinct trees in $\mathcal{S}_i$,

$$\mathcal{T}|Z = \mathcal{T}'|Z.$$

Each of the sets $\mathcal{S}_i$ corresponds to some tree in $\mathcal{P}$, and so we may assume that $\mathcal{S}_i = \mathcal{T}_i|\mathcal{Y}_Z$ for all $i \in \{1, \ldots, k\}$.

It suffices now to show that for each three-element subset $W$ of $[n]-Z$, the tree $\mathcal{T}_i|(Z \cup W)$ is uniquely determined by members of $\mathcal{P}|\mathcal{Y}$. Let $W \in \binom{[n]-Z}{3}$, and choose some $w \in W$. Then there is some $z \in Z$ so that $Z' = (Z-z) \cup w$ separates $\mathcal{P}$. Let $\mathcal{Y}_{Z'}$ be the set

$$\mathcal{Y}_{Z'} = \{Y \in \mathcal{Y} : Z' \subset Y\}.$$

As before, $\mathcal{P}|\mathcal{Y}_{Z'}$ can be partitioned into $k$ disjoint sets $\mathcal{S}'_1, \ldots, \mathcal{S}'_k$. Moreover, as $\mathcal{Y}_Z$ and $\mathcal{Y}_{Z'}$ have a non-empty intersection, and both $Z$ and $Z'$ separate $\mathcal{P}$, each $\mathcal{S}_i$ has a non-empty intersection with exactly one $\mathcal{S}'_j$. Thus, by symmetry, we may assume without loss of generality that $\mathcal{S}_i \cap \mathcal{S}'_i$ is non-empty, and that

$$\mathcal{S}_i \cup \mathcal{S}'_i \subseteq \mathcal{T}_i|\mathcal{Y}.$$

For some $x \in Z \cap Z'$, we have

$$\begin{aligned}\mathcal{Q}(\mathcal{S}_i) \cup \mathcal{Q}(\mathcal{S}'_i) &\supseteq \{q \in \mathcal{Q}(\mathcal{T}_i|(Z \cup W)) : |\mathcal{L}(q) \cap Z| \geq 2 \text{ or } |\mathcal{L}(q) \cap Z'| \geq 2\} \\ &= \{q \in \mathcal{Q}(\mathcal{T}_i|(Z \cup W)) : x \in \mathcal{L}(q)\},\end{aligned}$$

which defines $\mathcal{T}_i|Z\cup W$. By using this argument over all three-element subsets $W$ of $[n]-Z$, we can reconstruct each of the $\mathcal{T}_i$ uniquely, completing the proof. □

This last theorem raises some interesting questions. Most notably, are either of the functions $D(k)$ or $S(k)$ bounded above by some fixed integer $N > 0$, or can they grow arbitrarily large. Secondly, is $D(k)$ always larger than $S(k)$? We have seen that this is so when $k = 2$ ($D(2) = 6$, whereas $S(2) = 4$). In fact, is it true perhaps that $D(k) = S(k) + 2$ for all $k \geq 2$?

This next lemma demonstrates quite easily that $S(k)$ is unbounded.

**Lemma 6.2.9.** *For any positive integer $m$, there is some $k_0 > 0$ such that*

$$S(k) > m$$

*for all $k > k_0$.*

*Proof.* Consider the set of trees $\mathscr{T}_m$, and let $k_0 = |\mathscr{T}_m|$. Let $\mathcal{P} \subseteq \mathscr{T}_n$ be a collection of $k$ distinct trees for some $k > k_0$, and let $Y \subseteq [n]$ separate $\mathcal{P}$. We may assume that $Y = [m]$. Since $[m]$ now separates $\mathcal{P}$, there must be at least $k$ distinct binary trees in $\mathscr{T}_m$. This contradiction finishes the proof. □

While Lemma 6.2.9 shows that $S(k)$ is in fact unbounded, the lower bound that we get by inverting

$$m \leq S((2m-5)!!)$$

grows extremely slowly. In the case of the disentangling number, we can construct an explicit example that proves an asymptotically better lower bound.

**Lemma 6.2.10.** *If $k$ is some positive integer, then*

$$D(2^{k-1}) \geq 3k.$$

*Proof.* Let $k$ be a positive integer, and let $\mathcal{T} \in \mathscr{T}_k$ be some binary leaf-labelled tree. Further, let $A = \alpha_1, \ldots, \alpha_k$ be a binary sequence with $\alpha_i \in \{0, 1\}$ for

all $i \in \{1, \ldots, k\}$. We construct the binary leaf-labelled tree $\mathcal{T}_A$ from $\mathcal{T}$ by replacing each leaf $i$ by three new leaves $a_i, b_i, c_i$ so that, for all $x \notin \{a_i, b_i, c_i\}$,

(i) if $\alpha_i = 0$, then $a_i b_i | c_i x \in \mathcal{Q}(\mathcal{T}_A)$; and

(ii) if $\alpha_i = 1$, then $x a_i | b_i c_i \in \mathcal{Q}(\mathcal{T}_A)$.

We specify the *weight* $w(A)$ of a sequence of zeroes and ones to be the number of ones. That is, for $A = \alpha_1, \ldots, \alpha_k$, where $\alpha_i \in \{0, 1\}$,

$$w(A) = \sum_{i=1}^{k} \alpha_i.$$

Let the sets of trees $\mathscr{T}^{\text{even}}$ and $\mathscr{T}^{\text{odd}}$ be defined by

$$\mathscr{T}^{\text{even}} = \{\mathcal{T}_A : w(A) \text{ is even}\},$$
$$\mathscr{T}^{\text{odd}} = \{\mathcal{T}_A : w(A) \text{ is odd}\},$$

where $A$ ranges over all binary sequences of length $k$. That is, both of $\mathscr{T}^{\text{even}}$ and $\mathscr{T}^{\text{odd}}$ contain exactly $2^{k-1}$ trees.

It remains to show now that taking $\mathcal{Y}$ to be all $3k-1$-element subsets of $X$ yields

$$\mathscr{T}^{\text{even}}|\mathcal{Y} = \mathscr{T}^{\text{odd}}|\mathcal{Y},$$

thus proving the lemma. It suffices to show that, for any $Y \in \mathcal{Y}$ and any $\mathcal{T}' \in \mathscr{T}^{\text{even}}$, there is some $\mathcal{T}'' \in \mathscr{T}^{\text{odd}}$ so that

$$\mathcal{T}'|Y = \mathcal{T}''|Y.$$

Due to symmetry, we may assume that $\mathcal{T}' = \mathcal{T}_{A'} \in \mathscr{T}^{\text{even}}$, where $A'$ consists entirely of zeroes, and that $Y = X - c_k$. If we let $A''$ have $\alpha_k = 1$ as the single non-zero entry, then the tree $\mathcal{T}'' = \mathcal{T}_{A''}$ satisfies our requirements, and the proof is complete. □

**Corollary 6.2.11.** *For all $k \geq 2$, there exists some $c > 0$ such that*

$$c \log k \leq D(k) \leq 3k.$$

*Proof.* The lower bound is a consequence of Lemma 6.2.10, while the upper bound follows by combining Lemma 6.2.7 and Theorem 6.2.8. □

## 6.3 Further Ideas

We conclude this chapter with two fairly general conjectures.

**Conjecture 6.3.1.** *The following statements are equivalent:*

*(i)* $\mathcal{Y}$ *disentangles* $\binom{\mathscr{T}_n}{k}$*;*

*(ii)* $\mathcal{Y}$ *disentangles* $\bigcup_{j\leq k}\binom{\mathscr{T}_n}{j}$*.*

The implication (ii)⇒(i) is trivial, but the other direction is not so obvious. A proof of Conjecture 6.3.1 would bring the results of this chapter in line with the work in Matsen *et al.* ([33]). They denote the set of all binary trees on $X$ by $B(X)$, and the subsets of $B(X)$ of size at most $k$ by $B(X,k)$. That is,

$$B(X,k) = \bigcup_{j\leq k}\binom{B(X)}{j}.$$

Theorem 18 from this paper states that $B(X,2)$ can be disentangled by the subsets of $X$ of size at most six, which is not quite the same as our assertion that $D(2) = 6$. If the previous conjecture were shown to be true, then this would make the two notions equivalent.

As a final comment, the following conjecture seems intuitive in some ways, but again neither a proof nor a counterexample has as yet been found.

**Conjecture 6.3.2.** *If* $\binom{[n]}{j}$ *disentangles some* $\mathscr{P}$*, and* $\mathcal{Y}$ *also disentangles* $\mathscr{P}$*, then*

$$\left\{Z \in \binom{Y}{j} : Y \in \mathcal{Y}\right\}$$

*disentangles* $\mathscr{P}$*.*

The essence of this conjecture is that we need only consider sets $\mathcal{Y}$ in which every member is of the same size; not only that, but every disentangling set $\mathcal{Y}$ for a family $\mathscr{P}$ can be reduced to a minimal disentangling set.

# Chapter 7

# Ramsey Theory and Leaf-Labelled Trees

## 7.1 Introduction

The essence of Ramsey theory ([37]) is that it is impossible to have complete disorder within a structure. That is, as we increase the size of some object of interest, the randomness of the object cannot prevent the appearance of certain highly ordered substructures.

The following simple but deep theorem which first appeared in [23] illustrates a Ramsey-type result, and will come in useful for proving some of the key results in this chapter.

**Theorem 7.1.1** (Erdős-Szekeres Theorem)**.** *If $A$ is a sequence of $n^2 + 1$ distinct integers, then $A$ contains a monotonic subsequence of length $n + 1$. Moreover, there is a sequence $A$ of $n^2$ distinct integers that contains no monotonic subsequence of length $n + 1$.*

As an example, let the sequence $A$ be some permutation of the first ten positive integers. Then the Erdős-Szekeres Theorem (Theorem 7.1.1) tells us that there are four elements of $A$ that occur either in strictly increasing order or in strictly decreasing order. However, if $A$ is instead a permutation of $\{1, \ldots, 9\}$, then we can choose $A$ so that the longest monotonic subsequence contains at most three elements. One such example would be $3, 2, 1, 6, 5, 4, 9, 8, 7$.

A key aspect of Ramsey theory is that while certain highly regular substructures may be shown to exist, the proofs are often non-constructive. As a

result, large numbers and fast growing functions are a commonly encountered phenomenon in this area of mathematics.

Given two or more leaf-labelled trees on overlapping leaf sets, a problem of interest is to find the size of the largest common subtree. We will instead approach this problem from a Ramsey theory perspective, with the goal being to show that a collection of trees on $X$ must have a common subtree on an arbitrarily-sized leaf set provided $X$ is chosen to be large enough. There is some overlap between this chapter and results from [28, 44].

## 7.2 Common Subtrees

We begin with the simplest non-trivial case. Suppose that we have two binary leaf-labelled trees on the same leaf-set of size $n$. If $n$ is large enough, can we guarantee that the trees have a common quartet? That is, is there some $n$ such that $\mathcal{Q}(\mathcal{T}_1) \cap \mathcal{Q}(\mathcal{T}_2)$ is non-empty for all pairs $\mathcal{T}_1, \mathcal{T}_2 \in \mathscr{T}_n$? And if so, how large does $n$ need to be?

It turns out that $n = 6$ is enough to ensure that two members of $\mathscr{T}_n$ share an induced quartet. To demonstrate this, set $X = [n]$ and let $Y_1 \subset X$ be a cherry of $\mathcal{T}_1$ and $Y_2 \subset X$ be a cherry of $\mathcal{T}_2$. If $Y_1 \cap Y_2$ is empty, then $Y_1|Y_2$ is a common quartet for the two trees. Hence if the trees do not share a quartet then they must each have exactly two cherries and therefore both are caterpillars. Now for $i \in \{1, 2\}$, there is some $Z_i \subset X$ of size three such that $Z_i|X - Z_i \in \Sigma(\mathcal{T}_i)$. Without loss of generality, $Z_1 \cap Z_2$ contains at least two elements, and so the non-trivial split $Z_1 \cap Z_2|X - (Z_1 \cup Z_2)$ is common to both trees, implying the existence of a common quartet.

Let us now define the notation that we will be using. The arrow notation is borrowed from mainstream Ramsey theory, although we have adapted it to fit the model we are working within.

**Definition 7.2.1.** For $k, m > 0$, we write

$$n \to (m)_k$$

if any set of $k$ trees $\mathcal{T}_1, \ldots, \mathcal{T}_k \in \mathscr{T}_n$ share a common $m$-leafed subtree.

The following lemma, which we include for the sake of completeness with-

out formal proof, highlights a trivial consequence of Definition 7.2.1.

**Lemma 7.2.2.** *Let $k' \leq k, m' \leq m$ and $n' \geq n$ be positive integers such that $n \to (m)_k$. Then*

$$n' \to (m')_{k'}.$$

The arrow notation gives a compact way of expressing specific Ramsey theoretic results. In general though, we are more interested in the behaviour of the function that, for each $k, m > 0$, outputs the minimal value of $n$ for which $n \to (m)_k$. From Lemma 7.2.2, we know that this behaviour is monotonic with respect both to $k$ and $m$.

**Definition 7.2.3.** Let $k, m > 0$ be integers. The function $\tau_k(m)$ denotes the smallest integer $n$ such that

$$n \to (m)_k.$$

Our earlier argument shows that $6 \to (4)_2$. It is easily verified (see Fig. 7.1), that there are two trees in $\mathscr{T}_5$ that do not have a common quartet, and so it follows that $\tau_2(4) = 6$.


Figure 7.1: Two trees in $\mathscr{T}_5$ that have no quartet in common.

**Theorem 7.2.4.** *The function $\tau_k(m)$ is well-defined, That is, for all integers $k, m > 0$, there exists some $N > 0$ such that for all $n \geq N$*

$$n \to (m)_k.$$

Note that in Definition 7.2.1, we have already restricted ourselves to binary trees. If we were to allow trees that have vertices of arbitrary degree

then $\tau_k(m)$ would no longer be well-defined. To illustrate this, we point out that for all $n$, the star tree on $[n]$ and a fully resolved tree in $\mathscr{T}_n$ have no four-leafed subtree in common. We do remark, however, that Theorem 7.2.4 may be upgraded to sets of leaf-labelled trees that have bounded degree.

In order to prove Theorem 7.2.4, we first consider the much simpler case where the trees in question are all caterpillars. To this end, we require some further definitions.

**Definition 7.2.5.** For $k, m > 0$, we write

$$n \underset{c}{\to} (m)_k$$

if any set of $k$ caterpillars $\mathcal{C}_1, \ldots, \mathcal{C}_k \in \mathscr{C}_n$ share a common $m$-leafed subtree.

**Definition 7.2.6.** Let $k, m > 0$ be integers. The function $\kappa_k(m)$ denotes the smallest integer $n$ such that

$$n \underset{c}{\to} (m)_k.$$

**Lemma 7.2.7.** *$\kappa_k(m) \leq \tau_k(m)$ for all $k, m > 0$.*

The above lemma is an immediate consequence of the fact that $\mathscr{C}_n \subseteq \mathscr{T}_n$ for all positive $n$. The proof is straightforward and the details are therefore omitted.

**Theorem 7.2.8.** *The function $\kappa_k(m)$ is well-defined, That is, for all integers $k, m > 0$, there exists some $N > 0$ such that for all $n \geq N$*

$$n \underset{c}{\to} (m)_k.$$

*Proof.* We begin with the case $k = 2$. Fix some $m > 0$ and let $n \geq (m-1)^2 + 1$. We may assume that $\mathcal{C}_1 \in \mathscr{C}_n$ has the canonical labelling $[1, \ldots, n]$. Since the labelling of $\mathcal{C}_2 \in \mathscr{C}_n$ is some permutation of $[n]$, it suffices to show that any such permutation has a monotonic subsequence of length $m$. This is guaranteed by the Erdős-Szekeres Theorem (Theorem 7.1.1), and so

$$(m-1)^2 + 1 \underset{c}{\to} (m)_2.$$

Now suppose that $k > 2$ and again fix some $m > 0$. Then there is some $n > 0$ such that

$$n \underset{c}{\to} (\kappa_2(m))_{k-1},$$

from which it follows that $n \underset{c}{\to} (m)_k$, completing the proof. □

**Lemma 7.2.9.** *For all $l > 0$ there exists some $n > 0$ such that any tree $\mathcal{T} \in \mathscr{T}_n$ has a subtree with $l$ leaves that is a caterpillar.*

*Proof of Theorem 7.2.4.* By Lemma 7.2.9, for any $l > 0$ we can find some $n > 0$ such that, for any $\mathcal{T}_1, \ldots, \mathcal{T}_k \in \mathscr{T}_n$, there is a leaf set $Y$ of size $l$ so that the restriction of $\mathcal{T}_i$ to $Y$ is a caterpillar for all $i \in \{1, \ldots, k\}$. If we take $l = \kappa_k(m)$ then the result follows by Theorem 7.2.8. □

As mentioned previously, the non-constructive nature of many Ramsey theoretic proofs can lead to bounds that are extremely fast growing functions. Embedded in the proofs of Theorem 7.2.4 and the preceeding results leading up to it, we have the following corollary. Note that $\mathbb{Z}^+$ denotes the set of strictly positive integers.

**Corollary 7.2.10.** *Define the functions $f, g : \mathbb{Z}^+ \to \mathbb{Z}^+$ by*

$$f(x) = \begin{cases} 3 \cdot 2^{\frac{x-4}{2}} + 1 & \text{if } x \text{ is even,} \\ 2^{\frac{x-1}{2}} + 1 & \text{if } x \text{ is odd,} \end{cases}$$

*and*

$$g(x) = x^2 - 2x + 2.$$

*For all $k \geq 2$ and all $m \geq 4$,*

$$\kappa_k(m) \leq g^{k-1}(m), \tag{7.1}$$

*and*

$$\tau_k(m) \leq f^k \circ g^{k-1}(m). \tag{7.2}$$

*Proof.* Suppose we have a binary leaf-labelled tree $\mathcal{T}$ with $f(m)$ leaves. Then there is a subtree of $\mathcal{T}$ on at least $m$ leaves which is a caterpillar (see for example [44, Lemma 3.3]). Suppose instead that $\mathcal{C}_1, \mathcal{C}_2 \in \mathscr{C}_n$, where $n \geq g(m)$. Then $\mathcal{C}_1$ and $\mathcal{C}_2$ share a common subtree on at least $m$ leaves. That is,

$$\kappa_2(m) \leq g(m)$$

by the Erdős-Szekeres Theorem (Theorem 7.1.1), and

$$\tau_2(m) = f^2 \circ g(m).$$

Let $\mathcal{C}_1, \ldots, \mathcal{C}_k \in \mathscr{C}_n$ be a collection of caterpillars, where $n = g^{k-1}(m)$ leaves. Since $\kappa_2(m) \leq g(m)$, we can find a common caterpillar for $\mathcal{C}_{k-1}, \mathcal{C}_k$ that has at least $g^{k-2}(m)$ leaves. Let $\mathcal{C}_i'$ be the restriction of $\mathcal{C}_i$ to the leaf-set of this common caterpillar. Then there is some permutation of the leaves so that $\mathcal{C}_1', \ldots, \mathcal{C}_{k-1}' \in \mathscr{C}_{n'}$, where $n' \geq g^{k-2}(m)$, and so (7.1) holds by induction.

Let $\mathcal{T}_1, \ldots, \mathcal{T}_k \in \mathscr{T}_n$ be a collection of trees, where $n = f^k \circ g^{k-1}(m)$ leaves. There is some subtree of $\mathcal{T}_1$ that is a caterpillar and has $f^{k-1} \circ g^{k-1}(m)$ leaves. Let $\mathcal{T}_i'$ be the restriction of $\mathcal{T}_i$ to the leaf-set of this caterpillar. We can permute the leaves so that $\mathcal{T}_1', \ldots, \mathcal{T}_k' \in \mathscr{T}_{n'}$, where $n' = f^{k-1} \circ g^{k-1}(m)$. By iterating this over all $i \in \{1, \ldots, k\}$, there is some leaf-set $Y$ of size $g^{k-1}(m)$ such that the restriction $\mathcal{T}_i'' = \mathcal{T}_i | Y$ is a caterpillar for all $i \in \{1, \ldots, k\}$. Using (7.1), we can now verify that (7.2) holds. □

Calculating upper bounds on $\tau_2(m)$ using Corollary 7.2.10 gives $\tau_2(4) \leq 2049$ and $\tau_2(5) \leq 9223372036854775809$. Since we showed earlier that in fact $\tau_2(4) = 6$, there is clearly much room for improvement and to this end we conclude this section with a conjecture.

**Conjecture 7.2.11.** *$\tau_k(m) = \kappa_k(m)$ for all $k, m > 0$.*

## 7.3 Numerical Bounds

We turn our attention now to bounding the size of $\kappa_2(m)$, which as we have already shown is well-defined and grows at most quadratically. An idea

that will come in useful through this section is that of *pattern avoidance*. Roughly speaking, pattern avoidance is concerned with sequences that have no subsequence isomorphic in some sense to another given sequence. Let us formalise this.

**Definition 7.3.1.** For some $n > 0$, let $p = p_1, \ldots, p_n$ and $q = q_1, \ldots, q_n$ be sequences of distinct positive integers. We say that $p$ and $q$ have the same *pattern* if $p_i < p_j$ implies $q_i < q_j$ for all $i \neq j$.

**Definition 7.3.2.** For some $n > m > 0$, let $p = p_1, \ldots, p_n$ and $q = q_1, \ldots, q_m$ be sequences of distinct positive integers. We say that $p$ *avoids* $q$ if no subsequence of $p$ has the same pattern as $q$.

As a simple example, if $p$ avoids the pattern $1, 2$, then $p$ is necessarily a monotonically decreasing sequence. Slightly more involved, if $p$ avoids both of the patterns $1, \ldots, n$ and $n, \ldots, 1$, then the length of $p$ is at most $(n-1)^2$. This second example is essentially a restatement of the Erdős-Szekeres Theorem (Theorem 7.1.1).

The problem of finding how large two caterpillars may be so that don't share a common subtree of a given size can also be rephrased in terms of pattern avoidance. Let $T(m)$ be the set of all sequences $t = t_1, \ldots, t_m$ that are permutations of $[m]$ and satisfy either

(i) $t_1, t_2 < t_3 < \cdots < t_{m-2} < t_{m-1}, t_m$; or

(i) $t_1, t_2 > t_3 > \cdots > t_{m-2} > t_{m-1}, t_m$.

Thus the set $T(m)$ is the set of label orderings of the caterpillars that are isomorphic to the caterpillar labelled $[1, \ldots, n]$. The proof of this next lemma follows immediately from definitions that have already been given, and as such we omit it.

**Lemma 7.3.3.** *Let $n > m \geq 4$. There is some permutation $S$ of $[n]$ that avoids every pattern in $T(m)$ if and only if there are two caterpillars $\mathcal{C}_1, \mathcal{C}_2 \in \mathscr{C}_n$ that share no common $m$-leafed subtree.*

With this in mind, we may concentrate on permutations of $[n]$ for some $n$ and disregard the underlying tree structure. For two sequences $p = p_1, \ldots, p_k$

and $q = q_1, \ldots, q_l$, the *concatenation* $pq$ of $p$ and $q$ is the sequence

$$pq = p_1, \ldots, p_k, q_1, \ldots, q_l,$$

and the *reverse* $r(p)$ of $p$ is the sequence

$$r(p) = p_k, \ldots, p_1.$$

We denote by $c(i, j, k)$ the sequence that has $i$ as the first element, and that is the concatenation of $k$ increasing sequences of $j$ consecutive integers, with the increasing sequences placed in decreasing order. Thus

$$\begin{aligned} c(i,j,k) =& i, i+1, \ldots, i+j-1, \\ & i-j, i-j+1, \ldots, i-1, \ldots, \\ & i-(k-1)j, i-(k-1)j+1, \ldots, i-(k-2)j-1. \end{aligned}$$

A simple example of this is $c(5, 2, 3) = 5, 6, 3, 4, 1, 2$. Slightly more complicated, the central block of the pattern in Fig. 7.2 corresponds to the reverse of $c(i, m-1, m-5)$ for some $i$.

Using this notation, for all $m \geq 5$ we define $S(m)$ to be the sequence

$$\begin{aligned} S(m) =& 2m-4, c(2m-6, 2, m-3), 1, \\ & r(c(m^2-5m+4, m-1, m-5)), \\ & m^2-2m-3, c(m^2-2m-5, 2, m-3), m^2-4m+3. \end{aligned}$$

Now, $S(m)$ is a permutation of $[m^2-2m-3]$, and a depiction of this sequence for $m \geq 5$ is shown in Fig. 7.2. Using this illustration as an aid, we prove the following lemma.

**Lemma 7.3.4.** *For all $m \geq 5$, the permutation $S(m)$ of $[m^2-2m-3]$ avoids every pattern in $T(m)$.*

*Proof.* Fix some $m \geq 5$. We remind the reader that the general pattern of $S(m)$ is shown in Fig. 7.2. Suppose that $S(m)$ does not avoid $t = t_1, \ldots, t_m$ for some $t \in T(m)$, and let $S' = s'_1, \ldots, s'_m$ be a subsequence of $S(m)$ that has the same pattern as $t$. Then $s'_2, \ldots, s'_{m-1}$ is monotonic, and must either

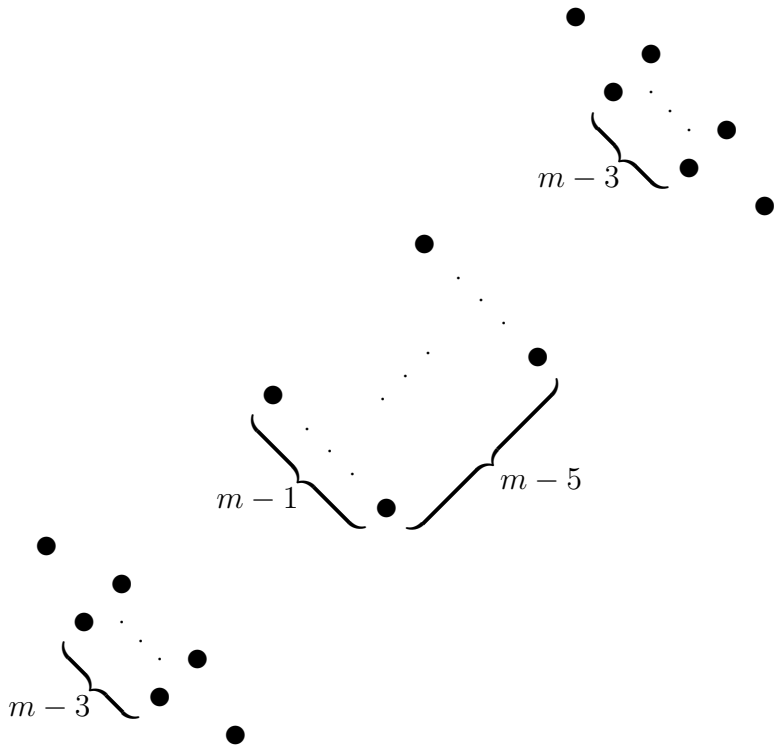


Figure 7.2: The pattern generated by $S$ for $m \geq 5$.

be contained in one of the decreasing segments of $S(m)$, or contain at most one element from each of the decreasing segments. However, in neither case can we extend $s'_2, \ldots, s'_{m-1}$ to some subsequence of $S(m)$ that has the same pattern as any member of $T(m)$, contradicting our assumption. □

Applying Lemma 7.3.3 yields a lower bound on $\kappa_2(m)$.

**Theorem 7.3.5.** *$\kappa_2(m) > m^2 - 2m - 3$ for all $m \geq 4$.*

*Proof.* For $m = 4$, the result follows from $\kappa_2(4) = 6$, while for $m \geq 5$ we combine Lemmas 7.3.3 and 7.3.4. □

We remark here that, combining Corollary 7.2.10 and Theorem 7.3.5 gives

$$m^2 - 2m - 3 < \kappa_2(m) \leq m^2 - 2m + 2,$$

with the upper bound being precisely the sharp bound found in the Erdős-Szekeres Theorem (Theorem 7.1.1). This disproves Conjecture 1 in [28],

which may be phrased in our notation as

$$\kappa_2(m) \leq m^2 - 4m + 6.$$

Intuitively we would expect the real value of $\kappa_2(m)$ to be strictly smaller than $m^2 - 2m + 1$, given we have slightly stronger constraints on patterns to avoid when dealing with trees as opposed to sequences. This is certainly true for $m = 4$.

We continue by finding an exact value for $\kappa_2(5)$.

**Theorem 7.3.6.** $\kappa_2(5) = 13$.

*Proof.* Let $\mathcal{C}_1$ be the caterpillar labelled $[1, \ldots, n]$. Consider some permutation $\sigma$ of $[n]$, and suppose that this is the label ordering for some caterpillar $\mathcal{C}_2$. For each $x \in [n]$ we define the following quantities:

$$\begin{aligned} l_a(x) &= |\{y < x : \sigma^{-1}(y) < \sigma^{-1}(x)\}| \\ l_b(x) &= |\{y > x : \sigma^{-1}(y) < \sigma^{-1}(x)\}| \\ r_a(x) &= |\{y < x : \sigma^{-1}(y) > \sigma^{-1}(x)\}| \\ r_b(x) &= |\{y > x : \sigma^{-1}(y) > \sigma^{-1}(x)\}| \end{aligned}$$

The first of these, $l_a(x)$, represents the number of elements in $[n]$ that are smaller than $x$ and also appear before $x$ in $\sigma$. The remaining quantities may be described similarly. We can find some simple but useful relationships between these:

$$\begin{aligned} l_b(x) &= \sigma^{-1}(x) - l_a(x) - 1 \\ l_b(x) &= n - x - r_b(x) \\ r_a(x) &= x - l_a(x) - 1 \\ r_a(x) &= n - \sigma^{-1}(x) - r_b(x) \end{aligned}$$

Assume that $\mathcal{C}_1, \mathcal{C}_2$ do not share a five-leafed subtree, and suppose that for some $x \in [n]$, we have $l_a(x) \geq 2$ and $r_b(x) \geq 2$. Then we can find a

common five-leafed subtree for $\mathcal{C}_1, \mathcal{C}_2$ having $x$ as the leaf not appearing in a cherry. Hence one of $l_a(x), r_b(x)$ is at most one. Similarly, one of $l_b(x)$ and $r_a(x)$ is at most one.

Now suppose that $l_a(x) \leq 1$. Then since $\min(l_b(x), r_a(x)) \leq 1$, we have either $x \leq 3$ or $\sigma^{-1}(x) \leq 3$. There are at most six choices for $x$ that will satisfy one of these conditions, and hence at most six different $x$ for which $l_a(x) \leq 1$.

Suppose instead that $r_b(x) \leq 1$. Then using the same reasoning as above we have either $x \geq n-2$ or $\sigma^{-1}(x) \geq n-2$. Again, there are at most six $x$ that can satisfy this. That is, $13 \underset{c}{\rightarrow} (5)$.

To complete the proof, it suffices to give a permutation of $[12]$ so that $\mathcal{C}_1$ and $\mathcal{C}_2$ don't share a five-leafed subtree. An example of such a permutation is $[6, 4, 5, 2, 3, 1, 12, 10, 11, 8, 9, 7]$, and hence $\kappa_2(5) = 13$. □

This result can be used to start the induction for a more general construction.

**Lemma 7.3.7.** *$\kappa_2(m) \leq m^2 - m - 7$ for all $m \geq 5$.*

*Proof.* Let $S = s_1, \ldots, s_n$ be a permutation of $[n]$ where $n > 4$. For all $i \in [n]$, let $\alpha_i$ be the length of the longest monotonically increasing subsequence $a_1, \ldots, a_r$ of $S$ such that

(i) $a_1 = s_i$;

(ii) there exist $j, k < i$ such that $s_j, s_k < s_i$; and

(iii) if $a_r = s_l$, then there exist $j, k > l$ such that $s_j, s_k > s_l$.

We define $\beta_i$ similarly for decreasing subsequences. That is, for $i \in [n]$, we set $\beta_i$ to be the length of the longest monotonically decreasing subsequence $b_1, \ldots, b_r$ of $S$ such that

(i) $b_1 = s_i$;

(ii) there exist $j, k < i$ such that $s_j, s_k > s_i$; and

(iii) if $b_r = s_l$, then there exist $j, k > l$ such that $s_j, s_k < s_l$.

Suppose that, for some $i < j$, we have $\alpha_i = \alpha_j > 0$. Then $s_i > s_j$, for otherwise $\alpha_i \geq \alpha_j + 1$. That is, for all $c > 0$ the subsequence $S_c^\alpha$ of $S$ consisting of all elements $s_i$ such that $\alpha_i = c$ is monotonically decreasing. Similarly, if the subsequence $S_c^\beta$ of $S$ included only those elements $s_i$ of $S$ such that $\beta_i = c$, then $S_c^\beta$ is monotonically increasing.

Now, choose some $m \geq 6$ and assume that the theorem holds for all smaller values of $m$. Suppose that $\mathcal{C}_1, \mathcal{C}_2 \in \mathscr{C}_n$, where $n = m^2 - m - 7$, are two caterpillars that share no $m$-leafed subtree. We can assume that $\mathcal{C}_1$ has the canonical labelling $[1, \ldots, n]$ and that $\mathcal{C}_2$ has the labelling $[s_1, \ldots, s_n]$. For all $i \in [n]$, we have $\alpha_i, \beta_i < m - 4$, for otherwise there is a common subtree with $m$ leaves.

Let $I \subseteq [n]$ be a set of size $\kappa_2(m-1)$. Then for some $i \in I$, either $\alpha_i$ or $\beta_i$ is at least $m-5$. That is, there are at least $n - \kappa_2(m-1) + 1 \geq 2m-1$ distinct $i \in [n]$ for which either $\alpha_i = m - 5$ or $\beta_i = m - 5$. So by the pigeonhole principle, one of the subsequences $S_{m-5}^\alpha, S_{m-5}^\beta$ has length $m$. However, this means that we have a monotonic sequence of length $m$ in $S$, and hence $\mathcal{C}_1, \mathcal{C}_2$ share a common subtree with $m$ leaves. □

This last result is an improvement on Corollary 7.2.10 only for $m < 9$, while for $m > 9$ the corollary remains the best known bound. The table below provides a summary of the results from this section.

| $m$ | $\kappa_2(m)$ |
|---|---|
| 4 | 6 |
| 5 | 13 |
| 6 | $[22, 23]$ |
| 7 | $[33, 35]$ |
| 8 | $[46, 49]$ |
| $\geq 9$ | $[m^2 - 2m - 2, m^2 - 2m + 2]$ |

Table 7.1: Known bounds on $\kappa_2(m)$ for $m \geq 4$.

# PART III

# TREE REARRANGEMENT OPERATIONS

In the late sixties, Robinson introduced *nearest neighbour interchange* (NNI) as a measure for comparing two leaf-labelled trees that share the same leaf-set ([39]). The underlying notion of NNI is that one tree is transformed into the other through a sequence of edge deletions and insertions. That is, we cut a tree into two pieces by deleting an edge, and then put these pieces back together by inserting a new edge.

The NNI operation is the precursor to a number of other tree rearrangement operations. Two that we study in particular in this thesis are *subtree prune and regraft* (SPR) and *tree bisection and reconnection* (TBR), with the latter being the primary focus of our results. Both SPR and TBR are generalisations of the basic NNI move we described above. Formal definitions of all three operations are given in Chapter 8.

Each of these rearrangement operations induces a metric on the space of binary leaf-labelled trees with $n$ leaves. That is, there is a distance defined between any pair of trees in $\mathscr{T}_n$ in each of these metrics. A natural question to consider is why this distance may be taken as a measure of similarity between the trees in question. The nature of the operations means that any single operation preserves most of the structural information in the original tree.

Rearrangments of rooted trees are not studied in this thesis, although we make brief mention of them here as motivation for studying tree rear-

rangment in general. The rooted version of SPR is often used to model the effects of recombination within an evolutionary system, by which we mean any hereditary process that passes information from one type to another other than through purely tree-like evolution ([3]). Examples of this may be found in linguistics with the genesis of creole languages and in stemmatology when a scribe copies from two or more manuscripts simultaneously.

The usefulness of tree rearrangement can also be seen in algorithmic applications. Quantitative measures of how well a specific tree fits a set of data are commonly used as optimisation criteria for selecting the most likely tree to underlie an evolutionary system. Given the size of the tree space for any practical application, it is unrealistic to calculate how well every single tree models a given data set and to then choose an optimal tree. Using the assumption that the collection of the most optimal trees share a degree of similarity, we can implement a search starting at some tree in the space, and then iterate by choosing the optimal tree that lies within one operation of the original tree, thus dramatically reducing the search space at any one iteration. In the light of this method, there are two factors which affect the efficiency of the algorithm. Firstly, how many trees are within a single operation of a given tree, and secondly, how far apart can two trees be. The first of these questions has been fully answered for both the NNI and SPR metrics, while upper and lower bounds are known for the maximum distance between a pair of trees under each of the three metrics.

The key results in this part of the thesis are all improvements on previous authors' work. In Chapter 8, we find an exact expression for the size of the TBR unit neighbourhood of a tree, and at the same time reprove the known analogue for SPR. We then continue by characterising the trees that respectively maximise and minimise the size of this neighbourhood. This work was carried out in collaboration with Taoyang Wu. Chapter 9 was researched jointly with Stefan Grünewald, and is concerned with finding the maximum distance between a pair of trees in both the SPR and the TBR metrics. While an exact result is not achieved, we improve on both the current best known upper and lower bounds.

# Chapter 8

# The TBR Unit Neighbourhood

## 8.1 Introduction

As we have already outlined, tree rearrangement operations may be of use in heuristic algorithms for finding a tree that optimally explains a given data set. The problem of quantifying the agreement between a tree and a data set is not addressed here. For the purposes of this discussion, we will instead impose a hypothetical measure $\mu$ on the implied tree space, so that if $\mu(\mathcal{T}) > \mu(\mathcal{T}')$ then it is understood that $\mathcal{T}$ is a more optimal tree than $\mathcal{T}'$.

The basic method is to choose, either randomly or intelligently, a tree $\mathcal{T}$ to serve as the initial input for the algorithm, and to calculate $\mu(\mathcal{T})$. The measure $\mu(\mathcal{T}')$ is then calculated for all trees $\mathcal{T}'$ that are exactly one rearrangement operation from $\mathcal{T}$. If $\mu(\mathcal{T}) \geq \mu(\mathcal{T}')$ for all such trees, then the algorithm outputs $\mathcal{T}$. Otherwise, the tree with the highest known measure is fed back in to the algorithm, and the same procedure is followed until a tree is returned. This guarantees to output a tree that locally maximises $\mu$ within the metric induced by the chosen rearrangement operation.

We now give formal definitions for each of the tree rearrangement operations of interest, namely NNI (nearest neighbour interchange), SPR (subtree prune and regraft) and TBR (tree bisection and reconnnection). Although NNI was the point of departure for the study of these operations, we begin by defining TBR, being the most general of the three.

A TBR operation on a binary leaf-labelled tree $\mathcal{T}$ involves deleting some edge $e$ from $\mathcal{T}$ (the *bisection*), and subsequently inserting a new edge $f$ so that the resultant tree $\mathcal{T}'$ is distinct from $\mathcal{T}$ (the *reconnection*). Since we require $\mathcal{T}'$ to be binary, it is necessary to subdivide an edge in one (in the case

that the other component is an isolated labelled vertex) or both components created in the bisection stage before inserting the new edge. An example is given in Fig. 8.1. We can transform $\mathcal{T}_1$ into $\mathcal{T}_2$ by first deleting the edge $e$

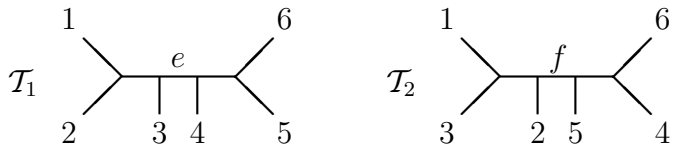


Figure 8.1: Two trees $\mathcal{T}_1, \mathcal{T}_2 \in \mathscr{T}_6$ that are one TBR operation apart.

from $\mathcal{T}_1$, and then adding the new edge $f$. To check that there has been no other change to the tree's structure, note that deleting $e$ from $\mathcal{T}_1$ gives the same forest as deleting $f$ from $\mathcal{T}_2$.

For a binary tree $\mathcal{T}$, we define the set $\mathcal{O}_{\text{TBR}}(\mathcal{T})$ to be all possible TBR operations $\theta$ that can be applied to the tree $\mathcal{T}$. An important point to note here is that for distinct $\theta_1, \theta_2 \in \mathcal{O}_{\text{TBR}}$, we may have $\theta_1(\mathcal{T}) = \theta_2(\mathcal{T})$. The reason for this is that an operation $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ is not specified solely by the output tree $\theta(\mathcal{T})$, but also by the edge $e$ that is deleted from $\mathcal{T}$ in the bisection stage of $\theta$.

Observe that for any two distinct trees $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$, there is a TBR operation $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ for which $\theta(\mathcal{T}) = \mathcal{T}'$ if and only if there is some split $X_1|X_2 \in \Sigma(\mathcal{T}) \cap \Sigma(\mathcal{T}')$ such that $\mathcal{T}|X_i = \mathcal{T}'|X_i$ for all $i \in \{1, 2\}$. To demonstrate this, if the edges $e$ and $f$ have respectively been deleted and inserted in the TBR operation that changes $\mathcal{T}$ into $\mathcal{T}'$, then the forest obtained by deleting $e$ from $\mathcal{T}$ must be identical to the forest obtained by deleting $f$ from $\mathcal{T}'$. This provides not only the common bipartition of the leaf set, but also the common subtrees induced by each part of this bipartition[1].

SPR is a special case of TBR in which there is less freedom at the reconnection stage. Let $\mathcal{T}$ be a binary tree, and let $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ be a TBR operation on $\mathcal{T}$ in which the edge $e$ is deleted, and let $X_1|X_2$ be the split of $\mathcal{T}$ induced by $e$. Then $\theta$ is an SPR operation for $\mathcal{T}$ if and only if, without

[1]The natural extension of this idea to arbitrary partitions is called an *agreement forest*. These are discussed in Chapter 9.

loss of generality, $\mathcal{T}|X_2 \cup x_1 = \theta(\mathcal{T})|X_2 \cup x_1$ for some $x_1 \in X_1$. Moreover, if this holds then in fact the same property holds for all $x_1 \in X_1$.

The significance of this condition is that one of the components formed in the bisection of $\mathcal{T}$, in this case $\mathcal{T}|X_2$, is treated as a rooted subtree, and is then regrafted so that this rooting is preserved with respect to the other component. We say that we have pruned $\mathcal{T}|X_2$ from $\mathcal{T}$, and regrafted it to form $\mathcal{T}'$.

The previous example (refer to Fig. 8.1) does not represent an SPR operation, since neither component obtained by deleting $e$ from $\mathcal{T}_1$ can be regrafted to the other to form $\mathcal{T}_2$. By making a subtle change, in particular by exchang-

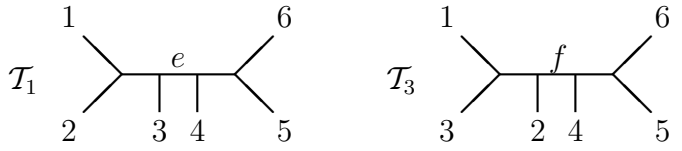


Figure 8.2: Two trees $\mathcal{T}_1, \mathcal{T}_3 \in \mathscr{T}_6$ that are one SPR operation apart.

ing the labels 4 and 5 on $\mathcal{T}_2$, we get a tree $\mathcal{T}_3$ that can be obtained from $\mathcal{T}_1$ by a single SPR operation. This example is depicted in Fig. 8.2.

NNI operations are TBR operations in which the reconnection is still more restricted than for SPR. Let $\mathcal{T}$ be a leaf-labelled tree, and let $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ be an SPR operation in which $\mathcal{T}|Y$ is pruned from $\mathcal{T}$ and regrafted to form $\mathcal{T}' = \theta(\mathcal{T})$. We say that $\theta$ is an NNI operation if and only if there is some cluster $Z \neq Y$ of $\mathcal{T}$ such that we can form $\mathcal{T}'$ from $\mathcal{T}$ by swapping the subtrees $\mathcal{T}|Y$ and $\mathcal{T}|Z$. In this case, $\mathcal{T}|Y$ and $\mathcal{T}|Z$ can be seen as adjacent in some sense, as shown by the schematic diagram in Fig. 8.3. Although

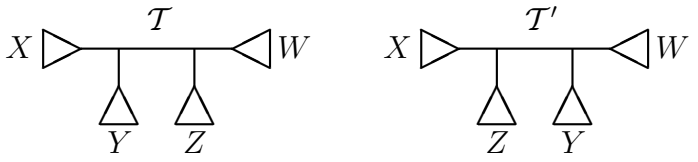


Figure 8.3: A tree illustrating the simplest case of Lemma 8.3.3.

it may not be immediately obvious, the example in Fig. 8.2 shows an NNI operation in which the leaves 2 and 3 have been swapped. Alternatively, the same outcome is reached by interchanging the subtrees labelled by $\{1\}$ and $\{4, 5, 6\}$ respectively. The possibility that two distinct operations can result in the same tree lies behind the main lemma (Lemma 8.2.1) in Section 8.2.

Extending our earlier notation for TBR to both SPR and NNI, we have

$$\mathcal{O}_{\text{NNI}}(\mathcal{T}) \subseteq \mathcal{O}_{\text{SPR}}(\mathcal{T}) \subseteq \mathcal{O}_{\text{TBR}}(\mathcal{T})$$

for any tree $\mathcal{T}$. For each $\vartheta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$, the $\vartheta$ *unit neighbourhood* of $\mathcal{T}$ is the set

$$N_\vartheta(\mathcal{T}) = \{\theta(\mathcal{T}) : \theta \in \mathcal{O}_\vartheta(\mathcal{T})\}.$$

That is, $N_\vartheta(\mathcal{T})$ is the set of all trees that are precisely one $\vartheta$ rearrangement operation from $\mathcal{T}$. Clearly, the elements in these neighbourhoods are dependent on the operation in question, and we have the corresponding nesting property as above. More explicitly,

$$N_{\text{NNI}}(\mathcal{T}) \subseteq N_{\text{SPR}}(\mathcal{T}) \subseteq N_{\text{TBR}}(\mathcal{T}).$$

Our interest in this chapter is in the size of these neighbourhoods, primarily the size of the TBR unit neighbourhood. For a tree $\mathcal{T} \in \mathscr{T}_n$, with $n \geq 4$, Robinson showed in [39] that the NNI unit neighbourhood has size exactly equal to $2n - 6$, while Allen and Steel ([2]) proved that

$$|N_{\text{SPR}}(\mathcal{T})| = 2(n-3)(2n-7).$$

It was also demonstrated in [2] that the size of the TBR unit neighbourhood is dependent on the shape of $\mathcal{T}$. More recently, the bounds

$$cn^2 \log n + O(n^2) \leq |N_{\text{TBR}}(\mathcal{T})| \leq \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2$$

were shown to hold for all $n \geq 4$, with the upper bound being met with equality if and only if $\mathcal{T}$ is a caterpillar ([27], see Appendix C).

The rest of this chapter is divided into two sections. In the first of these (Section 8.2), we relate the sizes of $\mathcal{O}_\vartheta(\mathcal{T})$ and $N_\vartheta(\mathcal{T})$ for SPR and TBR, and then use this to both reprove Allen and Steel's ([2]) result for the SPR neighbourhood and to obtain an expression for the TBR neighbourhood dependent on the tree shape. In Section 8.3, we characterise the trees that respectively maximise and minimise the size of $N_{\text{TBR}}(\mathcal{T})$ for all binary tree spaces $\mathscr{T}_n$. This is then extended to reprove the tight upper bound given in [27], and to further prove a tight asymptotic lower bound.

## 8.2 Neighbourhood Sizes

The approach used by Allen and Steel in [2] to determine both the size of the SPR unit neighbourhood and the upper bound on the size of the TBR unit neighbourhood was to count directly the number of trees that can be obtained from $\mathcal{T}$ via a single operation. While this seems the most natural approach, there is a fundamental barrier to performing this enumeration that we alluded to briefly in the introduction for this chapter. This is the fact that some operations in $\mathcal{O}_{\text{TBR}}(\mathcal{T})$ may be redundant. That is, there may be distinct elements $\theta_1, \theta_2 \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ for which

$$\theta_1(\mathcal{T}) = \theta_2(\mathcal{T}).$$

This potentially leads to counting some trees more than once. If we can determine precisely which operations in $\mathcal{O}_{\text{TBR}}(\mathcal{T})$ output the same tree, then we can relate the size of the TBR unit neighbourhood to the number of legitimate operations on $\mathcal{T}$.

It transpires, as the next lemma shows, that the only redundant TBR operations are all NNI operations.

**Lemma 8.2.1.** *Let $\theta, \theta' \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ be distinct* TBR *operations. If $\theta(\mathcal{T}) = \theta'(\mathcal{T})$, then $\theta \in \mathcal{O}_{\text{NNI}}(\mathcal{T})$.*

*Proof.* Suppose that $A|B$ is the split of $\mathcal{T}$ induced by $\theta$, and that $A'|B'$ is the split induced by $\theta'$. We may assume that $A \subset A'$ and $B' \subset B$. Since $\mathcal{T}|A' = \theta(\mathcal{T})|A'$, we have immediately that $\theta \in \mathcal{O}_{\text{SPR}}(\mathcal{T})$. Let $A_0 = A, A_1, \ldots, A_k = A'$ be clusters of $\mathcal{T}$ such that

(i) $A_i|B'$ is a partial split of $\mathcal{T}$; and

(ii) $A_{i+1}$ is a minimal cluster of $\mathcal{T}$ that contains $A_i$.

The generic structure of $\mathcal{T}$ is depicted in Fig. 8.4. If $k = 1$, then it must be

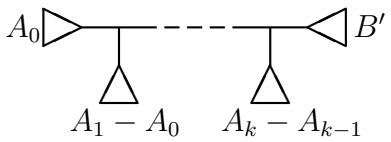


Figure 8.4: The tree $\mathcal{T}$ in Lemma 8.2.1.

that $\mathcal{T} = \theta(\mathcal{T})$. On the other hand, if $k \geq 3$ then in order for $\mathcal{T}|A' = \theta(\mathcal{T})|A'$ to hold, we must regraft the pruned subtree $\mathcal{T}|A$ in the same place so that again $\mathcal{T} = \theta(\mathcal{T})$. Hence $k = 2$, and so swapping the subtrees $\mathcal{T}|A$ and $\mathcal{T}|B'$ produces $\theta(\mathcal{T})$ from which it follows that $\theta$ is an NNI operation. □

As a consequence of Lemma 8.2.1, we can express the sizes of both the SPR and the TBR unit neighbourhoods in terms of the number of each operation for a tree and the size of the NNI neighbourhood.

**Lemma 8.2.2.** *For $\mathcal{T} \in \mathscr{T}_n$, $n \geq 4$, we have*

$$|N_\vartheta(\mathcal{T})| = |\mathcal{O}_\vartheta(\mathcal{T})| - 3|N_{\text{NNI}}(\mathcal{T})|,$$

*where $\vartheta \in \{\text{SPR}, \text{TBR}\}$.*

*Proof.* The proof follows from Lemma 8.2.1 and the observation that, if $\theta$ is an NNI operation for $\mathcal{T}$, then there are precisely four distinct operations $\theta' \in \mathcal{O}_{\text{NNI}}(\mathcal{T})$ such that $\theta(\mathcal{T}) = \theta'(\mathcal{T})$. □

Lemma 8.2.2 forms the basis of the two key results for this section. Both the number of distinct SPR operations and the number of distinct TBR operations for any given tree can be found relatively easily. We proceed with the SPR case first.

**Theorem 8.2.3.** *For a tree $\mathcal{T} \in \mathscr{T}_n$ where $n \geq 4$, we have*

$$|\mathcal{O}_{\text{SPR}}(\mathcal{T})| = 4(n-2)(n-3).$$

*Proof.* We consider two possible SPR operations on $\mathcal{T}$, firstly those that induce a trivial split on $\mathcal{T}$, and secondly those that induce a non-trivial split. In the first case, there are $n$ possible leaves that can be pruned from $\mathcal{T}$, and for each leaf $x$ there are $2n-6$ edges in $\mathcal{T}-x$ to which we can reconnect it so that the resulting tree is different from $\mathcal{T}$.

In the second case, suppose that the non-trivial split is $A|B$, with $|A| = a$ and $|B| = b$. If we choose $\mathcal{T}|A$ to be the pruned subtree, then there are $2b-3$ edges to which we can regraft $\mathcal{T}|A$. However, one of these results in the same tree as we began with, namely $\mathcal{T}$. Thus there are $2b-4$ such distinct operations. Similarly, if we choose $\mathcal{T}|B$ as the pruned subtree, then there are $2a-4$ possible SPR operations. Thus there are $2n-8$ distinct SPR operations for each of the $n-3$ non-trivial splits of $\mathcal{T}$. Hence

$$\begin{aligned} |\mathcal{O}_{\text{SPR}}(\mathcal{T})| &= n(2n-6) + (n-3)(2n-8) \\ &= 4(n-2)(n-3). \end{aligned}$$

□

As a corollary to this theorem, we obtain the result of Allen and Steel's ([2]) for the size of the SPR unit neighbourhood. The proof is omitted, as it follows trivially from Lemma 8.2.2 and Theorem 8.2.3.

**Corollary 8.2.4** (Theorem 2.1, [2])**.** *For $\mathcal{T} \in \mathscr{T}_n$ where $n \geq 4$, we have*

$$|N_{\text{SPR}}(\mathcal{T})| = 2(n-3)(2n-7).$$

We require one further idea before tackling the TBR problem. For a binary tree $\mathcal{T}$, we define $\Gamma(\mathcal{T})$ by

$$\Gamma(\mathcal{T}) = \sum |A| \cdot |B|,$$

where the sum is taken over all non-trivial splits $A|B$ of $\mathcal{T}$.

**Theorem 8.2.5.** *For a tree $\mathcal{T} \in \mathscr{T}_n$ where $n \geq 4$, we have*

$$|\mathcal{O}_{\text{TBR}}(\mathcal{T})| = 4\Gamma(\mathcal{T}) - 4(n-2)(n-3).$$

*Proof.* We consider two possible TBR operations on $\mathcal{T}$, firstly those that induce a trivial split on $\mathcal{T}$, and secondly those that induce a non-trivial split. The argument in the first case is identical to that given in the proof of Theorem 8.2.3, and gives $n(2n-6)$ distinct TBR operations.

Now, let $A|B$ be some non-trivial split of $\mathcal{T}$ induced by the edge $e$. Then when we bisect $\mathcal{T}$ by deleting $e$, there are $2|A|-3$ edges in one component of the resulting forest and $2|B|-3$ edges in the other. Hence, there are $(2|A|-3)(2|B|-3)$ ways to choose an edge from each of $\mathcal{T}|A$ and $\mathcal{T}|B$. Precisely one of these results in re-forming $\mathcal{T}$. Hence, by taking a sum over all non-trivial splits $A|B$ of $\mathcal{T}$, we get

$$\begin{aligned} |\mathcal{O}_{\text{TBR}}(\mathcal{T})| &= n(2n-6) + \sum [(2|A|-3)(2|B|-3) - 1] \\ &= 4\Gamma(\mathcal{T}) - 4(n-2)(n-3). \end{aligned}$$

□

This brings us to the main result of the chapter. While the following corollary gives the size of the TBR neighbourhood for $\mathcal{T}$ in terms of $\Gamma(\mathcal{T})$, calculating this quantity is straightforward. Also, as we will see in Section 8.3, Corollary 8.2.6 gives enough traction for us to characterise the trees in a space that respectively maximise and minimise the size of the neighbourhood.

**Corollary 8.2.6.** *For $\mathcal{T} \in \mathscr{T}_n$ where $n \geq 4$, we have*

$$|N_{\text{TBR}}(\mathcal{T})| = 4\Gamma(\mathcal{T}) - (4n-2)(n-3).$$

## 8.3 Characterisations of the Extremal Cases

Since the size of the TBR unit neighbourhood for $\mathcal{T}$ is dependent on both the number of leaves in $\mathcal{T}$ and the shape of $\mathcal{T}$, it makes sense to characterise which tree shapes give the extreme values for this size. As a consequence of Corollary 8.2.6, it suffices to determine which tree shapes maximise and

minimise the size of $\Gamma(\mathcal{T})$ over all trees in $\mathscr{T}_n$ for some $n$. We begin with the easier case, that is, finding the trees that maximise $\Gamma(\mathcal{T})$.

**Lemma 8.3.1.** *Let $\mathcal{T} \in \mathscr{T}_n$ be a tree such that $\Gamma(\mathcal{T}) \geq \Gamma(\mathcal{T}')$ for all $\mathcal{T}' \in \mathscr{T}_n$. Then $\mathcal{T}$ is a caterpillar.*

*Proof.* Suppose that $\{x_1, x_2\}$ and $\{x_3, x_4\}$ are cherries of $\mathcal{T}$, and let the sets $Y_1, \ldots, Y_k$ partition the remaining leaves so that $\mathcal{T}$ can be represented as in Fig. 8.5. Setting $y_i = |Y_i|$, it will suffice to show that $y_i = 1$ for all $i$. For

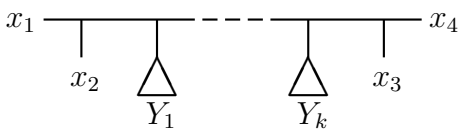


Figure 8.5: The tree $\mathcal{T}$ in the proof of Lemma 8.3.1.

some $i \in \{1, \ldots, k\}$, we form a second tree $\mathcal{T}'$ by moving the subtree $\mathcal{T}|Y_i$ to the position adjacent to $x_1$. The tree $\mathcal{T}'$ is shown in Fig. 8.6. Since $y_j \geq 1$

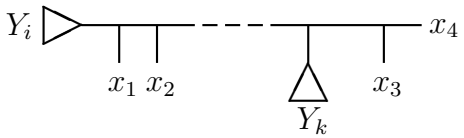


Figure 8.6: The tree $\mathcal{T}'$ in the proof of Lemma 8.3.1.

for all $j$, we have the inequality

$$y_i + (i - 1) \leq n - 4,$$

from which $n - y_i - i - 2$ is strictly positive. Now, calculating the difference

between $\Gamma(\mathcal{T})$ and $\Gamma(\mathcal{T}')$, we find that

$$\begin{aligned}\Gamma(\mathcal{T}) - \Gamma(\mathcal{T}') &= \sum_{j=0}^{i-1}(j+2)(n-j-2) - \sum_{j=0}^{i-1}(y_i+j+1)(n-y_i-j-1) \\ &= i(1-y_i)(n-y_i-i-2).\end{aligned}$$

If $y_i > 1$, then $\Gamma(\mathcal{T}) < \Gamma(\mathcal{T}')$, and so in fact $y_i = 1$ for all $i \in \{1, \ldots, k\}$. Thus $\mathcal{T}$ is a caterpillar. □

Recall from Section 8.1 that the previous best upper bound on the size of $N_{\text{TBR}}(\mathcal{T})$ for a tree $\mathcal{T} \in \mathscr{T}_n$ was $2n^3 + O(n^2)$. Theorem 8.3.2 confirms that the tight upper bound is a cubic function of $n$.

**Theorem 8.3.2.** *The tree $\mathcal{T} \in \mathscr{T}_n$ maximises the size of the* TBR *unit neighbourhood over $\mathscr{T}_n$ if and only if $\mathcal{T}$ is a caterpillar. Moreover, if $\mathcal{T}$ is a caterpillar then*

$$|N_{\text{TBR}}(\mathcal{T})| = \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2.$$

*Proof.* The first part of the theorem follows from Lemma 8.3.1. To find the size of the neighbourhood, we apply Corollary 8.2.6 from which we have

$$\begin{aligned}|N_{\text{TBR}}(\mathcal{T})| &= 4\Gamma(\mathcal{T}) - (4n-2)(n-3) \\ &= 4\sum_{i=2}^{n-2} i(n-i) - (4n-2)(n-3) \\ &= \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2.\end{aligned}$$

□

The characterisation of those trees that minimise the size of the TBR neighbourhood relies heavily on the next lemma (Lemma 8.3.3). Before proving this, we give an example of the simplest case of this lemma. Suppose that, in Fig. 8.7, the sizes of the pendant subtrees labelled by $X_1, \ldots, X_4$ are $x_1, \ldots, x_4$ respectively. If this tree has a minimal value for $\Gamma(\mathcal{T})$, then since

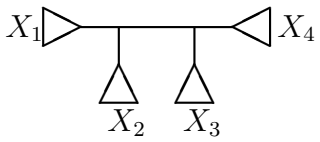


Figure 8.7: A tree illustrating the simplest case of Lemma 8.3.3.

$\Gamma(\mathcal{T})$ is the sum of $|A| \cdot |B|$ over all non-trivial splits $A|B$, we must have

$$(x_1 + x_2)(x_3 + x_4) \leq \min\{(x_1 + x_3)(x_2 + x_4), (x_1 + x_4)(x_2 + x_3)\}.$$

Assuming without loss of generality that $x_1$ is the smallest of the four quantities, it is easy to show that $x_2$ is the next smallest. Lemma 8.3.3 extends this observation to a more general result.

**Lemma 8.3.3.** *Let $X = \{1, \ldots, n\}$, and let $\mathcal{T} \in \mathscr{T}_n$ be such that $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}')$ for all $\mathcal{T}' \in \mathscr{T}_n$. Further, for some $k \geq 0$ let $X_1, \ldots, X_4, Y_1, \ldots, Y_k$ partition $X$ such that the following hold:*

*(i) $X_i|X - X_i \in \Sigma(\mathcal{T})$ for all $i \in \{1, \ldots, 4\}$;*

*(ii) $Y_i|X - Y_i \in \Sigma(\mathcal{T})$ for all $i \in \{1, \ldots, k\}$; and*

*(iii) $A_i|X - A_i \in \Sigma(T)$ for all $i \in \{0, \ldots, k\}$, where $A_0 = X_1 \cup X_2, A_i = A_{i-1} \cup Y_i$.*

*Then without loss of generality we have $x_1 \leq x_2 \leq x_3 \leq x_4$, where $x_i = |X_i|$.*

*Proof.* Without loss of generality, we can assume $x_1 \leq x_2 \leq x_3$. Supposing that the lemma is false, we have $x_2 > x_4$. Then either $x_1 = x_3$, and so $x_1 \geq x_2 \geq x_3 \geq x_4$, contradicting our assumption that the lemma is false, or $x_1 < x_3$.

Figure 8.8 shows the general structure of a tree $\mathcal{T}$ that satisfies the conditions of the lemma. Let $\mathcal{T}_1$ be the tree obtained from $\mathcal{T}$ by swapping the subtrees labelled by $X_1$ and $X_3$, and let $\mathcal{T}_2$ be similarly obtained by swapping

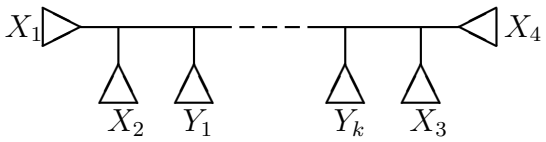


Figure 8.8: The tree $\mathcal{T}$ in Lemma 8.3.3.

the subtrees $\mathcal{T}|X_2$ and $\mathcal{T}|X_4$. Let $y_i = |Y_i|$, and $b_0 = 0, b_i = b_{i-1} + y_i$. Then we have

$$\begin{aligned}\Gamma(\mathcal{T}) - \Gamma(\mathcal{T}_1) &= \sum_{j=0}^{k}(x_1 + x_2 + b_j)(n - x_1 - x_2 - b_j) \\ &\quad - \sum_{j=0}^{k}(x_2 + x_3 + b_j)(n - x_2 - x_3 - b_j) \\ &= (x_3 - x_1)\left[2\sum_{j=0}^{k} b_j - (k+1)(n - x_1 - 2x_2 - x_3)\right].\end{aligned}$$

Since we assume that $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}_1)$, and that both $x_1 < x_3$ and $x_2 > x_4$ hold, we get

$$\begin{aligned}\Gamma(\mathcal{T}) - \Gamma(\mathcal{T}_2) &= (x_4 - x_2)\left[2\sum_{j=0}^{k} b_j - (k+1)(n - 2x_1 - x_2 - x_4)\right] \\ &> (x_4 - x_2)\left[2\sum_{j=0}^{k} b_j - (k+1)(n - x_1 - 2x_2 - x_3)\right] \\ &= \frac{x_4 - x_2}{x_3 - x_1}\left(\Gamma(\mathcal{T}) - \Gamma(\mathcal{T}_1)\right) \\ &\geq 0,\end{aligned}$$

contradicting the fact that $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}_2)$. □

Applying Lemma 8.3.3, we can completely characterise those trees $\mathcal{T}$ that minimise the size of $\Gamma(\mathcal{T})$, and therefore those trees that minimise the size of the TBR unit neighbourhood.

**Lemma 8.3.4.** *Let* $X = [n]$ *for some* $n = \sum_{i=0}^{k} \alpha_i 2^i$, *where* $\alpha_i \in \{0, 1\}$ *for* $0 \leq i < k$ *and* $\alpha_k = 1$. *Let* $\beta_j = \frac{1}{2^j} \sum_{i=j}^{k} \alpha_i 2^i$. *Let* $\mathcal{T} \in \mathscr{T}_n$ *such that* $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}')$ *for all* $\mathcal{T}' \in \mathscr{T}_n$. *Then for all* $0 \leq j \leq k-1$ *there is a partition* $X_1, \ldots, X_{\beta_j}$ *of* $X$ *into* $\beta_j$ *disjoint subsets such that following properties hold:*

*(i)* $X_p | X - X_p \in \Sigma(\mathcal{T})$ *for all* $1 \leq p \leq \beta_j$; *and*

*(ii)* $|X_p| = 2^j$ *for all* $1 \leq p < \beta_j$.

*Proof.* For $j = 0$, this holds trivially. We assume that for some $0 \leq j < k-1$, the partition $X_1, \ldots, X_{\beta_j}$ of $X$ satisfies the conditions of the lemma.

Suppose that for $1 \leq p < q < \beta_j$, there is no set $Y$ that contains either $X_p$ or $X_q$ such that $Y | X - Y \in \Sigma(\mathcal{T})$ and $|Y| = 2^{j+1}$. Then we can apply Lemma 8.3.3 to find a tree $\mathcal{T}'$ for which $\Gamma(\mathcal{T}') < \Gamma(\mathcal{T})$. Hence, for $m$ such that $2m < \beta_j$, there are disjoint subsets $X'_1, \ldots, X'_m$ of $X$ such that $X'_p | X - X'_p \in \Sigma(\mathcal{T})$ and $|X'_p| = 2^{j+1}$.

There are two cases to consider. Suppose firstly that $2m = \beta_j - 2$. Then there is some $1 \leq p < \beta_j$ such that $X_p$ is not contained in some $Y$, where $Y | X - Y \in \Sigma(\mathcal{T})$ and $|Y| = 2^{j+1}$. We can then use Lemma 8.3.3 again to show that if $X'_{\beta_{j+1}} = X_p \cup X_{\beta_j}$, then $X'_{\beta_{j+1}} | X - X'_{\beta_{j+1}} \in \Sigma(\mathcal{T})$. Since $m + 1 = \beta_{j+1}$, we have the required partition.

On the other hand, if $2m = \beta_j - 1$ then we can use Lemma 8.3.3 to show that there is some $1 \leq p \leq m$ such that, if $X'_{\beta_{j+1}} = X_{\beta_j} \cup X'_p$, then $X'_{\beta_{j+1}} | X - X'_{\beta_{j+1}} \in \Sigma(\mathcal{T})$. Again, this gives the required partition, completing the induction. □

The question now is what these trees look like. In some sense, the trees that minimise the size of $\Gamma(\mathcal{T})$ are maximally balanced, although we must define carefully what we mean by this. The only sizes of $n$ for which an unrooted binary tree can be truly balanced, or *perfect*, are $n = 2^k$ or $n = 3 \cdot 2^k$, where we have either two-fold symmetry about an interior edge of the tree or three-fold symmetry about an interior vertex, and the tree is vertex-transitive with respect to the leaves. For values of $n$ other than those which admit a perfect tree, we necesssarily lose the property of leaf-transitivity as a global structural property.

A tree $\mathcal{T} \in \mathscr{T}_n$, where $3 \cdot 2^k \leq n < 3 \cdot 2^{k+1}$ is *complete* if and only if

(i) there is a cluster $Y$ of $\mathcal{T}$ with $|Y| = 2^{k+1}$; and

(ii) for all clusters $Y$ with $2 \leq |Y| \leq 2^{k+1}$, there is a bipartition $Y_1, Y_2$ of $Y$ such that both of $Y_1, Y_2$ are clusters of $\mathcal{T}$, and $|Y_1| = 2^j$ and $2^{j-1} \leq |Y_2| < 2^{j+1}$ for some $j$.

That is, for each such cluster $Y$, the pendant subtree $\mathcal{T}|Y$ has minimal depth, and one half of this pendant subtree is perfectly balanced. The trees in Lemma 8.3.4 are precisely the complete trees in the space $\mathscr{T}_n$, from which we obtain the next theorem. The proof is routine and omitted.

**Theorem 8.3.5.** *The tree $\mathcal{T} \in \mathscr{T}_n$ minimises the size of the* TBR *unit neighbourhood over $\mathscr{T}_n$ if and only if $\mathcal{T}$ is complete.*

Let us continue towards finding the size of the TBR unit neighbourhood for complete trees.

**Lemma 8.3.6.** *Let $\mathcal{T} \in \mathscr{T}_n$ be a complete tree for some $n = \sum_{i=0}^{k} \alpha_i 2^i$, where $\alpha_i \in \{0, 1\}$ for $0 \leq i < k$ and $\alpha_k = 1$. Then*

$$\Gamma(\mathcal{T}) = \sum_{j=1}^{k-1} \left( \sum_{i=j}^{k} \alpha_i 2^i - 2^j \right) \left( 2n - \sum_{i=j}^{k} \alpha_i 2^i \right)$$

*if $\alpha_{k-1} = 1$, and*

$$\Gamma(\mathcal{T}) = \sum_{j=1}^{k-2} \left( \sum_{i=j}^{k} \alpha_i 2^i - 2^j \right) \left( 2n - \sum_{i=j}^{k} \alpha_i 2^i \right) + 2^{k-1}(n - 2^{k-1})$$

*if $\alpha_{k-1} = 0$.*

*Proof.* We use the proof of Lemma 8.3.4 to obtain this result. For each of the partitions $X_1, \ldots, X_{\beta_j}$, we take the sum of $|X_p| \cdot (n - |X_p|)$. We consider the family of complete trees on $n$ leaves, where $\alpha_{k-1} = 1$ following the notation

of Lemma 8.3.4. This gives

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1}\left[\sum_{p=1}^{\beta_j}|X_p|\cdot(n-|X_p|)\right]\\ &= \sum_{j=1}^{k-1}\left[2^j(\beta_j-1)(n-2^j)+|X_{\beta_j}|\cdot(n-|X_{\beta_j}|)\right].\end{aligned}$$

Now, we also have from Lemma 8.3.4 that

$$|X_{\beta_j}| = n - 2^j(\beta_j - 1),$$

so incorporating this into the above expression we find

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1}2^j(\beta_j-1)(2n-2^j\beta_j)\\ &= \sum_{j=1}^{k-1}\left(\sum_{i=j}^{k}\alpha_i 2^i - 2^j\right)\left(2n-\sum_{i=j}^{k}\alpha_i 2^i\right).\end{aligned}$$

In the case that $\alpha_{k-1}=0$, the partition $X_1,\ldots,X_{\beta_{k-1}}$ is a bipartition of the leaf set of $\mathcal{T}$, and so we need only take the product $|X_1|\cdot(n-|X_1|)$ once in the sum above. That is

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-2}\left[\sum_{p=1}^{\beta_j}|X_p|\cdot(n-|X_p|)\right]+2^{k-1}(n-2^{k-1})\\ &= \sum_{j=1}^{k-2}\left(\sum_{i=j}^{k}\alpha_i 2^i - 2^j\right)\left(2n-\sum_{i=j}^{k}\alpha_i 2^i\right)+2^{k-1}(n-2^{k-1}).\end{aligned}$$

□

We conclude this chapter with two corollaries that give firstly an exact value for the size of the TBR unit neighbourhood for perfect trees, and secondly an asymptotic lower bound on the size of this neighbourhood for complete trees. Both proofs follow from Lemma 8.3.6 and Corollary 8.2.6.

**Corollary 8.3.7.** *Let $\mathcal{T} \in \mathscr{T}_n$ be a perfect tree. Then*

$$|N_{\text{TBR}}(\mathcal{T})| = n^2\left(4k - \frac{32}{3}\right) + 22n - 6$$

*if $n = 3 \cdot 2^{k-1}$ for some $k$, and*

$$|N_{\text{TBR}}(\mathcal{T})| = n^2(4k - 13) + 22n - 6$$

*if $n = 2^k$ for some $k$.*

*Proof.* In the first case, where $n = 3 \cdot 2^k$, we have

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1} n(n - 2^j) \\ &= n^2(k-1) - n(2^k - 2) \\ &= n^2\left(k - \frac{5}{3}\right) + 2n,\end{aligned}$$

and the result follows by applying Corollary 8.2.6. On the other hand, if $n = 2^{k+1}$ then

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-2} n(n - 2^j) + \frac{n^2}{4} \\ &= n^2\left(k - \frac{7}{4}\right) - n(2^{k-1} - 2) \\ &= n^2\left(k - \frac{9}{4}\right) + 2n,\end{aligned}$$

and again applying Corollary 8.2.6 gives the required result. □

**Corollary 8.3.8.** *Let $\mathcal{T} \in \mathscr{T}_n$ be a complete tree. Then*

$$|N_{\text{TBR}}(\mathcal{T})| = 4n^2\lfloor \log_2 n \rfloor + O(n^2).$$

*Proof.* The proof is similar in nature to that for the previous corollary. In

the first case, where $3 \cdot 2^{k-1} \leq n < 2^{k+1}$ for some $k \geq 1$, we have

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1} \left( n - \sum_{i=0}^{j-1} \alpha_i 2^i - 2^j \right) \left( n + \sum_{i=0}^{j-1} \alpha_i 2^i \right) \\ &= n^2(k-1) - n(2^k - 2) - \sum_{j=1}^{k-1} \left( \sum_{i=0}^{j-1} \alpha_i 2^i \right)^2 .\end{aligned}$$

However, we can obtain a bound for the final term of this expression by assuming that $\alpha_i = 1$ for all $i \in \{0, \ldots, k-2\}$, giving

$$\begin{aligned}\sum_{j=1}^{k-1} \left( \sum_{i=0}^{j-1} \alpha_i 2^i \right)^2 &< \sum_{j=1}^{k-1} 2^{2j} \\ &= \frac{2}{3} \left( 2^{2k-1} - 1 \right) \\ &= O(n^2).\end{aligned}$$

The second case, where $2^k \leq n < 3 \cdot 2^{k-1}$, follows in a similar manner and we complete the proof by Corollary 8.2.6. □

# Chapter 9

# Agreement Forests

## 9.1 Introduction

In the previous chapter, we defined three different tree rearrangement operations that may be used to transform one leaf-labelled tree into another. We refer the reader back to Section 8.1 for the definitions of these operations. The purpose of Chapter 8 was to more accurately determine the size of the unit neighbourhood for some tree under TBR. That is, how many trees are obtainable from a given tree by a single TBR operation.

One of the primary motivations for studying tree rearrangements is that they can be used to quantify the level of similarity inherent in two trees in the same tree space. For $\vartheta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$, we define the $\vartheta$ *distance* between $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$ to be the fewest number of $\vartheta$ operations required to change $\mathcal{T}$ into $\mathcal{T}'$. The notation we use for the $\vartheta$ distance is

$$d_\vartheta(\mathcal{T}, \mathcal{T}') = d_\vartheta(\mathcal{T}', \mathcal{T}) = k,$$

where $k$ is the smallest non-negative integer such that there is a sequence of trees $\mathcal{T}_0 = \mathcal{T}, \mathcal{T}_1, \ldots, \mathcal{T}_k = \mathcal{T}'$ that satisfies

$$\mathcal{T}_i \in N_\vartheta(\mathcal{T}_{i-1})$$

for all $i \in \{1, \ldots, k\}$.

With the distance defined in this way, if

$$d_\vartheta(\mathcal{T}_1, \mathcal{T}_2) < d_\vartheta(\mathcal{T}_1, \mathcal{T}_3)$$

for trees $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3 \in \mathscr{T}_n$, then we would expect $\mathcal{T}_1$ and $\mathcal{T}_2$ to be more alike in some structural sense than $\mathcal{T}_1$ and $\mathcal{T}_3$ are.

The $\vartheta$ distance also induces a metric on the tree space $\mathscr{T}_n$ for each $\vartheta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$ ([2],[39]). In this chapter, we will introduce a graph that realises this metric for the various tree rearrangment operations we are considering. The *adjacency graph* $G_\vartheta(n)$ for the $\vartheta$ operation in $\mathscr{T}_n$ has the elements of $\mathscr{T}_n$ as its vertex set, and an edge between any two trees if and only if each tree lies in the $\vartheta$ unit neighbourhood of the other (note that $\mathcal{T}' \in N_\vartheta(\mathcal{T})$ forces $\mathcal{T} \in N_\vartheta(\mathcal{T}')$). That is, the edge set $E$ of $G_\vartheta(n)$ is

$$E = \{\{\mathcal{T}, \mathcal{T}'\} : \mathcal{T}, \mathcal{T}' \in \mathscr{T}_n, \mathcal{T}' \in N_\vartheta(\mathcal{T})\}.$$

Alternatively, the condition that $\mathcal{T}'$ is in $N_\vartheta(\mathcal{T})$ may now be replaced by

$$d_\vartheta(\mathcal{T}, \mathcal{T}') = 1,$$

as the two are equivalent. Also, as $G_\vartheta(n)$ is a direct representation of the metric that $\vartheta$ induces on $\mathscr{T}_n$, the $\vartheta$ distance between any two trees in $\mathscr{T}_n$ is the same as the length of the shortest path between the two corresponding vertices in the adjacency graph.

As a simple example, the graph $G_\vartheta(4)$ is isomorphic to $K_3$, the complete graph with three vertices, for each $\vartheta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$. For $n = 5$, the adjacency graphs $G_{\text{SPR}}(n)$ and $G_{\text{TBR}}(n)$ are again identical since any TBR operation on a tree in $\mathscr{T}_5$ is also an SPR operation. Rather than directly describe $G_{\text{TBR}}(5)$, it is easiest to begin with a description of the complementary graph, which we will denote by $H$. Suppose that $\mathcal{T}_1 \in \mathscr{T}_5$ is the tree shown in Fig. 9.1. If $\mathcal{T}'$ is a tree that is not in $N_{\text{TBR}}(\mathcal{T}_1)$, then $\mathcal{T}_1, \mathcal{T}'$ do

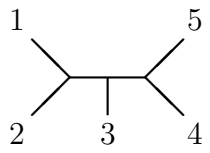


Figure 9.1: A tree $\mathcal{T}_1 \in \mathscr{T}_5$.

not share a quartet. A quick check will confirm that there are only two trees in $\mathscr{T}_5$ that satisfy this, namely the trees $\mathcal{T}_2, \mathcal{T}_3$ shown in Fig. 9.1. From

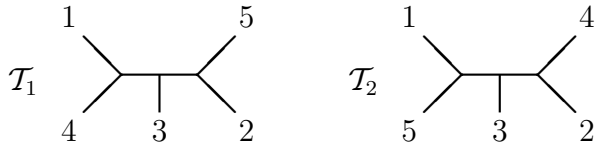


Figure 9.2: Two trees in $\mathcal{T}_2, \mathcal{T}_3 \in \mathscr{T}_5$ that are not in the TBR unit neighbourhood of the tree $\mathcal{T}_1$ from Fig. 9.1.

the symmetry, we deduce that $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ form a clique in $H$. Extending this across the remainder of $\mathscr{T}_5$, we can see that $H$ consists of five disconnected copies of $K_3$. Hence $G_{\text{TBR}}(5)$ is precisely isomorphic to the complete multipartite graph $K_{3,3,3,3,3}$, which alternatively is $K_{15}$ with the fifteen edges of five vertex-disjoint triangles missing.

These graphs give us a useful way to visualise the tree space as a highly connected object, as the example of $G_{\text{TBR}}(5)$ above shows. The results from Chapter 8 about the size of a tree's TBR unit neighbourhood can now alternatively be viewed in a graph theoretic sense as the degree of the corresponding vertex in some adjacency graph $G_{\text{TBR}}(n)$.

We turn our attention now to the problem of determining how far apart two trees in $\mathscr{T}_n$ may be under the TBR metric. This is equivalent to finding the diameter of the TBR adjacency graph for $\mathscr{T}_n$. Upper and lower bounds as functions of $n$ have been established in [2, 29] for the diameter of $G_\vartheta(n)$ for each $\vartheta \in \{\text{NNI}, \text{SPR}, \text{TBR}\}$, some of which are asymptotically tight. If we denote the diameter of $G_\vartheta(n)$ by $\Delta_\vartheta(n)$, then the best current known bounds are

$$\frac{n}{4}\log_2 n - o(n \log n) \leq \Delta_{\text{NNI}}(n) \leq n \log_2 n + O(n),$$
$$\frac{n}{2} - o(n) \leq \Delta_{\text{SPR}}(n) \leq n - 3,$$

and

$$\frac{n}{4} - o(n) \leq \Delta_{\text{TBR}}(n) \leq n - 3.$$

Using a result from Chapter 7, we can already improve on the upper bound for both SPR and TBR. As a special case of Theorem 7.2.4 we have the following lemma, which partially restates the theorem in a form more appropriate to the current context.

**Lemma 9.1.1.** *For all $m \geq 4$, there is some $N > 0$ such that for all $n \geq N$, any pair of distinct trees $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$ share a common subtree that has at least $m$ leaves.*

It is a short step from there to prove Theorem 9.1.2, although we remark that this result is further improved upon in this chapter.

**Theorem 9.1.2.** *For all positive integers $m \geq 4$, there is some $N > 0$ such that for all $n \geq N$,*

$$\Delta_{\text{SPR}}(n) \leq n - m.$$

*Proof.* Given two trees $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$, it suffices to construct a sequence of $n-m$ SPR operations that transforms $\mathcal{T}$ into $\mathcal{T}'$ By Lemma 9.1.1, the trees $\mathcal{T}, \mathcal{T}'$ share a common subtree with $m$ leaves. Let $Y$ be the leaf set of some such common subtree, and let $x \notin Y$ be some other leaf. Then we can perform an SPR operation by pruning $x$ from $\mathcal{T}$ and regrafting it to form a tree $\mathcal{T}''$ so that $\mathcal{T}''|Y \cup x = \mathcal{T}'|Y \cup x$. Now $\mathcal{T}''$ and $\mathcal{T}'$ share a common subtree on $m+1$ leaves. We induct this over each of the $n-m$ leaves in $[n]-Y$. This requires $n-m$ SPR operations, completing the proof of the theorem. □

As a further corollary to this, we have the same result holding for TBR. This is provable, for example, as a consequence of the SPR adjacency graph being a subgraph of the TBR adjacency graph.

**Corollary 9.1.3.** *For all positive integers $m \geq 4$, there is some $N > 0$ such that for all $n \geq N$,*

$$\Delta_{\text{TBR}}(n) \leq n - m.$$

We remind the reader that it was conjectured in Chapter 7 that

$$\tau_k(m) = \kappa_k(m) = \Theta(m^2)$$

for all $k, m$ (Conjecture 7.2.11). In particular, if this conjecture were true for $k = 2$, then the same construction used in the proof of Theorem 9.1.2 can be applied to show that

$$\Delta_{\text{SPR}}(n) \leq n - \Theta(\sqrt{n}).$$

In Section 9.2, we define agreement forests and explain how they relate to tree rearrangement operations, or to the TBR operation to be more specific. These ideas are then applied in Section 9.3 in showing that

$$\Delta_\vartheta(n) = n - \Theta(\sqrt{n})$$

for $\vartheta \in \{\text{SPR}, \text{TBR}\}$.

## 9.2 Agreement Forests

In the introduction to this chapter, we demonstrated how the idea of finding a common subtree for a pair of trees can be used to improve on the upper bound for the diameter of both the SPR and TBR adjacency graph. In this section, we show how this idea can be extended. This basic premise behind this extension is to break a pair of trees up into a collection of disjoint common subtrees. For TBR, this approach suffices, although for SPR we need to be a little more careful.

Suppose that $\mathcal{T}_1, \mathcal{T}_2 \in \mathscr{T}_n$ for some $n \geq 4$, and that the partition $X_0, \ldots, X_k$ of $X = [n]$ satisfies

(i) $\mathcal{T}_1|X_j = \mathcal{T}_2|X_j$ for all $j \in \{0, \ldots, k\}$; and

(ii) the subtrees $\mathcal{T}_i|X_j$, where $j \in \{0, \ldots, k\}$, are vertex disjoint subtrees of $\mathcal{T}_i$ for $i \in \{1, 2\}$.

Then we say that the forest

$$\mathcal{F} = \{\mathcal{T}_i | X_j : j \in \{0, \ldots, k\}\}$$

is an *agreement forest* for $\mathcal{T}_1, \mathcal{T}_2$. Further, if $k$ is the smallest such integer for which such a partition exists, then we call $\mathcal{F}$ a *maximum agreement forest* for $\mathcal{T}_1, \mathcal{T}_2$, and write

$$m(\mathcal{T}_1, \mathcal{T}_2) = k.$$

Note that $m(\mathcal{T}_1, \mathcal{T}_2) = |\mathcal{F}| - 1$. Allen and Steel proved in [2] that the size of a maximum agreement forest for two trees is directly related to the TBR distance between them. We will use this result later, and so include it here as a lemma.

**Lemma 9.2.1** (Theorem 2.4, [2])**.** *Let $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$ for some $n$. Then*

$$d_{\text{TBR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}').$$

The proof of [2, Theorem 2.4] (Lemma 9.2.1) involves showing that an agreement forest induces a sequence of TBR moves between two trees, and vice-versa, based on the observation that the construction of either a TBR sequence or an agreement forest requires repeated bipartitioning of the leaf set for the trees.

With all this in mind, agreement forests give us additional traction on the problem of bounding the diameter of the TBR adjacency graph. They allow us to generalise the proof method we used for Theorem 9.1.2, where we essentially constructed an agreement forest, albeit a relatively trivial one, by taking a large common subtree for the two trees and allowing all other leaves to be isolated components. It must be pointed out that as a general rule, agreement forests are of limited use in studying SPR-based problems. As an exception to this rule, we give an example of a result relating agreement forests to SPR in Lemma 9.2.2. This lemma is in fact used in Section 9.3 in proving an upper bound on $\Delta_{\text{SPR}}(n)$.

**Lemma 9.2.2.** *Let $\mathcal{F} = \{t_0, \ldots, t_k\}$ be an agreement forest for two bi-*

*nary trees* $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_X$ *such that* $|\mathcal{L}(t_i)| \leq 2$ *for all* $i \in \{1, \ldots, k\}$. *Then* $d_{\text{SPR}}(\mathcal{T}, \mathcal{T}') \leq k$.

*Proof.* We use an induction argument to construct a sequence of $k$ SPR operations to transform $\mathcal{T}'$ into $\mathcal{T}$. Suppose firstly that $k = 1$. Then we perform an SPR operation on $\mathcal{T}'$ by pruning $t_1$ from $\mathcal{T}'$, and then regrafting it to $t_0$ to form $\mathcal{T}$. That is, $d_{\text{SPR}}(\mathcal{T}, \mathcal{T}') = 1$, completing the basis for the induction.

Now suppose that $k > 2$, and let $X_i = \mathcal{L}(t_i)$ so that $X_0, \ldots, X_k$ forms a partition of $X$. Then there is some part $X_k$ say, such that by setting $X'_0 = X_0 \cup X_k$ and $X'_i = X_i$ for all $i \geq 1$, the partition of $X$ into $X'_0, \ldots, X'_{k-1}$ satisfies the agreement forest condition that the collection of induced subtrees be vertex-disjoint with respect to $\mathcal{T}$. Let $Y$ be the minimal cluster of $\mathcal{T}'$ that contains $X_k$ but not $X_0$. We perform the SPR operation that prunes the subtree $\mathcal{T}|Y$ from $\mathcal{T}$, and subsequently regrafts it to $\mathcal{T}|X - Y$ so that, in the resulting tree $\mathcal{T}''$ we have

(i) $\mathcal{T}''|X'_i = \mathcal{T}|X'_i$ for all $i \in \{0, \ldots, k-1\}$; and

(ii) the subtrees $\mathcal{T}''|X'_i$, where $i \in \{0, \ldots, k-1\}\}$, are vertex disjoint subtrees of $\mathcal{T}''$.

By definition, the forest $\mathcal{F}' = \{\mathcal{T}|X'_i : i \in \{0, \ldots, k-1\}\}$ is an agreement forest for $\mathcal{T}, \mathcal{T}''$. Also, $\mathcal{F}'$ satisfies the conditions of the lemma for $k-1$, and as $d_{\text{SPR}}(\mathcal{T}', \mathcal{T}'') = 1$ this completes the induction. □

Since our primary interest with agreement forests in this chapter is to bound the diameter of the SPR and TBR adjacency graphs, we require a further definition. For all $n \geq 4$, we let $m(n)$ be the least positive integer $k$ such that

$$m(\mathcal{T}, \mathcal{T}') \leq k$$

for all pairs of trees $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$. It is easy to check by Lemma 9.2.1 that $m(n) = \Delta_{\text{TBR}}(n)$. As a further consequence of the same lemma, we have one further result to end the section. The proof is routine.

**Lemma 9.2.3.** *For all $n \geq 4$, we have*

$$m(n) \leq m(n+1) \leq m(n)+1.$$

*Proof.* Let $\mathcal{T}, \mathcal{T}'$ be trees in $\mathscr{T}_{n+1}$, and let $X = [n], x = n+1$. Suppose firstly that $\mathcal{F}$ is an agreement forest for $\mathcal{T}, \mathcal{T}'$. Then $\mathcal{F}|X$ is an agreement forest for $\mathcal{T}|X, \mathcal{T}'|X \in \mathscr{T}_n$, proving the first inequality. If instead we suppose that $\mathcal{F}$ is an agreement forest for $\mathcal{T}|X, \mathcal{T}'|X$, then $\mathcal{F} \cup \{\mathcal{T}|x\}$ is an agreement forest for $\mathcal{T}, \mathcal{T}'$, from which the second inequality follows. □

## 9.3 Main Results

The central aim of this chapter is to improve the bounds on the diameter of both the SPR and the TBR adjacency graphs. The main theorem is stated below:

**Theorem 9.3.1.** *For all $n \geq 4$,*

$$\Delta_\vartheta(n) = n - \Theta(\sqrt{n}),$$

*where $\vartheta \in \{\text{SPR}, \text{TBR}\}$.*

We devote this section to a proof of Theorem 9.3.1, firstly showing that there is some constant $c > 0$ such that

$$\Delta_{\text{TBR}}(n) \geq n - c\sqrt{n} + O(1).$$

This is done constructively by proving that, for positive integers $k, l \geq 2$, there is a pair of caterpillars in $\mathscr{T}_{kl}$ for which any agreement forest contains a large number of isolated vertices as components.

**Lemma 9.3.2.** *Let $k, l, n \geq 2$ be positive integers such that $k \leq l$ and $n = kl$, and let $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$ be caterpillars such that $\mathcal{T}$ has the label ordering $[1, \ldots, kl]$ and $\mathcal{T}'$ has the label ordering $[1, k+1, \ldots, k(l-1)+1, 2, k+2, \ldots, k(l-1), kl]$. If $\mathcal{F}$ is a maximum agreement forest for $\mathcal{T}, \mathcal{T}'$, then either there are $k$ consecutive leaves in $\mathcal{T}$ of which $k-1$ are isolated in $\mathcal{F}$ or there are $l$ consecutive leaves in $\mathcal{T}'$ of which $l-1$ are isolated in $\mathcal{F}$.*

*Proof.* We assume throughout that the lemma is false. The case $k = 1$ is trivial, so we may assume that $2 \leq k \leq l$. There are two leaves $a, b \in \mathcal{L}(t_0)$ for some $t_0 \in \mathcal{F}$, where $1 \leq a < b \leq k$. Suppose that these are the smallest such $a$ and $b$. Now, in $\mathcal{T}'$ the leaves $\{jk + a : 1 < j < l\}$ lie between $a$ and $b$. If all of these are isolated then we have a contradiction, so $\mathcal{L}(t_0) \cap \{jk + a : 0 < j < l\} \neq \emptyset$. However, we can only add a single leaf of the form $jk + a$ to $t_0$, thus forming a maximal common subtree for $\mathcal{T}, \mathcal{T}'$ since $a, b$ must remain adjacent in $t_0$.

If we suppose that there is some $c$ such that $a < c < b$, then all the leaves in the set $\{jk + c : 0 \leq j < l\}$ are isolated in $\mathcal{F}$, which is again a contradiction. Thus $b = a + 1$. Further, the third leaf of $t_0$ must be $k + a$, for otherwise there are $k - 1$ leaves in $\{a + 1, \ldots, k + a\}$ that are isolated in $\mathcal{F}$. Hence $\mathcal{L}(t_0) = \{a, a + 1, k + a\}$.

Suppose next that $k + a + 1$ is isolated in $\mathcal{F}$. Then $k - 1$ leaves in $\{a+2, \ldots, k+a+1\}$ are isolated in $\mathcal{F}$, which again contradicts our assumption. We can follow a similar argument as above to show that there must be a component $t_i \in \mathcal{F}$ with $\mathcal{L}(t_i) = \{ik + a + i, ik + a + i + 1, (i + 1)k + a + i\}$ for all $0 \leq i \leq k - a - 1$.

Let $j = k - a - 1$, and consider the leaf $x = (j + 1)k + a + (j + 1)$. Note that $x \leq kl$, and hence $x \in [n]$. If $x$ is isolated in $\mathcal{F}$, then $k - 1$ of the leaves in $\{jk+a+(j+2), \ldots, (j+1)k+a+(j+1)\}$ are isolated in $\mathcal{F}$; contradiction. Otherwise, there must be some leaf $y$, where $(j + 1)k + a + (j + 1) < y \leq (j + 2)k + a + j$, such that $x, y$ are in the same component $t \in \mathcal{F}$. However, all such leaves, if they exist, are of the form $mk + c$ where $c < k$. Hence the path connecting $x$ and $y$ in $\mathcal{T}'$ must cross the path connecting $jk + a + j$ and $(j + 1)k + a + j$. That is, $t$ and $t_j$ are not vertex disjoint and cannot both be components of $\mathcal{F}$. This final contradiction completes the proof of the lemma. □

This lemma enables us to apply an inductive argument to construct a maximum agreement forest for two caterpillars with the specified labelling.

**Theorem 9.3.3.** *Let $k, l, n \geq 2$ be positive integers such that $k \leq l$ and $n = kl$, and let $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$ be caterpillars such that $\mathcal{T}$ has the label ordering $[1, \ldots, kl]$ and $\mathcal{T}'$ has the label ordering $[1, k + 1, \ldots, k(l - 1) + 1, 2, k +$*

$2, \ldots, k(l-1), kl]$. *Then*

$$m(\mathcal{T}, \mathcal{T}') = (k-1)(l-1).$$

*Proof.* To establish $(k-1)(l-1)$ as an upper bound for $m(\mathcal{T}, \mathcal{T}')$, it is sufficient to construct an agreement forest $\mathcal{F}$ of size $(k-1)(l-1)+1$. If we let

$$Y = \{1, \ldots, k\} \cup \{jk : 1 \leq j \leq l\},$$

then $\mathcal{T}_Y = \mathcal{T}'|Y$. Since $|Y| = k+l-1$, the forest $\mathcal{F}$ with $\mathcal{T}|Y$ as the unique non-trivial component is an agreement forest for $\mathcal{T}, \mathcal{T}'$ with

$$|\mathcal{F}| = kl - (k+l-1) + 1$$

as required.

We use induction to complete the theorem. When $k = 1$, the result is straightforward. Suppose that for some $m > 2$ the theorem holds for all pairs $k \leq l$ such that $k + l = m$. Suppose now that $k + l = m + 1$, and let $\mathcal{T}_1, \mathcal{T}_2$ satisfy the conditions of the theorem for $k, l$. Further, let $\mathcal{F}$ be a maximum agreement forest for $\mathcal{T}_1, \mathcal{T}_2$. By Lemma 9.3.2, there is either a set of $k$ consecutive leaves in $\mathcal{T}_1$ of which $k-1$ are isolated in $\mathcal{F}$, or a set of $l$ consecutive leaves in $\mathcal{T}_2$ of which $l-1$ are isolated in $\mathcal{F}$. Let $Y$ be such a set of consecutive leaves, and let $\mathcal{T}_1' = \mathcal{T}_1|X-Y$, $\mathcal{T}_2' = \mathcal{T}_2|X-Y$.

If $|Y| = l$, then under some permutation of $X-Y$, the trees $\mathcal{T}_1', \mathcal{T}_2'$ satisfy the conditions of the theorem for the pair of positive integers $k-1 \leq l$. Now, by our induction hypothesis, any maximum agreement forest $\mathcal{F}'$ for $\mathcal{T}_1', \mathcal{T}_2'$ contains precisely $(k-2)(l-1)+1$ components. Thus

$$|\mathcal{F}| \geq (k-1)(l-1) + 1,$$

as $l-1$ of the leaves in $Y$ are isolated in $\mathcal{F}$. The same argument follows if $|Y| = k$, noting that in the case $k = l$ we may exchange $k$ and $l$ to complete the induction. □

By setting $k, l \approx \sqrt{n}$, we obtain as a corollary to this last theorem the

fact that

$$\Delta_{\text{TBR}}(n) \geq n - \Theta(\sqrt{n}),$$

providing a lower bound on the diameter of the TBR adjacency graph. A more formal proof of this is given later. For now, we move on to finding an upper bound for $\Delta_{\text{SPR}}(n)$.

In the introduction to the chapter, we commented that a consequence of proving Conjecture 7.2.11 would be the result that

$$\Delta_{\text{SPR}}(n) \leq n - \Theta(\sqrt{n}),$$

which would suffice to complete a proof of Theorem 9.3.1. This observation was based around the algorithmic proof to Theorem 9.1.2, in which our current upper bound for $\Delta_{\text{TBR}}(n)$ is embedded. Explicitly,

$$\Delta_{\text{SPR}}(n) \leq n - m \leq n - 3$$

for all $m \geq 3$ and all $n \geq \tau(m)$. From the point of view of agreement forests, this approach involves the construction of a forest that has a single non-trivial component with $m$ leaves and many isolated leaves. The following lemma allows us to construct a set of vertex-disjoint subtrees of a specified minimum size.

**Lemma 9.3.4.** *Let $k, m, n > 0$ be positive integers such that $n \geq 2(k-1)(m-1)+m$, and let $\mathcal{T} \in \mathscr{T}_n$. Then there is a collection $t_1, \ldots, t_k$ of vertex-disjoint subtrees of $\mathcal{T}$ such that $|\mathcal{L}(t_i)| \geq m$ for all $i \in \{1, \ldots, k\}$.*

*Proof.* Let $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$, and let $X = [n]$. We will make repeated use of the fact that, for a tree $\mathcal{T} \in \mathscr{T}_n$ and for all $m \leq n$, there is a cluster $Y$ of $\mathcal{T}$ such that $m \leq |Y| \leq 2m-2$. Let $\mathcal{T}_1 = \mathcal{T}$. We define the collection $t_1, \ldots, t_k$ of subtrees of $\mathcal{T}$ recursively as follows. For each $i \in \{1, \ldots, k-1\}$, let $Y_i$ be some cluster of $\mathcal{T}_i$ with $m \leq |Y_i| \leq 2m-2$. Let $t_i = \mathcal{T}_i|Y_i$, and let $\mathcal{T}_{i+1} = \mathcal{T}_i|\mathcal{L}(\mathcal{T}_i) - Y_i$. Since each subtree $t_1, \ldots, t_{k-1}$ contains at most $2m-2$

leaves, the tree $\mathcal{T}_k$ must have at least

$$n - (k-1)(2m-2) = m$$

leaves. We then set $t_k = \mathcal{T}_k$, completing the proof. □

Using this, we can construct an agreement forest for an arbitrary pair of trees on $n$ leaves that contains $O(\sqrt{n})$ non-trivial components. This allows us to then complete the proof of Theorem 9.3.1.

**Theorem 9.3.5.** *For any two trees $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$ where $n \geq 4$, we have*

$$d_{\text{SPR}}(\mathcal{T}, \mathcal{T}') \leq n - \left\lfloor \frac{1}{2}\sqrt{n} \right\rfloor.$$

*Proof.* Let $\mathcal{T}, \mathcal{T}' \in \mathscr{T}_n$, and let $X = [n]$. By Lemma 9.3.4, if we set $k = \lfloor \frac{1}{2}\sqrt{n} \rfloor$, then there is a partition $X_1, \ldots, X_k$ of $X$ such that $|X_i| \geq 2k$ for all $i \in \{1, \ldots, k\}$, and such that $\{\mathcal{T}|X_i\}$ is a collection of vertex-disjoint subtrees of $\mathcal{T}$. We aim now to show that there is a second partition $Y_1, \ldots, Y_k$ of $X$ such that $|X_i \cap Y_i| \geq 2$ for all $i \in \{1, \ldots, k\}$, and such that $\{\mathcal{T}'|Y_i\}$ is a collection of vertex-disjoint subtrees of $\mathcal{T}'$.

Let $Y_1$ be a minimal cluster of $\mathcal{T}'$ such that $|X_i \cap Y_1| \geq 2$ for some $i \in \{1, \ldots, k\}$. We may also assume without loss of generality that $i = 1$. Since $Y_1$ is a minimal cluster satisfying this, we know that $Y_1$ has at most two leaves in common with each of $X_1, \ldots, X_k$. Now, suppose that for some $i \in \{2, \ldots, k\}$, we have already defined the subsets $Y_1, \ldots, Y_{i-1}$ of $X$, and let

$$Z_{i-1} = \bigcup_{j=1}^{i-1} (X_j \cup Y_j).$$

We let $Y_i$ be a minimal cluster of $\mathcal{T}'|X - Z_{i-1}$ such that $|X_h \cap Y_i| \leq 2$ for some $h \in \{i, \ldots, k\}$. We may assume without loss of generality that $h = i$, and we let

$$Z_i = \bigcup_{j=1}^{i} (X_j \cup Y_j).$$

If $i < k$, then by the minimality of each cluster $Y_1, \ldots, Y_i$, the set $X - Z_i$ contains at least $2(k-i)$ elements in common with each of $X_{i+1}, \ldots, X_k$, and so it follows that $Y_1, \ldots, Y_k$ is a partition of $X$ that satisfies our requirements.

Let $W = X - \bigcup_{j=1}^{k}(X_j \cap Y_j)$, and let $\mathcal{F}$ be the forest

$$\mathcal{F} = \{\mathcal{T}|(X_i \cap Y_i) : i \in \{1, \ldots, k\}\} \cup \{\mathcal{T}|w : w \in W\}.$$

It follows from the construction of the partitions $X_1, \ldots, X_k$ and $Y_1, \ldots, Y_k$ that $\mathcal{F}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ in which each component has at most two leaves. The result is now a consequence of Lemma 9.2.2. □

We conclude this chapter with a proof of the main theorem. The proof uses the two theorems we have already established in this section, and the inequality

$$d_{\text{TBR}}(\mathcal{T}, \mathcal{T}') \leq d_{\text{SPR}}(\mathcal{T}, \mathcal{T}'),$$

which holds for all pairs of trees $\mathcal{T}, \mathcal{T}'$ on the same leaf set ([2]).

*Proof of Theorem 9.3.1.* Let $k = l = \lceil \sqrt{n} \rceil$. Since $kl \geq n$, we can combine Lemma 9.2.3 and Theorem 9.3.3 to obtain

$$\begin{aligned} \Delta_{\text{TBR}}(n) &\geq (k-1)(l-1) - (kl - n) \\ &= n - k - l + 1 \\ &= n - 2\lceil \sqrt{n} \rceil + 1. \end{aligned}$$

Using the inequality between SPR and TBR stated above along with Theorem 9.3.5 yields

$$n - 2\lceil \sqrt{n} \rceil + 1 \leq \Delta_{\text{TBR}}(n) \leq \Delta_{\text{SPR}}(n) \leq n - \left\lfloor \frac{1}{2}\sqrt{n} \right\rfloor,$$

completing the proof. □

As a closing comment, we further remark that the approach taken in this chapter gives bounds of the same nature for the analogue of the TBR

operation in the space of rooted trees, namely the *rooted subtree prune and regraft* (rSPR) operation. This result is given without proof as Corollary 9.3.6.

**Corollary 9.3.6.** *For all* $n \geq 3$*,*

$$\Delta_{r\text{SPR}}(n) = n - \Theta(\sqrt{n}).$$

Interested readers may refer to [25, 43] for background to and a more formal treatment of rooted tree rearrangement operations.

# References


[1] Allen, B.L. (1998). *Subtree Transfer Operations and their Induced Metrics on Evolutionary Trees*. MSc thesis. University of Canterbury, Christchurch, NZ.

[2] Allen, B.L. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, **5**, 1–15.

[3] Baroni, M.C. (2004). *Hybrid Phylogenies: A graph-based approach to represent reticulate evolution*. PhD thesis. University of Canterbury, Christchurch, NZ.

[4] Berdichevsky, N. (2004). *Nations, language and citizenship*. Jefferson, NC: McFarland.

[5] Bininda-Emonds, O.R.P., Gittleman, J.L. and Steel, M.A. (2002). The (super)tree of life: procedures, problems and prospects, *Annual Reviews of Ecology and Systematics*, **33**, 265–289.

[6] Bininda-Emonds, O.R.P. (ed.) (2004). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*. Computational Biology Series. Dortrecht, NL: Kluwer.

[7] Blažek, V. (2005). On the internal classification of Indo-European languages: survey. *Linguistica Online*.

[8] Bocker, S., Dress, A. and Steel, M. (1999). Patching up $X$-trees, *Annals of Combinatorics*, **3**, 1–12.

[9] Bodlaender, H.L., Fellows, M.R. and Warnow, T.J. (1993). Two strikes against perfect phylogeny. In: *Proceedings of the International Colloquium on Automata, Languages and Programming*, Lecture Notes in Computer Science, **623**, 273–283. Berlin: Springer-Verlag.

[10] Bollobás, B. (1998). *Modern Graph Theory*. New York: Springer.

[11] Bona, M. (2002). *A Walk Through Combinatorics*. River Edge, NJ: World Scientific.

[12] Bona, M. (2004). *Combinatorics of Permutations*. Boca Raton: Chapman and Hall.

[13] Bordewich, M., Huber, K.T. and Semple, C. (2005). Identifying phylogenetic trees. *Discrete Mathematics*, **300**, 30–43.

[14] Bordewich, M. and Semple, C. (2005). Private communication.

[15] Bryant, D. and Steel, M. (1995). Extension operations on sets of leaf-labelled trees. *Advances in Applied Mathematics*, **16**, 425–453.

[16] Bryant, D. (1997). *Building Trees, Hunting for Trees, and Comparing Trees*. PhD thesis. University of Canterbury, Christchurch, NZ.

[17] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In: *Mathematics in the archaeological and historical sciences* (ed. F.R. Hodson, D.G. Kendall and P. Tautu), 387–395. Edinburgh: Edinburgh University Press.

[18] Buneman, P. (1974). A characterization of rigid circuit graphs. *Discrete Mathematics*, **9**, 205–212.

[19] Dekker, M.C.H. (1986). *Reconstruction methods for derivation trees*. Unpublished Masters thesis. Vrije Universiteit, Amsterdam, The Netherlands.

[20] Deutscher, G. (2005). *The Unfolding of Language: an evolutionary tour of mankinds greatest invention*. New York: Metropolitan Books.

[21] Dress, A. and Erdős, P. (2003). $X$-trees and weighted quartet systems. *Annals of Combinatorics*, **7**, 155–169.

[22] Embleton, S. (1986). *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

[23] Erdős, P. and Szekeres, G. (1935). A combinatorial problem in geometry. *Compositio Mathematica*, **2**, 463–470.

[24] Grünewald, S., Humphries, P. and Semple, C. (2008). Quartet compatibility and the quartet graph. *Electronic Journal of Combinatorics*, **15**, R103.

[25] Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, **36**, 369–405.

[26] Huber, K., Moulton, V., Semple, C. and Steel, M. (2004). Recovering a phylogenetic tree using pairwise closure operations. *Applied Mathematics Letters*, **18**, 361–366.

[27] Humphries, P. (2008). Bounds on the size of the TBR unit-neighbourhood. *Annals of Combinatorics*, in press.

[28] Kubicka, E., Kubicki, G. and McMorris, F.R. (1992). On agreement subtrees of two binary trees. *Congressus Numerantium*, **88**, 217–222.

[29] Li, M., Tromp, J. and Zhang, L. (1996). On the nearest neighbour interchange distance between evolutionary trees. *Journal of Theoretical Biology*, **182**, 463–467.

[30] Lipo, C.P., OBrien, M.J., Collard, M. and Shennan, S.J. (ed.) (2005). *Mapping Our Ancestors: Phylogenetic Approaches in Anthropology and Prehistory*. New York: Aldine Transaction.

[31] Maddison, D.R. (1991). The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology*, **43**, 315–328.

[32] Matsen, F.A. and Steel, M. (2007). Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, **56**, 767–775.

[33] Matsen, F.A., Mossel, E. and Steel, M. (2008). Mixed-up trees: the structure of phylogenetic mixtures. *Bulletin of Mathematical Biology*, in press.

[34] Meacham, C.A. (1983). Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. In: *Numerical Taxonomy* (ed. J. Felsenstein), NATO ASI Series, Vol. G1, 304–314. Berlin: Springer-Verlag.

[35] Mossel, E. and Steel, M. (2003). A phase transition for a random cluster model on phylogenetic trees. *Mathematical Biosciences*, **187**, 189–203.

[36] Ramat, A.G. and Ramat, P. (1998). *The Indo-European languages*. London: Routledge.

[37] Ramsey, F.P. (1930). On a problem of formal logic. *Proceedings of the London Mathematical Society*, series 2, **30**, 264–286.

[38] van Reenen, P. and van Mulken, M. (ed.) (1996). *Studies in Stemmatology*. Amsterdam: Benjamins.

[39] Robinson, D.F. (1971). Comparison of Labeled Trees with Valency Three. *Journal of Combinatorial Theory*, **11**, 105–119.

[40] Semple, C. and Steel, M. (2001). Tree reconstruction via a closure operation on partial splits. In: *Computational Biology: First International Conference on Biology, Informatics, and Mathematics, JOBIM 2000, Montepellier, France, May 3–5, 2000: Selected papers* (ed. O. Gascuel and M.-F. Sagot), 126–134. New York: Springer.

[41] Semple, C. and Steel, M. (2002). A characterization for a set of partial partitions to define an $X$-tree. *Discrete Mathematics*, **247**, 169–186.

[42] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford: Oxford University Press.

[43] Song, Y.S. (2003). On the combinatorics of rooted binary phylogenetic trees. *Annals of Combinatorics*, **7**, 365–379.

[44] Steel, M. (1989). *Distributions on bicoloured evolutionary trees*. PhD thesis. Massey University, Palmerston North, NZ.

[45] Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, **9**, 91–116.

# APPENDICES

These appendices contain three mathematical papers on very different topics. They are included here because all three were written during the course of this thesis, even though only one of them has a direct connection to it.

The first paper (Geelen and Humphries, 2006) was written while visiting the University of Waterloo in Ontario, and concerns a problem in matroid theory. The second (Humphries, 2007) was inspired by reading a biography of Ramanujan, and the third (Humphries, *in press*) is a precursor to Chapter 8 of this thesis.

## List of Publications


[A] Geelen, J. and Humphries, P. (2006). Rota's basis conjecture for paving matroids. *SIAM Journal of Discrete Mathematics*, **4**, 1042–1045.

[B] Humphries, P. (2007). Nesting polynomials in infinite radicals. *Bulletin of the Korean Mathematical Society*, **44**, 331–336.

[C] Humphries, P. (*in press*). Bounds on the size of the TBR unit-neighbourhood. *Annals of Combinatorics*.

# Appendix A
# Rota's basis conjecture for paving matroids

Jim Geelen and Peter J. Humphries

Abstract. Rota conjectured that, given $n$ disjoint bases of a rank-$n$ matroid $M$, there are $n$ disjoint transversals of these bases that are all bases of $M$. We prove a stronger statement for the class of paving matroids.

## A.1 Introduction

We prove the following theorem.

**Theorem A.1.1.** Let $B_1, \ldots, B_n$ be disjoint sets of size $n \geq 3$ and let $M_1, \ldots, M_n$ be rank-$n$ paving matroids on $\bigcup_i B_i$ such that $B_i$ is a basis of $M_i$ for each $i \in \{1, \ldots, n\}$. Then there exist $n$ disjoint transversals $A_1, \ldots, A_n$ of $(B_1, \ldots, B_n)$ such that $A_i$ is a basis of $M_i$ for each $i \in \{1, \ldots, n\}$.

A *paving matroid* $M$ is a matroid in which each circuit has size $r(M)$ or $r(M)+1$, where $r(M)$ is the rank of $M$. Theorem A.1.1 implies Rota's basis conjecture for paving matroids.

**Conjecture A.1.2** (Rota)**.** Given $n$ disjoint bases $B_1, \ldots, B_n$ in a rank-$n$ matroid $M$, there exist $n$ disjoint transversals $A_1, \ldots, A_n$ of $(B_1, \ldots, B_n)$ that are all bases of $M$.

For $n = 2$, Conjecture A.1.2 follows immediately from basis exchange in matroids. Chan [2] proved the conjecture for $n = 3$. Wild [9] proved a stronger conjecture for the class of strongly base-orderable matroids, while more recently a slightly weaker result was proved for a general matroid (Pono-marenko [8]). Further partial results may be found in [1], [3], [4], [5] and [9].

Theorem A.1.1 fails for both $n = 2$ and matroids in general. When $n = 2$, if we take $\mathcal{B}(M_1) = \{\{e, f\}, \{e, g\}, \{f, h\}, \{g, h\}\}$ and $\mathcal{B}(M_2) =$

$\{\{e,f\},\{e,h\},\{f,g\},\{g,h\}\}$, then $\{e,f\},\{g,h\}$ is the only pair of disjoint bases. In the second instance, if $r_{M_1}(E - B_1) = 0$, then there are no $M_1$-independent transversals of $(B_1,\ldots,B_n)$.

The remainder of this paper is taken up with the proof of the theorem. In Section A.2, we prove that Theorem A.1.1 holds when $n = 3$. This result is used, in Section A.3, as the base case of an inductive proof of Theorem A.1.1. The induction argument is surprisingly straightforward and can be read independently of Section A.2.

## A.2 The case $n = 3$

For basic concepts in matroid theory, the reader is referred to Oxley [7]. We follow the same notation as Oxley throughout this paper.

A closed set in a matroid is commonly known as a flat. We will primarily be interested in rank-2 flats, or *lines*. In the proof of Theorem A.2.1, we make frequent use of the fact that if $r_M(X) = r_M(Y) = 2$ and $|X \cap Y| \geq 2$, then $X$ and $Y$ are contained in the same line in $M$.

**Theorem A.2.1.** Theorem A.1.1 holds for $n = 3$.

*Proof.* Assume that the theorem is false. Then there exist bases $B_1 = \{a_1,a_2,a_3\}, B_2 = \{b_1,b_2,b_3\}, B_3 = \{c_1,c_2,c_3\}$ of rank-3 paving matroids $M_1, M_2, M_3$ respectively, with common ground set $E = B_1 \cup B_2 \cup B_3$, that provide a counterexample. The rank of a set $X$ in $M_i$ will be denoted by $r_i(X)$ and the closure by $\mathrm{cl}_i(X)$. A three-element subset of $E$ will be called a *transversal* if it meets each of $B_1$, $B_2$, and $B_3$. Note that we may assume that every non-trivial line in each matroid contains a transversal, since all non-trivial lines not containing a transversal may be relaxed to provide an alternative counterexample (see [7], Section 1.5, Exercise 3).

**A.2.1.1.** Let $X \subseteq E$ be a set that meets each of $B_1, B_2, B_3$. If $r_i(X) = 3$, then $X$ contains an $M_i$-independent transversal.

*Subproof.* Let $T \subseteq X$ be a transversal, and suppose that $T$ is $M_i$-dependent. Then since $r_i(X) = 3$, there is some $e \in X$ such that $e \notin \mathrm{cl}_i(T)$. Without loss of generality, $e \in B_1$, so let $f$ be the unique element in $T \cap B_1$. Then $r_i((T - f) \cup e) = 3$, and we are done. □

**A.2.1.2.** If no $M_1$-dependent transversal contains both $a_1$ and $b_1$, then there exists $e \in B_3$ such that $r_2(E - \{a_1,b_1,e\}) = 2$.

*Subproof.* For each $a \in B_1$ and $b \in B_2$, there exists $c \in B_3$ such that $\{a,b,c\}$ is $M_3$-independent (since $r_3(B_3) = 3$). In particular, there exist $e,f,g \in B_3$ such that $\{a_2,b_3,e\}$, $\{a_3,b_3,f\}$, and $\{a_2,b_2,g\}$ are $M_3$-independent. Then,

by A.2.1.1, $\{a_3, b_2\} \cup (B_3 - \{e\})$, $\{a_2, b_2\} \cup (B_3 - \{f\})$, and $\{a_3, b_3\} \cup (B_3 - \{g\})$ all have rank 2 in $M_2$ (since otherwise we would find the required partition into transversals). The second and third of these sets both have two points in common with the first, and so they are all contained in a common line in $M_2$. □

Suppose that $M_1$ has a line $L$ containing at least seven elements. Since $r_1(B_1) = 3$, $|L - B_1| \geq 5$. Up to symmetry, we may assume that $b_1, b_2, c_1, c_2, c_3 \in L$ and that $a_1 \notin \mathrm{cl}_1(L)$. Now neither $\{a_1, b_1\}$ nor $\{a_1, b_2\}$ is in an $M_1$-dependent transversal. So by A.2.1.2 $r_2(\{a_2, a_3, b_2, b_3\}) = r_2(\{a_2, a_3, b_1, b_3\}) = 2$, contradicting the fact that $r_2(B_2) = 3$. Thus none of $M_1$, $M_2$, and $M_3$ contain a line on seven or more elements.

**A.2.1.3.** Every pair $e \in B_i, f \notin B_i$ is contained in some $M_i$-dependent transversal.

*Subproof.* Suppose that no $M_1$-dependent transversal contains both $a_1$ and $b_1$. Then, by A.2.1.2 and symmetry, we may assume that $r_2(E - \{a_1, b_1, c_1\}) = 2$. Let $X = E - \{a_1, b_1, c_1\}$ and $Y = X - B_1$. Each transversal in $\{a_2, a_3, b_2, b_3, c_1\}$ is $M_2$-independent, for otherwise $E - \{a_1, b_1\}$ is a seven-point line in $M_2$. Since each transversal in $\{a_1, b_1, c_2, c_3\}$ is $M_1$-independent, there is no $M_3$-independent transversal in $X$; thus $r_3(X) = 2$. Similarly, since each transversal in $\{a_2, a_3, b_1, c_2, c_3\}$ is $M_2$-independent and each transversal in $\{a_2, a_3, b_2, b_3, c_1\}$ is $M_3$-independent, we conclude that $r_1(Y \cup \{a_1\}) = 2$. Without loss of generality, $a_2 \notin \mathrm{cl}_1(Y)$, and so both $\{a_2, b_2, c_2\}$ and $\{a_2, b_3, c_3\}$ are $M_1$-independent. This means that $\{a_1, b_1, c_2\}$ and $\{a_1, b_1, c_3\}$ are $M_2$-dependent, for otherwise we again have three disjoint transversals that are independent in their respective matroids. Thus $r_2(\{a_1, b_1, c_2, c_3\}) = 2$ and $E - \{c_1\}$ is an eight-point line in $M_2$, which is a contradiction. □

Assume that $B_2$ is dependent in $M_1$. Thus, some line $L$ in $M_1$ contains $B_2$; we may assume that $L$ also contains $a_1$ and $c_1$, since any non-trivial line contains a transversal. There must be some element $a_3$, say, of $B_1$ that is not in $\mathrm{cl}_1(L)$, but then no transversal containing both $a_3$ and $c_1$ is dependent in $M_1$, leading to a contradiction by A.2.1.3. Thus each of $B_1, B_2$, and $B_3$ is independent in all three matroids. This provides additional symmetry since we may now permute $(B_1, B_2, B_3)$.

Suppose next that $M_1$ contains a five- (or six-) point line $L$. By the conclusion of the last paragraph, we may assume that $a_1, b_1, b_2, c_1, c_2 \in L$ and that $a_3 \notin \mathrm{cl}_1(L)$. Now, since there is an $M_1$-dependent transversal containing $a_3, b_1$, we have that $\{a_3, b_1, c_3\}$ must be $M_1$-dependent. Likewise $\{a_3, b_2, c_3\}$ is $M_1$-dependent, and thus $r_1(\{a_3, b_1, b_2, c_3\}) = 2$, contradicting the fact that $a_3 \notin \mathrm{cl}_1(L)$. Hence, none of $M_1$, $M_2$, and $M_3$ have lines containing more than four points.

We suppose now that the transversal $\{a_3, b_3, c_3\}$ is $M_2$-independent and $M_3$-dependent. Since $r_1(E-\{a_3, b_3, c_3\}) = 3$, we may assume that $\{a_1, b_1, c_1\}$ is $M_1$-independent, and also that $r_3(\{a_2, b_2, c_2\}) = 2$ for otherwise we have the required disjoint bases. Now, at most one of $a_3$, $b_3$, and $c_3$ may be contained in $\mathrm{cl}_3(\{a_2, b_2, c_2\})$, so without loss of generality both $\{a_2, b_3, c_2\}$ and $\{a_3, b_2, c_2\}$ are $M_3$-independent. Then $\{a_3, b_2, c_3\}$ and $\{a_2, b_3, c_3\}$ are both $M_2$-dependent. The transversal $\{a_2, b_2, c_3\}$ must now be $M_2$-independent, for otherwise we get a line in $M_2$ containing $\{a_3, b_3, c_3\}$. Thus $r_3(\{a_3, b_3, c_2\}) = 2$, and further $r_3(\{a_3, b_3, c_2, c_3\}) = 2$. Then both of $\{a_2, b_2, c_3\}$ and $\{a_3, b_2, c_3\}$ are $M_3$-independent, for otherwise there is a line in $M_3$ that contains $E - \{a_1, b_1, c_1\}$. So we have $r_2(\{a_3, b_3, c_2\}) = r_2(\{a_2, b_3, c_2\}) = 2$. This, together with the dependence of $\{a_3, b_2, c_3\}$ and $\{a_2, b_3, c_3\}$ in $M_2$, further implies that $\{a_3, b_3, c_3\}$ is $M_2$-dependent, which is a contradiction.

From now on, we may assume that $M_1$, $M_2$, and $M_3$ are the same matroid $M$, since they share the same set of independent transverals. Suppose that $M$ contains the four-point line $\{a_3, b_3, c_2, c_3\}$. Without loss of generality, we may assume that $\{a_1, b_1, c_1\}$ is independent in $M$, but then both $\{a_2, b_3, c_3\}$ and $\{a_3, b_2, c_2\}$ are also independent in $M$, so we are done.

Thus, the rank-2 flats in $M$ each contain at most three points. Let $\{a_3, b_3, c_3\}$ be a dependent transversal of $M$. By A.2.1.1, the set $\{a_3, b_2, c_1, c_2\}$ contains a transversal that is independent in $M$. Suppose without loss of generality that $\{a_3, b_2, c_2\}$ is such a transversal. Then, again by A.2.1.1, the set $\{a_1, a_2, b_1, c_1\}$ contains an $M$-independent transversal, $\{a_1, b_1, c_1\}$ say. Finally, $\{a_2, b_3, c_3\}$ is also independent, for otherwise we get a four-point line, and we have the three required transversals. □

## A.3 Proof of Theorem A.1.1

Before proving Theorem A.1.1, we require two further lemmas. These allow us to apply induction with Theorem A.2.1 as the base case. Let $\mathcal{B}(M)$ denote the set of bases of a matroid $M$.

**Lemma A.3.1.** Let $B_1 \in \mathcal{B}(M_1), B_2 \in \mathcal{B}(M_2)$ be disjoint bases of rank-$n$ paving matroids on the same ground set, where $n \geq 3$. Let $X$ be a two-element subset of $B_1$. Then there is some $x \in X, y \in B_2$ such that $(B_1 - x) \cup y \in \mathcal{B}(M_1)$ and $(B_2 - y) \cup x \in \mathcal{B}(M_2)$.

*Proof.* Since $M_1, M_2$ are paving matroids, $(B_1 - X) \cup y$ is $M_1$-independent for all $y \in B_2$. Suppose that both $(B_1 - x) \cup y$ and $(B_1 - x') \cup y$ are circuits in $M_1$, where $x, x'$ are distinct elements of $X$. Then by circuit elimination, $B_1$ is also a circuit of $M_1$. Hence for each $y \in B_2$, at least one of $(B_1 - x) \cup y$ and $(B_1 - x') \cup y$ must be a basis of $M_1$.

Let $y_1, y_2, y_3$ be distinct elements of $B_2$. Then without loss of generality $(B_1 - x) \cup y_1, (B_1 - x) \cup y_2 \in \mathcal{B}(M_1)$. Also, one of $(B_2 - y_1) \cup x$ and $(B_2 - y_2) \cup x$ is a basis of $M_2$, so we are done. □

**Lemma A.3.2.** Let $B_1, \ldots, B_n$ be disjoint sets of size $n \geq 3$ and let $M_1, \ldots, M_n$ be rank-$n$ paving matroids on $\bigcup_i B_i$ such that $B_i$ is a basis of $M_i$ for each $i \in \{1, \ldots, n\}$. Then there is an ordering of the elements of $B_1$ as $a_1, \ldots, a_n$ and a transversal $\{b_2, \ldots, b_n\}$ of $(B_2, \ldots, B_n)$ such that for all $j \in \{2, \ldots, n\}$, the set $(B_1 - \{a_2, \ldots, a_j\}) \cup \{b_2, \ldots, b_j\}$ is a basis of $M_1$ and $(B_j - b_j) \cup a_j$ is a basis of $M_j$.

*Proof.* For $j = 2$, the lemma follows immediately from Lemma A.3.1. Suppose now that the lemma holds for some $j \in \{2, \ldots, n-1\}$, so that $B' = (B_1 - \{a_2, \ldots, a_j\}) \cup \{b_2, \ldots, b_j\} \in \mathcal{B}(M_1)$. Then $|B_1 \cap B'| \geq 2$, and so by Lemma A.3.1 there is some element $a_{j+1} \in B_1 \cap B'$ and some $b_{j+1} \in B_{j+1}$ such that $(B' - a_{j+1}) \cup b_{j+1} \in \mathcal{B}(M_1)$ and $(B_{j+1} - b_{j+1}) \cup a_{j+1} \in \mathcal{B}(M_{j+1})$, thus proving the lemma. □

Lemma A.3.2 is stated for $j \in \{2, \ldots, n\}$ to simplify the induction process. We only need the result for $j = n$ to prove main theorem of this paper.

*Proof of Theorem A.1.1.* Assume that the theorem is true for some $m \geq 3$, and take $n = m+1$. Let $B_1 = \{a_1, \ldots, a_n\}$ and $b_i \in B_i$ for each $i \in \{2, \ldots, n\}$. By Lemma A.3.2 we may assume that $A_1 = \{a_1, b_2, \ldots, b_n\}$ is a basis of $M_1$ and that $B_i' = (B_i - b_i) \cup a_i$ is a basis of $M_i$ for each $i \in \{2, \ldots, n\}$.

Now let $X = E - (B_1 \cup A_1)$ and $M_i' = (M_i / a_i)|X$ for each $i \in \{2, \ldots, n\}$. Then each $M_i'$ is a rank-$m$ paving matroid having $B_i - b_i$ as a basis. By our induction hypothesis, there are disjoint transversals $A_2', \ldots, A_n'$ of these $m$ bases such that $A_i'$ is a basis of $M_i'$. Hence $A_i = A_i' \cup a_i$ is a basis of $M_i$ for each $i \in \{2, \ldots, n\}$. Moreover, the bases $A_1, \ldots, A_n$ are disjoint transversals of $(B_1, \ldots, B_n)$ as required. □

## References for Appendix A


[1] Aharoni, R. and Berger, E. (2006). The intersection of a matroid and a simplicial complex. *Transactions of the American Mathematical Society*, **358**, 4895–4917.

[2] Chan, W. (1995). An exchange property of matroid. *Discrete Mathematics*, **146**, 299–302.

[3] Chow, T. (1995). On the Dinitz conjecture and related conjectures. *Discrete Mathematics*, **145**, 73–82.

[4] Drisko, A.A. (1997). On the number of even and odd Latin squares of order $p + 1$. *Advances in Mathematics*, **128**, 20–35.

[5] Drisko, A.A. (1998). Proof of the Alon-Tarsi conjecture for $n = 2^r p$. *Electronic Journal of Combinatorics*, **5**, R28.

[6] Huang, R. and Rota, G.-C. (1994). On the relations of various conjectures on Latin squares and straightening coefficients. *Discrete Mathematics*, **128**, 225–236.

[7] Oxley, J.G. (1992). *Matroid Theory*. New York: Oxford University Press.

[8] Ponomarenko, V. (2004). Reduction of jump systems. *Houston Journal of Mathematics*, **30**, 27–33.

[9] Wild, M. (1994). On Rota's problem about $n$ bases in a rank $n$ matroid. *Advances in Mathematics*, **108**, 336–345.

# Appendix B

# Nesting polynomials in infinite radicals

Peter J. Humphries

Abstract. We consider infinite nested radicals in which the arguments are positive polynomial sequences. It is shown that the evaluation of such a nesting is always finite, and we prove necessary and sufficient conditions for the evaluation to be a finite polynomial.

## B.1 Introduction

A famous problem posed by Ramanujan asks for the evaluation of the infinite nested radical

$$\sqrt{1+2\sqrt{1+3\sqrt{1+4\sqrt{1+\cdots}}}}$$

If we instead try to evaluate a more general expression, where we replace the increasing sequence by an arithmetic progression in $x$, namely

$$L(x)=\sqrt{1+x\sqrt{1+(x+1)\sqrt{1+(x+2)\sqrt{1+\cdots}}}}$$

then it can be seen that $L(x)$ satisifes the functional equation

$$L(x)^2=1+xL(x+1)$$

The solution to this is $L(x)=x+1$, giving the evaluation of Ramanujan's example correctly as 3. In fact, this numerical example is merely a special case of a more complicated identity in three variables (see the end of Section B.2).

Several identities concerning infinite nested radicals may be found in [1], [2] and [3]. In [1], nested radicals involving arithmetic sequences in $n$-th roots are considered. The purpose of the current paper is to study the case where the radicals have two polynomials as their arguments.

Throughout this paper, we denote the natural numbers (without zero) and the real numbers by $\mathbb{N}$ and $\mathbb{R}$ respectively. The ring of polynomials in $x$ with real coefficients will be specified by $\mathbb{R}[x]$, and we note further that any use of square roots automatically implies a positive square root. A sequence $a_n$ of positive real numbers is called a *positive polynomial sequence* if there exists a polynomial $a(x) \in \mathbb{R}[x]$ such that $a_i = a(i)$ for all $i \in \mathbb{N}$.

To remove the possibility of any ambiguity, we formalise the concept of evaluating an infinite nested radical

$$\sqrt{a_1 + b_1\sqrt{a_2 + b_2\sqrt{a_3 + \dots}}}$$

to be the limit

$$\lim_{n\to\infty} \sqrt{a_1 + b_1\sqrt{a_2 + \dots + b_{n-1}\sqrt{a_n + b_n}}} \tag{B.1}$$

where $a_n, b_n$ are sequences of real numbers.

In Section B.2, we characterise when an infinite nested radical involving polynomials from $\mathbb{R}[x]$ has a simple closed form as another polynomial in $\mathbb{R}[x]$. Section B.3 is devoted to proving that, for all positive polynomial sequences $a_n, b_n$, the limit in (B.1) exists and is finite.

## B.2 Identities involving nested radicals

The following lemma does not require proof, being a consequence of viewing the infinite nested radical as being a limit of an infinite sequence.

**Lemma B.2.1.** Let $L(x), p(x), q(x)$ be polynomials in $\mathbb{R}[x]$. Then

$$L(x) = \sqrt{p(x) + q(x)\sqrt{p(x+d) + q(x+d)\sqrt{p(x+2d) + \cdots}}}$$

if and only if

$$L(x) = \sqrt{p(x) + q(x)L(x+d)}$$

An analogous statement can be made for higher-order roots, and we may further replace the ring $\mathbb{R}[x]$ by any class of function in one or more variables. However, for the purposes of this paper we are primarily interested in

polynomials in one variable.

From the above lemma, we get any number of results. More importantly, though, given a nested radical to evaluate, we can now concentrate on solving the non-linear functional equation

$$L(x)^2 = p(x) + q(x)L(x+d) \tag{B.2}$$

rather than on the radical itself, where $L(x)$ is assumed to take positive values on the domain of interest.

Given $L(x)$, $q(x)$ and $d$, we can always find a $p(x)$ that satisfies equation (B.2). That is, $p(x) = L(x)^2 - q(x)L(x+d)$. A more interesting problem is, given $p(x)$, $q(x)$ and $d$, to find the function $L(x)$. In particular, we want to find some $L(x)$ that is a polynomial of finite degree.

This is not always possible, as the following example shows. If we take $p(x) = 1$, $q(x) = x$ and $d = 2$, then we wish to find some $L(x)$ that satisfies

$$L(x)^2 = 1 + xL(x+2)$$

It can be seen that the degree of $L(x)$ must be one, and moreover the linear term will be $x$. However, if we try to evaluate a constant term $a$, we run into problems:

$$\begin{aligned}(x+a)^2 &= 1 + x(x+2+a)\\ ax + a^2 &= 2x + 1\end{aligned}$$

Comparing the linear coefficients gives $a = 2$, but the constant terms give the solution $a = \pm 1$.

Our aim is to characterise when an infinite nested radical with polynomial arguments has a polynomial solution. That is, for what combinations of $p(x), q(x)$ and $d$ can we find some $L(x) \in \mathbb{R}[x]$ satisfying equation (B.2). It is known ([3]) that if both $p(x)$ and $q(x)$ are constants, $p$ and $q$ say, then $L(x)$ is also constant, and solves the quadratic equation $L^2 - qL - p = 0$.

Let $\deg(f)$ denote the degree of a polynomial $f(x)$, and $[x^i]f(x)$ denote the coefficient of $x^i$ in the function $f(x)$. Then we have the following two lemmas, which both follow from equation (B.2):

**Lemma B.2.2.** If $L(x) \in \mathbb{R}[x]$ solves equation (B.2) for some $p(x), q(x) \in \mathbb{R}[x]$, then $\deg(L) = \max\{\frac{\deg(p)}{2}, \deg(q)\}$.

**Lemma B.2.3.** Let $p(x), q(x)$ be polynomials in $\mathbb{R}[x]$, and let

$$F(x) = L(x)^2 - p(x) - q(x)L(x+d)$$

where $L(x) = a_k x^k + \ldots + a_0$. Then there exist $a_0, \ldots, a_k \in \mathbb{R}$ such that

$L(x)$ solves equation (B.2) if and only if there exist $a_0, \ldots, a_k \in \mathbb{R}$ such that $[x^i]F(x) = 0$ for all $i \geq 0$.

This now allows us to find a solution to equation (B.2) by comparing coefficients of $F(x)$. While in the last lemma it is stated that $[x^i]F(x)$ must be zero for all $i \geq 0$, it suffices by Lemma B.2.2 for this to hold only for values of $i$ not exceeding the maximum of $\deg(p)$ and $2\deg(q)$. While, for $L(x)$ of degree $k$, this could potentially involve solving up to $2k+1$ simultaneous polynomials in the $k+1$ coefficients of $L(x)$, we can use the next lemma to find the solution systematically by solving only one quadratic equation (taking the positive root) and at most $k$ linear equations.

**Lemma B.2.4.** Let $p(x), q(x)$ be polynomials in $\mathbb{R}[x]$, and let

$$F(x) = L(x)^2 - p(x) - q(x)L(x+d)$$

where $L(x) = a_k x^k + \ldots + a_0$, and $k = \max\{\frac{\deg(p)}{2}, \deg(q)\}$. Then

(i) $[x^{2k}]F(x)$ is quadratic in $a_k$;

(ii) $[x^j]F(x)$ is linear in $a_{j-k}$ for all $k \leq j < 2k$; and

(iii) $[x^j]F(x)$ is independent of $a_i$ for all $i < j - k$ where $k \leq j \leq 2k$.

*Proof.* The coefficient $[x^{2k}]F(x)$ is given by

$$\begin{aligned}[x^{2k}]F(x) &= ([x^k]L(x))^2 - [x^{2k}]p(x) - ([x^k]q(x))([x^k]L(x+d)) \\ &= a_k^2 - a_k[x^k]q(x) - [x^{2k}]p(x)\end{aligned}$$

proving part (i). Similarly, the coefficient $[x^j]F(x)$, where $k \leq j < 2k$ is

$$\begin{aligned}[x^j]F(x) = &\sum_{i=j-k}^{k} ([x^i]L(x))([x^{j-i}]L(x)) - [x^j]p(x) \\ &- \sum_{i=j-k}^{k} ([x^i]q(x))([x^{j-i}]L(x+d)) \\ &= 2a_{j-k}a_k - a_{j-k}[x^k]q(x) - [x^j]p(x) + g(a_{j-k+1}, \ldots, a_k)\end{aligned}$$

where $g(a_{j-k+1}, \ldots, a_k)$ takes care of the extra terms in the summations. This proves (ii), and (iii) follows directly from the expansions above. □

We can now prove the main result of this section.

**Theorem B.2.5.** Let $p(x), q(x)$ be polynomials of degree $s, t$ respectively in $\mathbb{R}[x]$, both with positive leading coefficients. Then there are $\max\{\frac{s}{2}, t\}$ equalities that must be satisfied by $d$ and the coefficients of $p(x), q(x)$ in order for some $L(x) \in \mathbb{R}[x]$ that solves equation (B.2) to exist. Moreover, if these equalities are satisfied, then there is a general solution for $L(x)$ in terms of $d$ and the coefficients of $p(x), q(x)$.

*Proof.* We take $L(x), F(x)$ as in Lemma B.2.4. Then, by the same lemma, we can find a positive $a_k \in \mathbb{R}$ that solves $[x^{2k}]F(x) = 0$. Further, given $a_i, \ldots, a_k$, where $i > 0$, we can find $a_{i-1}$ that solves $[x^{i+k-1}]F(x) = 0$. That is, we can find $a_0, \ldots, a_k$ that simultaneously solve $[x^i]F(x) = 0$ for all $k \leq i \leq 2k$.

Now, by Lemma B.2.3, for $L(x) \in \mathbb{R}[x]$ to exist, we need $[x^i]F(x) = 0$ for all $i \geq 0$. Since we have this equality for $k \leq i \leq 2k$, we need the remaining $k$ equations to be satisfied. That is, $[x^i]F(x) = 0$ for $0 \leq i < k$. Hence there are $k$ constraints on $d$ and the coefficients of $p(x), q(x)$.

We complete the proof of the theorem by noting that, by Lemma B.2.3 $L(x) = a_k x^k + \ldots + a_0$ solves equation (B.2) if and only if all of the $k$ constraints are met with equality. □

We illustrate the theorem with a more concrete example. If $p(x)$ and $q(x)$ are both linear, then we wish to find $L(x) \in \mathbb{R}[x]$ such that

$$L(x)^2 = (p_1 x + p_0) + (q_1 x + q_0) L(x + d)$$

where we assume that both $p_1$ and $q_1$ are non-zero. In this case, it can be seen that $L(x)$ is of the form $a_1 x + a_0$, and that in fact $a_1 = q_1$. So we have

$$\begin{aligned}(q_1 x + a_0)^2 &= (p_1 x + p_0) + (q_1 x + q_0)(q_1 x + q_1 d + a_0) \\ a_0 q_1 x + a_0^2 &= (q_1 q_0 + q_1^2 d + p_1)x + (q_1 q_0 d + a_0 q_0 + p_0)\end{aligned}$$

By comparing the linear terms, we get $a_0 = q_0 + q_1 d + \frac{p_1}{q_1}$, which on substitution into the constant terms gives

$$0 = (q_1^2 d + p_1)^2 + q_1(p_1 q_0 - p_0 q_1) \tag{B.3}$$

That is, the solution $L(x) \in \mathbb{R}[x]$ exists if and only if equation (B.3) holds, in which case

$$L(x) = q_1 x + \left( q_0 + q_1 d + \frac{p_1}{q_1} \right)$$

The identity of Ramanujan's, which we alluded to in the introduction, is

$$x + n + a = \sqrt{ax + (n + a)^2 + x\sqrt{a(x + n) + (n + a)^2 + \ldots}}$$

where $p(x) = ax+(n+a)^2$, $q(x) = x$ and $d = n$. Applying the results we have just derived we find that the constraint in equation (B.3) is indeed satisfied, and the evaluation of the nested radical is $x + n + a$ as expected.

## B.3 Convergence of nested radicals

At this point, we introduce a more compact notation for nested radicals. For two sequences $a_n, b_n$ of positive real numbers, we define the operator $\mathbf{R}$ by

$$\mathbf{R}_{i=1}^{n}(a_i, b_i) = \sqrt{a_1 + b_1\sqrt{a_2 + \ldots + b_{n-1}\sqrt{a_n + b_n}}}$$

It was proved by Herschfeld ([2]) that $\mathbf{R}_{i=1}^{n}(a_i, 1)$ converges if $a_n^{2^{-n}}$ has a finite upper limit as $n$ tends to infinity.

Let $p_n, q_n$ be positive polynomial sequences, and let the sequence $r_n$ be given by

$$r_n = \mathbf{R}_{i=1}^{n}(p_i, q_i)$$

Then we wish to find whether or not $r_n$ converges. The next lemma will be of use.

**Lemma B.3.1.** Let $u_n, v_n, y_n, z_n$ be sequences of positive real numbers such that $u_i \leq y_i, v_i \leq z_i$ for all $i \in \mathbb{N}$. Then for all $n \in \mathbb{N}$

$$\mathbf{R}_{i=1}^{n}(u_i, v_i) \leq \mathbf{R}_{i=1}^{n}(y_i, z_i)$$

*Proof.* The result is a straight-forward consequence of the sequences being strictly positive. □

**Theorem B.3.2.** Let $p_n, q_n$ be positive polynomial sequences. Then the sequence $r_n = \mathbf{R}_{i=1}^{n}(p_i, q_i)$ converges.

*Proof.* Let $p(x), q(x) \in \mathbb{R}[x]$ be polynomials such that $p(i) = p_i, q(i) = q_i$ for all $i \in \mathbb{N}$, and let $m \in \mathbb{N}$ be such that $2m > \deg(p), m > \deg(q) + 1$. Then let $L(x) = x^m$ and $v(x) = x^{m-1}$, and define $u(x)$ by

$$\begin{aligned} u(x) &= L(x)^2 - v(x)L(x+1) \\ &= x^{2m} + O(x^{2m-1}) \end{aligned}$$

We further define the sequences $u_n, v_n$ by $u_i = u(i), v_i = v(i)$. Then there is some $k \in \mathbb{N}$ such that $p_j \leq u_j, q_j \leq v_j$ for all $j \geq k$. Hence, by Lemmas B.2.1

and B.3.1 , we have

$$\begin{aligned}\lim_{n\to\infty} \mathbf{R}_{i=k}^{n}(p_i, q_i) &\leq \lim_{n\to\infty} \mathbf{R}_{i=k}^{n}(u_i, v_i) \\ &= L(k) \\ &= k^m\end{aligned}$$

This provides a finite upper bound on $r_n$ by applying Lemma B.3.1 again with the finite sequences $\langle p_1, \ldots, p_{k-1}\rangle$ and $\langle q_1, \ldots, q_{k-2}, k^m q_{k-1}\rangle$.

Now, there is also some $k \in \mathbb{N}$ such that $p_j + q_j > 1$ for all $j \geq k$. That is

$$\mathbf{R}_{i=1}^{j}(p_i, q_i) \leq \mathbf{R}_{i=1}^{j+1}(p_i, q_i)$$

for all $j \geq k$, and hence $r_n$ converges to some finite limit. □

## References for Appendix B

[1] Borwein, J.M. and de Barra, G. (1991). Nested radicals. *American Mathematical Monthly*, **98** (8), 735–739.

[2] Herschfeld, A. (1935). On Infinite Radicals. *American Mathematical Monthly*, **42** (7), 419–429.

[3] McGuffin, M. and Wong, B. *The Museum of Infinite Nested Radicals.* `http://www.dgp.toronto.edu/~mjmcguff/math/nestedRadicals.html`

# Appendix C

# Bounds on the size of the TBR unit-neighbourhood

Peter J. Humphries

Abstract. In this paper, we study the unit-neighbourhood of the *tree bisection and reconnection* operation on unrooted binary phylogenetic trees. Specifically, we provide a recursive method to calculate the size of the unit-neighbourhood for any tree in the space $\mathscr{T}_n$ of unrooted binary phylogenetic trees with $n$-leaves. We also give both upper and lower bounds on this size for all trees in $\mathscr{T}_n$, and characterise those trees for which the stated upper bound is sharp.

## C.1 Introduction

Phylogenetic (evolutionary) trees are used to display the relationships between a set of objects. The techniques from phylogenetics are most commonly applied to computational biology to determine how different species or sets of species are interrelated. Due in part to the incompleteness of biological data, uncertainty often arises as to the 'true' tree that describes the speciation process.

Making local changes to a phylogenetic tree is referred to as a *tree rearrangement* operation. These operations were first introduced by Robinson [4] as a measure of the similarity between two unrooted trees having the same set of leaf labels. Since then, this notion has been extended in a number of ways for both rooted and unrooted trees [1, 2, 3]. Our focus in this paper is solely on the *tree bisection and reconnection* (TBR) operation.

A primary use of tree rearrangment operations in evolutionary biology is in modelling the effects of recombinantion or horizontal gene transfer. Perhaps more importantly for our purposes, the TBR operation induces a metric on the space of unrooted trees, and is used as the basis for heuristic

algorithms that search this space for the best tree under some optimisation constraints [2, 3].

In view of the algorithmic applications of TBR, an important question is how many different trees can be obtained by performing exactly one TBR operation on a given unrooted tree $\mathcal{T}$. That is, what is the size of the TBR *unit-neighbourhood* for $\mathcal{T}$. Upper and lower bounds for this problem may be found in [1]. The main results of this paper provide a sharp upper bound on the size of a TBR unit-neighbourhood and a characterisation of all trees satisfying this upper bound, as well as an improvement on the current best known lower bound. This work fills a gap in the literature, as the size of unit-neighbourhoods using other well-known tree rearrangement operations has been completely solved [1, 4].

The organisation of this paper is as follows. In Section C.2, we formalise the basic concepts used in the remainder of the paper, and then state the main theorems. Section C.3 presents the proofs of these theorems.

## C.2 Definitions and results

For the purposes of this paper, we are interested primarily in unrooted binary phylogenetic trees. That is, bijectively leaf-labelled trees without a specified root in which every interior vertex has degree three. We denote by $\mathscr{T}_n$ the set of all unrooted binary phylogenetic trees with the leaf set $\{1, \ldots, n\}$. Two trees in $\mathscr{T}_7$ are shown in Figure 1. Some use is also made of rooted binary phylogenetic trees, or bijectively leaf-labelled rooted binary trees.

A *cherry* is a pair of leaves $\{x, y\}$ adjacent to the same interior vertex. For example, in Figure 1, $\{3, 4\}$ is the only cherry common to both trees. If a tree $\mathcal{T} \in \mathscr{T}_n$ has exactly two cherries, where $n \geq 4$, then we refer to $\mathcal{T}$ as a *caterpillar*. For example, the tree on the right in Figure 1 is a seven-leafed caterpillar.

The TBR operation consists of two steps, namely the bisection and the reconnection. Given an unrooted binary phylogenetic tree $\mathcal{T}$, we remove any edge $e$ of $\mathcal{T}$ to give two subtrees $t_1, t_2$, contracting any vertices of degree two so that $t_1, t_2$ are both binary. We then reconnect these two subtrees by adding a new edge $f$ between the midpoints of some edge of $t_1$ and some edge of $t_2$. If either $t_1$ or $t_2$ consists of a single leaf, then this new edge $f$ is incident with the leaf. Figure C shows an example of a TBR operation on a seven-leafed tree.

The concept of the TBR unit-neighbourhood was mentioned in the preceding section. More completely, the TBR unit-neighbourhood of a tree $\mathcal{T} \in \mathscr{T}_n$ is the set $N(\mathcal{T})$ of all trees $\mathcal{T}' \in \mathscr{T}_n$ that can be obtained from $\mathcal{T}$ by a single TBR operation. We now have enough to state the main results of this paper.

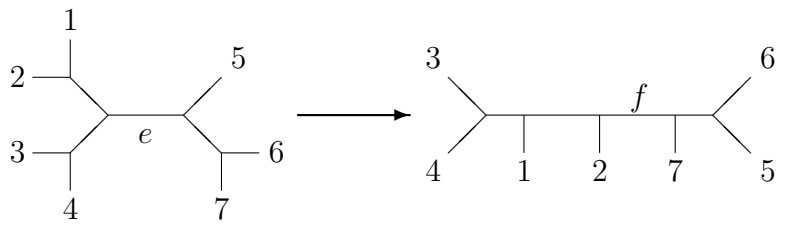


Figure C: An example of a TBR operation, where $e$ is the edge removed and $f$ is the edge added.

**Theorem C.2.1.** For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have

$$|N(\mathcal{T})| \leq \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2$$

Moreover, for $n \geq 6$, equality holds if and only if $\mathcal{T}$ is a caterpillar.

It was shown in [1] proved that the size of $N(\mathcal{T})$ for a tree in $\mathscr{T}_n$ is bounded above by $(2n-3)(n-3)^2$, which is of the same order as the expression in the above theorem. It was further shown in [5] that for rooted binary phylogenetic trees, the size of the unit-neighbourhood under the rooted subtree prune and regraft operation, which is considered to be the rooted analogue of TBR, is minimised when the tree is a caterpillar. This contrasts directly with Theorem C.2.1, and so we would expect in turn that for unrooted trees, the size of the TBR unit-neighbourhood is minimised for trees that are as balanced as possible. As yet, however, we have been unable to find a tight lower bound for the size of this unit-neighbourhood.

**Theorem C.2.2.** For all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$, we have

$$|N(\mathcal{T})| \geq 2n^2 - 8n + 2 + 2\sum_{i=1}^{n-4} l(i)$$

where $l(1) = 0$, and $l(i) = 2 + 4\sum_{j=2}^{i-1}\lfloor \log_2 j \rfloor$ otherwise.

Asymptotically, this bound is strictly larger than the current best known bound of $2(n-3)(2n-7)$ stated in [1]. The details of this may be found in Corollary C.3.5.

## C.3 Proofs of results

To prove our previously stated bounds on the size of the TBR unit neighbourhood, we exploit some ideas similar to those used by Song [4]. However,

without a root as a point of reference to orient unrooted trees, there are some modifications required. Instead, we apply an induction around a cherry of the tree in question.

For a rooted binary phylogenetic tree $\mathcal{R}$, let $\xi(\mathcal{R})$ denote the number of distinct ways to prune a subtree from $\mathcal{R}$ and root it to form $\mathcal{R}'$ so that $\mathcal{R} \neq \mathcal{R}'$.

**Lemma C.3.1.** Let $\mathcal{R}$ be a rooted binary phylogenetic tree with $n$ leaves, and let $\mathcal{R}_1$ and $\mathcal{R}_2$ be the two rooted trees obtained by deleting the root of $\mathcal{R}$. Then

$$\xi(\mathcal{R}) = \xi(\mathcal{R}_1) + \xi(\mathcal{R}_2) + 2n - 2.$$

*Proof.* There are three options to consider. We may take the rerooted subtree $\mathcal{R}'$ to be a rerooted subtree of either $\mathcal{R}_1$ or $\mathcal{R}_2$, which contributes exactly $\xi(\mathcal{R}_1) + \xi(\mathcal{R}_2)$. Alternatively, we may choose $\mathcal{R}' = \mathcal{R}_1$ or $\mathcal{R}' = \mathcal{R}_2$. Finally, we may reroot $\mathcal{R}$ on some edge not incident with the root of $\mathcal{R}$. The result follows by summing over these possibilities. □

**Theorem C.3.2.** Let $\mathcal{T}$ be an unrooted binary phylogenetic tree on $n$ leaves, $\{x, y\}$ be a cherry of $\mathcal{T}$, and let $\mathcal{T}'$ be the tree obtained by deleting $x$ from $\mathcal{T}$. Let $v$ be the unique interior vertex of $\mathcal{T}$ at distance 2 from $x$, and $\mathcal{R}_1, \mathcal{R}_2$ be the rooted trees not containing $x$ obtained by deleting $v$ from $\mathcal{T}$. Then

$$|N(\mathcal{T})| = |N(\mathcal{T}')| + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$$

*Proof.* We consider four classes of TBR operations on $\mathcal{T}$.

(1) Perform any TBR operation on $\mathcal{T}$ that retains $\{x, y\}$ as a cherry. There are precisely $|N(\mathcal{T}')|$ such TBR operations.

(2) Cut $x$ from $\mathcal{T}$ and reattach to any edge not adjacent to $y$. There are $2n - 6$ possible edges, and in no case is $\{x, y\}$ a cherry on the new tree.

(3) Cut $y$ from $\mathcal{T}$ and reattach to any edge not adjacent to $v$. If we attach $y$ to the edge $\{x, v\}$, we get a tree already formed in (1), and if we attach it to either of the other two edges incident with $v$ we get a tree formed in (2). There are $2n - 8$ possibilities here.

(4) Prune and reroot a subtree from either $\mathcal{R}_1$ or $\mathcal{R}_2$ to form $\mathcal{R}$, and then connect the root of $\mathcal{R}$ to either the edge adjacent to $x$ or the edge adjacent to $y$. None of these trees are formed by any of the above operations (1)-(3), and there are exactly $2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2)$ possibilities.

This covers all possible TBR operations on $\mathcal{T}$, and the result follows by summing the four quantities. □

This theorem can be used recursively to find the exact size of the TBR unit neighbourhood for any unrooted tree $\mathcal{T}$. However, we are primarily

interested in finding bounds on the size of $N(\mathcal{T})$ for a tree $\mathcal{T} \in \mathscr{T}_n$. To do so, we first find bounds on $\xi(\mathcal{R})$. Let $u(n)$ and $l(n)$ be the maximum and minimum values of $\xi(\mathcal{R})$ respectively over all rooted binary phylogenetic trees $\mathcal{R}$ with $n$ leaves. Applying Lemma C.3.1, we get

$$\begin{aligned} u(n) &= \max\{u(i) + u(n-i) : 1 \leq i \leq n-1\} + 2n - 2 \\ l(n) &= \min\{l(i) + l(n-i) : 1 \leq i \leq n-1\} + 2n - 2 \end{aligned}$$

with initial condition $u(1) = l(1) = 0$. We ideally want a closed form for both $u(n)$ and $l(n)$.

**Lemma C.3.3.** $u(n) = n^2 - n$ for all $n \geq 1$.

*Proof.* The lemma is true for $n = 1$, so assume $n > 1$ and that the lemma holds for all values strictly less than $n$. Then

$$\begin{aligned} u(n) &= \max\{u(i) + u(n-i) : 1 \leq i \leq n-1\} + 2n - 2 \\ &= u(n-1) + 2n - 2 \\ &= n^2 - n \end{aligned}$$

proving the lemma. □

**Lemma C.3.4.** $l(n) = 2 + 4\sum_{i=2}^{n-1} \lfloor \log_2 i \rfloor$ for all $n \geq 2$.

*Proof.* The lemma holds trivially for $n = 2$, so we assume that $n > 2$ and that it is true for all values less than $n$. We note that $l(j) > l(j-1)$, and also that $l(j) - l(j-1) \geq l(j-1) - l(j-2)$. Hence, if $l(i) + l(n-i)$ is to be minimised, we want $i$ and $n-i$ to be equal or to differ by one. Let $m = \lfloor \frac{n}{2} \rfloor$. Then

$$\begin{aligned} l(n) &= \min\{l(i) + l(n-i) : 1 \leq i \leq n-1\} + 2n - 2 \\ &= l(m) + l(n-m) + 2n - 2 \\ &= l(m) + l(n-m-1) + 2n - 2 + 4\lfloor \log_2(n-m-1) \rfloor \\ &= l(n-1) + 4 + 4\lfloor \log_2(n-m-1) \rfloor \\ &= 2 + 4\sum_{i=2}^{n-2} \lfloor \log_2 i \rfloor + 4\lfloor \log_2(n-1) \rfloor \\ &= 2 + 4\sum_{i=2}^{n-1} \lfloor \log_2 i \rfloor \end{aligned}$$

since $l(2) = 2$ is our boundary condition. We remark that the penultimate line in the working above follows from the definition of $m$. □

These two bounds now allow us to prove Theorems C.2.1 and C.2.2. Let $U(n)$ and $L(n)$ denote upper and lower bounds on the size of $N(\mathcal{T})$, where $\mathcal{T}$ is an unrooted binary phylogenetic tree with $n$ leaves. That is, $U(n) = \max\{|N(\mathcal{T})| : \mathcal{T} \in \mathscr{T}_n\}$ and $L(n) = \min\{|N(\mathcal{T})| : \mathcal{T} \in \mathscr{T}_n\}$.

*Proof of Theorem C.2.1.* It can be checked that the theorem holds for all $n \leq 6$, so assume that $n \geq 7$ and let $\mathcal{T}$ be an $n$-leafed caterpillar and $\mathcal{T}'$ be an $(n-1)$-leafed caterpillar. Then, by Theorem C.3.2 and Lemma C.3.3,

$$\begin{aligned}|N(\mathcal{T})| &= |N(\mathcal{T}')| + 4n - 14 + 2u(1) + 2u(n-3) \\ &= U(n-1) + 4n - 14 + 2\left((n-3)^2 - (n-3)\right) \\ &= \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2\end{aligned}$$

Suppose instead that $\mathcal{T} \in \mathscr{T}_n$ is not a caterpillar, and that $\mathcal{T}'$ is a tree obtained by deleting a single leaf from a cherry of $\mathcal{T}$. Note that we can do this in such a way that $\mathcal{T}'$ is not a caterpillar, and further that $u(i) + u(n-i-2)$ is maximised if $i = 1$. Then

$$\begin{aligned}|N(\mathcal{T})| &\leq |N(\mathcal{T}')| + 4n - 14 + 2u(1) + 2u(n-3) \\ &< U(n-1) + 4n - 14 + 2u(1) + 2u(n-3) \\ &= U(n)\end{aligned}$$

completing the proof of the theorem. □

*Proof of Theorem C.2.2.* The theorem holds when $n \leq 5$, so we assume that $n \geq 6$. Let $\mathcal{T} \in \mathscr{T}_n$ be a tree for which the size of the unit-neighbourhood is minimised, and let $\{x, y\}$ be a cherry of $\mathcal{T}$ that is at the end of a longest path. Taking $\mathcal{R}_1, \mathcal{R}_2$ as in Theorem C.3.2, we may assume that $\mathcal{R}_1$ has at most two leaves. Applying Theorem C.3.2 and Lemma C.3.4 we have

$$\begin{aligned}|N(\mathcal{T})| &\geq L(n-1) + 4n - 14 + 2\xi(\mathcal{R}_1) + 2\xi(\mathcal{R}_2) \\ &\geq 2n^2 - 8n - 2 + 2\sum_{i=1}^{n-5} l(i) + 2\min\{l(1) + l(n-3), l(2) + l(n-4)\}\end{aligned}$$

However, for all $n \geq 6$ we have $l(1) + l(n-3) \geq l(2) + l(n-4)$, and so

$$|N(\mathcal{T})| \geq 2n^2 - 8n + 2 + 2\sum_{i=1}^{n-4} l(i)$$

thus proving the theorem. □

Asymptotically, Theorem C.2.2 gives a better lower bound on the size of the unit-neighbourhood. Recall that the previous lower bound on the size of $N(\mathcal{T})$ for $\mathcal{T} \in \mathscr{T}_n$ was quadratic in $n$. Corollary C.3.5 improves this to a function of order $O(n^2 \log n)$.

**Corollary C.3.5.** There is some $c > 0$ such that, for all $n \geq 4$ and all $\mathcal{T} \in \mathscr{T}_n$,

$$|N(\mathcal{T})| \geq cn^2 \log n.$$

*Proof.* From Theorem C.2.2, the size of the unit-neighbourhood $N(\mathcal{T})$ is bounded by a double sum of logarithms, so that for some $c > 0$,

$$\begin{aligned} |N(\mathcal{T})| &\geq 2c \sum_{i=1}^{n} \sum_{j=1}^{i} \log j \\ &> 2c \int_1^n \int_1^x \log y dy dx \\ &= cn^2 \log n + O(n^2). \end{aligned}$$

□

We remark that the result may also be proved by way of generating functions. Let $G_l(x)$ be the ordinary generating function for the sequence $l(n)$, and let $G_f(x)$ be the ordinary generating function for the sequence $f(n)$, where $f(4) = 2$ and $f(n) = 2n^2 - 8n + 2 + 2\sum_{i=1}^{n-4} l(i)$ for $n \geq 5$. Then it can be shown that

$$G_l(x) = \frac{1}{(1-x)^2} \left[ 2x^2(1+x) + 4x \sum_{k=2}^{\infty} x^{2^k} \right],$$

provided we define $l(0) = 0$, and further that

$$\begin{aligned} G_f(x) &= \frac{2x^4}{1-x} \left[ \frac{1 + 3x - 2x^2}{(1-x)^2} + G_l(x) \right] \\ &= \frac{2x^4}{(1-x)^3} \left[ 1 + 3x - 2x^4 + 4x \sum_{k=2}^{\infty} x^{2^k} \right]. \end{aligned}$$

Extracting the coefficient of $x^n$ in $G_f(x)$ gives the same asymptotic bound on the size of the unit-neighbourhood as we established in Corollary C.3.5.

## References for Appendix C

[1] Allen, B.L. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, **5**, 1–15.

[2] Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination, *Journal of Molecular Evolution*, **36**, 369–405.

[3] Maddison, D. R. (1991). The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology*, **43** (3), 315–328.

[4] Robinson, D.F. (1971). Comparison of Labeled Trees with Valency Three. *Journal of Combinatorial Theory*, **11**, 105–119.

[5] Song, Y.S. (2003). On the Combinatorics of Rooted Binary Phylogenetic trees. *Annals of Combinatorics*, **7**, 365–379.

# Index