An Algebraic View of the Relation between Largest Common Subtrees and Smallest Common Supertrees

Rosselló, F. and Valiente, G.

Research Report LSI-04-60-R

Departament de Llenguatges i Sistemes Informàtics

UNIVERSITAT POLITÈCNICA DE CATALUNYA

An Algebraic View of the Relation between Largest Common Subtrees and Smallest Common Supertrees

Francesc Rosselló *,[1]

Department of Mathematics and Computer Science, Research Institute of Health Science (IUNICS), University of the Balearic Islands, E-07122 Palma de Mallorca

Gabriel Valiente [2]

Department of Software, Technical University of Catalonia, E-08034 Barcelona

Abstract

The relationship between two important problems in tree pattern matching, the largest common subtree and the smallest common supertree of two trees, is established by means of simple constructions, which allow one to obtain the largest common subtree from the smallest common supertree, and vice versa. These constructions are given for the problems of isomorphic, homeomorphic, topological, and minor embeddings. They can be implemented by a straightforward extension of any algorithm that solves one of the two problems, and the extension only takes time linear in the size of the trees.

Key words:
Tree pattern matching; subtree isomorphism; subtree homeomorphism, topological embedding; minor containment; largest common subtree; smallest common supertree; pushout; pullback.

* Corresponding author. Fax: +34-971-173-003.
Email addresses: cesc.rossello@uib.es (Francesc Rosselló), valiente@lsi.upc.es (Gabriel Valiente).

[1] Partially supported by the Spanish DGES and the EU program FEDER, project ALBIOM (BFM2003-00771).
[2] Partially supported by Spanish CICYT project MAVERISH (TIC2001-2476-C03-01) and by the Ministry of Education, Science, Sports and Culture of Japan through Grant-in-Aid for Scientific Research B-15300003 for visiting JAIST (Japan Advanced Institute of Science and Technology).

1 Introduction

Subtree isomorphism and the related problems of largest common subtree and smallest common supertree have practical applications in combinatorial pattern matching, pattern recognition, chemical structure search, computational molecular biology, and other areas of engineering and life sciences. In these areas, they are one of the most widely used techniques for comparing tree-structured data.

Largest common subtree is the problem of finding a largest tree that can be embedded in two given trees. A tree S can be embedded in another tree T in one of four different ways: S can be isomorphic to a subtree of T, it can be homeomorphic to a subtree of T, it can be topologically embedded in a subtree of T, and it can be contained in T as a minor.

A tree S is an embedded subtree of a tree T when there exists an injective mapping f from the nodes of S to the nodes of T satisfying the following condition: for every pair of nodes a and b of S, if there is an arc from a to b in S, then

isomorphic embedding there exists an arc from $f(a)$ to $f(b)$ in T;
homeomorphic embedding there exists a path from $f(a)$ to $f(b)$ in T with all intermediate nodes of out-degree 1 and with no intermediate node coming from S;
topological embedding there exists a path from $f(a)$ to $f(b)$ in T; and if there are arcs from a to two distinct nodes b and c in S, then the paths from $f(a)$ to $f(b)$ and to $f(c)$ in T have no node in common other than $f(a)$;
minor embedding there exists a path from $f(a)$ to $f(b)$ in T with no intermediate node coming from S.

These embedding problems for trees under different embedding relations, have been thoroughly studied in the literature. Their complexity is already settled: they are polynomial-time solvable for isomorphic, homeomorphic, and topological embeddings, and they are NP-complete for minor embeddings [4,10–12]. Efficient algorithms are known for subtree isomorphism [14,18], for subtree homeomorphism [2,19,20], for largest common subtree under isomorphic embeddings [18] and homeomorphic embeddings [13], and for both largest common subtree and smallest common supertree under isomorphic and topological embeddings [7]. The only (exponential) algorithm known for largest common subtree under minor embeddings is [15].

Particular cases of these embedding problems for trees have also been thoroughly studied in the literature. On ordered trees, they become polynomial-time solvable for isomorphic, homeomorphic, topological, and also minor embeddings. In this particular case, the largest common subtree problem un-

der homeomorphic embeddings is known as the maximum agreement subtree problem [1,3,17], the largest common subtree problem under minor embeddings is known as the tree edit problem [5,16,22], and the smallest common supertree problem under minor embeddings is known as the tree alignment problem [8,9,21]. The smallest common supertree problem under minor embeddings was also studied in [12] for trees of bounded degree.

In this paper, the relationship between the largest common subtree and the smallest common supertree of two trees is established by means of simple constructions, which allow one to obtain the largest common subtree from the smallest common supertree, and vice versa. These constructions are given for the problems of isomorphic, homeomorphic, topological, and minor embeddings. They can be implemented by a straightforward extension of any algorithm that solves one of the two problems, and the extension only takes time linear in the size of the trees.

Roughly, given two trees T_1 and T_2 and a largest common subtree T_μ of them, together with a pair of witness embeddings of T_μ into T_1 and into T_2, then a smallest common supertree T_σ of T_1 and T_2 can be obtained from the disjoint union of T_1 and T_2 by first merging all those nodes coming from a same node of T_μ and then, removing all parallel arcs and all arcs subsumed by paths. Conversely, given two trees T_1 and T_2 and a smallest common supertree T of them, together with a pair of witness embeddings of T_1 and T_2 into T, then a largest common subtree T_p of T_1 and T_2 can be obtained from T by just removing all nodes not coming from both T_1 and T_2. However, the justification for these simple constructions is rather intricate, and differs substantially for isomorphic, homeomorphic, topological, and minor embeddings.

Beyond their theoretical interest, these constructions provide an efficient solution to the smallest common supertree problem under homeomorphic embeddings, for which no previous algorithm is known. The solution extends the largest common homeomorphic subtree algorithm of [13], which in turn extended the subtree homeomorphism algorithm of [19,20].

In a similar vein, these constructions also provide a solution to the smallest common supertree problem under minor embeddings, for which no previous algorithm is known, either. The solution extends the unordered tree edit algorithm of [15].

This is, to the best of our knowledge, the first unified construction showing the relation between largest common subtrees and smallest common supertrees for the four embedding problems studied in the literature: isomorphic, homeomorphic, topological, and minor embeddings. A similar correspondence between largest common subgraphs and smallest common supergraphs under isomorphic embeddings was studied in [6].

The rest of the paper is organized as follows. Basic notions and notation are introduced in Section 2, together with some results about isomorphic, homeomorphic, topological, and minor embeddings. The construction of largest common subtrees from smallest common supertrees is studied in Section 3, while the reverse construction of smallest common supertrees from largest common subtrees is studied in Section 4. Algorithms for implementing both constructions in time linear in the size of the trees are discussed in Section 5. Finally, some conclusions are outlined in Section 6.

2 Preliminaries

There are many equivalent definitions of tree. In this paper we take the following one. A *tree* is a directed finite graph $T = (V, E)$ with V either empty or containing a distinguished node $r \in V$, called the *root*, such that for every other node $v \in V$ there exists one, and only one, path from the root r to v. Recall that every node in a tree has in-degree 1, except the root that has in-degree 0.

Henceforth, and unless otherwise stated, given a tree T we shall denote its set of nodes by $V(T)$ and its set of arcs by $E(T)$. The *order* of a tree T is the cardinal $|V(T)|$ of its set of nodes and its *size* the cardinal $|E(T)|$ of its set of arcs.

The *children* of a node v in a tree T are those nodes w such that $(v, w) \in E(T)$: in this case we also say that v is the *parent* of its children. The only node without a parent is the root, and the nodes without children are the *leaves* of the tree.

Given a path $(v_0, v_1, \ldots, v_k)$ in a tree T, its *origin* is v_0, its *end* is v_k, and its *intermediate nodes* are $v_1, \ldots, v_{k-1}$. Such a path is *non-trivial* when $k \geqslant 1$.

A path $(v_0, v_1, \ldots, v_k)$ in a tree T is *elementary* when, for every $i = 1, \ldots, k-1$, v_{i+1} is the only child of v_i; in other words, when all its intermediate nodes have out-degree 1. In particular, an arc forms an elementary path.

Two non-trivial paths $(a, v_1, \ldots, v_k)$ and $(a, w_1, \ldots, w_\ell)$ in a tree T are said to *diverge* when the only node they have in common is their origin a. Notice that, by the uniqueness of paths in trees, it is equivalent to the condition $v_1 \neq w_1$. The definition of tree also implies that, for every two nodes b, c of a tree not connected by a path, there exists one, and only one, node a such that there exist divergent paths from a to b and to c.

A tree S is an *isomorphic subtree* of a tree T when there exists an injective

mapping $f : V(S) \to V(T)$ satisfying the following condition: for every $a, b \in V(S)$, if there is an arc from a to b in S, then there exists an arc from $f(a)$ to $f(b)$ in T. Such a mapping f is called an *isomorphic embedding* $f : S \to T$.

A tree S is a *homeomorphic subtree* of a tree T when there exists an injective mapping $f : V(S) \to V(T)$ satisfying the following condition: for every $a, b \in V(S)$, if there is an arc from a to b in S, then there exists an elementary path from $f(a)$ to $f(b)$ in T with no intermediate node in $f(V(S))$. In this case, the mapping f is said to be a *homeomorphic embedding* $f : S \to T$.

Example 1 *Let T_1 be a tree with $V(T_1) = \{r_1, x_1, y_1\}$ and $E(T_1) = \{(r_1, x_1), (r_1, y_1)\}$ and let T_2 be a tree with $V(T_2) = \{r_2, x_2, y_2\}$ and $E(T_2) = \{(r_2, x_2), (x_2, y_2)\}$. The mapping $f : V(T_1) \to V(T_2)$ defined by $f(r_1) = r_2$, $f(x_1) = x_2$, and $f(y_1) = y_2$ satisfies that, for every $a, b \in V(T_1)$, if there is an arc from a to b in T_1, then there exists an elementary path from $f(a)$ to $f(b)$ in T_2, but it is not a homeomorphic embedding.*

A tree S is a *topological subtree* of a tree T when there exists an injective mapping $f : V(S) \to V(T)$ satisfying the following two conditions: for every $a, b \in V(S)$, if there is an arc from a to b in S, then there exists a path from $f(a)$ to $f(b)$ in T; and if $(a, b), (a, c) \in E(S)$ with $b \neq c$, then the paths from $f(a)$ to $f(b)$ and from $f(a)$ to $f(c)$ in T diverge. Such a mapping f is called a *topological embedding* $f : S \to T$.

In the definition of homeomorphic embedding (as well as in the definition of minor embedding below) we explicitly impose the condition on the image of an arc to be a path (of some specific type) such that no intermediate node in it comes from the source tree, and Example 1 shows that this extra condition was independent of the general one. This is not the case for topological embeddings, as the following lemma shows.

Lemma 2 *Let $f : S \to T$ be a topological embedding. For every $a, b \in V(S)$, if $(a, b) \in E(S)$, then the path from $f(a)$ to $f(b)$ in T does not contain any intermediate node in $f(V(S))$.*

PROOF. To begin with, notice that, since arcs in S are transformed into paths in T, it is clear that paths in S are also transformed into paths in T. Now, let $a, b \in V(S)$ be such that $(a, b) \in E(S)$ and let $c \in V(S)$ be such that $f(c)$ is an intermediate node in the path from $f(a)$ to $f(b)$ in S. Then, there cannot exist paths from c to a or from b to c in S: their images under f would build up paths in T with the paths from $f(a)$ to $f(c)$ and from $f(c)$ to $f(b)$, respectively.

Assume that there exists a path from a to c in S, say $(a, v_1, \ldots, c)$. Since it cannot cross b, we have that $v_1 \neq b$. Therefore the paths from $f(a)$ to $f(b)$

and to $f(v_1)$ diverge, and hence the paths from $f(a)$ to $f(b)$ and to $f(c)$ also diverge. This contradicts the assumption that $f(c)$ is an intermediate node of the path from $f(a)$ to $f(b)$.

So, a and c are not connected by a path. This implies that there exist in S a node x and paths $(x, v_1, \ldots, v_k, a)$ and $(x, w_1, \ldots, w_\ell, c)$ such that $w_1 \neq v_1$. But then the images of the arcs (x, v_1) and (x, w_1) are divergent paths in T, and this contradicts the existence in T of paths from $f(w_1)$ to $f(c)$ and from $f(v_1)$ to $f(c)$ (through $f(a)$). Therefore, c cannot exist. $\square$

A tree S is a *minor* of a tree T when there exists an injective mapping $f : V(S) \to V(T)$ satisfying the following condition: for every $a, b \in V(S)$, if there is an arc from a to b in S, then there exists a path from $f(a)$ to $f(b)$ in T with no intermediate node in $f(V(S))$. In this case, the mapping f is said to be a *minor embedding* $f : S \to T$.

Example 1 shows that a mapping that transforms arcs into paths need not be a minor embedding.

The following lemma will be used several times in the next sections.

Lemma 3 *Let $f : S \to T$ be a minor embedding. For every $a, b \in V(S)$, there exists a path from a to b in S if and only if there exists a path from $f(a)$ to $f(b)$ in T. Moreover, if the path from $f(a)$ to $f(b)$ in T is elementary, then the path from a to b in S is also so, and if there is an arc from $f(a)$ to $f(b)$, then there is an arc from a to b.*

PROOF. Since the arcs in S become paths in T with all intermediate nodes "new," it is obvious that a path from a to b in S becomes, under f, a path from $f(a)$ to $f(b)$ in T such that the intermediate nodes in this path that belong to $f(V(S))$ are exactly the images under f of the intermediate nodes of the path from a to b.

Assume now that there exists a path from $f(a)$ to $f(b)$ in T, and let r be the root of S. If $a = r$ or $a = b$, it is clear that there exists a path from a to b in S. If $a \neq r$ and $a \neq b$, then the images of the paths from r to a and to b in S are paths from $f(r)$ to $f(a)$ and to $f(b)$ in T. Now, the uniqueness of paths in T implies that the path from $f(r)$ to $f(b)$ splits into the path from $f(r)$ to $f(a)$ and the path from $f(a)$ to $f(b)$, and in particular that $f(a)$ is an intermediate node in the path from $f(r)$ to $f(b)$. Then, the fact that the only intermediate nodes in this path that belong to $f(V(S))$ are the images under f of the intermediate nodes of the path from r to b, together with the injectivity of f, imply that a is an intermediate node in the path from r to b: that there is a path from a to b in S.

Moreover, if a node in S has more than one children, then its image under f has also more than one children. This implies that if the path from $f(a)$ to $f(b)$ is elementary, then the path from a to b is elementary, too. Finally, if there is an arc from $f(a)$ to $f(b)$, then there is a path from a to b that cannot have any intermediate node: an arc. $\square$

We clearly have the following implications:

$$\begin{array}{ccccccc} \text{isomorphic} \\ \text{subtree} & \Rightarrow & \text{homeomorphic} \\ & & \text{subtree} & \Rightarrow & \text{topological} \\ & & & & \text{subtree} & \Rightarrow & \text{minor} \end{array}$$

In particular, the thesis of Lemma 3 is true if f is an isomorphic embedding, a homeomorphic embedding, or a topological embedding.

A *largest common isomorphic subtree* of two trees S and T is a tree that is an isomorphic subtree of both of them and has the largest order among all isomorphic subtrees of both of them. *Largest common homeomorphic subtrees*, *largest common topological subtrees*, and *largest common minors* are defined in a similar way: a *largest common homeomorphic subtree* of two trees S and T is a tree that is a homeomorphic subtree of both of them and has the largest order among all homeomorphic subtrees of both of them, and so on.

A *smallest common isomorphic supertree* of two trees S and T is a tree such that both S and T are isomorphic subtrees of it and has the least order among all trees with this property. *Smallest common homeomorphic supertrees*, *smallest common topological supertrees*, and *smallest common supertrees under minor embeddings* are again defined in a similar way: for instance, a *smallest common supertree under minor embeddings* of two trees S and T is a tree such that both S and T are minors of it and has the least order among all trees having S and T as minors.

We shall denote by Tree_{iso}, Tree_{hom}, Tree_{top}, and Tree_{min} the categories of trees with morphisms the isomorphic, homeomorphic, topological, and minor embeddings, respectively. Whenever we denote generically any one of these categories by Tree_*, we shall use the following notations. By a Tree_*-*embedding* we shall mean a morphism in the corresponding category. By a (*largest*) *common* Tree_*-*subtree* of two trees we shall mean a tree endowed with a Tree_*-embedding to these two trees (respectively, with the largest order among all trees with this property). By a (*smallest*) *common* Tree_*-*supertree* of two trees we shall mean a tree endowed with a Tree_*-embedding from these two trees (respectively, with the smallest order among all trees with this property). And by a Tree_*-*path* we shall understand an arc if Tree_* stands for Tree_{iso}, an elementary path if Tree_* denotes Tree_{hom}, and an arbitrary path if Tree_* means Tree_{top} or Tree_{min}. Notice in particular that a Tree_*-embedding always transforms an arc into a Tree_*-path, and that an arc is always a Tree_*-path.

The following lemma is a direct consequence of Lemma 3.

Lemma 4 *Let* Tree_* *denote any category* Tree_{iso}, Tree_{hom}, Tree_{top}, *or* Tree_{min}, *and let* $f : S \to T$ *be a* Tree_**-embedding. For every* $a, b \in V(S)$, *if there exists a* Tree_**-path from* $f(a)$ *to* $f(b)$ *in* T, *then there exists a* Tree_**-path from* a *to* b *in* S.

3 Common subtrees as pullbacks

Let $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$ be henceforth two minor embeddings. Without any loss of generality, and unless otherwise stated, we shall assume that $V(T_1), V(T_2) \subseteq V(T)$ and that these minor embeddings f_1 and f_2 are given by these set-theoretical inclusions. For simplicity, we shall denote thus the image of a node $a \in V(T_i)$ under the corresponding f_i again by a.

Let T_p be the graph with set of nodes $V(T_p) = V(T_1) \cap V(T_2)$ and set of arcs defined in the following way: for every $a, b \in V(T_1) \cap V(T_2)$, there is an arc from a to b in T_p if and only if there is a path from a to b in T_1 and in T_2 such that no intermediate node of this path belongs to $V(T_1) \cap V(T_2)$. We shall call this graph T_p the *intersection* of T_1 and T_2 obtained through f_1 and f_2.

This graph satisfies the following useful lemma.

Lemma 5 *For every* $a, b \in V(T_1) \cap V(T_2)$*:*

(i) If there exists a path in T_p *from* a *to* b, *then there are paths from* a *to* b *in* T_1 *and in* T_2.
(ii) If there exists a path in T_1 *or in* T_2 *from* a *to* b, *then there exists also a path in* T_p *from* a *to* b *whose intermediate nodes are exactly the intermediate nodes in the path from* a *to* b *in* T_1 *or in* T_2 *that belong to* $V(T_1) \cap V(T_2)$.

PROOF. Point (i) is a direct consequence of the fact that every arc in T_p corresponds to paths in T_1 and T_2.

As far as point (ii) goes, we shall prove that if there exists a path in T_1 going from a to b, then there exists also a path in T_p from a to b with intermediate nodes the intermediate nodes of the path in T_1 that belong to $V(T_1) \cap V(T_2)$, by induction on the number n of such intermediate nodes belonging to $V(T_1) \cap V(T_2)$.

If $n = 0$, then there exists a path in T_1 from a to b that does not cross any node in $V(T_1) \cap V(T_2)$. Since f_1 transforms arcs into paths with no intermediate node belonging to T_1, this implies that there exists a path in T from a to b

that does not cross any node in $V(T_1) \cap V(T_2)$, either. Then, by Lemma 3, this path is induced by a path in T_2 from a to b, and by the same reason this path does not contain any intermediate node in $V(T_1) \cap V(T_2)$. So, there are paths from a to b in T_1 and T_2 that do not contain any intermediate node in $V(T_1) \cap V(T_2)$, and therefore, by definition, there exists an arc from a to b in T_p.

As the induction hypothesis, assume that the claim is true for paths in T_1 with n intermediate nodes in $V(T_1) \cap V(T_2)$, and assume now that the path going from a to b has $n+1$ such nodes. Let a_0 be the first intermediate node in this path belonging to $V(T_1) \cap V(T_2)$. Then, by the case $n = 0$, there is an arc in T_p from a to a_0, and by the induction hypothesis there is a path from a_0 to b in T_p whose only intermediate nodes are the intermediate nodes in the path in T_1 from a_0 to b that belong to $V(T_1) \cap V(T_2)$; by concatenating them we obtain the path from a to b in T_p we were looking for. $\square$

The intersection of two minors need not be a tree, as the following example shows.

Example 6 *Let T be a tree with nodes a_1, a_2, b, c and arcs $(a_1, a_2), (a_2, b)$, (a_2, c), and let $T_1 = T_2$ be the tree with nodes a, b, c and arcs $(a, b), (a, c)$. Let $f_1 : T_1 \to T$ be the minor embedding defined by $f_1(a) = a_1$, $f_1(b) = b$ and $f_1(c) = c$, and $f_2 : T_2 \to T$ the minor embedding defined by $f_2(a) = a_2$, $f_2(b) = b$ and $f_2(c) = c$. In this case T_p is the graph with nodes b, c and no arc, and in particular it is not a tree.*

Now we have the following result.

Proposition 7 *T_p can be enlarged to a common minor of T_1 and T_2.*

PROOF. If T_p is empty, then it is a tree and its inclusions into T_1 and T_2 are clearly minor embeddings. In this case, T_p is a common minor of T_1 and T_2.

So, assume in the sequel that T_p is non-empty. If it has no node without parents, then it must contain a circuit and this implies the existence of circuits in the trees T_1 and T_2, which yields a contradiction. Therefore, T_p must contain some node without a parent. Now we must consider two possibilities.

(1) T_p has only one node r_p without a parent. Then every other node a in T_p can be reached from r_p through a path, because this graph does not contain any circuit (as we have seen) and hence it must contain a path from a node of in-degree 0 to a. To check that this path is unique, we shall prove that no node in T_p has in-degree greater than 1.

Indeed, assume that there are nodes $a, b, c \in V(T_p)$, with $b \neq c$, and arcs from b and c to a. This means that there are paths in T_1 and in T_2 from b and c to a that do not contain any intermediate node in $V(T_1) \cap V(T_2)$. But since, say, T_1 is a tree, if there exist paths in T_1 from b and c to a, one of the nodes b, c must be intermediate in the path from the other one to a, which yields a contradiction.

Therefore, in this case T_p is a tree. And by definition, for every $a, b \in V(T_p)$, if there is an arc from a to b in T_p then there are paths from a to b in T_1 and in T_2 without any intermediate node in $V(T_p)$. Therefore, the set-theoretical inclusions $V(T_p) \hookrightarrow V(T_i)$ induce minor embeddings $\iota_i : T_p \to T_i$, for $i = 1, 2$, and hence T_p is a common minor of T_1 and T_2.

(2) T_p contains more that one node without a parent, say $x_1, \ldots, x_k$. The same argument used in (1) shows in this case that every other node $a \in V(T_p)$ can be reached from one of these nodes x_i through a path in T_p, and that no node in T_p has in-degree greater than 1.

Let now $\widetilde{T}_p$ be the graph obtained by adding to T_p one node r and an arc from r to each x_i, $i = 1, \ldots, k$. Then, r is the only node without a parent in this graph and every node can be reached from r in $\widetilde{T}_p$ through a unique path. Indeed, each x_i can be reached from r through the new arc going from r to it, and then every other node in $\widetilde{T}_p$ can be reached from r by the path going from some x_i to it in T_p preceded by the corresponding arc from r to this x_i. And these paths are unique, because no node in $\widetilde{T}_p$ has in-degree greater than 1.

Therefore, $\widetilde{T}_p$ is a tree with root r.

Now, notice that there is no non-trivial path in either T_1 or T_2 from any node belonging to $V(T_1) \cap V(T_2)$ to any x_i: such a path, by Lemma 5, would induce a non-trivial path in T_p and therefore the node x_i would have a parent in T_p. In particular, neither the root of T_1 nor the root of T_2 belong to $V(T_1) \cap V(T_2)$.

Consider then the injective mappings $\widetilde{\iota}_i : \widetilde{T}_p \to T_i$, $i = 1, 2$, defined by the set-theoretical inclusions on $V(T_p)$ and sending r to the root of the corresponding T_i. It is clear that they are minor embeddings: on the one hand, arguing as in (1) above (and recalling that each $\widetilde{\iota}_i$ sends r to the root of the corresponding T_i), the restriction of each $\widetilde{\iota}_i$ to T_p sends every arc to a path in T_i without any intermediate node coming from $\widetilde{T}_p$; on the other hand, $\widetilde{\iota}_i$ sends every arc (r, x_ℓ) to the path in T_i going from its root to x_ℓ which, as we saw above, cannot contain any intermediate node in $V(T_1) \cap V(T_2)$.

Thus, $\widetilde{T}_p$ is a common minor of T_1 and T_2.

In all, we have proved that T_1 and T_2 have always a common minor that is either T_p or is obtained by adding a root to this graph. $\square$

Let us restrict now from minor embeddings to topological embeddings. In this

case we have the following result.

Proposition 8 *Let $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$ be topological embeddings. Then:*

(i) T_p is a tree.
(ii) The set-theoretical inclusions $V(T_p) \hookrightarrow V(T_i)$ induce topological embeddings $\iota_i : T_p \to T_i$, for $i = 1, 2$. Therefore, T_p is a common topological subtree of T_1 and T_2.

PROOF. (i) From the proof of Proposition 7, and specifically point (1) in it, we deduce that it is enough to prove that if $V(T_1) \cap V(T_2) \neq \emptyset$, then T_p has only one node without a parent.

So, assume that $a, b \in V(T_1) \cap V(T_2)$ have no parent in T_p. Then, neither T_1 nor T_2 contains any non-trivial path from some node in $V(T_1) \cap V(T_2)$ to one of these nodes: by Lemma 5, such a path would entail a non-trivial path in T_p finishing in a or b and hence one of these nodes would have a parent in T_p. In particular, there is no path connecting a and b.

Let $x_1 \in V(T_1)$ be the node such that there exist divergent paths from x_1 to a and to b in T_1; let v_1 and v_1' be the first nodes (after x_1) in these paths. Since the inclusion $f_1 : T_1 \to T$ is a topological embedding, there are divergent paths from x_1 to v_1 and to v_1', which are followed then by paths from v_1 to a and from v_1' to b. This means that x_1 is also the node in T such that there exist divergent paths from x_1 to a and to b in T.

Now, if $x_2 \in V(T_2)$ is the node such that there exist divergent paths from x_2 to a and to b in T_2, then, by symmetry, x_2 is also the node with this property in T. Therefore, $x_1 = x_2$ and both a and b can be reached from a node in $V(T_1) \cap V(T_2)$ through paths in T_1 and in T_2, which yields a contradiction.

(ii) We shall prove that ι_1 is a topological embedding. By point (1) in the proof of Proposition 7, we already know that ι_1 is a minor embedding. So, it remains to prove that if there are arcs from a to b and to c in T_p, then the paths from a to b and to c in T_1 diverge. The argument will be similar to the one used in (i).

To begin with, since the paths from a to b and to c have no intermediate node in $V(T_1) \cap V(T_2)$, neither b nor c appears in the path from a to the other one, and therefore there is no path connecting b and c.

Then, let $x_1 \in V(T_1)$ be the node such that there exist divergent paths from x_1 to b and to c in T_1. Since there are paths from a to b and to c in T_1, either $x_1 = a$ or there exists a path in T_1 from a to x_1. Now, arguing as in (i), this

node x_1 is also the node in T such that there exist divergent paths from x_1 to b and to c in T.

And then, also arguing as in (i), if $x_2 \in V(T_2)$ is the node such that there exist divergent paths from x_2 to b and to c in T_2, then $x_1 = x_2$. In particular, $x_1 \in V(T_1) \cap V(T_2)$. Since there cannot be any intermediate node in the paths from a to b and to c in T_1 belonging to $V(T_1) \cap V(T_2)$, we conclude that $a = x_1$ and the paths from a to b and to c in T_1 are divergent, as we wanted to prove. $\square$

We have similar results for homeomorphic and isomorphic embeddings.

Proposition 9 *Let $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$ be homeomorphic embeddings. Then, the set-theoretical inclusions $V(T_p) \hookrightarrow V(T_i)$ induce homeomorphic embeddings $\iota_i : T_p \to T_i$, for $i = 1, 2$. Therefore, in this case T_p is a common homeomorphic subtree of T_1 and T_2.*

PROOF. We already know (Proposition 8.(i)) that T_p is a tree. We shall prove now that $\iota_1 : T_p \to T_1$ is a homeomorphic embedding. By point (1) in the proof of Proposition 7, if there is an arc from a to b in T_p, then there is a path in T_1 from a to b without any intermediate node in $V(T_1) \cap V(T_2)$. It remains to prove that this path is elementary.

Now, if $a, b \in V(T_1) \cap V(T_2)$ and there exists a path from a to b in T_1, then every node in this path with more than one children also belongs to $V(T_1) \cap V(T_2)$. Indeed, if there is a path from a to b in T_1, then there is a path from a to b in T and then, by Lemma 3, there is also a path from a to b in T_2. Every node with more than one children in the path in T_1 has also more than one children in T. Now, since every arc in T_2 becomes, under the homeomorphic embedding $f_2 : T_2 \to T$, an elementary path with no intermediate node in $V(T_2)$, and no node with more than one children can be an intermediate node of an elementary path, we deduce that every node with more than one children in the path from a to b in T must belong to $V(T_2)$. Therefore, every node in the path from a to b in T_1 with more than one children belongs to $V(T_1) \cap V(T_2)$.

Thus, if there is an arc from a to b in T_p, then there is a path in T_1 without any intermediate node in $V(T_1) \cap V(T_2)$, and hence, as we have seen, without any intermediate node with more than one children.

Notice in particular that this argument shows that, for every $a, b \in V(T_1) \cap V(T_2)$, there is an arc from a to b in T_p if and only if there are elementary paths from a to b in T_1 and in T_2 without any intermediate node in $V(T_1) \cap V(T_2)$. $\square$

Proposition 10 *Let $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$ be isomorphic embeddings. Then, the set-theoretical inclusions $V(T_p) \hookrightarrow V(T_i)$ induce isomorphic embeddings $\iota_i : T_p \to T_i$, for $i = 1, 2$. Thus, in this case T_p is a common isomorphic subtree of T_1 and T_2.*

PROOF. We already know from Proposition 9 that T_p is a tree and that $\iota_1 : T_p \to T_1$ and $\iota_2 : T_p \to T_2$ are homeomorphic embeddings, and in particular that if there is an arc from a to b in T_p, then there are elementary paths in T_1 and in T_2 from a to b without any intermediate node in $V(T_1) \cap V(T_2)$.

Now, if there are non-trivial paths from a to b in T_1 and in T_2, the node b has a parent in each one of these trees; let these parents be c_1 and c_2, respectively. Then both c_1 and c_2 are parents of b in T, which implies that $c_1 = c_2 \in V(T_1) \cap V(T_2)$: the parents in T_1 and in T_2 of b are the same and they belong to $V(T_1) \cap V(T_2)$. We conclude that if the paths from a to b in T_1 and in T_2 have no intermediate node in $V(T_1) \cap V(T_2)$, then they must be arcs.

In all, this proves that if there is an arc from a to b in T_p, then there are arcs in T_1 and in T_2 from a to b.

Notice in particular that we have also proved that, for every $a, b \in V(T_1) \cap V(T_2)$, there is an arc from a to b in T_p if and only if there are arcs from a to b in T_1 and in T_2. $\square$

We have finally the following result.

Proposition 11 *Let Tree_* denote any category Tree_{iso}, Tree_{hom}, or Tree_{top}. For every pair of Tree_*-embeddings $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$,*

$$(T_p, \iota_1 : T_p \to T_1, \iota_2 : T_p \to T_2)$$

is a pullback of f_1 and f_2 in Tree_*.

PROOF. We know from the previous propositions that $\iota_1 : T_p \to T_1$ and $\iota_2 : T_p \to T_2$ are Tree_*-embeddings, and it is clear that $f_1 \circ \iota_1 = f_2 \circ \iota_2$. Let us check now the universal property of pullbacks in Tree_*.

Let S be any tree and let $g_1 : S \to T_1$ and $g_2 : S \to T_2$ be two Tree_*-embeddings such that $f_1 \circ g_1 = f_2 \circ g_2$. Then, at the level of nodes, there exists a unique mapping $g : V(S) \to V(T_1) \cap V(T_2)$ such that each g_i is equal to g followed by the corresponding set-theoretical inclusion $V(T_i) \hookrightarrow V(T)$. We must prove that this mapping g induces a Tree_*-embedding $g : S \to T_p$.

Let $a, b \in S$ be such that $(a, b) \in E(S)$. Then, there exist Tree_*-paths in T_1 and in T_2 from $g(a)$ to $g(b)$ without any intermediate node in $g(V(S))$. These paths induce, by Lemma 4, a Tree_*-path from $g(a)$ to $g(b)$ in T_p and it will not have any intermediate node in $g(V(S))$, either. This already shows that g is a Tree_*-embedding when Tree_* stands for Tree_{iso} or Tree_{hom}.

In the case of Tree_{top}, we must also prove that if $a, b, c \in V(S)$ are such that $(a, b), (a, c) \in E(S)$ and $b \neq c$, then the paths from $g(a)$ to $g(b)$ and to $g(c)$ in T_p diverge. But since g_1 is a topological embedding, the paths from $g(a)$ to $g(b)$ and to $g(c)$ in T_1 are divergent, and this clearly entails that the paths from $g(a)$ to $g(b)$ and to $g(c)$ in T_p are divergent, too. $\square$

Therefore, the categories Tree_{iso}, Tree_{hom}, and Tree_{top} have all pullbacks.

Remark 12 *Tree_{min} does not have all pullbacks. For instance, the minor embeddings f_1 and f_2 in Example 6 do not have a pullback. Indeed, let P, together with $g_1 : P \to T_1$ and $g_2 : P \to T_2$, be a pullback of them in Tree_{min}. Then, since $f_1 \circ g_1 = f_2 \circ g_2 : V(P) \to V(T)$, we have that $g_1(V(P)) \subseteq \{b, c\}$ and $g_2(V(P)) \subseteq \{b, c\}$ and hence, P being a tree and g_1 and g_2 being minor embeddings, it must happen that P consists of only one node, say $\{x\}$, and no arc, and that g_1 and g_2 send x to the same node, b or c, in T_1 and in T_2. To fix ideas, assume that $g_1(x) = g_2(x) = b$. But then, if we take the minor embeddings $h_1 : P \to T_1$ and $h_2 : P \to T_2$ given by $h_1(x) = h_2(x) = c$, there is no minor embedding $h : P \to P$ such that $h_1 = g_1 \circ h$ and $h_2 = g_2 \circ h$, which contradicts the definition of pullback.*

Anyway, the first part of the proof of Proposition 11 can also be used to prove that if $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$ are minor embeddings such that T_p is a tree, then $(T_p, \iota_1 : T_p \to T_1, \iota_2 : T_p \to T_2)$ is a pullback of f_1 and f_2 in Tree_{min}.

4 Common supertrees as pushouts

Let Tree_* be henceforth any one of the categories of trees Tree_{iso}, Tree_{hom}, Tree_{top}, or Tree_{min}.

Let T_1 and T_2 be two trees. Let T_μ be a largest common Tree_*-subtree of them, and let $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$ be any Tree_*-embeddings. Let $T_1 + T_2$ be the graph obtained as the disjoint sum of the trees T_1 and T_2. Let θ be the equivalence relation on $V(T_1) \uplus V(T_2)$ defined, up to symmetry, by the following condition: $(a, b) \in \theta$ if and only if $a = b$ or there exists some $c \in V(T_\mu)$ such that $a = m_1(c)$ and $b = m_2(c)$.

Let T_{po} be the quotient graph of $T_1 + T_2$ by this equivalence:

- its set of nodes $V(T_{po})$ is the quotient set $(V(T_1) \uplus V(T_2))/\theta$, with elements the equivalence classes $[a]_\theta$ of the nodes a of T_1 or T_2;
- its arcs are those induced by the arcs in T_1 or T_2, in the sense that $([a]_\theta, [b]_\theta) \in E(T_{po})$ if and only if there exist $a' \in [a]_\theta$, $b' \in [b]_\theta$ and some $i = 1, 2$ such that $(a', b') \in E(T_i)$.

Let $\ell_i : V(T_i) \to V(T_{po})$, $i = 1, 2$, denote the inclusion $V(T_i) \hookrightarrow V(T_1) \uplus V(T_2)$ followed by the quotient mapping $V(T_1) \uplus V(T_2) \to (V(T_1) \uplus V(T_2))/\theta$. It is straightforward to check that these mappings are injective and they define *morphisms of graphs* $\ell_i : T_i \to T_{po}$, $i = 1, 2$, in the sense that if $(a, b) \in E(T_i)$, then $(\ell_i(a), \ell_i(b)) \in E(T_{po})$.

We shall call this graph T_{po}, together with these injective morphisms $\ell_i : T_i \to T_{po}$, $i = 1, 2$, the *join* of T_1 and T_2 obtained through m_1 and m_2. It is well known that it is a pushout of $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$ in the usual category of graphs.

Notice that, since every node in T_1 and T_2 has in-degree at most 1, all nodes in T_{po} have in-degree at most 2. Beside this straightforward property of joins of trees, we shall need those given by the following lemma.

Lemma 13 *Let T_1 and T_2 be trees, let T_μ be a largest common Tree_*-subtree of T_1 and T_2, let $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$ be any Tree_*-embeddings, and let T_{po} be the join of T_1 and T_2 obtained through m_1 and m_2.*

(i) T_{po} has no circuit.
(ii) For every $v, w \in V(T_{po})$, if $(v, w) \in E(T_{po})$ and there is another path from v to w in T_{po}, then $v, w \in \ell_1(V(T_1)) \cap \ell_2(V(T_2))$, this path is unique, it is a Tree_*-path and it has no intermediate node in $\ell_1(V(T_1)) \cap \ell_2(V(T_2))$.

PROOF. (i) To begin with, notice that if T_{po} contains a path from $[m_i(x)]_\theta$ to $[m_i(y)]_\theta$, with $x, y \in V(T_\mu)$ and $i = 1, 2$, and all arcs in this path are induced by arcs in T_i, then, by Lemma 4, T_μ contains a path from x to y. This implies that if T_{po} contains any path from $[m_i(x)]_\theta$ to $[m_i(y)]_\theta$, with $x, y \in V(T_\mu)$ and $i = 1, 2$, then T_μ contains a path from x to y.

Now, assume that T_{po} contains a circuit. Not all arcs in this circuit can be induced by arcs in the same tree T_1 or T_2, because these trees do not contain circuits. Therefore, two different nodes in this circuit must belong to $\ell_1(m_1(V(T_\mu)))$. This implies that there exist $x, y \in V(T_\mu)$, $x \neq y$, such that T_{po} contains a path from $[m_1(x)]_\theta$ to $[m_1(y)]_\theta$ and a path from $[m_1(y)]_\theta$ to $[m_1(x)]_\theta$. The previous observation entails that T_μ contains a path from x to y and a path from y to x, and hence a circuit, which yields a contradiction.

(ii) If $(v, w) \in E(T_{po})$, then there exists, say, $a, b \in V(T_1)$ such that $v = [a]_\theta$,

$w = [b]_\theta$ and $(a, b) \in E(T_1)$. Assume now that there is another path from $[a]_\theta$ to $[b]_\theta$ in T_{po}. Since, by the previous argument, T_{po} does not contain circuits, the last intermediate node in this path must be different from $[a]_\theta$, and hence $[b]_\theta$ has in-degree 2 in T_{po}. This implies that b is identified in T_{po} with some node of T_2 that has in-degree 1. Therefore, there exists some $y \in V(T_\mu)$ such that $b = m_1(y)$ and there exists some $c \in V(T_2)$ such that $(c, m_2(y)) \in E(T_2)$.

If y is the root of T_μ then, by Lemma 4, neither a nor c have preimage under the corresponding Tree_*-embedding $m_i : T_i \to T_\mu$. In this case, we can enlarge T_μ by adding to it a new node x_0 and a new arc (x_0, y) and we can extend m_1 and m_2 by defining $m_1(x_0) = a$ and $m_2(x_0) = c$, and in this way we obtain a common Tree_*-subtree of T_1 and T_2 larger that T_μ, which gives a contradiction.

So, there exists some $x \in V(T_\mu)$ such that $(x, y) \in E(T_\mu)$. Then, there exist a Tree_*-path in T_1 from $m_1(x)$ to b without any intermediate node in $m_1(V(T_\mu))$ and a Tree_*-path in T_2 from $m_2(x)$ to $m_2(y)$ without any intermediate node in $m_2(V(T_\mu))$. Since their intermediate nodes do not come from T_μ, these paths generate in T_{po} two Tree_*-paths from $[m_1(x)]_\theta = [m_2(x)]_\theta$ to $[b]_\theta$ with all their intermediate nodes of in-degree 1 and not belonging to $\ell_1(V(T_1)) \cap \ell_2(V(T_2))$.

If $m_1(x) \neq a$, then $[a]_\theta$ is intermediate in the path from $[m_1(x)]_\theta$ to $[b]_\theta$. Since all nodes in T_{po} have in-degree at most 2, and hence $[b]_\theta$ has in-degree exactly 2, and all intermediate nodes in both paths from $[m_1(x)]_\theta = [m_2(x)]_\theta$ to $[b]_\theta$ have in-degree exactly 1, any path from $[a]_\theta$ to $[b]_\theta$ different from the arc $([a]_\theta, [b]_\theta)$ must cross $[m_1(x)]_\theta$. But such a path would mean a circuit in T_{po}, which by (i) is impossible.

Therefore, it must happen that $m_1(x) = a$. In this case we have obtained a Tree_*-path from $[a]_\theta$ to $[b]_\theta$ without any intermediate node belonging to $\ell_1(V(T_1)) \cap \ell_2(V(T_2))$. The facts that $[b]_\theta$ has in-degree 2, that all intermediate nodes in the new path from $[a]_\theta$ to $[b]_\theta$ we found have in-degree exactly 1, and that T_{po} contains no circuit, entail that this is the only other path from $[a]_\theta$ to $[b]_\theta$. □

Let now T_σ be the graph obtained from T_{po} by removing every arc that is subsumed by a path: we remove from T_{po} an arc (v, w) if there is a path in T_{po} with at least one intermediate node going from v to w. We shall call this graph the Tree_*-*sum* of T_1 and T_2 obtained through m_1 and m_2.

As a direct consequence of Lemma 13, we have that if $\mathsf{Tree}_* = \mathsf{Tree}_{iso}$, then $T_\sigma = T_{po}$ and that if $\mathsf{Tree}_* = \mathsf{Tree}_{hom}$, then T_σ is obtained from T_{po} by removing every arc that is subsumed by an elementary path.

Proposition 14 *For every two trees T_1 and T_2, any Tree_*-sum of T_1 and T_2 is a common Tree_*-supertree of them.*

PROOF. Let T_1 and T_2 be two trees, let T_μ be a largest common Tree_*-subtree of them and let $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$ be any Tree_*-embeddings. Let T_σ be the Tree_*-sum of T_1 and T_2 obtained through m_1 and m_2, and let $\ell_i : V(T_i) \to V(T_\sigma) = (V(T_1) \uplus V(T_2))/\theta$, $i = 1, 2$, stand for the corresponding restrictions of the quotient mappings.

Since T_σ is obtained by removing arcs from the join T_{po} of T_1 and T_2 that are subsumed by paths, when an arc is removed no existing path in T_{po} is broken, and therefore there is a path from x to y in T_{po} if and only if there is a path from x to y in T_σ. In particular, the only nodes in T_{po} than can possibly have no parent are the images of the roots of T_1 and T_2 and hence the same also happens in T_σ.

Now, if r_μ is the root of T_μ then $m_1(r_\mu)$ is the root r_1 of T_1 or $m_2(r_\mu)$ is the root r_2 of T_2: if neither of them is the corresponding root, then each one of these images has a parent in the corresponding tree, and in this case adding a parent to r_μ in T_μ and extending m_1 and m_2 to this new node by sending it to the parent of $m_1(r_\mu)$ and $m_2(r_\mu)$, respectively, we would always get a common Tree_*-subtree of T_1 and T_2 with strictly larger order. Therefore, the root of one of the trees is always identified in T_σ with a node of the other tree.

Now, we have two possibilities. First, if $m_1(r_\mu) = r_1$ and $m_2(r_\mu) = r_2$, then $[r_1]_\theta = [r_2]_\theta$ is the only node in T_σ with no parent, and every node v in T_{po} (as well as in T_σ, as we said) can be reached from this node through a path: if $v = [a_1]_\theta$ with $a_1 \in V(T_1)$, through the path induced by the path in T_1 going from r_1 to a_1, and if $v = [a_2]_\theta$ with $a_2 \in V(T_2)$, through the image of the path in T_2 going from r_2 to a_2.

Second, if, say, $m_1(r_\mu) = r_1$ but $m_2(r_\mu) \neq r_2$, then $[r_2]_\theta$ is the only node in T_σ with no parent and every node in T_σ can be reached from this node through a path: every node of the form $[a_2]_\theta$ with $a_2 \in V(T_2)$ through the path induced by the path in T_2 going from r_2 to a_2, and every node of the form $[a_1]_\theta$ with $a_1 \in V(T_1)$ through the path obtained by concatenating the image of the path in T_2 from r_2 to $m_2(r_\mu)$ and the image of the path in T_1 from r_1 to a_1.

In all, T_σ has one, and only one, node without parent, and every other node in T_σ can be reached from it through a path. It remains to prove that these paths are unique. To do it, it is enough to check that no node in T_σ has two different parents. So, assume that there exist arcs in T_σ from $[b]_\theta$ and $[c]_\theta$ to $[a]_\theta$ and that $[b]_\theta \neq [c]_\theta$. Then, to begin with, one of these arcs must be induced by one arc in T_1, and the other by one arc in T_2: if they are both induced by arcs in, say, T_1, then, since θ does not identify any pair of different elements in $V(T_1)$, there would be two arcs with the same target node in T_1, which is impossible. Therefore, we may assume that there exist $y \in V(T_\mu)$, $b \in V(T_1)$ and $c \in V(T_2)$ and an arc from b to $m_1(x)$ in T_1 and an arc from c to $m_2(y)$

in T_2.

There are now several possibilities.

- y has no parent in T_μ and hence it is its root. In this case, as we saw above, $m_1(y)$ or $m_2(y)$ should be the root of the corresponding tree, which contradicts the assumption that they have parents in the corresponding trees.
- y has a parent x in T_μ. Then, there is a Tree_*-path from $m_1(x)$ to $m_1(y)$ in T_1 and a Tree_*-path from $m_2(x)$ to $m_2(y)$ in T_2. This yields, up to symmetry, three possibilities:
  - If $m_1(x) = b$ and $m_2(x) = c$, then $[b]_\theta = [c]_\theta$. This is the only possible case if $\mathsf{Tree}_* = \mathsf{Tree}_{iso}$, and hence in the other two cases we understand that $\mathsf{Tree}_* \neq \mathsf{Tree}_{iso}$.
  - If $m_1(x) = b$ and $m_2(x) \neq c$, then the arc from $[b]_\theta$ to $[m_1(y)]_\theta$ is subsumed by the path from $[m_2(x)]_\theta$ to $[m_1(y)]_\theta$ coming from T_2, and it should have been removed from T_σ.
  - If $m_1(x) \neq b$ and $m_2(x) \neq c$, then there are Tree_*-paths from $m_1(x)$ to $m_1(y)$ in T_1 and from $m_2(x)$ to $m_2(y)$ in T_2 without intermediate nodes coming from $V(T_\mu)$, and b and c are the last intermediate nodes in the corresponding paths.

    In this case we can enlarge T_μ by adding a new node x_0 and replacing the arc from x to y by an arc from x to x_0 and an arc from x_0 to y, and we can extend m_1 and m_2 to this new node by sending it, respectively, to b and c. It is clear that in this way we obtain a tree with order larger than that of T_μ, and the extensions of m_1 and m_2 are Tree_*-embeddings: the new arc (x, x_0) is transformed under them into the Tree_*-paths —without intermediate nodes coming from $V(T_\mu)$— that go from $m_1(x)$ to b and from $m_2(x)$ to c, respectively; the new arc (x_0, y) is transformed under them into the arcs $(b, m_1(y))$ and $(c, m_2(y))$, respectively; and it is clear that if m_1 and m_2 were topological embeddings, then their extensions are still so. Thus, in this way we obtain a new common Tree_*-subtree of T_1 of T_2 with order larger than that of T_μ, which yields a contradiction.

In all, his proves that T_σ is a tree. Now we have to prove that $\ell_1 : T_1 \to T_\sigma$ and $\ell_2 : T_2 \to T_\sigma$ are Tree_*-embeddings. We shall prove that ℓ_1 is a Tree_*-embedding. Recall that the mapping $\ell_1 : V(T_1) \to V(T_{po}) = V(T_\sigma)$ is injective, and notice that, by Lemma 13, if $(a, b) \in E(T_1)$, then there is a Tree_*-path in T_σ from $\ell_1(a) = [a]_\theta$ to $\ell_1(b) = [b]_\theta$ that does not contain any intermediate node in $\ell_1(V(T_1) \cap \ell_2(V(T_2))$: either the arc induced by the arc in T_1 or the Tree_*-path that made this arc to be removed. Notice actually that in the proof of Lemma 13 we proved more: if the arc $([a]_\theta, [b]_\theta)$ induced by an arc $(a, b) \in E(T_1)$ is removed because of a second path, then all intermediate nodes in this path are equivalence classes of nodes in $V(T_2) - m_2(V(T_\mu))$: they do not belong to $\ell_1(V(T_1))$.

In all this proves that $(a,b) \in E(T_1)$, then there is a Tree_*-path in T_σ from $\ell_1(a)$ to $\ell_1(b)$ that does not contain any intermediate node in $\ell_1(V(T_1))$. If Tree_* stands for Tree_{iso}, Tree_{hom} or Tree_{min}, this shows that ℓ_1 is a Tree_*-embedding.

In the case of Tree_{top} we must prove that if $(a,b),(a,c) \in E(T_1)$, then the paths from $[a]_\theta$ to $[b]_\theta$ and to $[c]_\theta$ are divergent. By the injectivity of ℓ_1, the only case that needs to be discussed is when there are paths with intermediate nodes from $[a]_\theta$ to $[b]_\theta$ and to $[c]_\theta$.

From the proof of Lemma 13 we know that, in this case, there are $x,y,z \in V(T_\mu)$ such that $m_1(x) = a$, $m_1(y) = b$, $m_1(z) = c$, and $(x,y),(x,z) \in E(T_\mu)$, and the intermediate nodes in the paths from $[a]_\theta$ to $[b]_\theta$ and to $[c]_\theta$ are the equivalence classes of the intermediate nodes in the paths from $m_2(x)$ to $m_2(y)$ and to $m_2(z)$ in T_2, which do not belong to $m_2(V(T_\mu))$. But, since m_2 is a topological embedding, and hence these paths have no intermediate node in common, and θ does not identify any pair of nodes in $V(T_2)$, we deduce that no equivalence class of an intermediate node in the path from $m_2(x)$ to $m_2(y)$ is equal to an equivalence class of an intermediate node in the path from $m_2(x)$ to $m_2(z)$, and hence that the paths from $[a]_\theta$ to $[b]_\theta$ and to $[c]_\theta$ diverge, as we wanted to prove. $\Box$

The following result extends the last one by showing that Tree_*-sums not only yield common Tree_*-subtrees, but pushouts.

Theorem 15 *Let T_1 and T_2 be trees, let T_μ be a largest common Tree_*-subtree of T_1 and T_2, and let $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$ be any Tree_*-embeddings.*

Then, the Tree_*-sum T_σ of T_1 and T_2 obtained through m_1 and m_2, together with the Tree_*-embeddings $\ell_1 : T_1 \to T_\sigma$ and $\ell_2 : T_2 \to T_\sigma$, is a pushout in Tree_* of m_1 and m_2.

PROOF. It is clear that $\ell_1 \circ m_1 = \ell_2 \circ m_2$. Therefore, it remains to prove that T_σ, together with the Tree_*-embeddings $\ell_1 : T_1 \to T_\sigma$ and $\ell_2 : T_2 \to T_\sigma$, satisfies the universal condition of pushouts in Tree_*.

So, let $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$ be any Tree_*-embeddings such that $f_1 \circ m_1 = f_2 \circ m_2$. It is well-known that there exists one, and only one, injective mapping $f : (V(T_1) \uplus V(T_2))/\theta \to T$ such that $f \circ \ell_1 = f_1$ and $f \circ \ell_2 = f_2$: namely, the one defined by $f([a]_\theta) = f_1(a)$ if $a \in V(T_1)$ and $f([a]_\theta) = f_2(a)$ if $a \in V(T_2)$. We must prove that this mapping f is a Tree_*-embedding.

Let us prove first that it is injective. Assume that there exist $v,w \in V(T)$, $v \neq w$, such that $f(v) = f(w)$. Since f_1 and f_2 are injective, it is clear that they cannot be classes of nodes of the same tree T_i. Thus, there exist

$a \in V(T_1) - m_1(V(T_\mu))$ and $b \in V(T_2) - m_2(V(T_\mu))$ such that $v = [a]_\theta$ and $w = [b]_\theta$ and $f_1(a) = f_2(b)$.

As we saw in the proof of the previous proposition, the image under some m_i of the root of T_μ is the root of the corresponding T_i. This implies that there exists a path from the image of a node in T_μ to one of these nodes a or b in the corresponding tree. Assume that there is, say, some $x \in V(T_\mu)$ such that there exists a path in T_1 from $m_1(x)$ to a. Then there exists a path from $f_1(m_1(x)) = f_2(m_2(x))$ to $f_1(a) = f_2(b)$ in T and hence a path from $m_2(x)$ to b in T_2. By symmetry, if there exists some $x \in V(T_\mu)$ such that there exists a path from $m_2(x)$ to b in T_2, then there exists a path from $m_1(x)$ to a in T_1.

This shows that there exists a node $x \in V(T_\mu)$ such that there exist paths in T_1 from $m_1(x)$ to a and in T_2 from $m_2(x)$ to b, and moreover such that no child of it satisfies this property. These paths induce, through f_1 and f_2, two paths from $f_1(m_1(x)) = f_2(m_2(x))$ to $f_1(a) = f_2(b)$ that must be the same. Let now e be the first node after $m_1(x)$ in the path from this node to a in T_1, and d be the first node after $m_2(x)$ in the path from this node to b in T_2. Then $f_1(e)$ and $f_2(d)$ are intermediate in the path from $f_1(m_1(x)) = f_2(m_2(x))$ to $f_1(a) = f_2(b)$. Without any loss of generality, assume that $f_2(d)$ appears in the piece of this path before $f_1(e)$, or that it is equal to this node.

In this case we can enlarge $\widehat{T_\mu}$, by adding a new node y_0, a new arc from x to y_0, and splitting through this new arc all those arcs $(x, z) \in E(T_\mu)$ such that the path from $m_1(x)$ to $m_1(z)$ in T_1 crosses e. It is clear that the graph $\widehat{T}_\mu$ obtained in this way is a tree. And we can extend m_1 and m_2 to $\widehat{T}_\mu$ by defining $m_1(y_0) = e$ and $m_2(y_0) = d$.

Let us check that $m_1 : \widehat{T}_\mu \to T_1$ and $m_2 : \widehat{T}_\mu \to T_2$ are Tree_*-embeddings. On the one hand, the arc (x, y_0) is transformed under them into the arcs $(m_1(x), e)$ and $(m_1(x), d)$, respectively. Assume now that $\widehat{T}_\mu$ contains a new arc (y_0, z). This means that T_μ contained (x, z) and that e is the first intermediate node in the Tree_*-path without intermediate nodes in $m_1(V(T_\mu))$ from $m_1(x)$ to $m_1(z)$. This implies that there exists in T_1 a Tree_*-path without intermediate nodes in $m_1(V(\widehat{T}_\mu))$ from $e = m_1(y_0)$ to $m_1(z)$. As far as m_2 goes, notice that the arc (x, z) in T_μ induces under $f_1 \circ m_1$ a Tree_*-path from $f_1(m_1(x)) = f_2(m_2((x))$ to $f_1(m_1(z)) = f_2(m_2((z))$ that crosses $f_1(e)$, and therefore it also crosses $f_2(d)$. This entails that d is an intermediate node, actually the first one, in the Tree_*-path in T_2 without nodes in $m_2(T_\mu)$ that goes from $m_2(x)$ to $m_2(z)$, and therefore that there exists a Tree_*-path in T_2 without nodes in $m_2(\widehat{T}_\mu)$ that goes from $d = m_2(y_0)$ to $m_2(z)$.

This shows that $m_1 : \widehat{T}_\mu \to T_1$ and $m_2 : \widehat{T}_\mu \to T_2$ transform arcs into Tree_*-paths without any node coming from $\widehat{T}_\mu$, and hence that they are Tree_*-morphisms when $\mathsf{Tree}_* = \mathsf{Tree}_{hom}$ or $\mathsf{Tree}_* = \mathsf{Tree}_{min}$. To cover the case

$\mathsf{Tree}_* = \mathsf{Tree}_{top}$, we must also prove that they transform pairs of arcs with the same source node into divergent paths, but this is a direct consequence of the fact that $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$ did so.

In all, this proves that in this case there exists a common Tree_*-subtree $\widehat{T}_\mu$ of T_1 and T_2 with one more node that T_μ, which contradicts the assumption that T_μ is a largest common Tree_*-subtree of T_1 and T_2. Therefore, there cannot exist two different nodes $v, w \in V(T)$ such that $f(v) = f(w)$.

So, we know that $f : V(T_\sigma) \to V(T)$ is injective. Now, assume there is an arc from v to w in T_σ. Then, say, there exist $a, b \in V(T_1)$ such that $v = [a]_\theta$ and $w = [b]_\theta$ and there is an arc from a to b in T_1 that is not subsumed by any other arc in T_{po}. This implies that there is a Tree_*-path from $f(v) = f_1(a)$ to $f(w) = f_1(b)$ in T.

This already covers the case $\mathsf{Tree}_* = \mathsf{Tree}_{iso}$, and thus we shall assume henceforth that $\mathsf{Tree}_* \neq \mathsf{Tree}_{iso}$.

In this case, we must check that no intermediate node in this path from $f(v)$ to $f(w)$ belongs to $f(V(T_\sigma)) = f_1(V(T_1)) \cup f_2(V(T_2))$. Now, f_1 being a Tree_*-embedding, we already know that no intermediate node in this path belongs to $f_1(V(T_1))$, and therefore we only have to check that no such intermediate node belongs to $f_2(V(T_2))$. But, before proceeding, notice that we have already proved that f sends arcs to paths, and hence that this mapping transforms paths in T_σ into paths in T.

So, assume that there is some $c \in V(T_2)$ such that $f_2(c)$ is an intermediate node in the path from $f_1(a)$ to $f_1(b)$ in T. This prevents the existence of paths in T_σ from $[c]_\theta$ to $[a]_\theta$ and from $[b]_\theta$ to $[c]_\theta$: the image of such a path under f would be a path in T that would build up a circuit with the path from $f_1(a) = f([a]_\theta)$ to $f_2(c) = f([c]_\theta)$ or from $f_2(c) = f([c]_\theta)$ to $f_1(b) = f([b]_\theta)$, respectively, that we already know to exist. Moreover, $c \notin m_2(V(T_\mu))$, because if $c = m_2(x)$, then $f_2(c) = f_1(m_1(x)) \in f_1(V(T_1))$.

After excluding these possibilities, we still must discuss several cases:

- $a = m_1(x)$ and $b = m_1(y)$ for some $x, y \in V(T_\mu)$. In this case, Lemma 3 implies that there exists an arc from x to y in T_μ, which on its turn entails a path from $m_2(x)$ to $m_2(y)$ in T_2. But this path cannot have any intermediate node, because there is an arc in T_σ from $[m_2(x)]_\theta = [a]_\theta$ to $[m_2(y)]_\theta = [b]_\theta$. So, we conclude that there is an arc in T_2 from $m_2(x)$ to $m_2(y)$ and hence the path in T from $f_2(m_2(x)) = f_1(a)$ to $f_2(m_2(y)) = f_1(b)$ does not contain any intermediate node in $f(\ell_2(V(T_2))) = f_2(V(T_2))$, which contradicts the existence of c.
- $a = m_1(x)$ for some $x \in V(T_\mu)$ but $b \notin m_2(V(T_\mu))$. In this case, the existence of a Tree_*-path in T from $f_2(m_2(x)) = f_1(a)$ to $f_2(c)$ without any

intermediate node in $f_1(V(T_1))$ entails the existence of a Tree_*-path from $m_2(x)$ to c in T_2 without any intermediate node in $m_2(V(T_\mu))$.

But in this case we can enlarge T_μ by adding a new node y_0, a new arc from x to y_0, and splitting through this arc all those arcs $(x,z) \in E(T_\mu)$ such that the path from $m_1(x)$ to $m_1(z)$ crosses b. It is clear that the graph $\widehat{T}_\mu$ obtained in this way is a tree. And we can extend m_1 and m_2 to $\widehat{T}_\mu$ by defining $m_1(y_0) = b$ and $m_2(y_0)$ to be the first node d in the path $m_2(x)$ to c in T_2 after $m_2(x)$.

Let us check that $m_1 : \widehat{T}_\mu \to T_1$ and $m_2 : \widehat{T}_\mu \to T_2$ are Tree_*-embeddings. On the one hand, the arc (x, y_0) is transformed under them into the arcs (a, b) and (a, d), respectively. Assume now that $\widehat{T}_\mu$ contains a new arc (y_0, z). This means that T_μ contained (x, z) and that b is the first intermediate node in the Tree_*-path without intermediate nodes in $m_1(V(T_\mu))$ from $a = m_1(x)$ to $m_1(z)$. This implies that there exists in T_1 a Tree_*-path without intermediate nodes in $m_1(V(\widehat{T}_\mu))$ from $b = m_1(y_0)$ to $m_1(z)$. As far as m_2 goes, notice that the arc (x, z) in T_μ induces under $f_1 \circ m_1$ a Tree_*-path from $f_1(a) = f_1(m_1(x)) = f_2(m_2((x))$ to $f_1(m_1(z)) = f_2(m_2((z))$ that crosses $f_1(b)$, and therefore it also crosses $f_2(c)$ and hence $f_2(d)$, too. This entails that d is an intermediate node, actually the first one, in the Tree_*-path in T_2 without nodes in $m_2(T_\mu)$ that goes from $m_2(x)$ to $m_2(z)$, and therefore that there exists a Tree_*-path in T_2 without nodes in $m_2(\widehat{T}_\mu)$ that goes from $d = m_2(y_0)$ to $m_2(z)$.

This shows that $m_1 : \widehat{T}_\mu \to T_1$ and $m_2 : \widehat{T}_\mu \to T_2$ transform arcs into Tree_*-paths without any node coming from T_μ, and hence that they are Tree_*-morphisms when $\mathsf{Tree}_* = \mathsf{Tree}_{hom}$ or $\mathsf{Tree}_* = \mathsf{Tree}_{min}$. To cover the case $\mathsf{Tree}_* = \mathsf{Tree}_{top}$, we should also prove that they transform pairs of arcs with the same source node into divergent paths, but this is a direct consequence of the fact that $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$ did so.

In all, in this way we obtain a common Tree_*-subtree $\widehat{T}_\mu$ of T_1 and T_2 with one more node that T_μ, and hence the latter would not be the largest.

- $a \notin m_1(V(T_\mu))$. Since we know that $c \notin m_2(V(T_\mu))$, this case splits into the final two possibilities:
  - There exist some $x \in V(T_\mu)$ and paths in T_1 and T_2 from $m_1(x)$ to a and from $m_2(x)$ to c, respectively. Without any loss of generality we can take such an x satisfying moreover that no child of it satisfies this property.

    The fact that $f_1(a)$ is an intermediate node in the path from $f_1(m_1(x))$ to $f_2(c)$ implies that x has no child y such that there is a path from $m_1(y)$ to a *or* from $m_2(y)$ to c. Indeed, if there is a child y of x such that there is a path from $m_1(y)$ to a, then there is a path in T from $f_1(m_1(y))$ to $f_1(a)$ and then a path from $f_2(m_2(y)) = f_1(m_1(y))$ to $f_2(c)$ which, on its turn, entails a path from $m_2(y)$ to c in T_2. In a similar way, if there is a child y of x such that there is a path from $m_2(y)$ to c, then there is a path in T from $f_2(m_2(y))$ to $f_2(c)$. Since $f_2(m_2(y)) = f_1(m_1(y))$ cannot be an intermediate node in the path from $f_1(a)$ to $f_2(c)$, this entails that $f_1(a)$ is intermediate in the path from $f_2(m_2(y))$ to $f_2(c)$: that there is a path

from $f_1(m_1(y)) = f_2(m_2(y))$ to $f_1(a)$ which, finally, entails a path from $m_1(y)$ to a in T_1.

So, x has no child y such that there is a path from $m_1(y)$ to a or from $m_2(y)$ to c. Let e be the first node (after $m_1(x)$) in the path from $m_1(x)$ to a in T_1, and let d be the first node (after $m_2(x)$) in the path from $m_2(x)$ to c in T_2. The uniqueness of paths in T entails that the path from $f_2(m_2(x))$ to $f_2(c)$ (the image under f_2 of the path from $m_2(x)$ to c in T_2) is the concatenation of the path from $f_1(m_1(x))$ to $f_1(a)$ (the image under f_1 of the path from $m_1(x)$ to a) and the path from $f_1(a)$ to $f_2(c)$. In particular, both $f_1(e)$ and $f_2(d)$ are intermediate nodes in this path. But then, arguing as in the proof of the injectivity of $f : T_\sigma \to T$ we can build up a common Tree_*-subtree $\widehat{T}_\mu$ of T_1 and T_2 with one more node that T_μ, which contradicts the assumption that T_μ is a largest common Tree_*-subtree of T_1 and T_2.

- There is no $x \in V(T_\mu)$ such that there are paths in T_1 and T_2 from $m_1(x)$ to a and from $m_2(x)$ to c, respectively. Arguing as before we see that there is no $x \in V(T_\mu)$ such that there is a path in T_1 from $m_1(x)$ to a or a path in T_2 from $m_2(x)$ to c. But this is impossible, because the image under m_1 or under m_2 of the root of T_μ is the root of T_1 or T_2, respectively.

This proves the universal condition, and with it the statement, for the categories Tree_{iso}, Tree_{hom}, and Tree_{min}.

So assume we are in Tree_{top}. Since every topological embedding is a minor embedding, we already know that if there is an arc from v to w in T_σ, then there is a path from $f(v)$ to $f(w)$. Now assume there are arcs (v, w) and (v, u) in T_σ.

Consider first the case when these arcs are induced by arcs in the same tree, and to fix ideas assume that there exist $(a, b), (a, c) \in V(T_1)$ such that $v = [a]_\theta$, $w = [b]_\theta$ and $u = [c]_\theta$. Then, since f_1 is a topological embedding, the paths from $f(v) = f_1(a)$ to $f(w) = f_1(b)$ and to $f(u) = f_1(c)$ are divergent.

Now consider the case when each one of these arcs is induced by an arc in a different tree: there exist $x \in V(T_\mu)$, $b \in V(T_1)$ and $c \in V(T_2)$ such that $v = [m_1(x)]_\theta = [m_2(x)]_\theta$, $w = [b]_\theta$ and $u = [c]_\theta$, and there are arcs $(m_1(x), b) \in E(T_1)$ and $(m_2(x), c) \in E(T_2)$ that induce the arcs $(v, w), (v, u) \in E(T_\sigma)$.

Consider first the case when b or c has a preimage in T_μ. To fix ideas, assume that there exists $y \in V(T_\mu)$ such that $m_1(y) = b$. Then, there exists an arc from x to y in T_μ and hence a path from $m_1(x)$ to $m_2(y)$ in T_2. But since there is an arc from $[m_2(x)]_\theta = [m_1(x)]_\theta$ to $[m_2(y)]_\theta = [b]_\theta$, the path from $m_1(x)$ to $m_2(y)$ in T_2 must also be an arc, and hence both arcs (v, w) and (v, u) are induced by arcs in T_2.

Consider now the case when neither b nor c have a preimage in T_μ. There are

two possibilities to discuss.

- If there exists an arc $(x, z) \in V(T_\mu)$ such that b is the first intermediate node in the path from $m_1(x)$ to $m_1(z)$, then $[b]_\theta$ is also the first intermediate node in the path from $[m_1(x)]_\theta$ to $[m_1(z)]_\theta$. In particular, $[c]_\theta$ does not appear in this last path, which entails that the arc $(m_2(x), c)$ and the path from $m_2(x)$ to $m_2(z)$ are divergent. Since f_2 is a topological embedding, this implies that the paths in T from $f_2(m_2(x))$ to $f_2(c)$ and from $f_2(m_2(x)) = f_1(m_1(x))$ to $f_2(m_2(z)) = f_1(m_1(z))$ are also divergent. Since $f_1(b)$ is an intermediate node in this path, we finally deduce that the paths from $f(v) = f_2(m_2(x)) = f_1(m_1(x))$ to $f(u) = f_1(b)$ and to $f(w) = f_2(c)$ are divergent.

  The case when there exists an arc $(x, z) \in V(T_\mu)$ such that c is the first intermediate node in the path from $m_2(x)$ to $m_2(z)$ is solved in a similar way.
- If there is no arc (x, z) in T_μ such that b or c are intermediate nodes in the paths from $m_1(x)$ to $m_1(z)$ or from $m_2(x)$ to $m_2(z)$, respectively, then we can enlarge this tree by adding to it a new node y_0 and an arc (x, y_0), and we can extend m_1 and m_2 to this new tree by defining $m_1(y_0) = b$ and $m_2(y_0) = c$, and it is straightforward to prove that in this way we obtain a new topological subtree of T_1 and T_2, which contradicts the assumption that T_μ is a largest common topological subtree of T_1 and T_2.

This finishes the proof for Tree_{top}. □

5 Largest common subtrees and smallest common supertrees

Let Tree_* still denote any category Tree_{iso}, Tree_{hom}, Tree_{top}, or Tree_{min}. In this section we show that the constructions introduced in the last two sections can be used to obtain largest common Tree_*-subtrees and smallest common Tree_*-supertrees of pairs of trees. The key will be the following result.

Lemma 16 *Let T_1 and T_2 be two trees, and let T_μ be a largest common Tree_*-subtree of them. For every common Tree_*-supertree T of T_1 and T_2, we have that $|V(T)| \geqslant |V(T_1)| + |V(T_2)| - |V(T_\mu)|$.*

PROOF. Consider the Tree_*-embeddings $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$. Propositions 7, 8, 9, and 10 show that for every pair of Tree_*-embeddings $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$ there exists a common Tree_*-subtree T_0 of T_1 and T_2 with set of nodes containing $f_1(V(T_1)) \cap f_2(V(T_2))$: after a relabeling of the nodes (so that f_1 and f_2 are given by set-theoretical inclusions of the sets of nodes), it will be $\widetilde{T}_p$ in Tree_{min} and T_p in Tree_{iso}, Tree_{hom}, and Tree_{top}.

Then

$$|f_1(V(T_1)) \cap f_2(V(T_2))| \leqslant |V(T_0)| \leqslant |V(T_\mu)|$$

and hence

$$\begin{aligned} |V(T)| &\geqslant |f_1(V(T_1)) \cup f_2(V(T_2))| \\ &= |f_1(V(T_1))| + |f_2(V(T_2))| - |f_1(V(T_1)) \cap f_2(V(T_2))| \\ &\geqslant |V(T_1)| + |V(T_2)| - |V(T_0)| \\ &\geqslant |V(T_1)| + |V(T_2)| - |V(T_\mu)|, \end{aligned}$$

as we claimed. □

We start with smallest common Tree_*-supertrees.

Theorem 17 *For every pair of trees T_1 and T_2, any Tree_*-sum of T_1 and T_2 is a smallest common Tree_*-supertree of them.*

PROOF. By Proposition 14, any Tree_*-sum T_σ of T_1 and T_2 is a common Tree_*-supertree of them, and by construction

$$|V(T_\sigma)| = |V(T_1)| + |V(T_2)| - |V(T_\mu)|,$$

for some largest common Tree_*-subtree T_μ of them. Thus, T_σ achieves the lower bound established in Lemma 16 for common Tree_*-supertrees of T_1 and T_2, which entails that it is a smallest common Tree_*-supertree of them. □

Thus, for every pair of trees T_1 and T_2, the pushout in Tree_* of any Tree_*-embeddings from a largest common Tree_*-subtree of them yields a smallest common Tree_*-supertree of them.

Theorem 18 *For every two trees T_1 and T_2, any intersection of T_1 and T_2 obtained through Tree_*-embeddings into a smallest common Tree_*-supertree of them is a largest common Tree_*-subtree of T_1 and T_2.*

PROOF. Let T_1 and T_2 be two trees, let T'_σ be a smallest common Tree_*-supertree of T_1 and T_2, let $p_1 : T_1 \to T'_\sigma$ and $p_2 : T_2 \to T'_\sigma$ be any Tree_*-embeddings, and let T'_μ be any common Tree_*-subtree of T_1 and T_2 obtained by expanding the intersection T_p of T_1 and T_2 obtained through p_1 and p_2, which exists by Propositions 7, 8, 9, and 10.

Now, by Theorem 17 we have that, for any largest common Tree_*-subtree T_μ of T_1 and T_2,

$$|V(T'_\sigma)| = |V(T_1)| + |V(T_2)| - |V(T_\mu)|$$

and we know that

$$|p_1(V(T_1)) \cap p_2(V(T_2))| \leqslant |V(T'_\mu)| \leqslant |V(T_\mu)|.$$

Then

$$\begin{aligned}
&|V(T_1)| + |V(T_2)| - |V(T_\mu)| \\
&= |V(T'_\sigma)| \geqslant |p_1(V(T_1)) \cup p_2(V(T_2))| \\
&= |p_1(V(T_1))| + |p_2(V(T_2))| - |p_1(V(T_1)) \cap p_2(V(T_2))| \\
&\geqslant |V(T_1)| + |V(T_2)| - |V(T'_\mu)| \\
&\geqslant |V(T_1)| + |V(T_2)| - |V(T_\mu)|.
\end{aligned}$$

This implies that $|V(T'_\mu)| = |V(T_\mu)| = |p_1(V(T_1)) \cap p_2(V(T_2))|$: that T'_μ is also a largest common Tree_*-subtree of T_1 and T_2 and that $V(T'_\mu) = p_1(V(T_1)) \cap p_2(V(T_2))$: that $T'_\mu = T_p$, as we claimed. $\square$

Therefore, for every two trees T_1 and T_2, the pullback in Tree_* of any Tree_*-embeddings into a smallest common Tree_*-supertree of them yields a largest common Tree_*-subtree of them.

Corollary 19 *The problems of finding a largest common* Tree_**-subtree and a smallest common* Tree_**-supertree of two trees, in each case together with a pair of witness* Tree_**-embeddings, are reducible to each other in time linear in the size of the trees.*

PROOF. Given two trees T_1 and T_2, if we know a largest common Tree_*-subtree T_μ of them, together with a pair of witness Tree_*-embeddings $m_1 : T_\mu \to T_1$ and $m_2 : T_\mu \to T_2$, then the construction of the pushout

$$(T_\sigma, \ell_1 : T_1 \to T_\sigma, \ell_2 : T_2 \to T_\sigma)$$

of m_1 and m_2 described in Theorem 15 gives a smallest common Tree_*-supertree of T_1 and T_2, and this construction can be obtained in time linear in the size of T_1 and T_2, as follows.

First, make copies T'_1 and T'_2 of T_1 and T_2, with $\ell_1 : T_1 \to T'_1$ and $\ell_2 : T_2 \to T'_2$ identity mappings. Second, join T'_1 and T'_2 into a graph T_σ. Third, for each $a \in V(T_\mu)$, merge nodes $\ell_1(m_1(a))$ and $\ell_2(m_2(a))$, and remove all parallel arcs.

Next, remove from T_σ all arcs subsumed by paths, as follows. For each node $y \in V(T_\sigma)$ of in-degree 2, let $x, x' \in V(T_\sigma)$ be the source nodes of the two

arcs coming into y. Now, perform a simultaneous traversal of the paths of arcs coming into x and x', until reaching node x' along the first path or x along the second path. The simultaneous traversal of incoming paths may stop along either path, but continue along the other one, because a node of in-degree 0 or in-degree 2 is reached. Finally, remove from T_σ either arc (x', y), if node x' was reached along the first path, or arc (x, y), if node x was reached along the second path.

Conversely, if we know a smallest common Tree_*-supertree T of T_1 and T_2, together with a pair of witness Tree_*-embeddings $f_1 : T_1 \to T$ and $f_2 : T_2 \to T$, then, by Theorem 18, the pullback

$$(T_p, \iota_1 : T_p \to T_1, \iota_2 : T_p \to T_2)$$

of f_1 and f_2 described in Section 3 yields a largest common Tree_*-subtree of T_1 and T_2, and this construction can also be obtained in time linear in the size of T_1 and T_2, as follows.

First, make a copy T_p of T, with $g : T \to T_p$ the identity mapping. Second, for each $a \in V(T_1)$, mark $g(f_1(a))$ in T_p. Third, for each $a \in V(T_2)$, if $g(f_2(a))$ is already marked in T_p, double-mark it. Next, for each node of T_p which is not double-marked, add a new arc from its parent (if any) to each of its children (if any) in T_p, and remove the node not double-marked. Finally, set mappings $\iota_i : T_p \to T_i$ for $i = 1, 2$, as follows. For each $a \in V(T_i)$, if $g(f_i(a))$ is defined, set $\iota_i(g(f_i(a))) = a$.

An additional step is needed in the case of Tree_{min}-embeddings. If there is more than one node of in-degree 0 in T_p, say $x_1, \ldots, x_k$, add a new node r to T_p, add a new arc (r, x_i) for each $i = 1, \ldots, k$, and set $\iota_i(r)$ to the root of T_i for $i = 1, 2$. □

6 Conclusion

Subtree isomorphism and the related problems of largest common subtree and smallest common supertree count among the most widely used techniques for comparing tree-structured data, with practical applications in combinatorial pattern matching, pattern recognition, chemical structure search, computational molecular biology, and other areas of engineering and life sciences. Four different embedding relations are of interest in these application areas: isomorphic, homeomorphic, topological, and minor embeddings.

The complexity of the largest common subtree problem and the smallest common supertree problem under these embedding relations is already settled:

they are polynomial-time solvable for isomorphic, homeomorphic, and topological embeddings, and they are NP-complete for minor embeddings. Moreover, efficient algorithms are known for largest common subtree under isomorphic, homeomorphic, and topological embeddings, and for smallest common supertree under isomorphic and topological embeddings, and an exponential algorithm is known for largest common subtree under minor embeddings.

In this paper, we have established the relationship between the largest common subtree and the smallest common supertree of two trees by means of simple constructions, which allow one to obtain the largest common subtree from the smallest common supertree, and vice versa. We have given these constructions for isomorphic, homeomorphic, topological, and minor embeddings, and have shown their implementation in time linear in the size of the trees.

In doing so, we have filled the gap by providing a simple extension of previous largest common subtree algorithms for solving the smallest common supertree problem, in particular under homeomorphic and minor embeddings for which no previous algorithm is known.

Beside the practical interest of these extensions to previous algorithms, we have given a unified algebraic construction showing the relation between largest common subtrees and smallest common supertrees for the four different embedding problems studied in the literature: isomorphic, homeomorphic, topological, and minor embeddings. The unified construction shows that smallest common supertrees are pushouts and largest common subtrees are pullbacks.

References

[1] A. Amir, D. Keselman, Maximum agreement subtree in a set of evolutionary trees: Metrics and efficient algorithms, SIAM Journal on Computing 26 (6) (1997) 1656–1669.

[2] M.-J. Chung, $O(n^{2.5})$ time algorithms for the subgraph homeomorphism problem on trees, Journal of Algorithms 8 (1) (1987) 106–112.

[3] R. Cole, M. Farach-Colton, R. Hariharan, T. M. Przytycka, M. Thorup, An $O(n \log n)$ algorithm for the maximum agreement subtree problem for binary trees, SIAM Journal on Computing 30 (5) (2000) 1385–1404.

[4] A. Dessmark, A. Lingas, A. Proskurowski, Faster algorithms for subgraph isomorphism of k-connected partial k-trees, Algorithmica 27 (1) (2000) 337–347.

[5] S. Dulucq, L. Tichit, RNA secondary structure comparison: Exact analysis of the Zhang-Shasha tree edit algorithm, Theoretical Computer Science 306 (1–3) (2003) 471–484.

[6] M.-L. Fernández, G. Valiente, A graph distance measure combining maximum common subgraph and minimum common supergraph, Pattern Recognition Letters 22 (6–7) (2001) 753–758.

[7] A. Gupta, N. Nishimura, Finding largest subtrees and smallest supertrees, Algorithmica 21 (2) (1998) 183–210.

[8] J. Jansson, A. Lingas, A fast algorithm for optimal alignment between similar ordered trees, Fundamenta Informaticae 56 (1–2) (2003) 105–120.

[9] T. Jiang, L. Wang, K. Zhang, Alignment of trees—an alternative to tree edit, Theoretical Computer Science 143 (1) (1995) 137–148.

[10] P. Kilpeläinen, H. Mannila, Ordered and unordered tree inclusion, SIAM Journal on Computing 24 (2) (1995) 340–356.

[11] J. Matoušek, R. Thomas, On the complexity of finding isomorphisms and other morphisms for partial k-trees, Discrete Mathematics 108 (1–3) (1992) 343–364.

[12] N. Nishimura, P. Ragde, D. M. Thilikos, Finding smallest supertrees under minor containment, Int. Journal of Foundations of Computer Science 11 (3) (2000) 445–465.

[13] R. Y. Pinter, O. Rokhlenko, D. Tsur, M. Ziv-Ukelson, Approximate labelled subtree homeomorphism, in: Proc. 15th Annual Symp. Combinatorial Pattern Matching, Vol. 3109 of Lecture Notes in Computer Science, Springer-Verlag, 2004, pp. 55–69.

[14] R. Shamir, D. Tsur, Faster subtree isomorphism, Journal of Algorithms 33 (2) (1999) 267–280.

[15] D. Shasha, J. T.-L. Wang, K. Zhang, F. Y. Shih, Exact and approximate algorithms for unordered tree matching, IEEE Transactions on Systems, Man, and Cybernetics 24 (4) (1994) 668–678.

[16] D. Shasha, K. Zhang, Fast algorithms for the unit cost editing distance between trees, Journal of Algorithms 11 (4) (1990) 581–621.

[17] M. A. Steel, T. Warnow, Kaikoura tree theorems: Computing the maximum agreement subtree, Information Processing Letters 48 (2) (1993) 77–82.

[18] G. Valiente, Algorithms on Trees and Graphs, Springer-Verlag, Berlin, 2002.

[19] G. Valiente, Constrained tree inclusion, in: Proc. 14th Annual Symp. Combinatorial Pattern Matching, Vol. 2676 of Lecture Notes in Computer Science, Springer-Verlag, 2003, pp. 361–371.

[20] G. Valiente, Constrained tree inclusion, Journal of Discrete Algorithms 00 (00) (2004) 000–000, in press.

[21] L. Wang, J. Zhao, Parametric alignment of ordered trees, Bioinformatics 19 (17) (2003) 2237–2245.

[22] K. Zhang, D. Shasha, Simple fast algorithms for the editing distance between trees and related problems, SIAM Journal on Computing 18 (6) (1989) 1245–1262.

Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Research Reports - 2004

- LSI-04-1-R : *Automatic Generation of Polynomial Loop Invariants: Algebraic Foundations*, Rodríguez, E. and Kapur, D.
- LSI-04-2-R : *Comparison of Methods to Predict Ozone Concentration* , Orozco, J.
- LSI-04-3-R : *Towards the definition of a taxonomy for the cots product´s market* , Ayala, Claudia P.
- LSI-04-4-R : *Modelling Coalition Formation over Time for Iterative Coalition Games*, Mérida-Campos, C. and Willmott, S.
- LSI-04-5-R : *Illegal Agents? Creating Wholly Independent Autonomous Entities in Online Worlds*, Willmott, S.
- LSI-04-6-R : *An Analysis Pattern for Electronic Marketplaces*, Queralt, A. and Teniente, E.
- LSI-04-7-R : *Exploring Dopamine-Mediated Reward Processing through the Analysis of EEG-Measured Gamma-Band Brain Oscillations*, Vellido, A. and El-Deredy, W.
- LSI-04-8-R : *Studying Embedded Human EEG Dynamics Using Generative Topographic Mapping*, Vellido, A. and El-Deredy, W. and Lisboa, P.J.G.
- LSI-04-9-R : *Similarity and Dissimilarity Concepts in Machine Learning*, Orozco, J.
- LSI-04-10-R : *A Framework for the Definition of Metrics for Actor-Dependency Models*, Quer, C. and Grau, G. and Franch, X.
- LSI-04-11-R : *QM: A Tool for Building Software Quality Models*, Carvallo, J.P. and Franch, X. and Grau, G. and Quer, C.
- LSI-04-12-R : *COSTUME: A Method for Building Quality Models for Composite COTS-based Software Systems*, Carvallo, J.P. and Franch, X. and Grau, G. and Quer, C.
- LSI-04-13-R : *Enabling Collaboration in Virtual Reality Navigators*, Theoktisto, V. and Fairén, M. and Navazo, I.
- LSI-04-14-R : *DesCOTS: A Software System for Selecting COTS Components*, Carvallo, J.P. and Franch, X. and Grau, G. and Quer, C.
- LSI-04-15-R : *Evaluation and symmetrisation of alignments obtained with the Giza++ software*, Lambert, P. and Castell, N.
- LSI-04-16-R : *A note on the use of topology extensions for provoking instability in communication networks*, Blesa, M.J.
- LSI-04-17-R : *An ISO/IEC-compliant Quality Model for ER Diagrams*, Costal, D. and Franch, X.
- LSI-04-18-R : *A Case Study on Pruning General Ontologies for the Development of Conceptual Schemas* , Conesa, J.
- LSI-04-19-R : *Adding Efficient and Reliable Access Paths to the JCF*, Marco, J. and Franch, X.

- LSI-04-20-R : *Exploiting Simple Corporate Memory in Iterative Coalition Games*, Mérida-Campos, C. and Willmott, S.
- LSI-04-21-R : *On the Semantics of Operation Contracts in Conceptual Modeling* , Queralt, A. and Teniente, E.
- LSI-04-22-R : *Complexity issues on bounded restrictive H-coloring*, Díaz, J. and Serna, M. and Thilikos, D.M.
- LSI-04-23-R : *Chromatic number in random scaled sector graphs*, Díaz, J. and Sanwalani, V. and Serna, M. and Spirakis, P.
- LSI-04-24-R : *Bounds on the bisection width for random d-regular graphs*, Díaz, J. and Serna, M. and Wormald, N.C.
- LSI-04-25-R : *Open Source environment to define constraints in route planning for GIS-T*, Pérez, L. and Silveira, A. da M.
- LSI-04-26-R : *A basic repository of operations for the refinement of general ontologies*, de Palol, X.
- LSI-04-27-R : *Tetrahedral mesh subdivision based on underlying volume data*, Rodríguez, L. and Navazo, I. and Vinacua, A.
- LSI-04-28-R : *The Price of Connectedness in Expansions*, Fomin, F.V. and Fraigniaud, P. and Thilikos, D.M.
- LSI-04-29-R : *Smaller kernels for hitting set problems of constant arity*, Nishimura, N. and Ragde, P. and Thilikos, D.M.
- LSI-04-30-R : *Searching Spatial Sense in the Ontological World: Discovering Spatial Objects*, Morocho, V and Pérez, L. and Saltor, F.
- LSI-04-32-R : *Implementation considerations of an Expert System to assess Stream Water Quality management*, Cabanillas, D. and Willmott, S.
- LSI-04-33-R : *Multisided patches*, Pla, N. and Vigo, M. and Cotrina, J.
- LSI-04-34-R : *SVMTool: A general POS tagger generator based on Support Vector Machines*, Giménez, J. and Màrquez, Ll.
- LSI-04-35-R : *A distributed and mobile component system based on the ambient calculus*, Mylonakis, N. and Orejas, F.
- LSI-04-36-R : *Developing Competitive HMM PoS Taggers Using Small Training Corpora*, Padró, M. and Padró, Ll.
- LSI-04-37-R : *The AlignmentSet Toolkit*, Lambert, P.
- LSI-04-38-R : *Integración de Fuentes de Datos espaciales: análisis e implementación de una Ontología de términos espaciales: Primera Parte - - Creación de una Ontología*, Ramos, Erik G.
- LSI-04-39-R : *Integración de Fuentes de Datos espaciales: análisis e implementación de una Ontología de términos espaciales: Segunda Parte - - Evaluación de similaridades*, Ramos, Erik G.
- LSI-04-40-R : *Kernels on Structured Domains*, Valentín, L.

- LSI-04-41-R : *Determining the Structural Events that May Violate an Integrity Constraint*, Cabot, J. and Teniente, E.
- LSI-04-42-R : *Review of Statistical Word Alignment Techniques* , Lambert, P.
- LSI-04-43-R : *Algoritmos geneticos en el problema de la solucion deseadaOptimizacion de parametros*
  , Barreiro, E. and Joan-Arinyo, R. and Luzón, M.V.
- LSI-04-44-R : *Generative Topographic Mapping as a constrained mixture of Student t-distributions: Theoretical developments* , Vellido, A.
- LSI-04-45-R : *Adapting Agent Communication Languages for Web Service to Web Service Communication*, Willmott, S. and Fernández-Peña, F. O. and Mérida-Campos, C. and Con-stantinescu, I.
- LSI-04-49-R : *A brief on constraint solving*, Hoffmann, C.M. and Joan-Arinyo, R.
- LSI-04-50-R : *Missing data imputation through Generative Topographic Mapping as a mixture of t-distributions: Theoretical developments*, Vellido, A.
- LSI-04-51-R : *A two-tiered Methodology for Metamodel Extension Applied to UML 1.4*, Franch, X. and Ribó, J. M.
- LSI-04-52-R : *Virtual reality for prostate gland cryosurgery*, Joan-Arinyo, R.
- LSI-04-53-R : *High level communication functionalities for wireless sensor networks*, Àlvarez, C. and Díaz, J. and Petit, J. and Rolim, J. and Serna, M.
- LSI-04-54-R : *A MOF-Compliant Approach to Software Quality Modeling*, Burgués, X. and Franch, X. and Ribó, J. M.
- LSI-04-55-R : *Pure Nash equilibria in games with a large number of actions*, Àlvarez, C. and Gabarró, J. and Serna, M.
- LSI-04-56-R : *An Optimal Anytime Estimation Algorithm*, Gavaldà, R.
- LSI-04-60-R : *An Algebraic View of the Relation between Largest Common Subtrees and Smallest Common Supertrees*, Rosselló, F. and Valiente, G.

Hardcopies of reports can be ordered from:

Núria Sanchez
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Campus Nord, Mòdul C6
Jordi Girona Salgado, 1-3
03034 Barcelona,Spain
nurias@lsi.upc.es

See also the Departament WWW pages, http://www.lsi.upc.es/