Compatible split systems on a multiset

Vincent Moulton[a] Guillaume E. Scholz[b]

Submitted: Jan 14, 2022; Accepted: Nov 2, 2024; Published: Dec 17, 2024

Abstract

A split system on a multiset $\mathcal{M}$ is a multiset of bipartitions of $\mathcal{M}$. Such a split system $\mathfrak{S}$ is compatible if it can be represented by a tree in such a way that the vertices of the tree are labelled by the elements in $\mathcal{M}$, the removal of each edge in the tree yields a bipartition in $\mathfrak{S}$ by taking the labels of the two resulting components, and every bipartition in $\mathfrak{S}$ can be obtained from the tree in this way. Compatibility of split systems is a key concept in phylogenetics, and compatible split systems have applications to, for example, multi-labelled phylogenetic trees. In this contribution, we present a novel characterization for compatible split systems, and for split systems admitting a unique representation by a tree. In addition, we show that a conjecture on compatibility stated in 2008 holds for some large classes of split systems.

Mathematics Subject Classifications: 03E02,05C05

1 Introduction

Let $\mathcal{M}$ be a multiset with underlying set X. For $x \in X$, we denote by $\mathcal{M}(x) \geqslant 1$ the multiplicity of x in $\mathcal{M}$, and we put $\Delta(\mathcal{M}) = \sum_{x\in X}(\mathcal{M}(x) - 1)$. To ease notation, we sometimes write $a_1a_2 \dots a_n$ for a multiset $\{a_1, a_2, \dots, a_n\}$, and if an element a_i has multiplicity $k > 1$, then we also denote this by writing a_i^k. We denote by $\mathcal{M}^* \subseteq X$ the set of elements of X with multiplicity 1 in $\mathcal{M}$, that is, $\mathcal{M}^* = \{x \in X : \mathcal{M}(x) = 1\}$. Similarly, for $A \subseteq \mathcal{M}$, we denote by A^* the set of elements of A with multiplicity 1 in $\mathcal{M}$, that is, $A^* = A \cap \mathcal{M}^*$. Note that set operations and inclusion relations on multisets are defined in the usual way.

A *split* (or *bipartition*) S of $\mathcal{M}$ is a pair $\{A, B\}$ such that A, B are nonempty sub(multi)sets of $\mathcal{M}$, and the union $A \cup B$ is precisely $\mathcal{M}$. We usually write $S = A|B$ (or $S = B|A$, as the roles of A and B are symmetric). When the set $\mathcal{M}$ is clear from the

[a]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK (v.moulton@uea.ac.uk).
[b]Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, Universität Leipzig, D-04107 Leipzig, Germany (guillaume@bioinf.uni-leipzig.de).

context, we also sometimes write $A|\overline{A}$ instead of $A|B$, where $\overline{A} = \mathcal{M} - A$. A non-empty multiset $\mathfrak{S}$ of splits $A|B$ of $\mathcal{M}$ is called a *split system* on $\mathcal{M}$.

A *labeled tree* is a pair $\mathcal{T} = (T, \lambda)$ such that T is a tree (that is, an undirected, connected acyclic graph), and λ is a map from the vertex set $V(T)$ of T to $\mathcal{P}(\mathcal{M})$, the power set of $\mathcal{M}$. An $\mathcal{M}$-*tree* is a labeled tree (T, λ) satisfying the following two properties:

(M1) The union $\cup_{v \in V(T)} \lambda(v)$ is $\mathcal{M}$.

(M2) All vertices v of T of degree 1 or 2 satisfy $\lambda(v) \neq \emptyset$.

For $v \in V(T)$, we call $\lambda(v)$ the label of v, and we say that an element $a \in \mathcal{M}$ (*resp.* a subset $A \subseteq \mathcal{M}$) *labels* v if $a \in \lambda(v)$ (*resp.* $A \subseteq \lambda(v)$). Abusing terminology, we also sometimes call a vertex (*resp.* edge) of T a vertex (*resp.* edge) of $\mathcal{T}$. For example, the labeled tree depicted in Figure 1 is an $\mathcal{M}$-tree for $\mathcal{M} = \{a^2, b^2, c^2, x, y\}$. We say that two $\mathcal{M}$-trees $\mathcal{T} = (T, \lambda)$ and $\mathcal{T}' = (T', \lambda')$ are *isomorphic* if there exists a graph isomorphism $\phi : V(T) \to V(T')$ such that $\lambda'(\phi(v)) = \lambda(v)$ for all $v \in V(T)$.

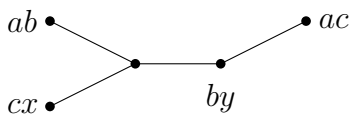

Figure 1: An $\mathcal{M}$-tree representing the split system $\{ab|\overline{ab}, ac|\overline{ac}, cx|\overline{cx}, abcx|abcy\}$ on $\mathcal{M} = \{a^2, b^2, c^2, x, y\}$. The labels of each vertex are indicated next to the vertex. The union of all labels is $\mathcal{M}$, so (M1) is satisfied. Moreover, all vertices of degree 1 or 2 have a nonempty label, so (M2) is satisfied.

Let $\mathcal{T} = (T, \lambda)$ be an $\mathcal{M}$-tree. Since T is a tree, the removal of an edge e from T results in a graph with exactly two connected components T_1 and T_2. Putting $A = \bigcup_{v \in V(T_1)} \lambda(v)$ and $B = \bigcup_{v \in V(T_2)} \lambda(v)$, it follows from (M1) that $A|B$ is a split of $\mathcal{M}$, which we call the split *induced* by e. The multiset $\mathfrak{S}(\mathcal{T})$ of splits obtained by taking the union of the splits associated to the edges of T is called the split system *represented* by $\mathcal{T}$. We say that a split system $\mathfrak{S}$ on $\mathcal{M}$ is *compatible* if there exists an $\mathcal{M}$-tree $\mathcal{T} = (T, \lambda)$ representing $\mathfrak{S}$, that is, satisfying $\mathfrak{S} = \mathfrak{S}(\mathcal{T})$. Otherwise, we say that $\mathfrak{S}$ is *incompatible*. For example, the split system $\{ab|\overline{ab}, ac|\overline{ac}, cx|\overline{cx}, abcx|abcy\}$ on $\mathcal{M} = a^2b^2c^2xy$ is compatible, and admits the $\mathcal{M}$-tree depicted in Figure 1 as a representation.

It is not difficult to see that if a split system $\mathfrak{S}$ is compatible, then it is *pairwise compatible*, that is, for any pair of splits $S_1, S_2 \in \mathfrak{S}$ there is some $A \in S_1$ and $B \in S_2$ such that $A \cap B = \emptyset$. In 1971, Buneman proved the following corner-stone result concerning compatibility [1]:

Theorem 1 ([1])**.** *Let $\mathfrak{S}$ be a set of splits of a set X. Then $\mathfrak{S}$ is compatible if and only if $\mathfrak{S}$ is pairwise compatible. Moreover, if $\mathfrak{S}$ is compatible, then there exists a unique (up to isomorphism) X-tree $\mathcal{T}$ representing $\mathfrak{S}$.*

However, as was remarked in [8, p.640], the equivalence stated in the last theorem does not necessarily hold for split systems on multisets. For example, the split system $\mathfrak{S} = \{ab|acd, ac|abd, ad|abc\}$ on $\mathcal{M} = \{a^2, b, c, d\}$ is pairwise compatible, but it is not compatible. Moreover, in [4, Fig. 5] it was shown that the uniqueness condition in the theorem can also fail to hold in general.

Despite these issues, in [8] it was shown that compatibility of a split system on a multiset $\mathcal{M}$ can in fact be characterised by taking into account the quantity $\Delta(\mathcal{M})$:

Theorem 2 ([8], Thm. 4.3)**.** *Let $\mathfrak{S}$ be a split system on a multiset $\mathcal{M}$. Then $\mathfrak{S}$ is compatible if and only if every submultiset of $\mathfrak{S}$ of size at most* $\max\{2\Delta(\mathcal{M}), \Delta(\mathcal{M}) + 2\}$ *is compatible.*

Note that the last result generalizes the first statement of Theorem 1, since in case $\mathcal{M}$ and $\mathfrak{S}$ are both sets, $\Delta(\mathcal{M}) = 0$. In addition, in [8, Remark 4.7] the authors asked whether or not the following statement holds:

Conjecture 3. Let $\mathfrak{S}$ be a split system on a multiset $\mathcal{M}$. Then,

($\bigstar$) $\mathfrak{S}$ is compatible if and only if every submultiset of $\mathfrak{S}$ of size at most $\Delta(\mathcal{M}) + 2$ is compatible.

Note that in [8, Remark 4.7], the authors remarked that they had checked that this conjecture holds for every multiset which contains at most three elements that have multiplicity greater than one.

This contribution is organized as follows. In Section 2, we introduce the split-containment graph $\Gamma(\mathfrak{S})$ of a split system $\mathfrak{S}$. We then use that graph to state a characterization of compatibility of a split system for the multiset case (Theorem 9). In addition, we show that $\Gamma(\mathfrak{S})$ can be used to count the number of non-isomorphic tree representations of a compatible split system. This gives rise to a characterization of compatible split systems admitting a unique representation (Corollary 10), thus providing an answer to the question raised in [8, Remark 4.5 (b)].

In Sections 3 to 5, we turn our attention to Conjecture 3 which remains open. More specifically, we show that for a split system $\mathfrak{S}$ on some multiset $\mathcal{M}$, the equivalence ($\bigstar$) in Conjecture 3 holds in case (1) the graph $\Gamma(\mathfrak{S})$ enjoys a sparsity property (Theorem 11), and (2) all of the splits in $\mathfrak{S}$ have the same size, where the size of a split $A|B$ equals $\min\{|A|, |B|\}$ (Theorem 12). In addition, we show that in case all of the splits in a split system $\mathfrak{S}$ have size at most 3, then $\mathfrak{S}$ satisfies a slightly weaker version of ($\bigstar$), in which $\Delta(\mathcal{M}) + 2$ is replaced by $\Delta(\mathcal{M}) + 3$ (Theorem 22).

Before proceeding, we remark that in case $\mathcal{M}$ is a set, $\mathcal{M}$-trees and their relationship with compatible split systems form a fundamental part of the underlying theory for the area of phylogenetics (see e.g. [14, Chapter 3]). Moreover, $\mathcal{M}$-trees for $\mathcal{M}$ a multiset are closely related to *multi-labeled phylogenetic trees* (or *MUL-trees* for short), that arise in the context of polyploid studies, tree-reconciliation and phylogenetic network theory (see e.g. [10, 12, 3, 13]).

Roughly speaking, MUL-trees are rooted trees equipped with a (not necessarily injective) function from their leaf set to some set X. Understanding mathematical properties

of MUL-trees is an active area of research; for example, see [16] for recent work on the relationship between MUL-trees with tree shapes, and [9] for interesting connections with so-called ploidy profiles. In addition, the compatibility problem raises interesting related algorithmic questions and results; see e.g. [5, 7] for recent work in this area. Other structures related to MUL-trees include *tangled trees* [11], introduced as a way to model host-parasite co-evolution, and *area cladograms* [4], that are used in biogeographical studies. The compatibility problem for split systems is also closely related to the so-called perfect phylogeny haplotyping problem [6] – see for example [2].

2 The split-containment graph

As we have seen in the introduction, pairwise compatibility of a split system does not necessarily imply compatibility. However, in this section we show that we can characterize compatibility in terms of a certain graph that we shall associate to a split system. As we shall see, this graph also yields a characterization for when a compatible split system is represented by a unique tree.

We begin by defining the graph. For a split system $\mathfrak{S} = \{S_1, \ldots, S_n\}$, $n \geqslant 1$, we define the *split-containment graph* $\Gamma(\mathfrak{S})$ of $\mathfrak{S}$ as follows. The vertex set of $\Gamma(\mathfrak{S})$ is the multiset $\{(A, S_i) : A \in S_i, 1 \leqslant i \leqslant n\}$, and the arc set of $\Gamma(\mathfrak{S})$ is the multiset of ordered pairs $((A, S_i), (B, S_j))$ satisfying $A \subsetneq B$ (as multisets) and $i, j \in \{1, \ldots, n\}$ distinct. Note that since $\mathfrak{S}$ is a multiset, $S_i = S_j$ may hold.

Clearly the graph $\Gamma(\mathfrak{S})$ is acyclic. Moreover, that graph satisfies the property that if $((A, S_i), (B, S_j))$ is an arc of $\Gamma(\mathfrak{S})$, then $((\overline{B}, S_j), (\overline{A}, S_i))$ is an arc of $\Gamma(\mathfrak{S})$. It follows that for any i, j distinct, $\Gamma(\{S_i, S_j\})$ either contains no arcs or it is isomorphic to one of the digraphs D_1 or D_2, where D_1 and D_2 are digraphs on four vertices $\{v, w, p, q\}$ such that D_1 has arcs $(v, p), (q, w)$ and D_2 has arcs $(v, p), (p, w), (v, q), (q, w)$ (see Figure 2).

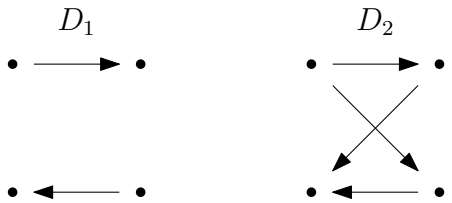

Figure 2: For $\mathfrak{S} = \{S_1, \ldots, S_n\}$ a split system and $i, j \in \{1, \ldots, n\}$ distinct, either $\Gamma(\{S_i, S_j\})$ has no arcs, or $\Gamma(\{S_i, S_j\})$ is isomorphic to one of the depicted graphs D_1 or D_2.

In view of these observations, we obtain the following lemma:

Lemma 4. *Suppose $\mathfrak{S} = \{S_1, S_2\}$ is a split system on $\mathcal{M}$. Then the following are equivalent:*

(i) $\mathfrak{S}$ is compatible.

(ii) $\Gamma(\mathfrak{S})$ *is isomorphic to* D_1 *or* D_2*.*

(iii) The arc set of $\Gamma(\mathfrak{S})$ *is non-empty.*

Moreover, if $\mathfrak{S}$ *is compatible, then there is a unique* $\mathcal{M}$*-tree representing* $\mathfrak{S}$ *if and only if* $\Gamma(\mathfrak{S})$ *is isomorphic to* D_1*.*

In light of the last lemma, we call a subgraph G of $\Gamma(\mathfrak{S})$ *thin* if $V(G) = V(\Gamma(\mathfrak{S}))$ and the restriction of G to $\{(A, S_i), (\overline{A}, S_i), (B, S_j), (\overline{B}, S_j)\}$ is isomorphic to D_1 for all $i, j \in \{1, \ldots, n\}$ distinct. We have:

Lemma 5. *If* G *is a thin subgraph of* $\Gamma(\mathfrak{S})$*, then for any two splits* $S_i = A|\overline{A}$*,* $S_j = B|\overline{B}$*,* $i \neq j$ *of* $\mathfrak{S}$*, there exists exactly one arc in* G *from an element of* $\{(A, S_i), (\overline{A}, S_i)\}$ *to an element of* $\{(B, S_j), (\overline{B}, S_j)\}$*.*

Proof. Since G is thin, the graph induced by G on the set $\{(A, S_i), (\overline{A}, S_i), (B, S_j), (\overline{B}, S_j)\}$ is isomorphic to D_1. By definition of G as a subgraph of $\Gamma(\mathfrak{S})$, there is no arc in G between (A, S_i) and $(\overline{A}, S_i)$, and no arc between (B, S_j) and $(\overline{B}, S_j)$. Hence, G contains exactly one arc from an element of $\{(A, S_i), (\overline{A}, S_i)\}$ to an element of $\{(B, S_j), (\overline{B}, S_j)\}$, and one arc from an element of $\{(B, S_j), (\overline{B}, S_j)\}$ to an element of $\{(A, S_i), (\overline{A}, S_i)\}$. □

We say that $\mathfrak{S}$ is *thin* if $\Gamma(\mathfrak{S})$ is a thin subgraph of itself. In particular, if $\mathcal{M}$ is a set, then a split system $\mathfrak{S}$ on $\mathcal{M}$ is thin if and only if $\mathfrak{S}$ is compatible. Note that there exist split systems $\mathfrak{S}$ on multisets that are thin but not compatible, and compatible but not thin. For example, the split system $\{xx|xyz, xy|xxz, xz|xxy\}$ on $\mathcal{M} = \{x^3, y, z\}$ is thin but not compatible, and the split system $\{ab|\overline{ab}, ac|\overline{ac}, cx|\overline{cx}, abcx|abcy\}$ on $\mathcal{M} = \{a^2, b^2, c^2, x, y\}$ represented by the $\mathcal{M}$-tree depicted in Figure 1 is not thin.

We now consider the structure of thin subgraphs of $\Gamma(\mathfrak{S})$ in more detail. For G a thin subgraph of $\Gamma(\mathfrak{S})$, we call an arc $((A, S_i), (B, S_j))$, $i \neq j$, of G *critical* if there does not exist a directed path from (A, S_i) to (B, S_j) in G other than the path formed by the single arc $((A, S_i), (B, S_j))$. Note that if $((A, S_i), (B, S_j))$ is a critical arc of G, then the corresponding arc $((\overline{B}, S_j), (\overline{A}, S_i))$ is also a critical arc of G. Moreover, we have:

Lemma 6. *Let* G *be a thin subgraph of* $\Gamma(\mathfrak{S})$*. If two vertices* (A, S_i)*,* (B, S_j)*,* $i \neq j$ *of* G *are joined by a directed arc, then there is a path from* (A, S_i) *to* (B, S_j) *that contains only critical arcs.*

Proof. If $((A, S_i), (B, S_j))$ is a critical arc, then the result holds. Otherwise, there exists a path P from (A, S_i) to (B, S_j) that does not contain the arc $((A, S_i), (B, S_j))$. By the same argument, each arc of P that is not critical can be replaced by a path, and one can recursively apply this arc-replacement process until all arcs on the resulting path are critical. □

Note however that the converse of Lemma 6 does not necessarily hold.

Next, we call a thin subgraph G of $\Gamma(\mathfrak{S})$ *consistent* if

$$\bigcup_{(B,S_j)\in\mathcal{C}_G((A,S_i))} B \subseteq A,$$

for all $(A, S_i) \in V(G)$, where $\mathcal{C}_G((A, S_i))$ is the set of all vertices (B, S_j) of G such that $((B, S_j), (A, S_i))$ is a critical arc of G.

Interestingly, consistent thin subgraphs are uniquely determined by their critical arcs. Indeed, we have:

Lemma 7. *Let $\mathfrak{S} = \{S_1, \ldots, S_n\}$, $n \geqslant 1$, be a split system on a multiset $\mathcal{M}$, and let G be a consistent thin subgraph of $\Gamma(\mathfrak{S})$. Then two vertices $(A, S_i), (B, S_j)$, $i, j \in \{1, \ldots, n\}$ distinct, of G are joined by an arc if and only if there is a path in G from (A, S_i) to (B, S_j) that contains only critical arcs.*

Proof. One direction comes directly from Lemma 6: If $((A, S_i), (B, S_j))$ is an arc of G, then either $((A, S_i), (B, S_j))$ is a critical arc of G, or there is a path in G from (A, S_i) to (B, S_j) of length two or more, such that all arcs on this path are critical arcs.

Conversely, suppose that there exists a path in G from (A, S_i) to (B, S_j) that contains only critical arcs. In particular, $A \subsetneq B$ holds. Since G is thin, Lemma 5 implies that exactly one of the ordered pairs $((A, S_i), (B, S_j))$, $((B, S_j), (A, S_i))$, $((\overline{A}, S_i), (B, S_j))$, and $((B, S_j), (\overline{A}, S_i))$ must be an arc of G. So it suffices to show that for the last three pairs this is impossible.

If $((B, S_j), (A, S_i))$ is an arc of G, then $B \subsetneq A$ holds, which is impossible given that $A \subsetneq B$ holds.

If $((\overline{A}, S_i), (B, S_j))$ is an arc of G, then in view of Lemma 6, there exists a path P_1 from $(\overline{A}, S_i)$ to (B, S_j) in G that contains only critical arcs. By assumption, there also exists a path P_2 in G from (A, S_i) to (B, S_j) that contains only critical arcs. Now, let (C, S_k) be the first vertex that is common to P_1 and P_2. Such a vertex must exist, since P_1 and P_2 have the same end-vertex (B, S_j). Since G is consistent, and the paths from $(\overline{A}, S_i)$ to (C, S_k) and from (A, S_i) to (C, S_k) are vertex-disjoint by choice of (C, S_k), it follows that $\mathcal{M} = \overline{A} \cup A \subseteq C$, which is impossible.

Finally, if $((B, S_j), (\overline{A}, S_i))$ is an arc of G, then in view of Lemma 6, there exists a path P_1 from (B, S_j) to $(\overline{A}, S_i)$ in G that contains only critical arcs. Since by assumption, there exists a path in G from (A, S_i) to (B, S_j) that contains only critical arcs, there must exist a path P_2 in G from $(\overline{B}, S_j)$ to $(\overline{A}, S_i)$ with the same property. Now, let (C, S_k) be the first vertex that is common to P_1 and P_2. Such a vertex must exist, since P_1 and P_2 have the same end-vertex $(\overline{A}, S_i)$. Since G is consistent, and the paths from (B, S_j) to (C, S_k) and from $(\overline{B}, S_j)$ to (C, S_k) are vertex-disjoint by choice of (C, S_k), it follows that $\mathcal{M} = B \cup \overline{B} \subseteq C$, which is impossible. □

In addition, the existence of certain pairs of critical arcs in a consistent thin subgraph forces the existence of further critical arcs, as summarized in Figure 3:

Lemma 8. *Let $\mathfrak{S} = \{S_1, \ldots, S_n\}$, $n \geqslant 1$, be a split system on a multiset $\mathcal{M}$, and let G be a consistent thin subgraph of $\Gamma(\mathfrak{S})$. If $i, j, k \in \{1, \ldots, n\}$ distinct are such that $((\overline{A}, S_i), (B, S_j))$ and $((\overline{C}, S_k), (B, S_j))$ are critical arcs of G, then $((\overline{A}, S_i), (C, S_k))$ is a critical arc of G.*

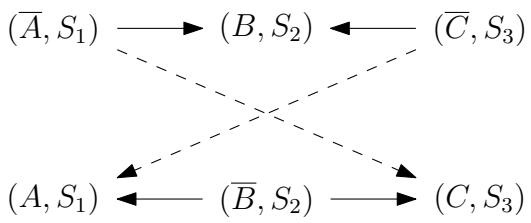

Figure 3: For $S_1 = A|\overline{A}$, $S_2 = B|\overline{B}$ and $S_3 = C|\overline{C}$, a thin subgraph G of $\Gamma(\mathfrak{S})$, where $\mathfrak{S} = \{S_1, S_2, S_3\}$. By Lemma 8, if G is consistent, and the four solid arcs are critical arcs of G, then the dashed arcs must be critical arcs of G.

Proof. We first show that $((\overline{A}, S_i), (C, S_k))$ is an arc of G. Since G is thin, it follows from Lemma 5 that exactly one of the ordered pairs $((\overline{A}, S_i), (C, S_k))$, $((\overline{A}, S_i), (\overline{C}, S_k))$, $((C, S_k), (\overline{A}, S_i))$ and $((\overline{C}, S_k), (\overline{A}, S_i))$ must be an arc of G. Clearly, $((\overline{A}, S_i), (\overline{C}, S_k))$ is not an arc of G, as in this case $((\overline{A}, S_i), (B, S_j))$ is not a critical arc of G. By symmetry, $((\overline{C}, S_k), (\overline{A}, S_i)$ is not an arc of G either. It is also impossible to have $((C, S_k), (\overline{A}, S_i))$ an arc in G. Indeed, $\mathfrak{S}$ is consistent, so B contains the union $\overline{A} \cup \overline{C}$. If $C \subseteq \overline{A}$ holds, then it follows that $\mathcal{M} = C \cup \overline{C} \subseteq A \cup \overline{C} \subseteq B$, which is impossible. Hence, $((\overline{A}, S_i), (C, S_k))$ is an arc of G.

Now, assume for contradiction that the arc $((\overline{A}, S_i), (C, S_k))$ is not critical. In view of Lemma 6, there exists a path from $(\overline{A}, S_i)$ to (C, S_k) that only contains critical arcs. Let (D, S_l) be the last vertex of that path before (C, S_k). In particular, $((D, S_l), (C, S_k))$ is a critical arc of G. Since G is thin, it follows from Lemma 5 that exactly one of $((B, S_j), (D, S_l))$, $((D, S_l), (\overline{B}, S_j))$, $((\overline{B}, S_j), (D, S_l))$ and $((D, S_l), (B, S_j))$ must be an arc of G. We next show that none of these arcs can be an arc of G.

Since G is consistent, C contains the multiset union $D \cup \overline{B}$. Hence, $B \subseteq D$ cannot hold, as this would imply $\mathcal{M} \subseteq C$, so $((B, S_j), (D, S_l))$ is not an arc of G.

If $((D, S_l), (\overline{B}, S_j))$ is an arc of G, then this contradicts the assumption that $((D, S_l), (C, S_k))$ is critical. Similarly, if $((\overline{B}, S_j), (D, S_l))$ is an arc of G, then this contradicts the assumption that $((\overline{B}, S_j), (C, S_k))$ is critical. So, neither $((D, S_l), (\overline{B}, S_j))$ nor $((\overline{B}, S_j), (C, S_k))$ are arcs of G.

Finally, suppose $((D, S_l), (B, S_j))$ is an arc of G. By choice of (D, S_l), there is a directed path in G from $(\overline{A}, S_i)$ to (D, S_l). Since $\Gamma(\mathfrak{S})$ is acyclic, this path does not contain the arc $((\overline{A}, S_i), (B, S_j))$. But this contradicts the assumption that $((\overline{A}, S_i), (B, S_j))$ is critical. Hence, $((D, S_l), (B, S_j))$ is not an arc of G.

In summary, the existence of a path from $(\overline{A}, S_i)$ to (C, S_k), that contains only critical arcs is impossible if $((\overline{A}, S_i), (C, S_k))$ is not critical itself. Hence, $((\overline{A}, S_i), (C, S_k))$ is a critical arc of G. □

We are now in a position to give a characterization for when a split system $\mathfrak{S}$ is compatible in terms of the existence of consistent thin subgraphs of $\Gamma(\mathfrak{S})$.

Theorem 9. *Let $\mathfrak{S}$ be a split system on a multiset $\mathcal{M}$. Then $\mathfrak{S}$ is compatible if and only if there exists a consistent thin subgraph G of $\Gamma(\mathfrak{S})$.*

Proof. Let $\mathfrak{S} = \{S_1, \dots S_n\}$, $n \geqslant 1$. Assume first that $\mathfrak{S}$ is compatible. Let $\mathcal{T} = (T, \lambda)$ be a representation of $\mathfrak{S}$. By definition, there exists a bijection $e : \mathfrak{S} \to E(T)$ such that for all $i \in \{1, \dots n\}$, the graph $T - e(S_i)$ has two connected components, each labeled with one part of S_i. We construct G from $\mathcal{T}$ as follows: The vertices of G are the vertices of $\Gamma(\mathfrak{S})$. The arcs of G are the arcs $((A, S_i), (B, S_j))$ such that the connected component of $T - e(S_i)$ labeled with A is a subgraph of the connected component of $T - e(S_j)$ labeled with B. As $\mathcal{T}$ is an $\mathcal{M}$-tree this definition implies that $A \subsetneq B$ for all arcs $((A, S_i), (B, S_j))$ of G, so G is a subgraph of $\Gamma(\mathfrak{S})$. Moreover, for any two $i, j \in \{1, \dots, n\}$ distinct, the subgraph of G induced by $\{(A, S_i), (\overline{A}, S_i), (B, S_j), (\overline{B}, S_j)\}$ contains exactly two arcs, so G is thin.

To see that G is consistent, let (A, S_i) be a vertex of G. Let T_A be the connected component of $T - e(S_i)$ labelled with A, and let v be the vertex of T_A adjacent to $e(S_i)$. We claim that an arc $((C, S_k), (A, S_i))$, $k \neq i$ of G is critical if and only if the edge $e(S_k)$ is adjacent to v.

To see that the claim holds, let $((C, S_k), (A, S_i))$ be an arc of G. By construction, $e(S_k)$ is an edge of T_A, and the connected component of $T - e(S_k)$ labelled with C is the connected component that does not contain v. Suppose first that $e(S_k)$ is not adjacent to v. Then there exists a path in T_A from $e(S_k)$ to v. Let e be an edge of that path, and let $S_j \in \mathfrak{S}$ be the split induced by $T - e$. Finally, let B be the part of S_j labelling the connected component of $T - e$ that does not contain v. Then by construction, $((C, S_k), (B, S_j))$ and $((B, S_j), (A, S_i))$ are both arcs of G, so $((C, S_k), (A, S_i))$ is not critical. Conversely, suppose that $((C, S_k), (A, S_i))$ is not critical. By Lemma 6, there exists a path from (C, S_k) to (A, S_i) in G that contains only critical arcs. Let (B, S_j) be the last vertex of that path before (A, S_i). Since $((B, S_j), (A, S_i))$ is an arc of G, $e(S_j)$ is an edge of T_A, and the connected component of $T - e(S_j)$ labelled with B is the connected component that does not contain v. Moreover, (C, S_k) is an ancestor of (B, S_j) in G, so $e(S_k)$ belongs to the connected component of $T - e(S_j)$ labelled with B. Therefore $e(S_k)$ and v belong to two distinct connected components of $T - e(S_j)$, so $e(S_k)$ is not adjacent to v. This completes the proof of the claim.

The claim being true, it follows that the label set of T_A is $(\bigcup_{(C,S_k)\in\mathcal{C}_G((A,S_i))} C) \cup \lambda(v)$. Since the label set of T_A is A, it follows that $\bigcup_{(C,S_k)\in\mathcal{C}_G((A,S_i))} C = A - \lambda(v) \subseteq A$, which proves that G is consistent.

Conversely, assume that $\Gamma(\mathfrak{S})$ has a consistent thin subgraph G. We define $\lambda_G : V(G) \to \mathcal{P}(\mathcal{M})$ by putting, for $(A, S_i) \in V(G)$, $\lambda(A, S_i) = A - \bigcup_{(C,S_k)\in\mathcal{C}_G((A,S_i))} C$. Since G is consistent, the union is contained in A, so λ_G is well defined.

Next, we define an equivalence relation $\sim_G$ on $V(G)$ as follows: for $(A, S_i), (B, S_j) \in V(G)$, we put $(A, S_i) \sim_G (B, S_j)$ if and only if (A, S_i) and (B, S_j) are the same vertex of G, or $((\overline{A}, S_i), (B, S_j))$ is a critical arc of G. By definition, $\sim_G$ is reflexive. It is also symmetric, as $((\overline{A}, S_i), (B, S_j))$ is critical if and only if $((\overline{B}, S_j), (A, S_i))$ is critical. Finally, transitivity is a direct consequence of Lemma 8.

We claim that for all pairs $(A, S_i), (B, S_j)$ of vertices of G with $(A, S_i) \sim_G (B, S_j)$, we have $\lambda_G(B, S_j) = \lambda_G(A, S_i)$. To see that, note that by Lemma 8, if $((C, S_k), (B, S_j))$ is a critical arc of G distinct from $((\overline{A}, S_i), (B, S_j))$, then $((C, S_k), (A, S_i))$ is a critical arc of

G. Since the roles of A and B are symmetric in this argument it follows that the sets

$$\{(C, S_k) \in V(G) - \{(\overline{A}, S_i)\} : (C, S_k) \in \mathcal{C}_G((B, S_j))\}$$

and

$$\{(C, S_k) \in V(G) - \{(\overline{B}, S_j)\} : (C, S_k) \in \mathcal{C}_G((A, S_i))\}$$

are equal. Denoting this set by $\mathcal{C}$, we have $\lambda_G((A, S_i)) = (A - \overline{B}) - \bigcup_{(C,S_k)\in\mathcal{C}} C$ and $\lambda_G((B, S_j)) = (B - \overline{A}) - \bigcup_{(C,S_k)\in\mathcal{C}} C$. Since $A - \overline{B} = B - \overline{A}$, it follows that $\lambda_G((A, S_i)) = \lambda_G((B, S_j))$ as claimed.

Now, we denote by T the undirected graph whose vertex set is the set of equivalence classes of $\sim_G$, where two equivalence classes u, v are joined by an edge if and only if there exists $i \in \{1, \ldots, n\}$ such that $(A, S_i) \in u$ and $(\overline{A}, S_i) \in v$. Note that by construction, the degree of a vertex u of T is precisely the size of the equivalence class u. In view of the above, λ_G trivially induces a map $\lambda : V(T) \to \mathcal{P}(\mathcal{M})$.

We next show that $\mathcal{T}(G) = (T, \lambda)$ is a representation of $\mathfrak{S}$. We do this by induction on $n = |\mathfrak{S}|$. If $n = 1$, this is trivial, as T is a single edge $\{u, v\}$ with $\lambda(u) = A$ and $\lambda(v) = \overline{A}$, where $A|\overline{A}$ is the unique element of $\mathfrak{S}$. Assume then that $n \geqslant 2$, and that the property holds for all split systems $\mathfrak{S}'$ with $|\mathfrak{S}'| < |\mathfrak{S}|$.

Let (A, S_i) be a vertex of indegree 0 in G (which exists as $\Gamma(\mathfrak{S})$ is acyclic), and let G' be the graph obtained from G by removing the vertices (A, S_i) and $(\overline{A}, S_i)$ and their adjacent arcs. Clearly, G' is a consistent thin subgraph of $\Gamma(\mathfrak{S} - \{S_i\})$. By our induction hypothesis $\mathcal{T}(G') = (T', \lambda')$ is a representation of $\mathfrak{S} - \{S_i\}$. Since (A, S_i) has indegree 0 in G, (A, S_i) is the unique element of its equivalence class under $\sim_G$. In particular, there exists a leaf u of T such that $\lambda(u) = A$. Moreover, (A, S_i) has outdegree at least 1 in G, so there exists a vertex (B, S_j) of G such that $((A, S_i), (B, S_j))$ is a critical arc of G. So, $(\overline{A}, S_i) \sim_G (B, S_j)$ holds. Finally, since (A, S_i) has indegree 0 in G (and by symmetry, $(\overline{A}, S_i)$ has outdegree 0 in G), it follows that an arc $((C, S_k), (B, S_j))$ of G' is a critical arc of G' if and only if $((C, S_k), (B, S_j))$ is a critical arc of G. Hence, the equivalence classes of $\sim_G$ are exactly the equivalence classes of $\sim_{G'}$, plus the equivalence class of (A, S_i). In particular, $V(T) = V(T') \cup \{u\}$, and T is obtained from T' by adding leaf u to the vertex w of T' corresponding to the equivalence class of $(\overline{A}, S_i)$ under $\sim_G$. Moreover, we have $\lambda(v) = \lambda'(v)$ for all $v \in V(T)$ distinct from u, w, and, as already stated, $\lambda(u) = A$. Finally, we have $\lambda(w) = B - \bigcup_{(C,S_k)\in\mathcal{C}_G((B,S_j))} C$ and $\lambda'(w) = B - \bigcup_{(C,S_k)\in\mathcal{C}_{G'}((B,S_j))} C$, so $\lambda(w) = \lambda'(w) - A$. In particular, $A \subseteq \lambda'(w)$ holds.

Putting these observations together, it follows that $\mathcal{T}(G)$ is obtained from $\mathcal{T}(G')$ by (i) attaching a new leaf labeled with A to the vertex w, and (ii) removing A from the label of w. Since $\mathcal{T}(G')$ is an $\mathcal{M}$-tree by our induction hypothesis, it follows that $\mathcal{T}(G)$ is an $\mathcal{M}$-tree if and only if w is not a leaf of $\mathcal{T}(G')$ with $\lambda'(w) = A$. To see that this is the case, we remark that the degree of w in $\mathcal{T}(G')$ is precisely the size of the equivalence class of $\sim_{G'}$ corresponding to w. In particular, if w is a leaf, then (B, S_j) is the unique element of its equivalence class under $\sim_{G'}$, and $\lambda'(w) = B$. Since $((A, S_i), (B, S_j))$ is an arc of $\Gamma(\mathfrak{S})$, we have $A \subsetneq B$, so $\lambda_{G'}(w) \neq A$ as claimed.

So, $\mathcal{T}(G)$ is an $\mathcal{M}$-tree, and by construction $\mathfrak{S}(\mathcal{T}(G)) = \mathfrak{S}(\mathcal{T}(G')) \cup \{S_i\} = \mathfrak{S}$. So, $\mathcal{T}(G)$ is a representation of $\mathfrak{S}$. □

If $\mathfrak{S}$ is a compatible split system and G is a consistent thin subgraph of $\Gamma(\mathfrak{S})$, we denote by $\mathcal{T}(G)$ the $\mathcal{M}$-tree constructed from G as in the proof of Theorem 9. As shown in that proof, $\mathcal{T}(G)$ is a representation of $\mathfrak{S}$. For example, for the split system $\mathfrak{S} = \{a|abbcd, abc|abd, b|aabcd\}$ on $\mathcal{M} = \{a^2, b^2, c, d\}$, the graph $\Gamma(\mathfrak{S})$ has four distinct consistent thin subgraphs, two of which we depict in Figure 4.

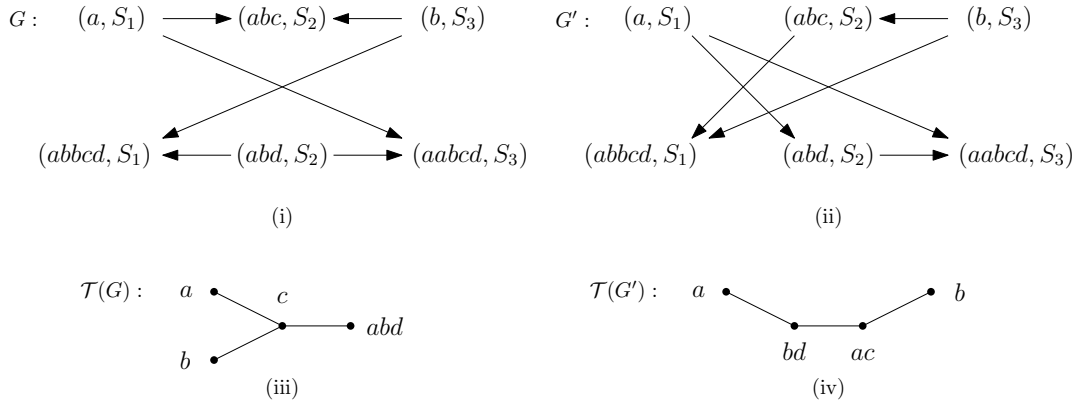

Figure 4: (i) and (ii) Two consistent thin subgraphs of $\Gamma(\mathfrak{S})$, where $\mathfrak{S} = \{a|abbcd, abc|abd, b|aabcd\}$. The other two consistent thin subgraphs of $\Gamma(\mathfrak{S})$ can be obtained by swapping the vertices (abc, S_2) and (abd, S_2) in G and in G'. (iii) and (iv) The $\mathcal{M}$-trees $\mathcal{T}(G)$ and $\mathcal{T}(G')$ constructed from G and G' respectively, as in the proof of Theorem 9. Both $\mathcal{T}(G)$ and $\mathcal{T}(G')$ are representations of $\mathfrak{S}$. The other two representations of $\mathfrak{S}$ can be obtained by swapping labels c and d in $\mathcal{T}(G)$ and in $\mathcal{T}(G')$.

Note that there are compatible split systems $\mathfrak{S}$ for which $\Gamma(\mathfrak{S})$ has two or more distinct consistent thin subgraphs G, G' such that $\mathcal{T}(G)$ is isomorphic to $\mathcal{T}(G')$. In particular, the map assuciating a tree $\mathcal{T}(G)$ to each thin subgraph G of $\Gamma(\mathfrak{S})$ is, in general, not a bijection.

The last example motivates the final result in this section. Suppose that $\mathfrak{S}$ is a compatible split system. We call two distinct thin subgraphs G and G' of $\mathfrak{S}$ *isomorphic* if there is a digraph isomorphism between G and G' which maps each vertex of the form (A, S_i) to one of the form (B, S_j), where $A = B$.

Corollary 10. *Let $\mathfrak{S}$ be a compatible split system on a multiset $\mathcal{M}$. Then the non-isomorphic representations $\mathcal{T}$ of $\mathfrak{S}$ are in bijective correspondence with the non-isomorphic consistent thin subgraphs of $\Gamma(\mathfrak{S})$. In particular, $\mathfrak{S}$ has a unique representation if and only if $\Gamma(\mathfrak{S})$ has a unique consistent thin subgraph up to isomorphism.*

Proof. Let $\mathfrak{S} = \{S_1, \ldots, S_n\}$, $n \geqslant 1$. As seen in the proof of Theorem 9, a representation $\mathcal{T}$ of $\mathfrak{S}$ trivially induces a consistent thin subgraph G of $\Gamma(\mathfrak{S})$. It is straightforward to see that $\mathcal{T} = \mathcal{T}(G)$ in that case.

Now, suppose that G_1 and G_2 are two distinct consistent thin subgraphs of $\Gamma(\mathfrak{S})$. If G_1 and G_2 are isomorphic, then it is straight-forward to check that $\mathcal{T}(G_1)$ is isomorphic to $\mathcal{T}(G_2)$.

So, suppose that G_1 and G_2 are not isomorphic. By Lemma 7, G_1 and G_2 are uniquely determined by their sets of critical arcs. Since G_1 and G_2 are not isomorphic, there must exist a vertex (A, S_i) of $\Gamma(\mathfrak{S})$ such that (i) the sets $\mathcal{C}_{G_1}((A, S_i))$ and $\mathcal{C}_{G_2}((A, S_i))$ are distinct, and (ii) no further vertex (B, S_j) of $\Gamma(\mathfrak{S})$ satisfies $B = A$ and $\mathcal{C}_{G_2}((B, S_j)) = \mathcal{C}_{G_1}((A, S_i))$. For $k \in \{1, 2\}$, we denote by v_k the vertex of $\mathcal{T}(G_k)$ corresponding to the equivalence class of (A, S_i) under $\sim_{G_k}$. By definition of $\mathcal{T}(G_1)$ and $\mathcal{T}(G_2)$, the set of splits induced by edges adjacent to v_1 in $\mathcal{T}(G_1)$ is $\mathcal{C}_{G_1}((A, S_i))$ and the set of splits induced by edges adjacent to v_2 in $\mathcal{T}(G_2)$ is $\mathcal{C}_{G_2}((A, S_i))$, which by (i) is distinct from $\mathcal{C}_{G_1}((A, S_i))$. Moreover, (ii) implies that there is no further vertex w of $\mathcal{T}(G_2)$ such that the multiset of splits induced by edges adjacent to w in $\mathcal{T}(G_2)$ is precisely $\mathcal{C}_{G_1}((A, S_i))$. In summary, $\mathcal{T}(G_1)$ contains a vertex, such that the multiset of splits induced by its adjacent edges is precisely $\mathcal{C}_{G_1}((A, S_i))$. Conversely, no vertex of $\mathcal{T}(G_2)$ satisfies this property. Thus, $\mathcal{T}(G_1)$ and $\mathcal{T}(G_2)$ are not isomorphic.

The rest of the proof follows from the observation that G_1 (*resp.* G_2) is precisely the graph obtained from $\mathcal{T}(G_1)$ (*resp.* $\mathcal{T}(G_2)$) using the construction described in the first part of the proof of Theorem 9. □

Note that by Corollary 10, if $\Gamma(\mathfrak{S})$ is a thin and consistent subgraph of itself, then $\mathfrak{S}$ has a unique representation. However, there exist non-thin split systems $\mathfrak{S}$ such that $\Gamma(\mathfrak{S})$ has a unique (up to isomorphism) thin and consistent subgraph, and so it is not necessary for $\mathfrak{S}$ to be thin in order for $\mathfrak{S}$ to have a unique representation. For example, the split system $\mathfrak{S}$ on $\mathcal{M} = \{a^2, b^2, c^2, x, y\}$ represented by the $\mathcal{M}$-tree $\mathcal{T}$ depicted in Figure 1 is not thin. However, $\Gamma(\mathfrak{S})$ has only one consistent thin subgraph, so $\mathcal{T}$ is the unique representation of $\mathfrak{S}$.

3 Thin split systems

In this section we shall show that Conjecture 3 holds for thin split systems and that, as a corollary, this is also the case for split systems $\mathfrak{S}$ in which $\min\{|A|, |\overline{A}|\} = \min\{|B|, |\overline{B}|\}$ holds for all pairs of splits in $A|\overline{A}, B|\overline{B}$ in $\mathfrak{S}$.

Theorem 11. *Suppose that $\mathfrak{S}$ is a thin split system on a multiset $\mathcal{M}$. Then $\mathfrak{S}$ is compatible if and only if every subset of $\mathfrak{S}$ of size at most $\Delta(\mathcal{M}) + 2$ is compatible.*

Proof. One direction is trivial: if $\mathfrak{S}$ is compatible, then all submultisets of $\mathfrak{S}$ are compatible.

To see the opposite direction, we show that if $\mathfrak{S}$ is thin and not compatible, then there exists a submultiset $\mathfrak{S}'$ of $\mathfrak{S}$ of size $\Delta(\mathcal{M}) + 2$ or less that is not compatible. So, suppose that $\mathfrak{S}$ is thin and not compatible. Since $\mathfrak{S}$ is thin, the only thin subgraph of $\Gamma(\mathfrak{S})$ is $\Gamma(\mathfrak{S})$. By Theorem 9, $\Gamma(\mathfrak{S})$ is not consistent. Thus, there exists $k + 1 \geqslant 3$ splits $S = A|\overline{A}, S_1 = A_1|\overline{A_1}, \ldots, S_k = A_k|\overline{A_k}$ in $\mathfrak{S}$ such that $((A_i, S_i), (A, S))$ is a critical arc of $\Gamma(\mathfrak{S})$ for all $i \in \{1, \ldots, k\}$, and $\bigcup_{i=1}^{k} A_i \not\subseteq A$. Note that since the arcs $((A_i, S_i), (A, S))$, $i \in \{1, \ldots, k\}$ are critical, the sets $A_1, \ldots, A_k$ satisfy:

(i) $A_i \subseteq A$ for all $i \in \{1, \dots, k\}$, and

(ii) there are no $i, j \in \{1, \dots, k\}$ such that $A_i \subsetneq A_j$.

Put $\mathfrak{S}' = \{S, S_1, \dots, S_k\}$. Clearly, $\mathfrak{S}'$ is a submultiset of $\mathfrak{S}$ such that $\Gamma(\mathfrak{S}')$ is thin and not consistent. Without loss of generality, we may assume that all proper submultisets $\mathfrak{S}''$ of $\mathfrak{S}'$ are such that $\Gamma(\mathfrak{S}'')$ is consistent.

Now, let $A' = \bigcup_{i=1}^{k} A_i$. Since, $A' \not\subseteq A$, there exists $x \in \mathcal{M}$ such that $m_{A'}(x) > m_A(x)$. Note that $m_{A'}(x) = \sum_{i=1}^{k} m_{A_i}(x)$. By minimality of $\mathfrak{S}'$, we have $m_{A_i}(x) \geqslant 1$ for all $i \in \{1, \dots, k\}$.

Property (ii), together with the fact that $\mathfrak{S}'$ is thin, implies that for all $i, j \in \{1, \dots, k\}$ distinct, we have $A_i \subsetneq \overline{A_j}$ and $A_j \subsetneq \overline{A_i}$. In particular, $((A_i, S_i), (\overline{A_j}, S_j))$ and $((A_j, S_j), (\overline{A_i}, S_i))$ are arcs of $\Gamma(\mathfrak{S}')$, and these arcs are critical in $\Gamma(\mathfrak{S}')$. Moreover, the graph $\Gamma(\{S_1, \dots, S_k\})$ is consistent by choice of $\mathfrak{S}'$. Thus, we have $\bigcup_{i=1}^{k-1} A_i \subseteq \overline{A_k}$. In particular, we have $\sum_{i=1}^{k-1} m_{A_i}(x) \leqslant m_{\overline{A_k}}(x)$. Since $m_{A_i}(x) \geqslant 1$ for all $i \in \{1, \dots, k\}$, we have $\sum_{i=1}^{k-1} m_{A_i}(x) \geqslant k-1$, and $m_{\overline{A_k}}(x) \leqslant m_{\mathcal{M}}(x) - 1$. Putting these inequalities together, $k \leqslant m_{\mathcal{M}}(x)$ follows. In particular, we have $|\mathfrak{S}'| = k + 1 \leqslant m_{\mathcal{M}}(x) + 1$. Moreover, we have $\Delta(\mathcal{M}) \geqslant m_{\mathcal{M}}(x) - 1$ by definition of $\Delta(\mathcal{M})$. Combining the last two inequalities, we obtain $|\mathfrak{S}'| \leqslant \Delta(\mathcal{M}) + 2$.

In summary, if $\mathfrak{S}$ is not compatible, then $\mathfrak{S}$ contains a submultiset $\mathfrak{S}'$ with $|\mathfrak{S}'| \leqslant \Delta(\mathcal{M}) + 2$ such that $\Gamma(\mathfrak{S}')$ is not consistent. Since $\mathfrak{S}$ is thin, $\mathfrak{S}'$ is thin, so $\Gamma(\mathfrak{S}')$ is the only thin subgraph of itself. By Theorem 9, $\mathfrak{S}'$ is not compatible, which concludes the proof. □

Note that the bound given in the last theorem is tight, as there exist thin split systems $\mathfrak{S}$ such that every subset of size at most $\Delta(\mathcal{M})+1$ is compatible but $\mathfrak{S}$ is not compatible. For example, let $n > m \geqslant 1$ and consider the split system $\mathfrak{S} = \{a_{i,1} \dots a_{i,m} x | \overline{a_{i,1} \dots a_{i,m} x} : i \in \{1, \dots, n\}\}$ on $\mathcal{M} = \{a_{i,j} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\} \cup \{x^{n-1}\}$, where all $a_{i,j}$ are pairwise distinct and distinct from x. Clearly, $\mathfrak{S}$ is thin and $|\mathfrak{S}| = n$. It is straight-forward to check that $\mathfrak{S}$ is not compatible, but all proper subsets of $\mathfrak{S}$ are compatible. Since $\Delta(\mathcal{M}) = n-2$, we have in particular that all subsets of $\mathfrak{S}$ of size $\Delta(\mathcal{M}) + 1$ are compatible.

Suppose that $S = A|\overline{A}$ is a split on $\mathcal{M}$. We say that A (*resp.* $\overline{A}$) is the *small part* of S if $|A| \leqslant |\overline{A}|$ (*resp.* $|\overline{A}| \leqslant |A|$). We define the *size* of S as the size of its small part, that is, $\min\{|A|, |\overline{A}|\}$ and we call a split of size $k \geqslant 1$ a *k-split*. As a consequence of Theorem 11 we immediately obtain the following generalization of [8, Lemma 4.6 (ii)].

Theorem 12. *Suppose that $\mathfrak{S}$ is a split system on a multiset $\mathcal{M}$ in which every split has the same size. Then $\mathfrak{S}$ is compatible if and only if every submultiset of $\mathfrak{S}$ of size at most $\Delta(\mathcal{M}) + 2$ is compatible.*

Proof. Suppose that every submultiset of $\mathfrak{S}$ of size at most $\Delta(\mathcal{M})+2$ is compatible. Then every pair of splits in $\mathfrak{S}$ is compatible. Since every split in $\mathfrak{S}$ has the same size it follows that $\mathfrak{S}$ is thin. Now we can apply Theorem 11 to see that $\mathfrak{S}$ is compatible. □

4 Superterminal sets

In this section, we shall introduce the concept of a superterminal set of a split system. We then use this concept to prove a technical result concerning such sets (Theorem 15) which, in turn, we will use to help prove the main result of the next section (Theorem 22).

For a split system $\mathfrak{S} = \{S_1, \ldots, S_n\}$, $n \geqslant 1$ on a multiset $\mathcal{M}$, we say that a set $A \subseteq \mathcal{M}$ is a *terminal set* of $\mathfrak{S}$ if $A|\overline{A} = S_i$ for some $i \in \{1, \ldots, n\}$, and the indegree of the vertex (A, S_i) in $\Gamma(\mathfrak{S})$ is 0. Equivalently, A is a terminal set of $\mathfrak{S}$ if for all splits $S_j = B|\overline{B}$, with $j \in \{1, \ldots, n\}$ distinct from i, neither $B \subsetneq A$ nor $\overline{B} \subsetneq A$ holds. We denote by $\mathfrak{S}_t$ the (multi)set of terminal sets of $\mathfrak{S}$. Note that since $\Gamma(\mathfrak{S})$ is finite and acyclic, $\mathfrak{S}_t$ is nonempty. As an example, for the split system $\mathfrak{S} = \{ab|\overline{ab}, ac|\overline{ac}, cx|\overline{cx}, abcx|abcy\}$ represented by the tree in Figure 1, we have $\mathfrak{S}_t = \{ab, ac, cx\}$. Terminal sets have a special place in representations of compatible split systems, as the following, straightforward to prove result shows:

Lemma 13. *Let $\mathfrak{S}$ be a split system on $\mathcal{M}$. If $\mathfrak{S}$ is compatible and $\mathcal{T} = (T, \lambda)$ is a representation of $\mathfrak{S}$, then for all $A \in \mathfrak{S}_t$, there exists a leaf x of T such that $\lambda(x) = A$.*

Proof. Since $A \in \mathfrak{S}_t$, we have $A|\overline{A} \in \mathfrak{S}$, so there exists a leaf x of T such that $\lambda(x) \subseteq A$. In particular, $\lambda(x)|\overline{\lambda(x)}$ is a split of $\mathfrak{S}$. Since $A \in \mathfrak{S}_t$, by definition of $\mathfrak{S}_t$, there is no split $A'|\overline{A'}$ of $\mathfrak{S}$ such that $A' \subsetneq A$. Therefore, the inclusion $\lambda(x) \subseteq A$ cannot be strict, that is, $\lambda(x) = A$ must hold. □

Note that if $\mathfrak{S}$ is compatible with representation $\mathcal{T} = (T, \lambda)$, the injection $\mathfrak{S}_t \to L(T)$ given by Lemma 13 is not necessarily a bijection. For example, the split system $\mathfrak{S} = \{ab|abcd, abc|abd\}$ is compatible and $\mathfrak{S}_t = \{ab\}$, but any representation $\mathcal{T} = (T, \lambda)$ of $\mathfrak{S}$ has two leaves, so one leaf of T has a label that is not in $\mathfrak{S}_t$.

Motivated by this fact, for a split system $\mathfrak{S} = \{S_1, \ldots, S_n\}$, $n \geqslant 1$, on a multiset $\mathcal{M}$, we call a terminal set A of $\mathfrak{S}$ a *superterminal set* of $\mathfrak{S}$ if $indegree(A, S_i) = 0$ and $outdegree(A, S_i) = |\mathfrak{S}| - 1$ in $\Gamma(\mathfrak{S})$. Equivalently, A is a superterminal set of $\mathfrak{S}$ if and only if for all splits $S_j = B|\overline{B}$, $j \neq i \in \{1, \ldots, n\}$, of $\mathfrak{S}$, neither $B \subsetneq A$ nor $\overline{B} \subsetneq A$ hold, and exactly one of $A \subsetneq B$ or $A \subsetneq \overline{B}$ holds. For example, the split system $\mathfrak{S} = \{ab|cxabcy, ac|bxabcy, cx|ababcy, abcx|abcy\}$ has precisely one superterminal set, that is cx. We denote by $\mathfrak{S}_t^{\times} \subseteq \mathfrak{S}_t$ the (multi)set of all superterminal sets of $\mathfrak{S}$. Moreover, for $A \in \mathfrak{S}_t^{\times}$ and $S = B|\overline{B} \in \mathfrak{S}$ distinct from $A|\overline{A}$ we denote by $S(A)$ the unique element in $\{B, \overline{B}\}$ satisfying $A \subseteq S(A)$. To extend this notation to $\mathfrak{S}$, we also put $S(A) = A$ for $S = A|\overline{A}$. Interestingly, for thin split systems, the collections of terminal sets and superterminal sets coincide.

Lemma 14. *If $\mathfrak{S}$ is thin, then $\mathfrak{S}_t^{\times} = \mathfrak{S}_t \neq \emptyset$.*

Proof. As already remarked above, $\mathfrak{S}_t \neq \emptyset$ always holds. To see that $\mathfrak{S}_t^{\times} = \mathfrak{S}_t$, let $A \in \mathfrak{S}_t$, and let $B|\overline{B} \in \mathfrak{S}$ distinct from $A|\overline{A}$. Since $\mathfrak{S}$ is thin, exactly one of $B \subsetneq A$, $\overline{B} \subsetneq A$, $A \subsetneq B$, $A \subsetneq \overline{B}$ holds. That $A \in \mathfrak{S}_t$ implies that the first two inclusions cannot hold. Hence, exactly one of $A \subsetneq B$ or $A \subsetneq \overline{B}$ holds, so $A \in \mathfrak{S}_t^{\times}$ follows. □

Now, suppose that $\mathfrak{S}$ is a split system, and that it has a superterminal set A_0. Our general aim is to remove the split $S_0 = A_0|\overline{A_0}$ from $\mathfrak{S}$ in a controlled manner, so that key properties of $\mathfrak{S}$ are preserved in the resulting split system.

To this end, we let a_0 be a new element that is not in $\mathcal{M}$ and define the split system $\mathfrak{S}^- = \mathfrak{S}^-_{A_0}$ on $\mathcal{M}^- = \mathcal{M}^-_{A_0} = (\mathcal{M} - A_0) \cup \{a_0\}$ as follows. For all $S \in \mathfrak{S} - \{S_0\}$, we add to $\mathfrak{S}^-$ the split $(S(A_0) - A_0) \cup \{a_0\}|\overline{S(A_0)}$ in case $S \neq S_0$, and the split $A_0|(\overline{A_0} - A_0) \cup \{a_0\}$ in case $S = S_0$. Note that the latter may indeed hold, since, if S_0 has multiplicity 2 or more in $\mathfrak{S}$, then we have $S_0 \in \mathfrak{S} - \{S_0\}$. In that case, the split system $\{S_0, S_0\} \subsetneq \mathfrak{S}$ is compatible, so $A_0 \subsetneq \overline{A_0}$ must hold. Roughly speaking, $\mathfrak{S}^-$ is obtained by merging, in all splits $S \in \mathfrak{S} - \{S_0\}$, the elements of A_0 in $S(A_0)$ (in case $S \neq S_0$) or in $\overline{A_0}$ (in case $S = S_0$) into a single element a_0. By definition, the multiset $\mathfrak{S}^-$ is a multiset of $n-1$ splits on $\mathcal{M}^-$. Note that by definition, $a_0|(\mathcal{M}^- - \{a_0\})$ is not a split of $\mathfrak{S}^-$.

We now prove the main result of this section, which shows that if A_0 is a superterminal set of an incompatible split system $\mathfrak{S}$ in which all proper subsets of $\mathfrak{S}$ are compatible, then $\mathfrak{S}^-_{A_0}$ has the same properties.

Theorem 15. *Let $\mathfrak{S}$ be an incompatible split system on $\mathcal{M}$ of size $k \geqslant 3$ with $\mathfrak{S}_t^{\times} \neq \emptyset$, and such that all proper subsets of $\mathfrak{S}$ are compatible. If $A_0 \in \mathfrak{S}_t^{\times}$, then $\mathfrak{S}^-_{A_0}$ is incompatible, and all proper subsets of $\mathfrak{S}^-_{A_0}$ are compatible. Moreover, $\Delta(\mathcal{M}^-) < \Delta(\mathcal{M})$.*

Proof. We begin by making a claim which is of independent interest:

Claim 16. *Let $\mathfrak{S}$ be a split system on $\mathcal{M}$ of size $n \geqslant 3$ such that all proper subsets of $\mathfrak{S}$ are compatible. If there is a split $S = A|\overline{A} \in \mathfrak{S}$ such that $A \subseteq \mathcal{M}^*$, then $\mathfrak{S}$ is compatible.*

Proof. Assume that there exists a split $S = A|\overline{A}$ in $\mathfrak{S}$ such that $A \subseteq \mathcal{M}^*$. We can choose S in $\mathfrak{S}$ such that A is minimal with respect to set inclusion. Note that the split system $\{S, S\}$ is not compatible. Indeed, if it were the case, then by Lemma 4, one of $A \subsetneq \overline{A}$ or $\overline{A} \subsetneq A$ would hold. However, since $A \subseteq \mathcal{M}^*$ both are impossible. Therefore, S must have multiplicity one in $\mathfrak{S}$. Thus, because $\mathfrak{S} - \{S'\}$ is compatible for all S' distinct from S, it follows that for all $B|\overline{B} \in \mathfrak{S}$, precisely one of $A \subseteq B$, $A \subseteq \overline{B}$ holds.

It follows that in any representation $\mathcal{T} = (T, \lambda)$ of $\mathfrak{S} - \{S\}$, there exists a vertex v of T with $A \subseteq \lambda(v)$. Otherwise, there exists an edge e in T and two elements $x, y \in A$ distinct such that the split $S_e = B|\overline{B}$ of $\mathfrak{S}$ associated to e satisfies (up to permutation) $x \in B$ and $y \in \overline{B}$. Since x and y have multiplicity one in $\mathcal{M}$, neither $A \subseteq B$ nor $A \subseteq \overline{B}$ holds, a contradiction.

Thus we can add a vertex u and edge $\{u, v\}$ to T, labelling u with A and removing A from the label of v. Because S has multiplicity 1 in $\mathfrak{S}$, v cannot be a leaf of $\mathcal{T}$ satisfying $\lambda(v) = A$, so the labeled tree $\mathcal{T}'$ obtained this way is an $\mathcal{M}$-tree. By construction, $\mathcal{T}'$ is a representation of $\mathfrak{S}$, so $\mathfrak{S}$ is compatible. □

Now, note that if A_0 is as in the statement of the theorem, it follows by Claim 16 that A_0 contains at least one element of multiplicity two or more in $\mathcal{M}$, and so $\Delta(\mathcal{M}^-) < \Delta(\mathcal{M})$.

To show that all subsets of $\mathfrak{S}^-$ of size $n-2$ are compatible, let $S^- \in \mathfrak{S}^-$ and let S be the corresponding split in $\mathfrak{S}$ (S is the split obtained from S^- by "replacing" a_0 with A_0). Since $\mathfrak{S} - \{S\}$ is compatible by assumption, there is a representation $\mathcal{T} = (T, \lambda)$ of $\mathfrak{S} - \{S\}$. Moreover, we have $S_0 \in \mathfrak{S} - \{S\}$, and $A_0 \in \mathfrak{S}_t^{\times} \subseteq \mathfrak{S}_t$, so by Lemma 13, there exists a leaf x of T with $\lambda(x) = A_0$. Thus it is not difficult to see that the tree T' obtained from T by changing the label of x from A_0 to a_0, and then collapsing the edge adjacent to x, is a representation of $\mathfrak{S}^- - \{S^-\}$.

To show that $\mathfrak{S}^-$ is not compatible, assume by contradiction that this is not the case, and let $\mathcal{T} = (T, \lambda)$ be a representation of $\mathfrak{S}^-$. Then there must exist a vertex v_0 of T such that $a_0 \in \lambda(v_0)$. By definition of $\mathfrak{S}^-$, the split $a_0|(\mathcal{M}^+ - \{a_0\})$ does not belong to $\mathfrak{S}^-$, so if v_0 is a leaf, $\lambda(v_0) - \{a_0\}$ is nonempty. By removing a_0 from the label of v_0 and adding a leaf x adjacent to v_0 labeled with A_0, we then obtain a representation of $\mathfrak{S}$. However, this is impossible, as $\mathfrak{S}$ is not compatible by assumption. So, $\mathfrak{S}^-$ is not compatible. □

5 2,3-split systems

In Theorem 12, we showed that Conjecture 3 holds for split systems in which every split has the same size. In this section we consider splits systems in which every split has size 2 or 3, which we shall call *2,3-split systems*. In particular we will prove in Theorem 22 that 2,3-split systems in which ever split has multiplicity 1 satisfy a slightly weaker version of (★).

We first investigate the structure of labelled trees $\mathcal{T}$ such that all splits in $\mathfrak{S}(\mathcal{T})$ have size 2 or 3.

Lemma 17. *Let $\mathcal{T} = (T, \lambda)$ be a $\mathcal{M}$-tree for some multiset $\mathcal{M}$ and let $\mathfrak{S} = \mathfrak{S}(\mathcal{T})$. If all splits in $\mathfrak{S}$ have size 2 or 3, and at least one split of $\mathfrak{S}$ has size 3, then there exists a vertex v^* of T such that:*

(i) v^* is adjacent to all edges of T that correspond to 3-splits of $\mathfrak{S}$.

(ii) All vertices of T distinct from v^* have degree 2 or less.

Proof. We first show that there exists a vertex v^* such that (i) holds. If $\mathfrak{S}$ contains exactly one split S of size 3, then (i) is true for both ends of the edge e of T associated to S in T. In this case, and for the purpose of showing (ii) later on, we choose v^* in such a way that the connected component of $T - e$ that does not contain v^* is labelled with the part of S of size 3.

Suppose now that $\mathfrak{S}$ contains at least two splits $S_1 = A|\overline{A}$ and $S_2 = B|\overline{B}$ of size 3. Without loss of generality, we can choose A and B such that $|A| = |B| = 3$. For $i \in \{1, 2\}$, let e_i be the edge of T associated to S_i in T, and let v_i be the vertex of e_i that belongs to the connected component of $T - e_i$ labelled with $\overline{A}$ or $\overline{B}$, respectively. We now claim that $v_1 = v_2$ must hold.

Since $S_1, S_2 \in \mathfrak{S}$, $\{S_1, S_2\}$ is compatible, and so Lemma 4 implies that $A \subsetneq \overline{B}$ and $B \subsetneq \overline{A}$, since all other inclusions are forbidden due to the respective size of the sets. In particular, e_1 belongs to the connected component of $T - e_2$ labelled with $\overline{B}$, since the converse would imply that one of $\overline{A} \subseteq B$ or $A \subseteq B$ held. In particular, v_1 belongs to the connected component of $T - e_2$ labelled with $\overline{B}$, and by symmetry, v_2 belongs to the connected component of $T - e_1$ labelled with $\overline{A}$. Now, suppose for contradiction that $v_1 \neq v_2$, and let e be an edge on the path between v_1 and v_2 in T. Clearly, v_1 and v_2 belong to distinct connected component of $T - e$. In view of the above observation, the connected component of $T - e$ containing v_1 has A in its label set, and the connected component containing v_2 has B in its label sets. Therefore, the split $S = C|\overline{C} \in \mathfrak{S}$ induced by e satisfies (up to permutation), $A \subseteq C$ and $B \subseteq \overline{C}$. By assumption, S has size at most three, so one of $A = C$ or $B = \overline{C}$ must hold. However, if $A = C$, then $\lambda(v_1) = \emptyset$, and v_1 must have degree 2 in T. This is impossible by definition of a $\mathcal{M}$-tree. Also, one can show using similar arguments that $B = \overline{C}$ cannot hold either. Hence, $v_1 = v_2$ must hold as claimed.

It follows that any two edges of T associated to a 3-split of $\mathfrak{S}$ must share a vertex. Since T is a tree, and therefore acyclic, it also follows that there exists a unique vertex v^* of T that is adjacent to all such edges. We thus pick this vertex v^*, as (i) clearly holds for this choice of v^*.

To see that (ii) holds, it suffices to remark that for either of the two choices of v^* above, after removal of v^* from T, all connected components have label set of size 2 or 3. In particular, if v is a vertex of T distinct from v^*, and e is the edge adjacent to v on the path between v and v^*, then the connected component of $T - e$ containing v has label set of size 2 or 3. If there exists two edges e_1 and e_2 adjacent to v and distinct from e, then this implies that (at least) one of the splits associated to e_1 or e_2 must have size 1. Since $\mathfrak{S}$ does not contain any split of size 1, this is impossible. Hence, v has degree at most 2 so (ii) holds. □

Next, we prove a useful technical lemma.

Lemma 18. *Let $\mathcal{M}$ be a multiset with underlying set X and let $A_1, \ldots, A_k$, $k \geqslant 2$ be a partition of $\mathcal{M}$. Let G be a graph whose vertex set is the set $\{A_i : 1 \leqslant i \leqslant k\}$, such that each edge $\{A_i, A_j\}$, $i \neq j$, of G satisfies $A_i \cap A_j \neq \emptyset$. Then, we have $\Delta(\mathcal{M}) \geqslant k - c$, where c is the number of connected components of G.*

Proof. Let F be the disjoint union of spanning trees with one taken in each component of G. Since F is acyclic, we have $|E(F)| = |V(F)| - c = k - c$, so it suffices to show that $\Delta(\mathcal{M}) \geqslant |E(F)|$.

To each edge $e = \{A, B\}$ of F, we can associate one element $p(e)$ in $A \cap B$. For x an element of X, we denote by $\pi(x)$ the number of edges $e \in E(F)$ such that $p(e) = x$. Since F is acyclic, the number of vertices A of F satisfying $x \in A$ is at least $\pi(x) + 1$. Since the vertices of F form a partition of $\mathcal{M}$, it follows that $\mathcal{M}(x) \geqslant \pi(x) + 1$. This implies $\Delta(\mathcal{M}) = \Sigma_{x \in X}(\mathcal{M}(x) - 1) \geqslant \Sigma_{x \in X}\pi(x) = |E(F)|$, as required. □

We now prove the first of two key propositions.

Proposition 19. *Suppose $\mathfrak{F}$ is an incompatible 2,3-split system on $\mathcal{M}$ containing $k \geqslant 3$ splits, in which every split has multiplicity 1 such that all proper subsets of $\mathfrak{F}$ are compatible. If $\mathfrak{F}_t^{\times} = \emptyset$, then $\Delta(\mathcal{M}) \geqslant k-3$.*

Proof. First, note that if all splits in $\mathfrak{F}$ have the same size, then $\mathfrak{F}$ is thin, so $\mathfrak{F}_t = \mathfrak{F}_t^{\times} \neq \emptyset$ holds by Lemma 14. Therefore, $\mathfrak{F}$ contains at least one 2-split and at least one 3-split. We begin by showing that $\mathfrak{F}$ enjoys the following property:

(*) If $A|\overline{A} \in \mathfrak{F}$ such that $|A| = 3$, then there exists some $B|\overline{B}, C|\overline{C} \in \mathfrak{F}$ with $B \neq C$, $|B| = |C| = 2$ and $B, C \subseteq A$.

To see this, first note that for all splits $B|\overline{B}$ of $\mathcal{F}$ with $|B| \leqslant |\overline{B}|$ and $B \neq A$, we have $|B| \leqslant |A| = 3 \leqslant |\overline{B}|$. Therefore, neither $A \subsetneq B$ nor $\overline{B} \subsetneq A$ can hold. Since any two splits of $\mathfrak{F}$ are pairwise compatible, it follows from Lemma 4 that at least one of $A \subsetneq \overline{B}$ or $B \subsetneq A$ holds. Since $\mathfrak{F}_t^{\times} = \emptyset$, $A \in \mathfrak{F}_t^{\times}$ is impossible, so there must exist some $B|\overline{B} \in \mathfrak{F}$ with $|B| = 2$ and $B \subsetneq A$. Now, suppose there does not exist some $C|\overline{C}$ in addition to $B|\overline{B}$. Then since $\mathfrak{F} - \{B|\overline{B}\}$ is compatible, there is some tree representing $\mathfrak{F} - \{B|\overline{B}\}$ with a leaf having label set A. But then $\mathfrak{F}$ is compatible, a contradiction.

Now, let S^* be a 2-split of $\mathfrak{F}$, and let $\{x, y\}$ be the part of S^* of size 2. Let $\mathfrak{S} = \mathfrak{F} - \{S^*\}$. Since $\mathfrak{S}$ is compatible, there is some $\mathcal{M}$-tree $\mathcal{T} = (T, \lambda)$ which represents $\mathfrak{S}$. Moreover, $\mathfrak{S}$ contains at least one 3-split, so Lemma 17 implies that T has a "central vertex" v^* that is adjacent to all edges that correspond to 3-splits of $\mathfrak{S}$, and all other vertices in T have degree 1 or 2. Let $\mathcal{M}' = \mathcal{M} - \lambda(v^*)$. Since $\mathcal{M}' \subseteq \mathcal{M}$, it follows that $\Delta(\mathcal{M}') \leqslant \Delta(\mathcal{M})$. We next proceed to show that $\Delta(\mathcal{M}') \geqslant k-3$.

Denote by V_0 (*resp.* V_1) the set of leaves of $\mathcal{T}$ (*resp.* the set of vertices of degree 2 of $\mathcal{T}$, excluding v^*). We also denote by $\lambda(V_0)$ (*resp.* $\lambda(V_1)$) the union of the multisets $\lambda(v)$, $v \in V_0$ (*resp.* $v \in V_1$). These sets form a partition $\lambda(V_0) \cup \lambda(V_1)$ of $\mathcal{M}'$. Note also that for all $v \in V_1$, we have $|\lambda(v)| = 1$, and that $|V_0| + |V_1| = |\mathfrak{S}|$.

Now, let $v \in V_1$ and let z_v be the unique element in $\lambda(v)$. If $z_v \notin \{x, y\}$, then by (*), there exists a leaf l of $\mathcal{T}$ such that $z_v \in \lambda(l)$. Otherwise, if $z_v \in \{x, y\}$, say $z_v = x$, then by (*), either there exists a leaf l such that $x \in \lambda(l)$, or the leaf l_v adjacent to v satisfies $y \in \lambda(l_v)$ (note that these two cases are not mutually exclusive). Put together, these observations imply that there is at most one element in $\mathcal{M}$ (which must be either x or y) that belongs to the label of some vertices in V_1 but does not belong to the label of any vertex of V_0. This means that $\Delta(\mathcal{M}') = \Delta(\lambda(V_0)) + |V_1|$ if there exists l_x and l_y in V_0 (necessarily distinct) such that $x \in \lambda(l_x)$ and $y \in \lambda(l_y)$, and $\Delta(\mathcal{M}') = \Delta(\lambda(V_0)) + |V_1| - 1$ otherwise.

We next focus our attention on the elements in the set V_0. We define the undirected graph $G(V_0)$ to be the graph with vertex set $\{\lambda(v) : v \in V_0\}$, in which two distinct sets $\lambda(u), \lambda(v), u, v \in V_0$ are joined by an edge if $\lambda(u) \cap \lambda(v) \neq \emptyset$. We claim that $G(V_0)$ has at most two connected components, and has two connected components only if there exists l_x and l_y in V_0 (necessarily distinct) such that $x \in \lambda(l_x)$ and $y \in \lambda(l_y)$.

To prove this claim, we first remark that since all vertices v of V_0 are leaves of $\mathcal{T}$, we have $\lambda(v_0)|\overline{\lambda(v_0)} \in \mathfrak{S}$, where $\lambda(v_0)$ is the small part of $\lambda(v_0)|\overline{\lambda(v_0)}$. Hence, $G(V_0)$ is a

subgraph of the graph $G(\mathfrak{S})$ with vertex set $\{A \subseteq \mathcal{M} : A \text{ is the small part of a split } S \in \mathfrak{S}\}$, and in which two sets A and B are joined by an edge if $A \cap B \neq \emptyset$. More precisely, $G(V_0)$ is obtained from $G(\mathfrak{S})$ by removing the set $\{x, y\}$ from $V(G(\mathfrak{S}))$, and all sets of size three that do not label leaves of $\mathcal{T}$.

We next show that $G(\mathfrak{S})$ is connected. Assume for contradiction that this is not the case. Then there exists a partition $\mathfrak{S}_1, \mathfrak{S}_2$ of $\mathfrak{S}$ such that for any two splits $S_1 \in \mathfrak{S}_1$, $S_2 \in \mathfrak{S}_2$, the small parts of S_1 and S_2 do not intersect. Since $\mathfrak{S}_1$ and $\mathfrak{S}_2$ are nonempty proper subsets of $\mathfrak{S}$, it follows that there exist two $\mathcal{M}$-trees $\mathcal{T}_1 = (T_1, \lambda_1)$, $\mathcal{T}_2 = (T_2, \lambda_2)$ representing $\mathfrak{S}_1$ and $\mathfrak{S}_2$, respectively. Moreover, there exist two vertices $v_1 \in V(T_1)$, $v_2 \in V(T_2)$ such that for $i, j \in \{1, 2\}$ distinct, $\bigcup_{v \in V(T_i) - \{v_i\}} \lambda_i(v) \subseteq \lambda_j(v_j)$. By identifying vertices v_1 and v_2, and defining the label of the newly created vertex as $\lambda_1(v_1) \cap \lambda_2(v_2)$, we obtain a representation of $\mathfrak{S}$, which is impossible since $\mathfrak{S}$ is not compatible. Hence, $G(\mathfrak{S})$ is connected.

To complete the proof of the claim, we now further consider the relationship between $G(V_0)$ and $G(\mathfrak{S})$. First, let G^- be the graph obtained from $G(\mathfrak{S})$ by removing all sets of size three that are not of the form $\lambda(v_0), v_0 \in V_0$. In view of (*), removing sets of size three from the vertex set of $G(\mathfrak{S})$ does not modify the number of connected components of the resulting graph, so G^- is connected. Recall that $G(V_0)$ is obtained from G^- by removing the vertex $\{x, y\}$. In particular $V_0 = V(G^-) - \{\{x, y\}\}$. By definition, each set adjacent to $\{x, y\}$ in G^- contains exactly one of x or y. Note that no vertex in $G(V_0)$ contains both x and y, as otherwise there exists a leaf l of $\mathcal{T}$ with $\{x, y\} \subsetneq \lambda(l)$, which is impossible since $\mathfrak{S}$ is not compatible. Since all vertices of $G(V_0)$ containing x (*resp.* all vertices of $G(V_0)$ containing y) form a clique in $G(V_0)$, it follows that $G(V_0)$ has at most two connected components. Moreover, if $G(V_0)$ has two connected components, then there exists at least one vertex adjacent to $\{x, y\}$ containing x, and one vertex adjacent to $\{x, y\}$ in G^- containing y. This means that there exists two leaves l_x, l_y of $\mathcal{T}$ such that $\lambda(l_x)$ contains x and $\lambda(l_y)$ contains y, which completes the proof of the claim.

To conclude the proof of the proposition, we distinguish between the cases where $G(V_0)$ has one and two connected components.

If $G(V_0)$ has one connected component, then it follows by Lemma 18 that $\Delta(\lambda(V_0)) \geqslant |V_0| - 1$. Since by the above we have $\Delta(\mathcal{M}') \geqslant \Delta(\lambda(V_0)) + |V_1| - 1$, it follows that $\Delta(\mathcal{M}') \geqslant |V_0| + |V_1| - 2 = |\mathfrak{S}| - 2 = k - 3$, which concludes the proof.

If $G(V_0)$ has two connected components, then it follows by Lemma 18 that $\Delta(\lambda(V_0)) \geqslant |V_0| - 2$. Moreover, the fact that $G(V_0)$ has two connected components also implies that there exists l_x and l_y in V_0 such that $x \in \lambda(l_x)$ and $y \in \lambda(l_y)$. By the above, this in turn implies that $\Delta(\mathcal{M}') = \Delta(\lambda(V_0)) + |V_1|$. Putting these two relationships together, we get $\Delta(\mathcal{M}') \geqslant |V_0| + |V_1| - 2 = |\mathfrak{S}| - 2 = k - 3$, which concludes the proof. □

Remark 20. There are examples of 2,3-split systems $\mathfrak{F}$ that satisfy the conditions of Proposition 19. For example, take

$$\mathfrak{F} = \{ab|\overline{ab}, ac|\overline{ac}, bc|\overline{bc}, cd|\overline{cd}, ce|\overline{ce}, de|\overline{de}, abc|\overline{abc}, cde|\overline{cde}\}$$

on $\mathcal{M} = \{a^2, b^2, c^5, d^2, e^2\}$. Note that in this example, $8 = \Delta(\mathcal{M}) > |\mathfrak{F}| - 3 = 5$.

We will also need the following proposition.

Proposition 21. *Suppose $\mathfrak{F}$ is an incompatible 2,3-split system on $\mathcal{M}$ containing $k \geqslant 3$ splits, in which every split has multiplicity 1 such that all proper subsets of $\mathfrak{F}$ are compatible. If $\mathfrak{F}_t^\times \neq \emptyset$, then $\Delta(\mathcal{M}) \geqslant k-3$.*

Proof. Note first that since $\Delta(\mathcal{M}) \geqslant 1$ always holds, the proposition trivially holds for split systems of size $k = 3$.

Suppose now that $\mathfrak{F}$ has size $k \geqslant 4$, and assume that the result holds for all $k' < k$. Let $A_0 \in \mathfrak{F}_t^\times$. In particular, $2 \leqslant |A_0| \leqslant 3$ must hold. Put $\mathfrak{F}^- = \mathfrak{F}^-_{A_0}$ and $\mathcal{M}^- = \mathcal{M}^-_{A_0}$. By Theorem 15, $\mathfrak{F}^-$ is incompatible, and every subset of $\mathfrak{S}^-$ of size $k-2$ is compatible.

We now show that $\mathfrak{S}^-$ is a 2,3-split system on $\mathcal{M}^-$ in which every split has multiplicity 1. For $S \in \mathfrak{F} - \{A_0|\overline{A_0}\}$, we denote by S^- the split of $\mathfrak{S}^-$ corresponding to S, that is, $S^- = (S(A_0) - A_0) \cup \{a_0\}|\overline{S(A_0)}$. By definition, we have $S^- = S'^-$ if and only if $S = S'$. Thus, since every split of $\mathfrak{F}$ has multiplicity 1, every split of $\mathfrak{F}^-$ has multiplicity 1. Moreover S is distinct from $A_0|\overline{A_0}$, so we have $2 \leqslant |(S(A_0)-A_0)\cup\{a_0\}| \leqslant S(A_0)-|A_0|+1$. Since $|S(A_0)| > |A_0|$, $|S(A_0)| > 2$ must hold, it follows that if $|S(A_0)| = 3$, then $|A_0| = 2$. Moreover, the latter inequality implies $|(S(A_0)-A_0)\cup\{a_0\}| = 2$. If otherwise, $|S(A_0)| > 3$, then since S has size 2 or 3, $2 \leqslant |\overline{S(A_0)}| \leqslant 3$ must hold. In both cases, if follows that S^- has a part of size 2 or 3, and so $\mathfrak{S}^-$ is a 2,3-split system.

To conclude the proof, note that if $(\mathfrak{S}^-)_t^\times = \emptyset$, then we can use Proposition 19, to conclude that $\Delta(\mathcal{M}^-) \geqslant k-4$. Otherwise, if $(\mathfrak{S}^-)_t^\times \neq \emptyset$, then by our induction hypothesis, we have $\Delta(\mathcal{M}^-) \geqslant k-3$. In both cases, since $\Delta(\mathcal{M}^-) < \Delta(\mathcal{M})$ holds by definition of $\mathcal{M}^-$, $\Delta(\mathcal{M}) \geqslant k-3$ follows. □

We now prove the main result of this section.

Theorem 22. *Suppose that $\mathfrak{S}$ is a 2,3-split system on a multiset $\mathcal{M}$ in which every split has multiplicity 1 (i.e. $\mathfrak{S}$ is a set). Then $\mathfrak{S}$ is compatible if and only if every submultiset of $\mathfrak{S}$ of size at most $\Delta(\mathcal{M})+3$ is compatible.*

Proof. The 'only if' direction is trivial. To see the 'if' direction, assume for contradiction that $\mathfrak{S}$ is not compatible. Let $\mathfrak{S}'$ be an incompatible subset of $\mathfrak{S}$ that is minimal with respect to set inclusion, that is, all proper subsets of $\mathfrak{S}'$ are compatible. Denoting by k the size of $\mathfrak{S}'$, we have by assumption that $k > \Delta(\mathcal{M})+3$. On the other hand, applying Proposition 19 or Proposition 21 to $\mathfrak{S}'$ implies that $\Delta(\mathcal{M}) \geqslant k-3$, a contradiction. □

6 Discussion

In this paper, we have shown that Conjecture 3 holds for some special classes of split systems, and that a slightly weaker version holds for $2,3$-split systems. In the special case of $2,3$-split systems, the arguments not only rely on the particular structure of the tree $\mathcal{T}$ representing such a split system (Lemma 17), but also on the fact that, for a $2,3$-split system $\mathfrak{S}$, the split system $\mathfrak{S}^-$ as defined in Section 4 is also a $2,3$-split system. These two properties make it impractical to easily extend the current proof in order to generalize

Theorem 22 to other classes of split systems with restricted split sizes. This being said, future work on trying to prove Conjecture 3 might start by considering other special types of split systems (for example, maybe it could be possible to somehow define and study split systems that are close to being thin).

Even though we would ideally like to show that Conjecture 3 holds in general, we note that for multisets $\mathcal{M}$ where $\Delta(\mathcal{M})$ is large, there is a substantial difference between the bound $2\Delta(\mathcal{M})$ guaranteed by Theorem 2, and the bound $\Delta(\mathcal{M}) + 2$ postulated by Conjecture 3. It might therefore be interesting to find ways to improve the former bound, even if such an improvement might not go as far as $\Delta(\mathcal{M}) + 2$ (such as the bound in Theorem 22).

Finally, as mentioned in the introduction, the work that we have presented has some links to the perfect phylogeny problem. With regards to that problem, a quite complex counterexample was recently found to a long-standing conjecture that is somewhat similar in nature to Conjecture 3 [15]. Bearing this in mind, it might also be worth to look into ways to systematically construct a counterexample to Conjecture 3. Indeed, even if there is no such counterexample, this approach might yield new ideas for tackling this intriguing problem.

Acknowledgements

GES would like to thank the streets of Norwich and the seaside town of Sheringham for inspirational autumn walks, during which ideas for some of the proofs arose. Both authors thank Katharina Huber for helpful discussions. Both authors would like to thank the anonymous referees for their careful reading of the first version of this manuscript and for their helpful suggestions for its improvement.

References

[1] P. Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the Archeological and Historical Sciences*, pages 387–395, 1971.

[2] Y. Cui, J. Jansson, and W.-K. Sung. Polynomial-time algorithms for building a consensus MUL-tree. *Journal of Computational Biology*, 19(9):1073–1088, 2012.

[3] M. Delabre, N. El-Mabrouk, K. T. Huber, M. Lafond, V. Moulton, E. Noutahi, and M. S. Castellanos. Evolution through segmental duplications and losses: a super-reconciliation approach. *Algorithms for Molecular Biology*, 15:1–15, 2020.

[4] G. Ganapathy, B. Goodson, R. Jansen, H.-S. Le, V. Ramachandran, and T. Warnow. Pattern identification in biogeography. *IEEE Trans. Comput. Biol. Bioinform.*, 3:334–346, 2006.

[5] M. Gascon, R. Dondi, and N. El-Mabrouk. MUL-tree pruning for consistency and optimal reconciliation-complexity and algorithms. *Theoretical Computer Science*, 937:22–38, 2022.

[6] D. Gusfield. Haplotyping as perfect phylogeny: conceptual framework and efficient solutions. In *Proceedings of the sixth annual international conference on Computational biology*, pages 166–175, 2002.

[7] C. Hampson, D. J. Harvey, C. S. Iliopoulos, J. Jansson, Z. Lim, and W.-K. Sung. MUL-tree pruning for consistency and compatibility. In *34th Annual Symposium on Combinatorial Pattern Matching (CPM 2023)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2023.

[8] K. T. Huber, M. Lott, V. Moulton, and A. Spillner. The complexity of deriving multi-labeled trees from bipartitions. *Journal of Computational Biology*, 15:639–651, 2008.

[9] K. T. Huber and L. J. Maher. The hybrid number of a ploidy profile. *Journal of Mathematical Biology*, 85(3):30, 2022.

[10] K. T. Huber and V. Moulton. Phylogenetic networks from multi-labelled trees. *Journal of Mathematical Biology*, 52(5):613–632, 2006.

[11] R. D. M. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58–77, 1994.

[12] M. Popp and B. Oxelman. Inferring the history of the polyploid silene aegaea (caryophyllaceae) using plastid and homoeologous nuclear dna sequences. *Molecular Phylogenetics and Evolution*, 20:474–481, 2001.

[13] C. J. Rothfels. Polyploid phylogenetics. *New Phytologist*, 230(1):66–72, 2021.

[14] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.

[15] L. van Iersel, M. Jones, and S. Kelk. A third strike against perfect phylogeny. *Systematic Biology*, 68(5):814–827, 2019.

[16] H. Wang. Split sizes and extremal tree shapes. *Advances in Applied Mathematics*, 104:135–164, 2019.