# Bounded Repairability for Regular Tree Languages

PIERRE BOURHIS, CNRS—CRIStAL UMR 9189
GABRIELE PUPPIS, CNRS—LaBRI, University of Bordeaux
CRISTIAN RIVEROS, Pontificia Universidad Católica de Chile
SŁAWEK STAWORKO, INRIA Lille-Nord Europe, University of Lille and LFCS, University of Edinburgh

We study the problem of bounded repairability of a given restriction tree language $R$ into a target tree language $T$. More precisely, we say that $R$ is bounded repairable with respect to $T$ if there exists a bound on the number of standard tree editing operations necessary to apply to any tree in $R$ to obtain a tree in $T$. We consider a number of possible specifications for tree languages: bottom-up tree automata (on curry encoding of unranked trees) that capture the class of XML schemas and document type definitions (DTDs). We also consider a special case when the restriction language $R$ is universal (i.e., contains all trees over a given alphabet).

We give an effective characterization of bounded repairability between pairs of tree languages represented with automata. This characterization introduces two tools—synopsis trees and a coverage relation between them—allowing one to reason about tree languages that undergo a bounded number of editing operations. We then employ this characterization to provide upper bounds to the complexity of deciding bounded repairability and show that these bounds are tight. In particular, when the input tree languages are specified with arbitrary bottom-up automata, the problem is coNExp-complete. The problem remains coNExp-complete even if we use deterministic nonrecursive DTDs to specify the input languages. The complexity of the problem can be reduced if we assume that the alphabet, the set of node labels, is fixed: the problem becomes PSpace-complete for nonrecursive DTDs and coNP-complete for deterministic nonrecursive DTDs. Finally, when the restriction tree language $R$ is universal, we show that the bounded repairability problem becomes Exp-complete if the target language is specified by an arbitrary bottom-up tree automaton and becomes tractable (P-complete, in fact) when a deterministic bottom-up automaton is used.

Categories and Subject Descriptors: H.2.3 [**Database Management**]: Languages—*Data manipulation languages*; H.2.5 [**Database Management**]: Heterogeneous Databases—*Data translation*

General Terms: Algorithms, Languages, Theory



P. Bourhis was supported by CPER Nord-Pas de Calais/FEDER DATA Advanced Data Science and Technologies 2015-2020. The research leading to these results was partly funded by EPSRC (UK) grant EP/G004021/1 and by EU project FOX (FP7-ICT-233599). G. Puppis was partially supported by the ExStream project (ANR-13-JS02-0010). C. Riveros was supported by CONICYT + PAI/Concurso Nacional Apoyo al Retorno de Investigadores/as desde el extranjero—Convocatoria 2013 + 821320001 and by the Millenium Nucleus Center for Semantic Web Research under grant NC120004. S. Staworko was supported by the EU FP7 DIACHRON project.
Authors' addresses: P. Bourhis, INRIA LILLE, Parc scientifique de la Haute Borne 40, avenue Halley. Bat. B, Park Plaza 59650 Villeneuve d'Ascq France; email: pierre.bourhis@univ-lille1.fr; G. Puppis, CNRS / LaBRI - Laboratoire Bordelais de Recherche en Informatique, 351 cours de la Liberation, F-33405 Talence cedex, France; email: gabriele.puppis@labri.fr; C. Riveros, Department of Computer Science, Pontificia Universidad Católica de Chile, Vicuña Mackenna 4860, Edificio San Agustin, 4to piso, Macul, Santiago, Chile; email: cristian.riveros@uc.cl; S. Staworko, INRIA Lille-Nord Europe, University of Lille and LFCS, University of Edinburgh, Office 5.37, Informatics Forum, 10 Crichton Street, EH8 9AB Edinburgh, United Kingdom; email: slawomir.staworko@inria.fr.

DOI: http://dx.doi.org/10.1145/2898995

Additional Key Words and Phrases: XML, XML schema, DTD, repair, edit distance

**ACM Reference Format:**
Pierre Bourhis, Gabriele Puppis, Cristian Riveros, and Sławek Staworko. 2016. Bounded repairability for regular tree languages. ACM Trans. Database Syst. 41, 3, Article 18 (June 2016), 45 pages.
DOI: http://dx.doi.org/10.1145/2898995

## 1. INTRODUCTION

A basic problem in data management is to ensure that data is valid, namely that it satisfies all integrity constraints associated with a schema [Bertossi 2011]. Validation of data with respect to a schema is crucial in any database system: if data does not satisfy the integrity constraints, then one cannot guarantee that the output produced by the system is correct. Nevertheless, when data does not satisfy constraints, a natural approach is to attempt a *repair*—that is, to modify the data minimally so that it becomes valid [Arenas et al. 1999; Afrati and Kolaitis 2009]. We may want to perform this transformation on the data, or we may be merely interested in knowing how difficult it is to perform the transformation in case of need—that is, determining how far a given collection of data is from satisfying the specification. For example, some applications may retain input data even when this contains a few errors, where "few" could be interpreted as a user-defined bound to the total number of errors or to the fraction of errors over the size of the input [Grahne and Thomo 2004].

On relational data, this problem has been extensively studied under the notion of constraint repair (e.g., see Arenas et al. [1999] and Afrati and Kolaitis 2009]): in this case, the specifications are given by relational integrity constraints, such as keys and foreign keys, and the problem asks to determine how much a database needs to be modified to satisfy a given constraint. This approach has been investigated for a variety of integrity constraints, starting with classical functional and inclusion dependencies [Arenas et al. 1999] and continuing with more expressive constraints such as tuple generating dependencies [Afrati and Kolaitis 2009]. In addition, several different repair operators have been considered in the relational case, including insertions, deletions, and modifications of tuples. Besides finding repairs of relations, this line of research also focuses on querying inconsistent documents via their minimal repairs.

In the XML framework, malformed or nonconformant documents are more the rule than the exception [Chen et al. 2005; Ofuonye et al. 2010]. Indeed, a recent study [Grijzenhout and Marx 2013] shows that although most XML documents are well formed (more than 85%), only 25% of them reference a downloaded schema. Even worse, in this study, it is shown that less than 10% of XML documents satisfy their downloaded schema. This means that most of the XML data on the Web can be read, but only a small part of it can be processed automatically. In this scenario, it is natural to look for automatic repairing XML data with respect to a target schema. The idea is that an automatic repair process receives an invalid XML document and produces the best sequence of edit operations that results in a document satisfying the target schema. The edit operations should respect the nested structure of the XML document and modify the document in a minimal way.

The notion of repair for XML data is defined in a natural way when considering documents as trees: in this case, a repair can be simply defined as the *tree edit distance* [Tai 1979] between the input tree and the repaired tree—that is, the number of atomic edit operations that are needed to get from one tree to another. An atomic edit operation here amounts to inserting, deleting, or modifying a single node in a tree [Bille 2005]. Edit distance can then be lifted to a measure of distance $\mathsf{dist}(t, T)$ of a tree $t$ from a specification $T$: this is nothing but the minimal distance of $t$ to any tree satisfying $T$. In our setting, $t$ can be seen as an XML document and $T$ as an XML Schema, and hence $\mathsf{dist}(t, T)$ measures how difficult it is to repair the data $t$ so as to satisfy

$T$. Furthermore, $\mathsf{dist}(t, T)$ can be computed efficiently when $T$ is a regular language (e.g., document type definition (DTD) or XML schema definition) specified by means of an automaton [Wagner 1974; Aho and Peterson 1972].

In this article, we take the next step in repairing XML documents or trees: given two regular specifications $S$ and $T$ over trees, we aim at calculating how difficult it is, in the worst case, to transform an object satisfying $S$ into an object satisfying $T$. The problem is motivated by considering $S$ to be a *source* (i.e., a constraint that the input is guaranteed to satisfy) and $T$ to be a *target* (i.e., a constraint that needs to be enforced). More precisely, we consider the worst case over all trees $t$ satisfying $S$ of the minimum number of edit operations needed to transform $t$ into some tree in $T$—that is,

$$\mathsf{cost}(S, T) = \sup_{t \in S} \mathsf{dist}(t, T) = \sup_{t \in S} \min_{t' \in T} \mathsf{dist}(t, t').$$

Of course, the preceding cost may be infinite. In this work, we isolate the pairs of schemas $S$ and $T$ such that $\mathsf{cost}(S, T)$ is finite, and we give optimal procedures to decide when this happens, namely when schema $S$ is "almost contained" in schema $T$. Specifically, we say that $S$ is *bounded repairable* into $T$ when $\mathsf{cost}(S, T) < \infty$—that is, when every document $t$ in $S$ can be repaired to a document $t'$ in $T$ by applying a finite, uniformly bounded number of edits.

The notion of bounded repair is also motivated by the schema matching problem [Rahm and Bernstein 2001]: we would like to identify whether two schemas are semantically related. In our setting, the semantic relation between two schemas is considered at a very low level, namely each schema is seen as a set of documents and not as a set of rules. Then the bounded repair problem states that a source schema $S$ is related to a schema $T$ if any XML document satisfying $S$ can be transformed with a finite, uniformly bounded number of operations into a document satisfying $T$. Here, it is important to notice that our repair operations are designed to only consider the structural part of the data. Further research needs to be done to include in the analysis the data itself, such as the constraints between attribute values in XML documents, as this is an important aspect to take into account when reasoning on and transforming XML documents.

The following examples give an account of some of the difficulties of telling whether one schema is bounded repairable into another.

*Example* 1.1. Recall that languages of unranked trees can be specified by means of DTDs—that is, by sets of rules of the form $a \to L_a$, where $L_a$ is a regular language describing the possible sequences of children of an $a$-labeled node. For the sake of brevity, we will often omit from DTDs the rules of the form $a \to \varepsilon$, which denote the fact that $a$-labeled nodes are leaves.

Consider the following DTDs:

$$\begin{array}{lll} S: & r \to d\,c^* & \qquad T: \; r \to a^*\,e \\ & d \to a^*\,b^* & \qquad\quad\; e \to b^*\,c^* \end{array}$$

The left-hand side schema $S$ defines the language of all trees of the form $r(d(a, \ldots, a, b, \ldots, b), c, \ldots, c)$, whereas the right-hand side schema $T$ defines the language of all trees of the form $r(a, \ldots, a, e(b, \ldots, b, c, \ldots, c))$. We claim that $S$ is repairable into $T$ with a uniformly bounded number of edit operations. Indeed, given a tree $r(d(a, \ldots, a, b, \ldots, b), c, \ldots, c)$ satisfying $S$, one can first delete the node labeled by $d$, obtaining the tree $r(a, \ldots, a, b, \ldots, b, c, \ldots, c)$, and then insert a new $e$-labeled node under the root, which adopts as children all nodes labeled by $b$ or $c$; this results in a tree $r(a, \ldots, a, e(b, \ldots, b, c, \ldots, c))$ that satisfies $T$.

*Example* 1.2. Consider the following DTDs:

$$S' : \begin{array}{l} r \to a \\ a \to b^* \end{array} \qquad T' : \begin{array}{l} r \to a \\ a \to b^* c \end{array}$$

It is easy to see that $S'$ is bounded repairable into $T'$: any tree $r(a(b, \ldots, b))$ in $S'$ can be modified into a tree in $T'$ by inserting a new $c$-labeled node as a right sibling of the nodes labeled by $b$. However, if we replace in both DTDs $S'$ and $T'$ the rule $r \to a$ with the rule $r \to a^*$, we obtain a new pair of languages $S''$ and $T''$ such that $S''$ is not bounded repairable into $T''$. This example suggests that bounded repairability depends on some interplay between the rules of DTDs and, more generally, between the specifications of the labelings of the nodes at different levels of the trees.

We will deal with the notion of bounded repairability for schemas that are more general than DTDs, such as schemas that are given by regular tree languages [Schwentick 2007], and which capture the structural part of the W3C's XML schema [Fallside and Walmsley 2004]. We will formalize the edit distance between regular tree languages, and from this we will define the *bounded repair problem*—that is, the problem of deciding bounded repairability between two given tree languages $S$ and $T$. Our main result is that it is decidable whether or not $S$ can be repaired into $T$ with a uniformly bounded number of edits.

For regular languages of words, the bounded repair problem was resolved in Benedikt et al. [2013]. There, it was shown that the problem is coNP-complete when the languages are represented by deterministic finite state automata, and a characterization of bounded repairability was given using a coverability relation between chains of connected components of the automata. In the case of tree languages, the problem turns out to be more complex, both in terms of complexity and in terms of proof techniques that are required to solve it. We will provide a characterization of bounded repairability that exploits a suitable notion of component of a stepwise tree automaton [Carme et al. 2004], a form of automaton that turns out to be particularly convenient for analyzing repairs. An additional complication for the tree case is that we need to consider structures of connected components of stepwise tree automata that take the form of trees rather than chains. Our characterization of the bounded repairability of $S$ into $T$ requires that every component structure of $S$ can be "covered" by a component structure of $T$. The notion of covering is subtle, and the proof that it captures bounded repairability requires lifting the notion of edit from the level of the individual trees to the level of the component trees associated with the automata for $S$ and $T$.

With an effective characterization at hand, we can decide the bounded repairability problem, and with some additional optimizations, we can give tight complexity bounds. It turns out that, differently from the string setting, the bounded repairability problem is equally complex no matter whether the tree languages are given by nondeterministic automata, deterministic automata, DTDs, or even nonrecursive DTDs. Indeed, for all of these cases, the bounded repair problem is coNExp-complete. We then look for tractable cases that are obtained by further restricting the tree specifications. For example, we show that the bounded repairability problem becomes much simpler when the source alphabet is fixed and the languages are given by deterministic DTDs, or when the source language is assumed to be trivial, namely the set of all trees over a given finite alphabet.

*New material in this article*. Preliminary versions of some of the results in this work appeared in Puppis et al. [2012]. However, this article contains substantial new material. We include a full proof of the main characterization result (Theorem 5.7). The upper bounds to the number of repairs (Lemma 5.5 and Proposition 5.8) are also new. As concerns the complexity of deciding bounded repairability, the article provides

new complexity bounds that are moreover tight. In the previous work, it was shown that the bounded repair problem for regular tree languages is decidable in $\Pi_2^{\text{EXP}}$ and is EXP-hard. Here, we show that the problem is actually coNEXP-complete (Theorems 7.3 and 7.4). We finally include new examples and proofs of other claims that were omitted in the previous work.

*Organization*. The article is organized as follows. In Section 2, we discuss some related work. In Section 3, we give some preliminaries on trees and regular tree languages, and in Section 4, we define the bounded repairability problem for tree languages. In Section 5, we give the formal statement of our main result—that is, a characterization of those pairs of schemas that are bounded repairable. Section 6 gives a detailed proof of the characterization. In Section 7, we analyze in detail the complexity of the bounded repairability problem. In Section 8, we give another, simple characterization of bounded repairability for the case where the source language is universal, and we accordingly derive new complexity results. Finally, in Section 9, we give our conclusions and future work.

## 2. RELATED WORK

Ever since their conception, computers required the input data to follow a set of strict structural and semantic rules, and the failure to do so typically resulted in operations producing unpredictable outputs, a well-known phenomenon of *Garbage In, Garbage Out* [Babbage 1864]. While initially this phenomenon was attributed to situations of erroneous manual data entry, with the ever-increasing number of applications exchanging data, the phenomenon has gained a new meaning, describing potential problems occurring when two applications attempt to communicate with incompatible protocols [Lidwell et al. 2010]. Although our research aims at solving the problems of the latter scenario, very early in the development of computer science we saw solutions to the problems resulting from erroneous manual data entry. One prominent example is the work on the error-correcting parser for context-free languages [Aho and Peterson 1972], where a malformed input string is repaired by applying a (minimal) number of editing operations that make it conform to the given grammar. Korn et al. [2013] consider a similar problem for XML, where a serialization of an XML document is not well formed (e.g., mismatching opening and closing tags or misspelled tag names) and is repaired to allow parsing into an XML tree. A slightly different variant of the problem is repairing well-formed XML with respect to a given schema in the form of a DTD [Suzuki 2005; Staworko and Chomicki 2006] or XML schema [Staworko et al. 2008]. The validity of the XML document is restored using a minimal set of editing operations (insertion, deletion, and renaming of nodes). Adding a move operation that can change the relative order of elements is challenging because of fundamental computational limitations [Cormode and Muthukrishnan 2007], and approximate measures have been studied for this operation [Boobna and de Rougemont 2004]. Furthermore, repairing XML documents with respect to analogues of classical relational constraints (key and inclusion dependencies) has also been studied [Flesca et al. 2005]. HTML documents often violate the syntactic rules of well formedness and the structural rules imposed by the HTML standard; consequently, repairing them requires methods that diligently combine the approaches of editing the textual serialization and editing the tree representation of the input document [Chen et al. 2005; Ofuonye et al. 2010].

The problem that we study is, however, more general than repairing a single input XML document with respect to a given schema, as we are interested in repairing any input document drawn from a possibly infinite set of documents with a number of operations that is independent on the size of the XML document. One could attempt to approximate the bound on the number of required editing operations by randomly

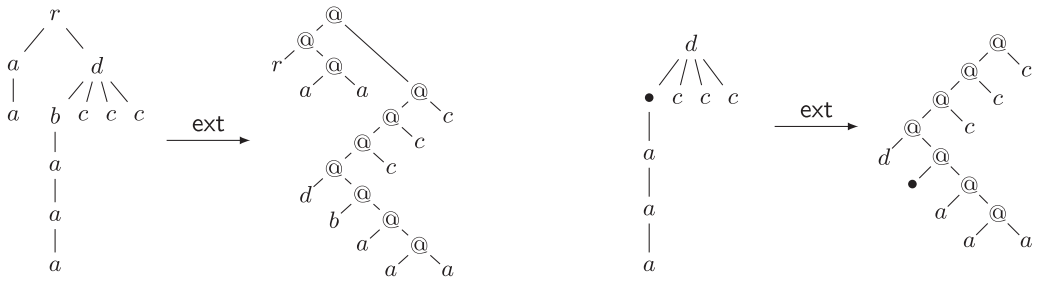



Fig. 1. Curry encodings of an unranked tree and a context.

generating the input document [Antonopoulos et al. 2013] and then computing their edit distance to the target regular language. Because in our setting the input document is drawn from a regular language and is repaired with respect to another regular language, the problem is in fact a generalization of a well-known and thoroughly studied problem of containment of two schemas [Comon et al. 2007; Colazzo et al. 2013; Martens et al. 2009]. Closely related is the problem of measuring similarity between two schemas based on a notion of embeddings studied in Fan and Bohannon [2008]. Yet there are significant differences: on the one hand, our semantic characterization of bounded reparability is stronger than the structural similarity determined with embeddings, but on the other hand, the framework of Fan and Bohannon [2008] introduces an additional challenge: it requires the embeddings to be information preserving—that is, for any query that can be evaluated on a document from the source schema, there exists an equivalent query over the corresponding document from the target schema.

The preservation of information expressed with queries is the essence of data exchange source-to-target mappings and finding (target) solutions for a given source document is a difficult problem in the context of XML [Arenas and Libkin 2008; Amano et al. 2009]. In this setting, the problem of absolute consistency [Bojańczyk et al. 2011], checking that a solution exists for any possible source instance, bears strong resemblance to the problem of reparability except it does not call for using editing operations, and consequently it does not impose any limit on the number of editing operations but merely inquires the possibility of always finding a solution. In fact, the authors propose a solution that uses a notion of a *kind* in a manner analogous to the connected components used in our approach.

## 3. REGULAR LANGUAGES OF TREES

In this article, we work with finite unranked ordered trees whose nodes are labeled over a finite alphabet $\Sigma$. Formally, the set of *finite unranked ordered trees over* $\Sigma$ (hereafter, simply *trees*) is inductively defined as follows: (1) every symbol $a \in \Sigma$ is a tree, and (2) if $a \in \Sigma$, $n \in \mathbb{N}$, and $t_1, \ldots, t_n$ are trees, then $a(t_1, \ldots, t_n)$ is a tree. A sequence of $\Sigma$-trees $t_1 \cdot \ldots \cdot t_n$ is called a *forest*. As an example, the left-hand side of Figure 1 shows a tree over the alphabet $\Sigma = \{r, a, b, c, d\}$. We denote by $\mathcal{T}_\Sigma$ the set of all trees over $\Sigma$. A (*tree*) *language over* $\Sigma$ is any subset $L$ of $\mathcal{T}_\Sigma$.

It is useful to identify nodes of an unranked tree with sequences of positive natural numbers. Given a tree $t$ of the form $a(t_1, \ldots, t_n)$, its domain is the subset of $\mathbb{N}^*$ that is formally defined as $\mathsf{nodes}(t) = \{\varepsilon\} \cup \{i \cdot x \mid x \in \mathsf{nodes}(t_i) \wedge 1 \leq i \leq n\}$. Note that the root of a tree is represented by $\varepsilon$. Finally, for every node $x \in \mathsf{nodes}(t)$, we denote by $t(x)$ the label of $x$ in $t$.

Given a tree $t$, we introduce two partial orders on the domain $\mathsf{nodes}(t)$, which are called *ancestor order* and *post-order* and are denoted by $\preccurlyeq_t^{\mathsf{anc}}$ and $\preccurlyeq_t^{\mathsf{post}}$, respectively.

The ancestor order $\preccurlyeq_t^{\mathsf{anc}}$ is nothing but the prefix order on the sequences of positive natural numbers that identify the nodes of $t$—that is, $x \preccurlyeq_t^{\mathsf{anc}} y$ if and only if $x$ is a prefix of $y$. The post-order $\preccurlyeq_t^{\mathsf{post}}$ is the total ordering on the nodes of $t$ (i.e., sequences of natural numbers) defined by $x \preccurlyeq_t^{\mathsf{post}} y$ if and only if $y$ is a prefix of $x$ or there exist $z, x', y' \in \mathbb{N}^*$ and $i, j \in \mathbb{N}$ such that $x = z \cdot i \cdot x'$, $y = z \cdot j \cdot y'$, and $i \leq j$.

*DTDs*. We manipulate regular languages of trees mainly by means of automaton-based specifications, which will be formally defined in the next paragraphs. However, we use less expressive specifications, such as XML document type definitions, to give examples of simple tree languages.

An *XML DTD* is defined as a tuple $D = (\Sigma, d, I)$, where $\Sigma$ is a finite alphabet, $d$ is a function that maps symbols from $\Sigma$ to regular expressions over $\Sigma$, and $I \subseteq \Sigma$ is the set of initial symbols [Comon et al. 2007]. A tree $t$ satisfies the DTD $D$ if $t(\varepsilon) \in I$ and, for every $x \in \mathsf{nodes}(t)$, the word $t(x \cdot 1) \cdot \ldots \cdot t(x \cdot n)$ belongs to the language defined by the regular expression $d(t(x))$, where $n$ is the number of children of $x$ in $t$. We denote by $\mathscr{L}(D)$ the language of trees satisfying the DTD $D$. We will often omit the rules of the form $a \to \varepsilon$, as well as the set of initial symbols from the definition of a DTD when this set is understood from the context (e.g., when it consists of a single symbol $r$). As an example, consider the following DTD:

$$\begin{array}{lllll} D: & r \to a\,d & & d \to b\,c^* \\ & a \to a + \varepsilon & & b \to a \end{array}$$

One can easily check that the left-hand side tree of Figure 1 satisfies the preceding DTD $D$.

It is known that DTDs define a proper subclass of regular tree languages [Martens et al. 2006; Martens and Niehren 2007]. Quite interestingly, most of our complexity lower bounds for the bounded repair problem hold even for languages defined by DTDs (see Section 7 for more details).

We will also consider languages defined by nonrecursive deterministic DTDs, which are often used in practice and have been studied extensively in Segoufin and Vianu [2002] and Segoufin and Sirangelo [2007]. Let us define the dependency graph of a DTD $D = (\Sigma, d, I)$ as the directed graph whose nodes are the letters in $\Sigma$ and whose edges connect any letter $a$ to a letter $b$ whenever $b$ occurs in the language specified by the regular expression $d(a)$. A DTD $D$ is called *nonrecursive* if its dependency graph is acyclic. A DTD $D$ is called *deterministic* if each regular expression $d(a)$ is one-unambiguous (namely, it can be equally seen as a deterministic finite state automaton) [Brüggemann-Klein and Wood 1998] and the set of initial symbols is a singleton.

*Curry encoding*. To ease the definitions of the automaton model and the reasoning on tree repairs, we introduce here the notion of curry encoding, also known as extension encoding, of a tree [Carme et al. 2004; Martens and Niehren 2007]. According to this encoding, any unranked tree over $\Sigma$ is seen as a binary tree with leaves labeled over $\Sigma$ and internal nodes labeled by a distinguished symbol @. Formally, the *curry encoding* is the function $\mathsf{ext}$ that injectively maps unranked trees to binary trees as follows:

$$\begin{aligned} &\mathsf{ext}(a) = a, \\ &\mathsf{ext}(a(t_1, \ldots, t_n)) = @(\mathsf{ext}(a(t_1, \ldots, t_{n-1}), \mathsf{ext}(t_n))). \end{aligned}$$

To ease readability, we use the symbol @ as a binary, infix, left-associative operator: for instance, $\mathsf{ext}(a(t_1, \ldots, t_n)) = a@\mathsf{ext}(t_1)@ \ldots @\mathsf{ext}(t_n)$. The left-hand side of Figure 1 illustrates the encoding of an unranked tree. The inverse $\mathsf{ext}^{-1}$ of the encoding is defined by providing the symbol @ with the semantics of the extension operator on unranked

trees and by evaluating the expression in a bottom-up fashion—that is, $\mathsf{ext}^{-1}(a) = a$ and $\mathsf{ext}^{-1}(a@t_1@\ldots@t_n) = a(\mathsf{ext}^{-1}(t_1), \ldots, \mathsf{ext}^{-1}(t_n))$.

We observe that there is a one-to-one correspondence between the nodes of an unranked tree and the leaves of the curried encoding. In particular, the root node of an unranked tree corresponds to the left-most leaf of its curry encoding. Moreover, the *yield* of a curried tree (i.e., the sequence of leaves taken from left to right) corresponds to the standard left-to-right preorder traversal of the corresponding unranked tree. Another observation follows from the semantics of the extension operator: the inner nodes of a curried tree, labeled with @, correspond to the edges of the unranked tree.

Hereafter, we will identify unranked trees with their curry encodings. In particular, by a slight abuse of notation, we will denote by $\mathcal{T}_\Sigma$ the set of all curry encodings of trees over $\Sigma$, and by $\mathsf{nodes}(t)$ the domain of the curry encoding of a tree $t$.

*Contexts*. We now fix another special symbol $\bullet \notin \Sigma$ that will be used as a placeholder for contexts. Formally, a (curried) *context* over $\Sigma$ is the curry encoding of a tree over the alphabet $\Sigma \cup \{\bullet\}$, with a single node labeled by $\bullet$ (note that in the curry encoding, the symbol $\bullet$ must occur in a leaf). We denote by $\mathcal{C}_\Sigma$ the set of all contexts over $\Sigma$. The *empty* context is the context $\bullet$ having exactly one node. The right-hand side of Figure 1 illustrates the encoding of a context. A context $C$ is *horizontal* if the placeholder $\bullet$ is the left-most leaf of $C$. We point out that a horizontal context has the form $\bullet@t_1@\ldots@t_n$ and represents the forest—that is, a sequence of trees—$(\mathsf{ext}^{-1}(t_1), \ldots, \mathsf{ext}^{-1}(t_n))$. Note that the empty context is horizontal.

For a context $C$ and a tree $t$, we denote by $C \circ t$ the tree obtained from the substitution of $\bullet$ by $t$ in $C$. Similarly, the composition $C_1 \circ C_2$ of two contexts $C_1$ and $C_2$ is obtained from the substitution of the placeholder in $C_1$ by $C_2$ (this results again in a context in $\mathcal{C}_\Sigma$). The composition of two horizontal contexts is also horizontal and corresponds to concatenation of the corresponding forests of unranked trees. Note, however, the difference in the order of context composition and the order of forest concatenation: if $C = \bullet@t_1@\ldots@t_n$ and $C' = \bullet@t'_1@\ldots@t'_m$, then $C \circ C' = \bullet@t'_1@\ldots@t'_m@t_1@\ldots@t_n$, which represents the forest $(\mathsf{ext}^{-1}(t'_1), \ldots, \mathsf{ext}^{-1}(t'_m), \mathsf{ext}^{-1}(t_1), \ldots, \mathsf{ext}^{-1}(t_n))$.

*Stepwise tree automata*. We use stepwise tree automata to specify regular tree languages. These are essentially bottom-up tree automata running on the curry encodings of trees [Carme et al. 2004; Martens and Niehren 2007; Comon et al. 2007]. Formally, a *stepwise automaton* is a tuple $\mathcal{A} = (\Sigma, Q, \delta, \delta_0, F)$, where

(1) $\Sigma$ is a finite set of labels,
(2) $Q$ is a finite set of states,
(3) $\delta : Q \times Q \to 2^Q$ is a transition function,
(4) $\delta_0 : \Sigma \to 2^Q$ is an assignment of initial states to labels, and
(5) $F \subseteq Q$ is a set of final states.

We say that the automaton $\mathcal{A}$ is *deterministic* if $\delta_0$ (respectively, $\delta$) can be described as a partial function from $\Sigma$ (respectively, $Q \times Q$) to $Q$. It is often convenient to represent $\delta_0$ and $\delta$ as a set of rules. For instance, we write $a \to q$ to indicate that $q \in \delta_0(a)$ and $q_1@q_2 \to q$ to indicate that $q \in \delta(q_1, q_2)$.

A *run* of a stepwise automaton $\mathcal{A} = (\Sigma, Q, \delta_0, \delta, F)$ on a tree $t \in \mathcal{T}_\Sigma$ is a function $\rho : \mathsf{nodes}(t) \to Q$ such that (1) for every leaf node $x$, $\rho(x) \in \delta_0(t(x))$, and (2) for every inner node $x$, $\rho(x) \in \delta(\rho(x{\cdot}1), \rho(x{\cdot}2))$ (recall that we represent $t$ with its curry encoding). A run $\rho$ is *accepting* if $\rho(\varepsilon) \in F$. The language recognized by $\mathcal{A}$, denoted by $\mathscr{L}(\mathcal{A})$, is the set of all trees $t \in \mathcal{T}_\Sigma$ on which $\mathcal{A}$ has an accepting run.

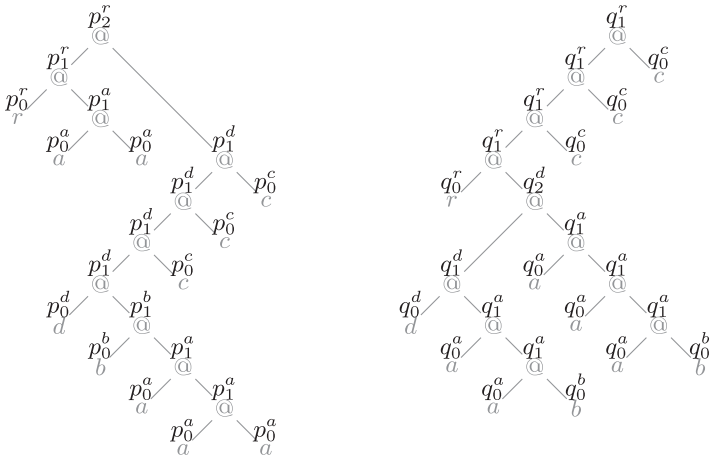


Fig. 2. Runs of two stepwise tree automata over curried trees.

*Example* 3.1. The following will serve as our running example. Consider two DTDs:

$$\begin{array}{lll} D: & r \to a\,d & \\ & a \to a + \varepsilon & \\ & d \to b\,c^* & \\ & b \to a & \end{array} \qquad \begin{array}{lll} D': & r \to d\,c^* & \\ & d \to a\,a & \\ & a \to a + b & \end{array}$$

The following two stepwise automata capture (modulo the curry encoding) the languages defined by the previous DTDs (the underlined states are final, and each rule with $q^a_?$ translates to two rules with $q^a_0$ and $q^a_1$):

$$\begin{array}{lll} \mathcal{S}: & r \to p^r_0 & p^r_0 \mathbin{@} p^a_? \to p^r_1 \\ & a \to p^a_0 & p^r_1 \mathbin{@} p^d_1 \to \underline{p^r_2} \\ & d \to p^d_0 & p^a_0 \mathbin{@} p^a_? \to p^a_1 \\ & b \to p^b_0 & p^d_0 \mathbin{@} p^b_1 \to p^d_1 \\ & c \to q^c_0 & p^d_1 \mathbin{@} p^c_0 \to p^d_1 \\ & & p^b_0 \mathbin{@} p^a_? \to p^b_1 \end{array} \qquad \begin{array}{lll} \mathcal{T}: & r \to q^r_0 & q^r_0 \mathbin{@} q^d_2 \to \underline{q^r_1} \\ & d \to q^d_0 & \underline{q^r_1} \mathbin{@} q^c_0 \to \underline{q^r_1} \\ & a \to q^a_0 & q^d_0 \mathbin{@} q^a_1 \to q^d_1 \\ & b \to q^b_0 & q^d_1 \mathbin{@} q^a_1 \to q^d_2 \\ & c \to q^c_0 & q^a_0 \mathbin{@} q^a_1 \to q^a_1 \\ & & q^a_0 \mathbin{@} q^b_0 \to q^a_1 \end{array}$$

Figure 2 presents the (accepting) runs of the automata $\mathcal{S}$ and $\mathcal{T}$ on some curried trees.

Stepwise automata capture exactly the class of regular (unranked) tree languages [Carme et al. 2004], and they are more succinct than other models of automata [Martens and Niehren 2007]. Even though other equivalent models of automata, such as unranked tree automata, are more frequently used in practice, these can be converted into stepwise tree automata in polynomial time. This means that algorithms for analyzing stepwise automata provide the same complexity bounds for unranked tree automata—in particular, all of our complexity results apply to both stepwise tree automata and unranked tree automata. The main advantage of using stepwise automata in our proofs is due to their ability to capture in a uniform way the "cyclic behavior" of a regular tree language (as we will see in Section 5, this cyclic behavior is defined in terms of strongly connected components of automata).

In the sequel, we will assume that our stepwise tree automata are trimmed, namely they contain only states that appear in valid accepting runs. Formally, a stepwise tree automaton $\mathcal{A} = (\Sigma, Q, \delta, \delta_0, F)$ is *trimmed* if for every state $q \in Q$, there exist $t \in \mathcal{T}_\Sigma$

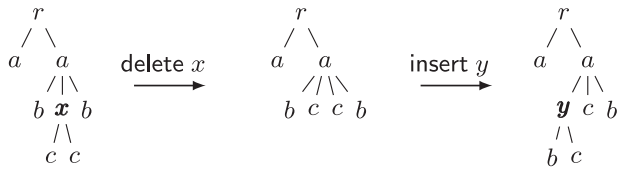



Fig. 3. Edit operations on unranked trees.

and an accepting run $\rho$ of $\mathcal{A}$ on $t$ such that $\rho(x) = q$ for some $x \in \mathsf{nodes}(t)$. Every stepwise tree automaton can be trimmed in linear time [Comon et al. 2007]. Given that all problems considered in this article are at least P-hard, the assumption that all stepwise tree automata are trimmed is without loss of generality.

As is usual for word automata, we extend the transition function $\delta$ of a stepwise automaton to trees in $\mathcal{T}_\Sigma$ and to contexts in $\mathcal{C}_\Sigma$. More precisely, we define the function $\delta^* : \mathcal{T}_\Sigma \to 2^Q$ such that $q \in \delta^*(t)$ if and only if there exists a run $\rho$ of $\mathcal{A}$ on $t$ and $\rho(\varepsilon) = q$. Similarly, we define the function $\delta^*_\bullet : Q \times \mathcal{C}_\Sigma \to 2^Q$ such that $q' \in \delta^*_\bullet(q, C)$ if and only if there exists a run $\rho$ of $\mathcal{A}_q = (\Sigma \cup \{\bullet\}, Q, \delta_0 \cup \{(\bullet, q)\}, \delta, Q)$ on $C$ and $\rho(\varepsilon) = q'$ (intuitively, we simulate some computation of $\mathcal{A}$ on $C$ under the assumption that the placeholder is assigned state $q$). In particular, we have $\delta^*_\bullet(q, \bullet) = \{q\}$. By an abuse of notation, we will denote $\delta^*$ and $\delta^*_\bullet$ simply by $\delta$.

## 4. THE BOUNDED REPAIR PROBLEM FOR TREES

We repair trees by using the standard set of edit operations over nodes [Tai 1979; Bille 2005]. We briefly recall the definitions of the standard edit operations on unranked trees, which are extensions of the edit operations over words. The first operation, called *deletion*, removes a distinguished (nonroot) node $x$ from a tree $t$ and promotes the subtrees $x$ as children of its parent. The second operation, called *insertion*, adds a new node $x$ in an unranked tree $t$, with a possible adoption of a list of consecutive children of the parent of $x$ whose original position immediately follows the position of $x$. Figure 3 provides an example of these two operations.

The last operation, called *relabeling*, modifies the label of a node $x$ to a new label in $\Sigma$. These three operations are the standard edit operations used to define the edit distance between trees (see Bille [2005] for a survey). We denote by $\mathsf{dist}(t, t')$ the minimum number of edits operations that are needed for transforming $t$ into $t'$ given two unranked trees $t$ and $t'$. Note that the operation of relabeling a node in an unranked tree, which is sometimes used as a standard edit operation, is subsumed by the insertion and deletion of nodes. Therefore, allowing or not the use of the operation of relabeling has an impact on the edit distance between two trees. However, the bounded repair problem is equivalent following that the relabeling is allowed or simulated by insertion and deletion operations.

We identify a close correspondence between the two basic edit operations on unranked trees and two operations on the corresponding curry encodings. Because the operations on unranked trees involve changing the parent of a sequence of consecutive subtrees (forests), the operations on curry encodings involve moving corresponding horizontal contexts.

Deleting an inner node $x$ in an unranked tree (Figure 4) corresponds to deleting the corresponding leaf node $x$ in the curry encoding, identifying the horizontal context $C$ that represents the sequence of children of the deleted node, and replacing the parent @ of $C$ by $C$. Note that the context $C$ is uniquely defined in the curry encoding and always has a parent @ because we apply the deletion operation to an inner node in the unranked tree.

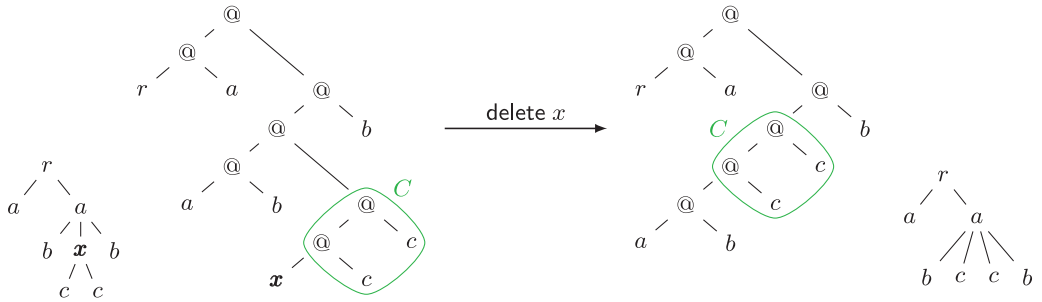



Fig. 4. Deleting a node in the curry encoding.

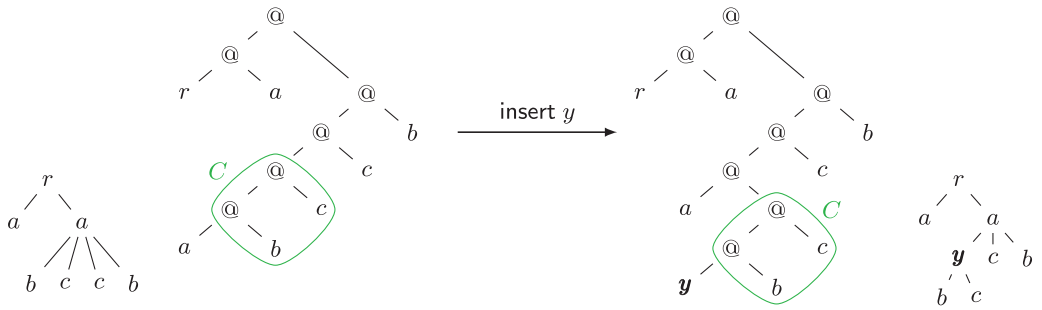



Fig. 5. Inserting a node in the curry encoding.

Conversely, inserting a node $y$ in an unranked tree (Figure 5) corresponds to identifying a context $C$ representing the forest of the adopted children, placing a new @ node in place of $C$ while attaching $C$ as the right child of the new node @, and substituting the placeholder of $C$ by the new node $y$. Here the context $C$ is also uniquely defined by the sequence of consecutive children adopted by the inserted node.

We are interested in studying the bounded repairability problem. This problem was studied for strings by Benedikt et al. [2013], so we extend their setting from strings to trees. We consider two finite alphabets $\Sigma$ and $\Delta$ and regular languages $S \subseteq \Sigma^*$ and $T \subseteq \Delta^*$, called the *source* and *target* languages, respectively. Furthermore, we define a *repair strategy* as any function from trees in $S$ to trees in $T$ for any tree languages $S$ and $T$.

We are now ready to introduce the problem in which we are mainly interested.

*Definition* 4.1. Given two regular tree languages $S$ and $T$, let

$$\mathsf{cost}(S, T) =^{\text{def}} \sup_{t \in S} \min_{t' \in T} \mathsf{dist}(t, t')$$

be the worst-case cost of repairing $S$ into $T$—note that this can be equally defined as the minimum of $\max_{t \in S} \mathsf{dist}(t, f(t))$ over all repair strategies $f$ from $S$ to $T$.

If $\mathsf{cost}(S, T)$ is finite, then we say that $S$ *is bounded repairable into* $T$, and we write $\mathsf{cost}(S, T) < \infty$ for short. Intuitively, this is equivalent to saying that there is a repair strategy $f$ transforming any tree $t \in S$ into a tree $f(t) \in T$ and having $\mathsf{dist}(t, f(t))$ uniformly bounded by a constant.

The bounded repair problem amounts to deciding, given two regular tree languages $S$ and $T$—specified by means of stepwise tree automata or DTDs—whether $S$ is bounded repairable into $T$.

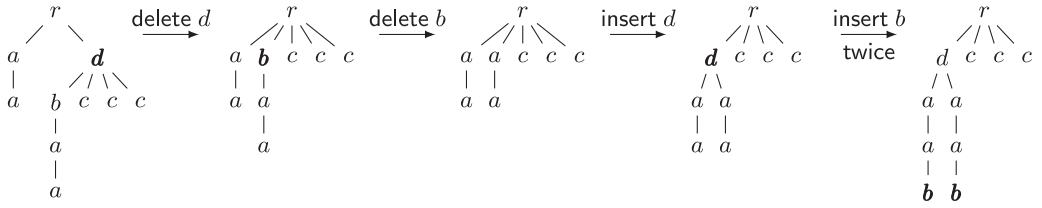



Fig. 6. Example of how to repair an unranked tree satisfying $D$ into one satisfying $D'$.

*Example* 3.1 (*continued*). Consider the two DTDs $D$ and $D'$ that we introduced in our running example. For the tree languages specified by $D$ and $D'$, one can check that $\mathscr{L}(D)$ is bounded repairable into $\mathscr{L}(D')$. Figure 6 shows how to repair a tree satisfying $D$ into one satisfying $D'$ with five edits: first, one removes the $d$-labeled node under the root and its $b$-labeled child; then, one adds a new $d$-labeled node above the two branches starting with $a$; and finally, one adds $b$-labeled leaves under the $r\,d\,a\,a$ branches. In fact, similar strategies with edit cost at most five can be used to repair any tree $t \in \mathscr{L}(D)$ into a tree $t' \in \mathscr{L}(D')$. In particular, this shows that $\mathsf{cost}(\mathscr{L}(D), \mathscr{L}(D')) < \infty$.

It is easy to verify that the bounded repairability relation $\mathsf{cost}(S, T) < \infty$ satisfies the following key properties, which shall be used later:

(1) Subset-subsumption—that is, $S \subseteq T$ implies $\mathsf{cost}(S, T) < \infty$;
(2) Transitivity—that is, $\mathsf{cost}(S, T) < \infty$ and $\mathsf{cost}(T, U) < \infty$ imply $\mathsf{cost}(S, U) < \infty$;
(3) Union-compatibility—that is, $\mathsf{cost}(S, T) < \infty$ and $\mathsf{cost}(S', T') < \infty$ imply $\mathsf{cost}(S \cup S', T \cup T') < \infty$.

The first property, subset-subsumption, is trivial to prove given that $\mathsf{cost}(S, T) = 0$ if and only if $S \subseteq T$. For the transitivity property, one can easily check that function $\mathsf{dist}$ is a metric over trees and then satisfies the triangle inequality (i.e., $\mathsf{dist}(t, t') \leq \mathsf{dist}(t, t'') + \mathsf{dist}(t'', t')$ for any $t, t', t'' \in \mathcal{T}_\Sigma$). This implies that $\mathsf{dist}(t, t'') \leq \mathsf{cost}(S, T) + \mathsf{cost}(T, U)$ given that $\mathsf{dist}(t, t') \leq \mathsf{cost}(S, T)$ and $\mathsf{dist}(t', t'') \leq \mathsf{cost}(T, U)$ for any $t \in S$, $t' \in T$, and $t'' \in U$. Thus, we conclude that $\mathsf{cost}(S, U)$ is also bounded and the transitivity property is proved. Finally, union-compatibility follows directly from the definition of worst-case cost of repairing a source into a target language. Indeed, if $\mathsf{cost}(S, T) < \infty$ and $\mathsf{cost}(S', T') < \infty$, then $\mathsf{cost}(S \cup S', T \cup T') \leq \max\{\mathsf{cost}(S, T), \mathsf{cost}(S', T')\}$, and we conclude that $\mathsf{cost}(S \cup S', T \cup T')$ is bounded as well.

## 5. CHARACTERIZATION OF BOUNDED REPAIRABILITY

In this section, we give an effective characterization of the bounded repairability relation between regular tree languages. Similarly to the string setting [Benedikt et al. 2013], this characterization is based on the notion of strongly connected component of the transition graph of a stepwise automaton. In the string case, a suitable coverability relation between chains of components is used to characterize bounded repairability. Because here we work with trees, we need to generalize the notion of coverability to a relation over the so-called synopsis trees—that is, full binary trees with nodes labeled by strongly connected components.

### 5.1. Components of Stepwise Automata

Given a stepwise automaton $\mathcal{A} = (\Sigma, Q, \delta, \delta_0, F)$, the *transition graph* of $\mathcal{A}$ is the graph $G_\mathcal{A} = (Q, E_h \cup E_v)$, where

$$E_h = \{(q_1, q) \in Q \times Q \mid \exists q_2.\ q \in \delta(q_1, q_2)\},$$
$$E_v = \{(q_2, q) \in Q \times Q \mid \exists q_1.\ q \in \delta(q_1, q_2)\}.$$

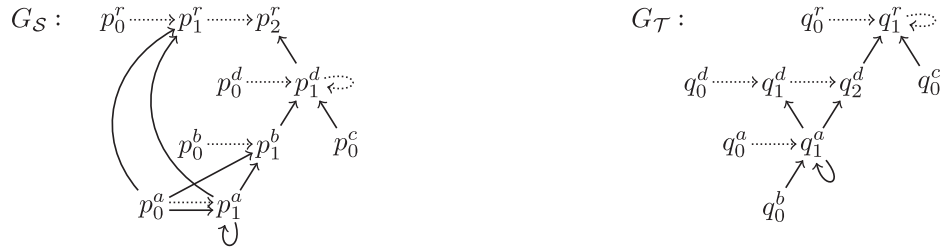


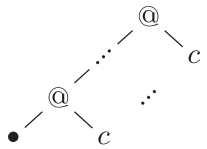

Fig. 7. Transition graphs (dashed and solid arrows represent horizontal and vertical edges, respectively).

We call the edges in $E_v$ *vertical* and the edges in $E_h$ *horizontal*. Note that an edge may be both vertical and horizontal. As an example, Figure 7 depicts the transition graphs of the automata $\mathcal{S}$ and $\mathcal{T}$ of Example 3.1 (dashed arrows represent horizontal edges, solid arrows represent vertical edges).

Recall that a *strongly connected component* of a graph (or simply a *component*) is a maximal set of nodes $X$ such that every two nodes $x, y$ in $X$ are connected by a direct path. By $\mathsf{SCC}(\mathcal{A})$, we denote the set of all strongly connected components in the transition graph of $\mathcal{A}$. In a way similar to the string setting, we associate with each component $X \in \mathsf{SCC}(\mathcal{A})$ the language $\mathscr{L}(\mathcal{A} \mid X)$ of contexts that are realizable within $X$:

$$\mathscr{L}(\mathcal{A} \mid X) = \{C \in \mathcal{C}_\Sigma \mid \exists p, q \in X.\, q \in \delta(p, C)\}.$$

For example, the contexts realizable within the component $\{p_1^d\}$ of the automaton of our running example (see also Figure 7) are all of the following form:




Because editing operations on unranked trees correspond to operations involving horizontal contexts in the curry encodings, we identify strongly connected components of an automaton that yield only horizontal contexts. A proper manipulation of those components translates to performing a fixed number of editing operations regardless of the contexts such components define, which is the basis of characterizing bounded repairability. Formally, a component $X \in \mathsf{SCC}(\mathcal{A})$ is *horizontal* if and only if $\mathscr{L}(\mathcal{A} \mid X)$ consists of horizontal contexts only. Similarly, we say that $X$ is *trivial* if and only if it realizes the empty context only—that is, $\mathscr{L}(\mathcal{A} \mid X) = \{\bullet\}$. Note that trivial components are horizontal.

As an example, consider again the transition graphs of Figure 7. All components except $\{p_1^d\}$, $\{p_1^a\}$, $\{q_1^r\}$, and $\{q_1^a\}$ are trivial. The components $\{p_1^d\}$ and $\{q_1^r\}$ are nontrivial horizontal, as they both realize the contexts $\bullet$, $\bullet@c$, $(\bullet@c)@c$, and so forth. The components $\{p_1^a\}$ and $\{q_1^a\}$ are nonhorizontal, as they both realize the contexts $\bullet$, $a@\bullet$, $a@(a@\bullet)$, and so forth.

### 5.2. Synopsis Trees

We now introduce a suitable structure that eases the characterization of bounded repairability, which we call the *synopsis tree*. The structure is a generalization of the chain of components that is used in Theorem 4.1 [Benedikt et al. 2013] to characterize bounded repairability between string languages. Formally, a *synopsis tree* of an automaton $\mathcal{A}$ is any full binary tree whose nodes are labeled with elements of $\mathsf{SCC}(\mathcal{A})$.

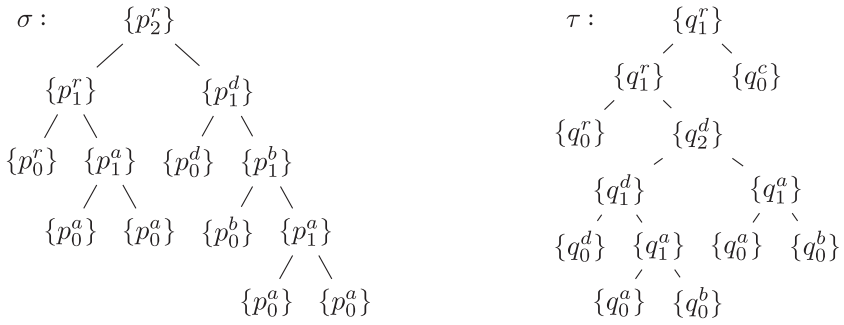



Fig. 8. Synopsis trees for stepwise tree automata $\mathcal{S}$ and $\mathcal{T}$.

The language $[\![\sigma]\!]_{\mathcal{A}}$ of curried trees that is induced by a synopsis tree $\sigma$ of $\mathcal{A}$ is defined recursively as follows:

$$[\![X]\!]_{\mathcal{A}} = \{C \circ a \mid C \in \mathscr{L}(\mathcal{A} \mid X),\ a \in \Sigma\},$$
$$[\![X(\sigma_1, \sigma_2)]\!]_{\mathcal{A}} = \{C \circ (t_1 @ t_2) \mid C \in \mathscr{L}(\mathcal{A} \mid X),\ t_1 \in [\![\sigma_1]\!]_{\mathcal{A}},\ t_2 \in [\![\sigma_2]\!]_{\mathcal{A}}\},$$

with $X \in \mathsf{SCC}(\mathcal{A})$. Intuitively, all trees in the language $[\![\sigma]\!]_{\mathcal{A}}$ are obtained by combining, in a suitable order that is compatible with the structure of $\sigma$, some contexts that are realizable within the components of $\sigma$. Figure 8 contains two synopsis trees $\sigma$ and $\tau$, respectively, for the source automaton $\mathcal{S}$ and the target automaton $\mathcal{T}$ of Example 3.1. An example of tree induced by the synopsis tree on the left is that of Figure 1.

Next we identify a family of synopsis trees that captures "closely enough" the language recognized by an automaton.

*Definition* 5.1. A *primitive synopsis tree* of an automaton $\mathcal{S} = (\Sigma, Q, \delta, \delta_0, F)$ is a synopsis tree $\sigma$ of $\mathcal{S}$ such that

(1) $\sigma$ respects the transition function of $\mathcal{S}$, namely for all nodes $x$, $x \cdot 1$, and $x \cdot 2$ in $\sigma$, there exist some states $q \in \sigma(x), q_1 \in \sigma(x \cdot 1)$, and $q_2 \in \sigma(x \cdot 2)$ such that $q \in \delta(q_1, q_2)$; and
(2) every internal node of $\sigma$ has label different from the labels of its children, namely for all nodes $x$, $x \cdot 1$, and $x \cdot 2$ in $\sigma$, $\sigma(x \cdot 1) \neq \sigma(x) \neq \sigma(x \cdot 2)$.

$\mathsf{PST}(\mathcal{S})$ denotes the set of all primitive synopsis trees of $\mathcal{S}$.

We observe that the second property stated in Definition 5.1 is equivalent to asking that every component appears at most once in every path of a primitive synopsis tree.

As an example, the tree $\sigma$ depicted to the left of Figure 8 is a primitive synopsis tree, and it corresponds to the run on the left-hand side of Figure 2 of the automaton $\mathcal{S}$ of Example 3.1. On the other hand, the synopsis tree $\tau$ depicted to the right is not primitive.

The idea underlying the notion of primitive synopsis tree is to capture the "cyclic behavior" of the components of the source automaton. This cyclic behavior has to be taken into account in the characterization of bounded repairability because it could generate arbitrary large fragments of trees that cannot be edited with uniformly bounded cost. Moreover, the use of primitive synopsis trees as a representation of the source language $\mathscr{L}(\mathcal{S})$ is *sound* in the sense that $\mathscr{L}(\mathcal{S})$ is contained in the union of the languages induced by primitive synopsis trees.

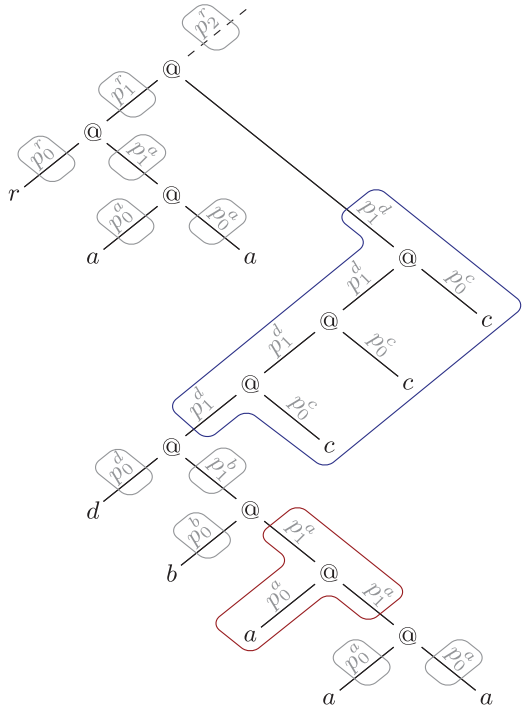



Fig. 9. Decomposition of a curry tree into contexts.

LEMMA 5.2. *For every stepwise tree automaton $\mathcal{S}$, we have $\mathscr{L}(\mathcal{S}) \subseteq \bigcup_{\sigma \in \mathsf{PST}(\mathcal{S})} [\![\sigma]\!]_{\mathcal{S}}$.*

Before entering the details of the proof of Lemma 5.2, we illustrate the main ideas on the example tree $t$ from Figure 1 accepted by the automaton $\mathcal{S}$ from Example 3.1. We use the accepting run of $\mathcal{S}$ on $t$ (see Figure 2), particularly the transitions in it that induce a change of component along both successors, to decompose $t$ into a binary tree structure, where each node represents a context realizable by some component of $\mathcal{S}$. We present this decomposition in Figure 9, where for a better visualization we annotate the states not on the nodes but on the edges above them (this requires adding a virtual edge entering the root).

PROOF OF LEMMA 5.2. Fix a curried tree $t \in \mathscr{L}(\mathcal{S})$ and an accepting run $\rho$ of $\mathcal{S}$ on $t$. We need to construct a primitive synopsis tree $\sigma$ such that $t \in [\![\sigma]\!]_{\mathcal{S}}$. Recall that a synopsis tree is a tree whose nodes are labeled with strongly connected components of $\mathcal{S}$. To construct $\sigma$, we first decompose $t$ into pieces: this will result in a tree-shaped arrangement of contexts, which we refer to as the context decomposition of $t$. We then show how to turn the context decomposition of $t$ into the desired primitive synopsis tree.

Formally, we represent a *context decomposition of* $t$ as a subset $D$ of the nodes of $t$ satisfying the following conditions:

(1) $D$ contains the root of $t$; and
(2) for every node $x$ in $D$, either $x$ is a leaf of $t$ or there is a descendant $y$ of $x$ in $t$ such that (i) the states $\rho(x)$ and $\rho(y)$ belong to the same strongly connected component of $\mathcal{S}$, (ii) both successors $y \cdot 1$ and $y \cdot 2$ belong to $D$, and (iii) for every node $z \in D$, if $z$ is a proper descendant of $x$, then $z$ is a proper descendant of $y$ as well.

Note that the node-to-child relation of $t$ induces an analogous structure on any set $D$ of the preceding form. In particular, we can think of a context decomposition $D$ as a full binary tree. We further associate with each internal node $x$ of $D$ the context $D(x)$ that is obtained by selecting the portion of the tree $t$ that lies between $x$ and the unique node $y$ such that $y \cdot 1$ and $y \cdot 2$ are children of $x$ in $D$—the node that corresponds to $y$ in the resulting context is labeled with the placeholder $\bullet$. The *flattening* $[\![D]\!]$ of a context decomposition $D$ is the language of trees that is inductively defined as follows:

$$[\![D]\!] = \Sigma \qquad \text{if } D \text{ has a single node,}$$
$$[\![D]\!] = \{C_0(t_1@t_2) \mid t_1 \in [\![D_1]\!],\ t_2 \in [\![D_2]\!]\} \qquad \text{if } D \text{ has more than one node,}$$

where, in the second line, $C_0$ is the context associated with the root of $D$ and $D_1$ and $D_2$ are the subtrees of $D$ ($D_1$ and $D_2$ can be seen as context decompositions of two disjoints subtrees of $t$). It is easy to see that $t \in [\![D]\!]$ for every context decomposition $D$ of $t$. We also associate with each decomposition $D$ of $t$ an induced synopsis tree $\sigma_D$ by replacing every context $C$ labeling an internal node $x$ of $D$ with the strongly connected component $X$ of the state $\rho(x)$ (note that $C \in \mathscr{L}(\mathcal{S} \mid X)$).

Due to the similarity between the definition of the flattening $[\![D]\!]$ and the definition of the language $[\![\sigma_D]\!]_{\mathcal{S}}$ induced by the synopsis tree $\sigma_D$, we have that $[\![D]\!] \subseteq [\![\sigma_D]\!]_{\mathcal{S}}$, from which $t \in [\![\sigma_D]\!]$. However, given a generic context decomposition $D$, there is no guarantee that $\sigma_D$ is a *primitive* synopsis tree—in particular, it may happen that two consecutive nodes in $\sigma_D$ are labeled with the same component. To overcome this problem, next we show how to construct a specific context decomposition $D$ of $t$ that satisfies the following additional property:

(3) If $x$ is an internal node of $D$ and $y \cdot 1$ and $y \cdot 2$ are the two immediate successors of $x$ in $D$, then the component of $\rho(x)$ is different from the components of $\rho(y \cdot 1)$ and $\rho(y \cdot 2)$.

Clearly, the additional property suffices to conclude that $\sigma_D$ is a primitive synopsis tree and to thus prove the lemma.

To construct a context decomposition $D$ of $t$ that satisfies properties (1) through (3), we follow maximal paths within the same component in the run $\rho$. More precisely, let $x$ be the root of $t$, and let $X$ be the component of $\rho(x)$. We distinguish two cases depending on whether or not there is a leaf $y$ of $\rho$ whose state belongs to the component $X$. If there is such a leaf $y$, then we define the context decomposition $D$ of $t$ to be the set containing only the two nodes $x$ and $y$. In this case, the corresponding synopsis tree $\sigma_D$ is clearly primitive. Otherwise, if all states associated with the leaves of $\rho$ are outside $X$, we choose any maximal path of $\rho$ that starts in $x$ and visits only states within the component $X$. Let $y$ be the last node of this path. Clearly, $y$ is not a leaf, and both states $\rho(y \cdot 1)$ and $\rho(y \cdot 2)$ are outside $X$. By exploiting a simple inductive argument, we can assume that the two subtrees of $t$ rooted at nodes $y \cdot 1$ and $y \cdot 2$ admit some context decompositions $D_1$ and $D_2$ satifsying (1) through (3). We can thus define our context decomposition $D$ of $t$ to be the set $\{x\} \cup D_1 \cup D_2$. It is routine to verify that $D$ satisfies properties (1) through (3) and induces a primitive synopsis tree $\sigma_D$.

Summing up, we constructed from $t$ and $\rho$ a suitable context decomposition $D$, and from this we derived the existence of a primitive synopsis tree $\sigma_D$ such that $t \in [\![D]\!] \subseteq [\![\sigma_D]\!]_{\mathcal{S}}$. This concludes the proof of the lemma. □

We also observe the following remark.

*Remark* 5.3. The height of a primitive synopsis tree of the source automaton $\mathcal{S}$ is bounded by the number of components in $G_{\mathcal{S}}$ and hence by the number of states of $\mathcal{S}$.

Consequently, $\mathsf{PST}(\mathcal{S})$ is a finite set and can be represented in exponential space with respect to the size of $\mathcal{S}$.

To represent the target language and the possible edited trees, one needs a relaxed version of primitive synopsis tree, called the *basic synopsis tree*, which enforces only the first condition of Definition 5.1. Basic synopsis trees are the analogs of chains of components over $\mathrm{dag}^*(\mathcal{T})$ that were used in Theorem 4.1 [Benedikt et al. 2013] to characterize bounded repairability between string languages.

*Definition* 5.4. A *basic synopsis tree* of an automaton $\mathcal{T}$ is a synopsis tree $\tau$ of $\mathcal{T}$ that respects the transition function of $\mathcal{T}$ (compare to first item of Definition 5.1). We denote by $\mathsf{BST}(\mathcal{T})$ the set of all basic synopsis trees of $\mathcal{T}$.

For example, the tree $\tau$ in Figure 8 is a basic synopsis tree that respects the transitions of the run of the automaton $\mathcal{T}$ depicted in the right-hand side of Figure 2.

Differently from primitive synopsis trees, basic synopsis trees may contain repeated occurrences of the same component. This implies that the set $\mathsf{BST}(\mathcal{T})$ of all basic synopsis trees of $\mathcal{T}$ is potentially infinite. However, this set can be finitely presented by means of a deterministic binary bottom-up tree automaton of size polynomial in the size of $\mathcal{T}$.

The following lemma shows that the language induced by a basic synopsis tree of $\mathcal{T}$ is bounded repairable into the language $\mathscr{L}(\mathcal{T})$.

LEMMA 5.5. *For every stepwise tree automaton* $\mathcal{T} = (\Delta, Q, \delta, \delta_0, F)$ *and every* $\tau \in \mathsf{BST}(\mathcal{T})$*, we have*

$$\mathsf{cost}\big([\![\tau]\!]_{\mathcal{T}}, \mathscr{L}(\mathcal{T})\big) \leq (4|\tau| + 1) \cdot 2^{|Q|},$$

*where* $[\![\tau]\!]_{\mathcal{T}}$ *is now seen as a language of unranked trees. In particular,* $\mathsf{cost}\big([\![\tau]\!]_{\mathcal{T}}, \mathscr{L}(\mathcal{T})\big) < \infty$.

PROOF. We begin by explaining the main ingredients of the proof. Any tree $t \in [\![\tau]\!]_{\mathcal{T}}$ can be seen as a composition of contexts, precisely a context $C_x$ for each node $x$ in $\tau$, with $C_x \in \mathscr{L}(\mathcal{T} \mid \tau(x))$. Every such context can be decorated by a partial run of $\mathcal{T}$ that justifies the fact that the context belongs to the language $\mathscr{L}(\mathcal{T} \mid \tau(x))$. These partial runs can be used to construct a complete run of $\mathcal{T}$, but only after the insertion of a small number of small pieces of runs, which provide the necessary connections between the partial run of a context $C_x$ and the partial runs of the successor contexts $C_{x\cdot 1}$ and $C_{x\cdot 2}$. The existence of these pieces is guaranteed by the fact that the synopsis tree $\tau$ respects the transition function of $\mathcal{T}$ and by the definition of strongly connected component. We now turn to a more detailed proof.

For every state $q$ of $\mathcal{T}$, we define the automaton $\mathcal{T}^q = (\Delta, Q, \delta, \delta_0, \{q\})$ that recognizes trees via runs that end with state $q$ at the root. We also define $\ell_{p,q} = \min\{|C| \mid q \in \delta(p, C)\}$ for each pair of states $p, q$, with $q$ reachable from $p$ in $G_{\mathcal{T}}$. We take the maximum of all these values (i.e., $\ell = \max_{p,q \in Q} \ell_{p,q}$), and we observe that $\ell \leq 2^{|Q|}$. We can associate with each pair $p, q \in Q$ such that $q$ is reachable from $p$ a context $C_{p,q}$ of size at most $\ell$ such that $q \in \delta(p, C_{p,q})$. Next we exploit an induction on the size of a basic synopsis tree $\tau$ to prove that for all states $q \in \tau(\varepsilon)$,

$$\mathsf{cost}\big([\![\tau]\!]_{\mathcal{T}}, \mathscr{L}(\mathcal{T}^q)\big) \leq 4|\tau| \cdot \ell.$$

Note that to get from the preceding statement to the claim of the lemma, it is sufficient to recall that $\mathcal{T}$ is trimmed, and hence any tree in $\mathscr{L}(\mathcal{T}^q)$ can be repaired into $\mathscr{L}(\mathcal{T})$ by simply inserting a context $C_{q,q'}$, where $q'$ is some state from $F$.

To prove the statement in the base case $\tau = X$, we consider a generic tree $t \in [\![X]\!]_{\mathcal{A}}$, and we observe that $t = C \circ a$ for some context $C \in \mathscr{L}(\mathcal{T} \mid X)$ and some letter $a \in \Delta$.

Clearly, there exist $p', q' \in X$ such that $q' \in \delta(p', C)$. Since $\mathcal{T}$ is trimmed, there exist a symbol $b \in \Delta$ and a state $q_0 \in \delta_0(b)$ such that $p'$ is reachable from $q_0$. We define the tree

$$t' = C_{q',q} \circ C \circ C_{q_0,p'} \circ b$$

and observe that $t'$ can be obtained from $t$ ($= C \circ a$) by replacing the leaf node $a$ with the tree $C_{q_0,p'} \circ b$ and by adding the context $C_{q',q}$ at the top. Clearly, $t' \in \mathscr{L}(\mathcal{T}^q)$ and the overall cost of transforming $t$ to $t'$ is at most $4\ell$.

Now, for the inductive step, suppose that $\tau = X(\tau_1, \tau_2)$ and consider $t \in [\![X(\tau_1, \tau_2)]\!]_{\mathcal{T}}$. By definition, we have $t = C \circ (t_1 @ t_2)$ for some context $C \in \mathscr{L}(\mathcal{T} \mid X)$ and some trees $t_1 \in [\![\tau_1]\!]_{\mathcal{T}}$ and $t_2 \in [\![\tau_2]\!]_{\mathcal{T}}$. Further, let $p', q' \in X$ be two states such that $q' \in \delta(p', C)$. Since the synopsis tree $\tau$ respects the transition function of $\mathcal{T}$, we know that there exist $p \in X$, $p_1 \in \tau_1(\varepsilon)$, and $p_2 \in \tau_2(\varepsilon)$ such that $p \in \delta(p_1, p_2)$. By inductive hypothesis, there exists $t'_1 \in \mathscr{L}(\mathcal{T}^{p_1})$ (respectively, $t'_2 \in \mathscr{L}(\mathcal{T}^{p_2})$) such that $\mathsf{dist}(t_1, t'_1) \leq 4|\tau_1| \cdot \ell$ (respectively, $\mathsf{dist}(t_1, t'_1) \leq 4|\tau_1| \cdot \ell$). We can then define

$$t' = C_{q',q} \circ C \circ C_{p,p'} \circ (t'_1 @ t'_2)$$

and claim that $t' \in \mathscr{L}(()\mathcal{T}^q)$. Moreover, the preceding tree $t'$ can be obtained from the original tree $t = C \circ (t_1 @ t_2)$ by first transforming the subtrees $t_1$ and $t_2$ into $t'_1$ and $t'_2$, respectively, then inserting the context $C_{p,p'}$ between $t'_1 @ t'_2$ and $C$, and finally adding the context $C_{q',q}$ at the top. Overall, this transformation costs at most $4|\tau_1| \cdot \ell + 4|\tau_2| \cdot \ell + 2\ell \leq 4|\tau| \cdot \ell$. □

### 5.3. Coverings

In the previous section, we introduced the concepts of primitive and basic synopsis trees, and we showed that they correspond roughly (i.e., up to boundedly many edits) to trees accepted by the source and target automata, respectively. The remaining part of the puzzle is to relate each primitive synopsis tree of the source automaton $\mathcal{S}$ to some basic synopsis tree of the target automaton $\mathcal{T}$, so as to characterize bounded repairability from $\mathscr{L}(\mathcal{S})$ to $\mathscr{L}(\mathcal{T})$. This is accomplished by the notion of covering between synopsis trees.

*Definition* 5.6. Given two stepwise tree automata $\mathcal{S}$ and $\mathcal{T}$ and two synopsis trees $\sigma$ of $\mathcal{S}$ and $\tau$ of $\mathcal{T}$, we say that $\sigma$ is *covered* by $\tau$ if and only if there is an *injective* mapping $\lambda$ from nontrivial nodes of $\sigma$ to nontrivial nodes of $\tau$ such that

(1) $\lambda$ maps components in a way that is compatible with the languages of contexts—that is, $\mathscr{L}(\mathcal{S} \mid \sigma(x)) \subseteq \mathscr{L}(\mathcal{T} \mid \tau(\lambda(x)))$ for every nontrivial node $x$ of $\sigma$;
(2) $\lambda$ preserves the post-order of nontrivial nodes—that is, $x \preccurlyeq_\sigma^{\mathsf{post}} y$ if and only if $\lambda(x) \preccurlyeq_\tau^{\mathsf{post}} \lambda(y)$ for any two nontrivial nodes $x, y$ of $\sigma$; and
(3) $\lambda$ preserves the ancestorship of nonhorizontal nodes—that is, $x \preccurlyeq_\sigma^{\mathsf{anc}} y$ if and only if $\lambda(x) \preccurlyeq_\tau^{\mathsf{anc}} \lambda(y)$ for every nonhorizontal node $x$ of $\sigma$ and every nontrivial node $y$ of $\sigma$.

We call the mapping $\lambda$ a *covering* from $\sigma$ to $\tau$, and we denote it shortly by $\lambda : \sigma \hookrightarrow \tau$.

Figure 10 presents a covering of a primitive synopsis tree $\sigma$ of $\mathcal{S}$ by a basic synopsis tree $\tau$ of $\mathcal{T}$, where the square boxes represent the nontrivial nodes and have double borders when the component is nonhorizontal.

We are now able to state the main characterization theorem of the article.

THEOREM 5.7. *Given two stepwise automata $\mathcal{S}$ and $\mathcal{T}$, the language $\mathscr{L}(\mathcal{S})$ is bounded repairable into the language $\mathscr{L}(\mathcal{T})$ if and only if every primitive synopsis tree $\sigma$ of $\mathcal{S}$ is covered by some basic synopsis tree $\tau$ of $\mathcal{T}$, namely*

$$\mathsf{cost}(\mathcal{S}, \mathcal{T}) < \infty \quad \textit{if and only if} \quad \forall \tau \in \mathsf{PST}(\mathcal{S}).\ \exists \sigma \in \mathsf{BST}(\mathcal{T}).\ \ \tau \hookrightarrow \sigma.$$

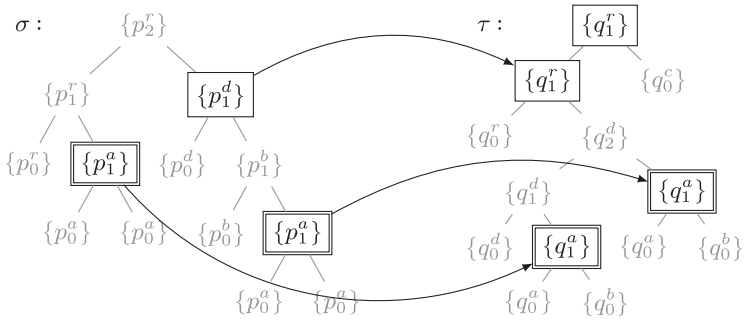



Fig. 10. Covering of a primitive synopsis tree by a basic synopsis tree.

The proof of the preceding result is given in Section 6. Here we briefly explain the main ideas underlying the definition of covering. We begin by observing that a reasonable strategy for repairing $\mathcal{S}$ into $\mathcal{T}$ with uniformly bounded cost applies the edit operations only at the "junctions" of the contexts realized by the nontrivial components. Indeed, since nontrivial components of $\mathcal{S}$ can realize arbitrary large repetitions of the same context, we have that either these repetitions do not need any editing at all or they need an arbitrary large amount of editing. This observation gives an intuitive account for the first condition of Definition 5.6, which enforces containment relationships between languages of contexts realizable within nontrivial components.

As for the other two conditions, it is worth looking at the effect of an edit operation on the curry encoding of an unranked tree $t \in \mathscr{L}(\mathcal{S})$. Let us consider a node $x$ in $t$ that is about to be deleted by the editing. There is a unique way to represent the curry encoding of $t$ together with the distinguished node $x$ as an expression of the form $C \circ (t' @ (C' \circ a))$, where $a$ is the label of $x$ and $C'$ is a *horizontal* context representing the forest of subtrees under $x$. The result of the deletion of node $x$ from $t$ is encoded by the curried tree $C \circ C' \circ t'$ (see Figure 4 for an example). Note that this operation does not allow the deletion of the left-most leaf node in the curried tree (this would correspond to deleting the root node in an unranked tree, an operation that is typically prohibited). The operation of inserting a new node $y$ in an unranked tree $t$ can be described in a symmetric way via curry encodings and transpositions of horizontal contexts—that is, given an unranked tree $t$ with curry encoding $C \circ C' \circ t'$, where $C'$ is a horizontal context, the curried tree $C \circ (t' @ (C' \circ a))$ represents the unranked tree that results from the insertion of a new $a$-labeled node $y$ in $t$ having as children the forest represented by $C'$.

We now observe that the transformations on curried trees that we just described satisfy two crucial properties: (i) they preserve the post-order of the nodes, and (ii) they preserve the ancestorship of nonhorizontal contexts (e.g., the context $C$ of Figure 4) with their descendants. These properties are precisely captured by the last two conditions of Definition 5.6.

We conclude the section by mentioning a strengthening of the "if" direction of Theorem 5.7, which gives an upper bound for the cost of an optimal repair strategy from $\mathscr{L}(\mathcal{S})$ to $\mathscr{L}(\mathcal{T})$.

PROPOSITION 5.8. *For all automata $\mathcal{S}$ and $\mathcal{T}$, if every primitive synopsis tree of $\mathcal{S}$ is covered by some basic synopsis tree of $\mathcal{T}$, then*

$$\mathsf{cost}(\mathscr{L}(\mathcal{S}), \mathscr{L}(\mathcal{T})) \in \mathcal{O}(|\mathsf{SCC}(\mathcal{T})| \cdot 2^{|Q|+|Q'|}),$$

*where $Q$ and $Q'$ are the set of states of $\mathcal{S}$ and $\mathcal{T}$, respectively.*

## 6. PROOF OF THE MAIN CHARACTERIZATION

The following sections are devoted to proving the two directions of the characterization.

### 6.1. From Covering to Repair

We begin with the proof of the "if" direction of Theorem 5.7. For the rest of the section, we fix two stepwise automata $\mathcal{S} = (\Sigma, Q, \delta, \delta_0, F)$ and $\mathcal{T} = (\Delta, Q', \delta', \delta'_0, F')$ recognizing the source and the target languages, respectively. We then assume that every primitive synopsis tree of $\mathcal{S}$ is covered by some basic synopsis tree of $\mathcal{T}$, and we show how to construct a repair strategy from $\mathscr{L}(\mathcal{S})$ to $\mathscr{L}(\mathcal{T})$ with uniformly bounded cost. The proof basically follows from a series of containments and bounded repairability relations between languages, which can be summarized as follows:

$$\underbrace{\mathscr{L}(\mathcal{S}) \subseteq \bigcup_{\sigma \in \mathsf{PST}(\mathcal{S})} [\![\sigma]\!]_{\mathcal{S}}}_{\text{Lemma 5.2}} \sqsubseteq \underbrace{\bigcup_{\tau \in \mathsf{BST}(\mathcal{T})} [\![\tau]\!]_{\mathcal{T}} \subseteq \mathscr{L}(\mathcal{T})}_{\text{Lemma 5.5}},$$

where $\sqsubseteq$ denotes the bounded repairability relation. The intermediate bounded repairability relation in the preceding chain is established by the following lemma.

LEMMA 6.1. *For every synopsis tree $\sigma$ of $\mathcal{S}$ and every synopsis tree $\tau$ of $\mathcal{T}$, if $\sigma$ is covered by $\tau$, then* $\mathsf{cost}\big([\![\tau]\!]_{\mathcal{S}}, [\![\sigma]\!]_{\mathcal{T}}\big) \leq 4|\tau| + 4|\sigma|$.

The proof of the preceding lemma is quite technical and will take the entire section.

Before entering the details, we briefly discuss how the "if" direction of Theorem 5.7 follows from it. By Lemma 5.2, the source language $\mathscr{L}(\mathcal{S})$ is contained in the union of the languages $[\![\sigma]\!]_{\mathcal{S}}$ induced by all primitive synopsis trees $\sigma \in \mathsf{PST}(\mathcal{S})$. By the hypothesis, each of these synopsis trees is covered by some basic synopsis tree $\tau$ of the target automaton $\mathcal{T}$. Thus, by Lemma 6.1, each language $[\![\sigma]\!]_{\mathcal{S}}$ can be repaired with uniformly bounded cost into the language $[\![\tau]\!]_{\mathcal{T}}$ for some $\tau \in \mathsf{BST}(\mathcal{T})$. By Lemma 5.5, each language $[\![\tau]\!]_{\mathcal{T}}$ can in turn be repaired with uniformly bounded cost into the target language $\mathscr{L}(\mathcal{T})$. The result now follows from the fact that there are only finitely many primitive synopsis trees and the fact that bounded repairability is a transitive relation that is moreover preserved by finite unions.

The rest of the section is devoted to the proof of Lemma 6.1. We begin by extending slightly the definition of synopsis tree and by allowing the use of special nodes labeled with $\varepsilon$ that represents *dummy* trivial components. The semantics is extended in the natural way by letting $\mathscr{L}(\mathcal{A} \mid \varepsilon) = \{\bullet\}$ (for any stepwise automaton $\mathcal{A}$). Because all trivial components have the same associated language $\{\bullet\}$, we shall often identify trivial components of automata with the dummy component $\varepsilon$.

For a technical reason (see the proof of Lemma 6.3 later), we also need to assume that the alphabet $\Sigma$ of the source automaton $\mathcal{S}$ is contained in the alphabet $\Delta$ of the target automaton $\mathcal{T}$. Note that this condition can be enforced without loss of generality—that is, without changing the recognized languages.

The first ingredient of the proof shows how to "interpolate" two synopsis trees $\sigma$ and $\tau$ by a third synopsis tree $\theta$ of $\mathcal{S}$ in such a way that

—$\theta$ has the same labels (i.e., components) as $\sigma$ on the nontrivial nodes, and it covers $\sigma$ via a *bijection* between nontrivial nodes of $\sigma$ and nontrivial nodes of $\theta$ that maps any nontrivial node of $\sigma$ with label $X \in \mathsf{SCC}(\mathcal{S})$ to a nontrivial node of $\theta$ with the same label $X$ (we say that $\sigma$ is *strongly covered* by $\theta$ and denote this by $\sigma \hookrightarrow\!\!\!\!\rightarrow \theta$); and
—$\theta$ has the same domain (i.e., set of nodes) as $\tau$, and it is covered by $\tau$ via the *identity* function between nontrivial nodes (we say that $\theta$ is *embedded* into $\tau$ and denote this by $\theta \leftrightharpoons \tau$).

It is not difficult to show that such an interpolating synopsis tree $\theta$ exists.

LEMMA 6.2. *We have that*

$$\sigma \hookrightarrow \tau \qquad \textit{implies} \qquad \exists \theta.\ \sigma \hookrightarrow \theta \leftrightharpoons \tau.$$

*In other words, if $\sigma$ is covered by $\tau$, then there is a synopsis tree $\theta$ that strongly covers $\sigma$ and that is embedded in $\tau$.*

PROOF. Let $\lambda$ be a covering function from $\sigma$ to $\tau$ and recall that $\lambda$ is injective. Let $\mathsf{range}(\lambda)$ denote the set of nodes of $\tau$ of the form $\lambda(x)$ for some $x \in \mathsf{nodes}(\sigma)$. Moreover, for every $y \in \mathsf{range}(\lambda)$, let $\lambda^{-1}(y)$ denote the unique node $x$ of $\sigma$ such that $\lambda(x) = y$. The synopsis tree $\theta$ has the same domain as $\tau$ and the same labels as $\sigma$—that is, for all $y \in \mathsf{nodes}(\theta) = \mathsf{nodes}(\tau)$,

$$\theta(y) = \begin{cases} \sigma(\lambda^{-1}(y)) & \text{if } y \in \mathsf{range}(\lambda), \\ \varepsilon & \text{otherwise.} \end{cases}$$

Note that $\lambda$ can be seen as a bijection between the nontrivial nodes of $\sigma$ and the nontrivial nodes of $\theta$ (the latter are precisely the nodes of $\tau$ that belong to $\mathsf{range}(\lambda)$). It follows that $\sigma$ is strongly covered by $\theta$ via the function $\lambda$ and that $\theta$ is embedded into $\tau$ via the identity function. □

A first advantage of considering an interpolating synopsis tree $\theta$ that is embedded into $\tau$ is that $\theta$ and $\tau$ have the same structure. As a consequence, we can claim that the language induced by $\theta$ is *contained* in (not just bounded repairable into) the language induced by $\tau$.

LEMMA 6.3. *If $\theta$ is a synopsis tree of $\mathcal{S}$, $\tau$ is a synopsis tree of $\mathcal{T}$, and $\theta$ is embedded into $\tau$, then $[\![\theta]\!]_{\mathcal{S}} \subseteq [\![\tau]\!]_{\mathcal{T}}$.*

PROOF. The proof is by structural induction on $\theta$ (or, equally, $\tau$). In the base case, where $\theta$ consists of a single node $x$, we consider the components $\theta(x) = X$ and $\tau(x) = Y$. If $X$ is a trivial component, then since the covering function from $\theta$ to $\tau$ is a bijection between nontrivial nodes, we deduce that $Y$ is also a trivial component, and hence $\mathscr{L}(\mathcal{S} \mid X) = \mathscr{L}(\mathcal{T} \mid Y) = \{\bullet\}$. Now, recall that we assumed that the source alphabet $\Sigma$ is contained in the target alphabet $\Delta$. From this, it follows that $[\![\theta]\!]_{\mathcal{S}} = \Sigma \subseteq \Delta = [\![\tau]\!]_{\mathcal{T}}$. Otherwise, if $X$ is a nontrivial component, then from the fact that $\theta$ is covered by $\tau$, we obtain $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y)$. As before, we conclude that $[\![\theta]\!]_{\mathcal{S}} \subseteq [\![\tau]\!]_{\mathcal{T}}$.

For the inductive step, we suppose that $\theta = X(\theta_1, \theta_2)$ and $\tau = Y(\tau_1, \tau_2)$. We consider a generic tree $t \in [\![\theta]\!]_{\mathcal{S}}$. By definition, we can write $t = C \circ (t_1 @ t_2)$ for some context $C \in \mathscr{L}(\mathcal{S} \mid X)$ and some trees $t_1 \in [\![\theta_1]\!]_{\mathcal{S}}$ and $t_2 \in [\![\theta_2]\!]_{\mathcal{S}}$. We know from the inductive hypothesis that $t_i \in [\![\tau_i]\!]_{\mathcal{T}}$ for both $i = 1$ and $i = 2$. If $X$ is a trivial component, then $C$ is necessarily the trivial context, which is also realizable within the component $Y$. Otherwise, if $X$ is a nontrivial component, then it must be mapped to the component $Y$ by the embedding function, and hence $C \in \mathscr{L}(\mathcal{T} \mid Y)$. In both cases, we conclude that $t \in [\![\tau]\!]_{\mathcal{T}}$. □

Now recall from Section 4 that the bounded repairability relation is transitive and it generalizes containment. In particular, the previous results reduce the statement of Lemma 6.1 to the problem of proving that $[\![\sigma]\!]_{\mathcal{S}}$ is bounded repairable into $[\![\theta]\!]_{\mathcal{S}}$. In proving the latter statement, we can take advantage of the fact that $\sigma$ is strongly covered by $\theta$. In particular, we observe that the strong coverability relation $\hookrightarrow$ is an *equivalence*: it is indeed reflexive, symmetric, and transitive (the last two properties follow from the fact that the function that witnesses strong coverability is a bijection between nontrivial nodes that preserves components).

To derive bounded repairability from strong coverability, we associate with each synopsis tree $\sigma$ of $\mathcal{S}$ a suitable *normal form* $\sigma^*$ that can be used as a *canonical representative*

of the equivalence class of $\sigma$ induced by the strong coverability relation. We will then prove that $[\![\sigma]\!]_{\mathcal{S}}$ is bounded repairable into $[\![\theta]\!]_{\mathcal{S}}$ first by repairing $[\![\sigma]\!]_{\mathcal{S}}$ into $[\![\sigma^*]\!]_{\mathcal{S}}$ and then by repairing $[\![\theta^*]\!]_{\mathcal{S}}$ $(= [\![\sigma^*]\!]_{\mathcal{S}})$ into $[\![\theta]\!]_{\mathcal{S}}$ (recall that $\sigma$ and $\theta$ strongly cover each other, and hence $\sigma^* = \theta^*$ by canonicity of the normal form). The repair strategy that witnesses bounded repairability between $[\![\sigma]\!]_{\mathcal{S}}$ and $[\![\sigma^*]\!]_{\mathcal{S}}$ (respectively, $[\![\theta^*]\!]_{\mathcal{S}}$ and $[\![\theta]\!]_{\mathcal{S}}$) can be read off the sequence of generic editing operations that takes $\sigma$ to its normal form $\sigma^*$ (respectively, $\theta$ to its normal form $\theta^*$).

In the sequel, we only manipulate synopsis trees of the source automaton $\mathcal{S}$. For this reason, we can omit the subscript $\mathcal{S}$ from notations like $[\![\sigma]\!]_{\mathcal{S}}$. Next we describe the structure of a synopsis tree in normal form.

*Definition* 6.4. A synopsis tree $\sigma$ is *in normal form* if one of the following cases holds:

(1) $\sigma = \varepsilon$, namely $\sigma$ consists of a single node labeled with a trivial component;
(2) $\sigma = X(\alpha, \varepsilon)$, where $X$ is a nontrivial horizontal component and $\alpha$ is a synopsis tree in normal form; and
(3) $\sigma = \varepsilon(\alpha, X(\beta, \varepsilon))$, where $X$ is a nonhorizontal component and $\alpha, \beta$ are synopsis trees in normal form.

We observe that the root of a synopsis tree in normal form is a horizontal (possibly trivial) node and its left subtree is also in normal form. In particular, this means that all components along the left-most branch of a synopsis tree in normal form are horizontal.

The following lemma shows that synopsis trees in normal form can be used as canonical representatives of the equivalence classes induced by the strong coverability relation.

LEMMA 6.5. *If $\sigma$ and $\sigma'$ are two synopsis trees in normal form that strongly cover each other, then $\sigma$ and $\sigma'$ are isomorphic.*

PROOF. Let $\sigma$ and $\sigma'$ be two synopsis trees in normal form, and let $\lambda$ be a bijection between the nontrivial nodes of $\sigma$ and the nontrivial nodes of $\sigma'$ that witnesses the fact that $\sigma \hookrightarrow\!\!\!\!\rightarrow \sigma'$. In the following, we often identify, for the sake of simplicity, the nodes of the synopsis trees $\sigma$ and $\sigma'$ with their labels. The proof is by structural induction and case analysis.

For the base case, suppose that $\sigma = \varepsilon$. Since $\sigma$ contains only trivial nodes and $\lambda$ is surjective over nontrivial nodes, $\sigma'$ contains only trivial nodes as well. Since $\sigma'$ is in normal form, it follows that $\sigma' = \varepsilon$.

For the inductive step, we distinguish two cases depending on whether $\sigma$ is of the form $X(\sigma_1, \varepsilon)$, where $X$ is a nontrivial horizontal component, or of the form $\varepsilon(\sigma_1, X(\sigma_2, \varepsilon))$, where $X$ is a nonhorizontal component.

In the former case (i.e., $\sigma = X(\sigma_1, \varepsilon)$), we recall that the mapping $\lambda$ is a bijection between nontrivial nodes that preserves the post-order. Because the root $X$ of $\sigma$ is nontrivial and is the maximal element with respect to the post-order relation, it must be mapped by $\lambda$ to the root $Y$ of $\tau$. Moreover, since $\lambda$ preserves the labels of nontrivial nodes, we have that $Y = \lambda(X) = X$. In particular, $Y$ is a nontrivial horizontal component. Since $\sigma'$ is in normal form, it follows that its right subtree is $\varepsilon$, and hence $\lambda$ maps the nontrivial nodes of the left subtree $\sigma_1$ of $\sigma$ to the nontrivial nodes of the left subtree $\sigma'_1$ of $\sigma'$. Finally, since both $\sigma_1$ and $\sigma'_1$ are synopsis trees in normal form, we obtain from the inductive hypothesis that $\sigma_1 = \sigma'_1$, and hence $\sigma = X(\sigma_1, \varepsilon) = Y(\sigma'_1, \varepsilon) = \sigma'$.

We consider the second case (i.e., $\sigma = \varepsilon(\sigma_1, X(\sigma_2, \varepsilon))$), where $X$ is a nonhorizontal component. For the sake of contradiction, suppose that the root $Y$ of $\sigma'$ is a nontrivial component. Since $Y$ is the last nontrivial node in the post-order traversal of $\sigma'$ and $\lambda$ preserves the post-order of nontrivial nodes, the pre-image $\lambda^{-1}(Y)$ in $\sigma$ would consist

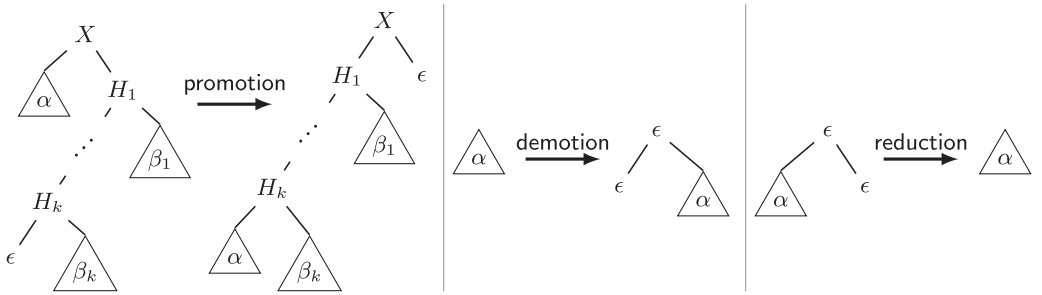



Fig. 11. Editing operations on synopsis trees.

of the last nontrivial node in the post-order traversal of $\sigma'$, from which $\lambda^{-1}(Y) = X$. However, since $X$ is a nonhorizontal component, this would be against the hypothesis that $\sigma'$ is in normal form (recall that the root of any synopsis tree in normal form is always a horizontal component). Knowing that $Y$ is a trivial node and $\sigma'$ is in normal form, we obtain $Y = \varepsilon$, and hence $\sigma'$ is of the form $\varepsilon(\sigma'_1, Z(\sigma'_2, \varepsilon))$ for some nonhorizontal component $Z$ and two synopsis trees $\sigma'_1$ and $\sigma'_2$ in normal form. Toward a conclusion, observe that in the post-order traversal of $\sigma$, the nontrivial nodes of $\sigma_1$ precede the nontrivial nodes of $\sigma_2$, and the latter are followed by the nonhorizontal node $X$. Since $\lambda$ is a bijection that preserves the post-order of nontrivial nodes and the ancestorship relation with nonhorizontal nodes, we conclude that $\lambda(X) = Z$ and that the synopsis trees $\sigma_1$ and $\sigma'_1$ (respectively, $\sigma_2$ and $\sigma'_2$) strongly cover each other. Using the inductive hypothesis, we finally conclude that $\sigma = \varepsilon(\sigma_1, X(\sigma_2, \varepsilon)) = \varepsilon(\sigma'_1, Z(\sigma'_2, \varepsilon)) = \sigma'$. □

Thanks to Lemma 6.5, we can define *the normal form* $\sigma^*$ of a synopsis tree $\sigma$ as the unique synopsis tree that is in normal form and that strongly covers $\sigma$, provided that this tree exists.

Our next goal is to prove that the normal form $\sigma^*$ of $\sigma$ indeed exists, and that it can be attained by a finite sequence of generic editing operations on synopsis trees. These operations are called *promotion*, *demotion*, and *reduction*, and they are presented in Figure 11. There, $\varepsilon$ represents a trivial component, $X$ represents an arbitrary component, $H_1, \ldots, H_k$ represent horizontal (possibly trivial) components, and $\alpha, \beta_1, \ldots, \beta_k$ represent arbitrary synopsis trees. Note that the figure describes the case where promotion, demotion, and reduction operations are applied at the root of a synopsis tree—in general, these operations can be applied to any subtree of a synopsis tree. We write $\sigma \to^*_{\mathsf{op}} \sigma'$ whenever $\sigma'$ can be obtained from $\sigma$ by applying a finite sequence of promotion, demotion, and reduction operations. To give further intuition about these operations, we remark an analogy between the operations of promotion, depicted in Figure 11, and deletion, depicted in Figure 4 (a similar correspondence holds between the operations of demotion and insertion of a new root). In this case, the root $X$ of the synopsis tree is acting as the context $C$ of the curried tree, the subtree $\alpha$ is acting as the curried subtree $t'$, and the subtree rooted at $H_1$ is acting as the horizontal context $C'$.

Notice that the editing operations on synopsis trees that we just described preserve the post-order of nontrivial nodes and the ancestorship of nonhorizontal nodes. From this, it follows that they also preserve the strong coverability relation. The following lemma shows that the normal form of a synopsis tree exists and can be obtained via a sequence of promotion, demotion, and reduction operations.

LEMMA 6.6. *For every synopsis tree $\sigma$, there is $\sigma^*$ in normal form such that $\sigma \to^*_{\mathsf{op}} \sigma^*$. Moreover, the number of operations needed to transform $\sigma$ into $\sigma^*$ is bounded by $2|\sigma|$.*

PROOF. The proof goes again by a structural induction on the synopsis tree $\sigma$. Intuitively, we first normalize the left and right subtrees of $\sigma$ separately using induction. Then we complete the normalization process by applying a suitable series of operations on the basis of the component at the root of $\sigma$: if this component is nonhorizontal, then we apply a promotion followed by a demotion; if it is horizontal and nontrivial, then we only apply a promotion operation; and if it is trivial, then we apply a promotion followed by a reduction operation.

We can assume without loss of generality that all leaves in the synopsis tree $\sigma$ are trivial and hence are labeled by $\varepsilon$ (indeed, we can append $\varepsilon$-labeled nodes to every leaf of $\sigma$ without changing its equivalence class). This assumption reduces the base case to the situation where the synopsis tree $\sigma$ consists of a single $\varepsilon$-labeled node. In this case, the synopsis tree is already in normal form, and hence the lemma is trivially satisfied by letting $\sigma^* = \varepsilon = \sigma$.

For the inductive step, we assume that $\sigma = X(\alpha, \beta)$. First, we transform the subtrees $\alpha, \beta$ of $\sigma$ into their corresponding normal forms $\alpha^*, \beta^*$ (this can be done since, by inductive hypothesis, $\alpha \to_{\mathsf{op}}^* \alpha^*$ and $\beta \to_{\mathsf{op}}^* \beta^*$). We consider the intermediate synopsis tree that we just obtained:

$$\sigma' \;=\; X(\alpha^*, \beta^*).$$

Since the right subtree $\beta^*$ is in normal form, its left-most branch consists of horizontal components only. We can thus write

$$\sigma' \;=\; X(\alpha^*,\; H_1(\ldots H_k(\varepsilon, \beta_k), \ldots, \beta_1))$$

for some horizontal (possibly trivial) components $H_1, \ldots, H_k$ and some synopsis trees $\beta_1, \ldots, \beta_k$. We can then perform a promotion operation at the root of $\sigma'$ and obtain the synopsis tree

$$\sigma'' \;=\; X(H_1(\ldots H_k(\alpha^*, \beta_k), \ldots, \beta_1), \varepsilon).$$

Using a simple induction on $i = k, \ldots, 1$ and the fact that the subtrees $H_i(\ldots H_k(\varepsilon, \beta_k), \ldots, \beta_1)$ of $\beta^*$ are in normal form, one can easily verify that the subtrees $H_i(\ldots H_k(\alpha^*, \beta_k), \ldots, \beta_1)$ of $\sigma''$ are also in normal form. In particular, this shows that the left subtree of $\sigma''$ is in normal form.

We now distinguish a few cases depending on whether the component $X$ is horizontal, trivial, or nonhorizontal:

(1) If the component $X$ is horizontal and nontrivial, then $\sigma''$ is already in normal form and we can simply let $\sigma'' = \sigma^*$.
(2) If the component $X$ is trivial, then we "lift" the left subtree of $\sigma''$ via a reduction operation. This results in a synopsis tree $\sigma^* = H_1(\ldots H_k(\alpha^*, \beta_k), \ldots, \beta_1)$ in normal form.
(3) If the component $X$ is nonhorizontal, then we apply a demotion operation to $\sigma''$ so as to obtain the synopsis tree

$$\sigma^* \;=\; \varepsilon(\varepsilon, X(H_1(\ldots H_k(\alpha^*, \beta_k), \ldots, \beta_1), \varepsilon).$$

We observe that both subtrees $\varepsilon$ and $H_1(\ldots H_k(\alpha^*, \beta_k), \ldots, \beta_1)$ are in normal form. Moreover, since $X$ is nonhorizontal, we know that $\sigma^*$ is also in normal form.

It remains to prove the upper bound on the minimum number $\#_{\mathsf{op}}(\sigma)$ of operations on synopsis trees that are required to transform $\sigma$ into $\sigma^*$. From previous constructions, we can easily verify that

$$\#_{\mathsf{op}}(\sigma) \;\leq\; \#_{\mathsf{op}}(\alpha) + \#_{\mathsf{op}}(\beta) + 2,$$

and hence $\#_{\mathsf{op}}(\sigma) \leq 2|\sigma|$. □

For the last ingredient of the proof, we show that if $\sigma \rightarrow_{\mathsf{op}}^{*} \sigma'$, then the two languages $[\![\sigma]\!]$ and $[\![\sigma']\!]$ are repairable one into the other by means of a sequence of editing operations of uniformly bounded length. In other words, a single promotion, demotion, or reduction operation to a synopsis tree $\sigma$ corresponds to a small amount of edits that are applicable to any generic tree in the language $[\![\sigma]\!]$. The proof of this result is via a simple analysis of the transformations on unranked trees that are induced by the operations of promotion, demotion, and reduction.

LEMMA 6.7. *If $\sigma'$ is a synopsis tree obtained from another synopsis tree $\sigma$ via a single promotion, demotion, or reduction operation, then* $\mathsf{cost}([\![\sigma]\!], [\![\sigma']\!]) \leq 2$ *and* $\mathsf{cost}([\![\sigma']\!], [\![\sigma]\!]) \leq 2$. *In particular, it follows by induction that* $\mathsf{cost}([\![\sigma]\!], [\![\sigma^*]\!]) \leq 4|\sigma|$ *and* $\mathsf{cost}([\![\sigma^*]\!], [\![\sigma]\!]) \leq 4|\sigma|$.

PROOF. It is sufficient to prove that $\mathsf{cost}([\![\sigma]\!], [\![\sigma']\!]) \leq 2$, as the symmetric bound $\mathsf{cost}([\![\sigma']\!], [\![\sigma]\!]) \leq 2$ follows from the fact that standard editing operations on trees can be reverted. In the sequel, we assume that all synopsis trees are related to a stepwise automaton $\mathcal{S}$. We apply a case distinction based on the type of operation that transforms $\sigma$ into $\sigma'$:

(1) Consider a promotion operation, which takes a synopsis tree of the form $\sigma = X(\alpha,\ H_1(\ldots H_k(\varepsilon, \beta_k), \ldots, \beta_1))$ and transforms it into the synopsis tree $\sigma' = X(H_1(\ldots H_k(\alpha, \beta_k), \ldots, \beta_1), \varepsilon)$. Consider also a generic tree $t \in [\![\sigma]\!]$. This can be written as

$$t = C \circ (s @ (C_1 \circ (\ldots C_k \circ (a @ s_k) \ldots @ s_1)))$$

for some $C \in \mathscr{L}(\mathcal{S} \mid X)$, $s \in [\![\alpha]\!]$, $C_i \in \mathscr{L}(\mathcal{S} \mid H_i)$, $s_i \in [\![\beta_i]\!]$, and $a \in \Sigma$. For the sake of brevity, define the horizontal context

$$\bar{C} = C_1 \circ (\ldots C_k \circ (\bullet @ s_k) \ldots @ s_1)$$

in such a way that we can write $t = C \circ (s @ (\bar{C} \circ a))$. After deleting the $a$-labeled node from $t$, we obtain the tree $t' = C \circ (\bar{C} \circ s)$, and after inserting a $b$-labeled node, we obtain the tree

$$t'' = C \circ ((\bar{C} \circ s) @ b) = C \circ ((C_1 \circ ((\ldots C_k \circ (s @ s_k) \ldots) @ s_1)) @ b),$$

which clearly belongs to $[\![\sigma']\!]$.

(2) Consider now a demotion operation, which takes a synopsis tree $\sigma$ and transforms it into the synopsis tree $\sigma' = \varepsilon(\varepsilon, \sigma)$. Let $t \in [\![\sigma]\!]$, and let $x$ be the left-most leaf in $t$ (note that this corresponds to the root of the unranked tree $\mathsf{ext}^{-1}(t)$). We can write $t = C \circ a$, where $a$ is the label of the left-most leaf $x$ of $t$ and $C$ is the horizontal context obtained from $t$ by relabeling $x$ with a placeholder. By applying an insertion operation to $t$, we obtain the tree $t' = b @ (C \circ a)$, which clearly belongs to $[\![\sigma']\!]$.

(3) We finally consider a reduction operation, which transforms a synopsis tree $\sigma = \varepsilon(\alpha, \varepsilon)$ into the synopsis tree $\sigma' = \alpha$. We can write any generic tree $t \in [\![\sigma]\!]$ as $t = s @ a$, where $s \in [\![\alpha]\!]$ and $a \in \Sigma$. In this case, it suffices to perform one deletion to obtain the tree $t' = s$, which clearly belongs to $[\![\sigma']\!]$.

We observe that the repair strategies defined previously can be lifted to trees under any given context $C$. More precisely, if a tree $t \in [\![\sigma]\!]$ is transformed with editing operations into a tree $t' \in [\![\sigma']\!]$, then the tree $C \circ t$ can be edited into $C \circ t'$ using the analogous strategy. This observation is important because the operations of promotion, demotion, and reduction may be applied at arbitrary nodes of synopsis trees.

To conclude the proof, we remark that, thanks to Lemma 6.6, the normal form $\sigma^*$ of any synopsis tree $\sigma$ can be obtained by applying a sequence of promotions,

demotions, and reductions of length at most $2|\sigma|$. A simple induction finally implies that $\mathsf{cost}(\llbracket\sigma\rrbracket, \llbracket\sigma^*\rrbracket) \leq 4|\sigma|$ and $\mathsf{cost}(\llbracket\sigma^*\rrbracket, \llbracket\sigma\rrbracket) \leq 4|\sigma|$. □

We have all of the ingredients now to prove Lemma 6.1.

Proof of Lemma 6.1. Let $\theta$ be the intermediate synopsis tree such that $\sigma \hookrightarrow\!\!\!\!\rightarrow \theta \leftrightharpoons \tau$, whose existence is shown in Lemma 6.2. Recall that Lemma 6.3 implies $\llbracket\theta\rrbracket_{\mathcal{S}} \subseteq \llbracket\tau\rrbracket_{\mathcal{T}}$. As $|\theta| = |\tau|$, it is sufficient to show that $\mathsf{cost}(\llbracket\sigma\rrbracket_{\mathcal{S}}, \llbracket\theta\rrbracket_{\mathcal{S}}) \leq 4|\sigma| + 4|\theta|$. This last claim can be proved using Lemma 6.7, which implies $\mathsf{cost}(\llbracket\sigma\rrbracket_{\mathcal{S}}, \llbracket\sigma^*\rrbracket_{\mathcal{S}}) \leq 4|\sigma|$ and $\mathsf{cost}(\llbracket\theta^*\rrbracket_{\mathcal{S}}, \llbracket\theta\rrbracket_{\mathcal{S}}) \leq 4|\theta|$, and the fact that $\sigma \hookrightarrow\!\!\!\!\rightarrow \theta$, which implies $\sigma^* = \theta^*$. □

We conclude the section by proving Proposition 5.8, which essentially gives an upper bound for the cost of an optimal repair strategy from $\mathscr{L}(\mathcal{S})$ to $\mathscr{L}(\mathcal{T})$. To prove this proposition, we need to analyze the minimum size of a basic synopsis tree of $\mathcal{T}$ that covers a given primitive synopsis tree of $\mathcal{S}$.

Lemma 6.8. *Given a primitive synopsis tree $\sigma$ of $\mathcal{S}$, if $\sigma$ is covered by some basic synopsis tree of $\mathcal{T}$, then it is covered by one such tree $\tau \in \mathsf{BST}(\mathcal{T})$ that has size at most $(4|\sigma| + 1) \cdot |\mathsf{SCC}(\mathcal{T})|$, where $|\sigma|$ is the number of nodes of $\sigma$ and $|\mathsf{SCC}(\mathcal{T})|$ is the number of components of $\mathcal{T}$.*

Proof. Let $\sigma \in \mathsf{PST}(\mathcal{S})$ and $\tau \in \mathsf{BST}(\mathcal{T})$ such that $\sigma \hookrightarrow \tau$, and let $\lambda$ be the injective function from nontrivial nodes of $\sigma$ to nontrivial nodes of $\tau$ that witnesses $\sigma \hookrightarrow \tau$. We begin by identifying those nodes of $\tau$ that belong to the range of $\lambda$. Formally, we say that a node $y$ of $\tau$ is *used* if $y \in \lambda(x)$ for some nontrivial node $x$ of $\sigma$. Next, we show how to restrict $\tau$ to a subset of its nodes having size at most $(4|\sigma| + 1) \cdot |\mathsf{SCC}(\mathcal{T})|$ and such that the induced subgraph is a basic synopsis tree of $\mathcal{T}$ that also covers $\sigma$.

We first define the set $V$ that only contains the following nodes:

(1) the root of $\tau$,
(2) the used nodes of $\tau$, and
(3) the nodes of $\tau$ whose both subtrees contain some used nodes of $\tau$.

We claim that the subgraph of $\tau$ induced by $V$ is a tree with at most two children on each node (note that some internal nodes in the induced subgraph $V$ may contain only one child). Consider two nodes $y_1, y_2$ in $V$. Let $y$ be the least common ancestor of $y_1$ and $y_2$ in $\tau$. As both subtrees of $y$ in $\tau$ contain used nodes—indeed, they contain $y_1$ and $y_2$, respectively—the node $y$ also belongs to $V$. This shows that $V$ is closed under the least common ancestor, and hence $\tau$ restricted to $V$ is a tree with out-degree at most 2.

We can also verify that the size of $V$ is at most $2|\sigma| + 1$. Indeed, the number of used nodes in $\tau$ is at most $|\sigma|$, and so is the number of nodes whose both subtrees contain used nodes.

Next we extend the set $V$ minimally in such a way that the induced subgraph of $\tau$ is a basic synopsis tree (in particular, it is a binary tree). Formally, we let $W$ be the set of the following nodes:

(4) the nodes in $V$,
(5) the nodes of $\tau$ with one subtree containing some used nodes and with a label *different* from that of its parent, and
(6) the immediate successors in $\tau$ of all previous nodes.

Clearly, the subgraph of $\tau$ induced by $W$, denoted $\tau|_W$, is a full binary tree, namely all internal nodes have exactly two children. Essentially, this holds because we included in $W$ the immediate successors of a set of nodes.

It is also easy to see that $\tau|_W$ is a basic synopsis tree. Indeed, consider a node $y \in W$ and its two successors $y_1$ and $y_2$ in the induced subgraph $\tau|_W$. If both $y_1$ and $y_2$ are

immediate successors of $y$ in $\tau$, then, since $\tau$ respects the transition function $\delta$ of $\mathcal{T}$, there exist some states $q \in \tau(y)$, $q_1 \in \tau(y_1)$, and $q_2 \in \tau(y_2)$ such that $q \in \delta(q_1, q_2)$. Otherwise, if $y_1$ is not an immediate successor of $y$, then all nodes of $\tau$ between $y$ and the parent of $y_1$ must have the same label $\tau(y)$ (otherwise, one of these nodes would belong to $W$, thus contradicting the fact that $y_1$ is a successor of $y$ in $\tau|_W$). From this, using similar arguments as in the previous case, we conclude that $q \in \delta(q_1, q_2)$ for some states $q \in \tau(y)$, $q_1 \in \tau(y_1)$, and $q_2 \in \tau(y_2)$. The case where $y_2$ is not an immediate successor of $y$ in $\tau$ is just symmetric. Overall, this proves that the induced subgraph $\tau|_W$ respects the transition function of $\mathcal{T}$, and hence it is a basic synopsis tree.

To prove that $\tau|_W$ covers $\sigma$, it suffices to recall that $W$ contains all used nodes of $\tau$—that is, $\lambda(x) \in W$ for all nontrivial nodes of $\sigma$, and hence the same function $\lambda$ that witnessed $\sigma \hookrightarrow \tau$ can be used to witness $\sigma \hookrightarrow \tau|_W$.

It remains to prove that $|W| \leq (4|\sigma| + 1) \cdot |\mathsf{SCC}(\mathcal{T})|$. It is easy to see that every node in $W \setminus V$ either has a descendant that is used or is the successor of a node with a used descendant. This means that any subset of $W \setminus V$ containing only nodes that are pairwise incomparable with respect to the ancestor relation has size at most twice the number of used nodes, hence at most $2|\sigma|$. Moreover, if we consider sets of nodes from $W \setminus V$ totally ordered with respect to the ancestor relation, then we observe that such a set has size at most $|\mathsf{SCC}(\mathcal{T})|$: indeed, every two nodes in this set that are consecutive in $\tau$ must be labeled with different components. Putting it all together and recalling that $|V| \leq 2|\sigma|+1$, we conclude that $|W| \leq 2|\sigma|+1+2|\sigma| \cdot |\mathsf{SCC}(\mathcal{T})| \leq (4|\sigma|+1) \cdot |\mathsf{SCC}(\mathcal{T})|$. □

We are now ready to derive an upper bound for the cost of an optimal repair strategy from $\mathscr{L}(\mathcal{S})$ to $\mathscr{L}(\mathcal{T})$ under the assumption that all primitive synopsis trees of $\mathcal{S}$ are covered by basic synopsis trees of $\mathcal{T}$.

Proof of Proposition 5.8. Let $f$ be a function that maps every primitive synopsis tree $\sigma$ of $\mathcal{S}$ to a basic synopsis tree $\tau$ of $\mathcal{T}$ that covers $\sigma$. Following the previous results, the strategy for repairing $\mathscr{L}(\mathcal{S})$ into $\mathscr{L}(\mathcal{T})$ can be obtained from a series of transformations between languages having the following costs:

$$\mathsf{cost}\Big(\mathscr{L}(\mathcal{S}),\ \bigcup_{\sigma \in \mathsf{PST}(\mathcal{S})} [\![\sigma]\!]_{\mathcal{S}}\Big) \ = \ 0, \qquad \text{(by Lemma 5.2)}$$

$$\mathsf{cost}\big([\![\sigma]\!]_{\mathcal{S}},\ [\![f(\sigma)]\!]_{\mathcal{T}}\big) \ \leq \ 4|\sigma| + 4|f(\sigma)|, \qquad \text{(by Lemma 6.1)}$$

$$\mathsf{cost}\big([\![f(\sigma)]\!]_{\mathcal{T}},\ \mathscr{L}(\mathcal{T})\big) \ \leq \ (4|f(\sigma)| + 1) \cdot 2^{|Q'|}, \qquad \text{(by Lemma 5.5)}$$

where $Q'$ is the set of states of $\mathcal{T}$. In particular,we get

$$\mathsf{cost}(\mathscr{L}(\mathcal{S}), \mathscr{L}(\mathcal{T})) \ \leq \ \max_{\sigma \in \mathsf{PST}(\mathcal{S})} \{4|\sigma| + 4|f(\sigma)| + (4|f(\sigma)| + 1) \cdot 2^{|Q'|}\}.$$

As we pointed out previously, any primitive synopsis tree $\sigma$ of $\mathcal{S}$ is of bounded size, precisely $|\sigma| \leq 2^{|Q|}$, where $Q$ is the set of states of $\mathcal{S}$. As concerns the minimum size of a basic synopsis tree $f(\sigma)$ that covers $\sigma$, by applying Lemma 6.8 we get that for every $\sigma \in \mathsf{PST}(\mathcal{S})$, there is $\tau \in \mathsf{BST}(\mathcal{T})$ such that $\sigma \hookrightarrow \tau$ and

$$|\tau| \ \leq \ (4|\sigma| + 1) \cdot |\mathsf{SCC}(\mathcal{T})|.$$

In particular, we can assume, without loss of generality, that $|f(\sigma)| \leq (2^{|Q|+2+1) \cdot |\mathsf{SCC}(\mathcal{T})|}$. Putting everything together, we obtain

$$\mathsf{cost}(\mathscr{L}(\mathcal{S}), \mathscr{L}(\mathcal{T})) \ = \ \mathcal{O}(|\mathsf{SCC}(\mathcal{T})| \cdot 2^{|Q|+|Q'|}). \quad \square$$

### 6.2. From Repair to Covering

In this section, we prove the "only if" direction of Theorem 5.7. We fix for the rest of the section two stepwise automata $\mathcal{S} = (\Sigma, Q, \delta, \delta_0, F)$ and $\mathcal{T} = (\Delta, Q', \delta', \delta'_0, F')$ recognizing

the source and the target languages, respectively. We assume that $\mathscr{L}(\mathcal{S})$ is repairable into $\mathscr{L}(\mathcal{T})$ with uniformly bounded cost, and we prove that every primitive synopsis tree of $\mathcal{S}$ is covered by some basic synopsis tree of $\mathcal{T}$.

The general idea is to associate with each primitive synopsis tree $\sigma$ of $\mathcal{S}$ a suitable tree $t_\sigma \in \mathscr{L}(\mathcal{S})$, called the *witness tree* of $\sigma$, such that from any optimal repair of $t_\sigma$ into $\mathscr{L}(\mathcal{T})$ one can extract a basic synopsis tree $\tau$ of $\mathcal{T}$ that covers $\sigma$. Intuitively, the witness tree $t_\sigma$ is obtained from the primitive synopsis tree $\sigma$ by replacing every nontrivial node $x$ with a sufficiently large number of repetitions of a special context in $\mathscr{L}(\mathcal{S} \mid \sigma(x))$, called the *fingerprint context*. The number of repetitions of each fingerprint context will depend on the worst-case repair cost $K = \mathsf{dist}(\mathscr{L}(\mathcal{S}), \mathscr{L}(\mathcal{T}))$. Using the definition of witness tree $t_\sigma$ and the assumption that $t_\sigma$ can be repaired into some tree $t'_\sigma \in \mathscr{L}(\mathcal{T})$ with at most $K$ edits, one can then argue that $t'_\sigma$ contains at least one copy of the fingerprint context associated with each nontrivial node $x$ of $\sigma$, and furthermore the arrangements of these fingerprint contexts inside $t_\sigma$ and inside $t'_\sigma$ are the same, both with respect to the post-order relation and with respect to the ancestorship of the nonhorizontal components. One finally looks at some run of $\mathcal{T}$ that accepts the tree $t'_\sigma$: this run, together with the structure of the fingerprints inside $t'_\sigma$, induces a basic synopsis tree $\tau$ of $\mathcal{T}$ and a coverability relation from $\sigma$ to $\tau$. Next we illustrate the various definitions and arguments in more detail. We divide up the proof into constructing the witness tree $t_\sigma$ and building the cover from its repair.

*Constructing the witness tree*. We begin by giving the following lemma, which defines the so-called fingerprint context of a component of $\mathcal{S}$. Basically, the lemma shows that given a component $X$ of $\mathcal{S}$, one can find a context $C_X$ that can be "pumped" inside the language $\mathscr{L}(\mathcal{S} \mid X)$ (i.e., $C_X \circ \ldots \circ C_X \in \mathscr{L}(\mathcal{S} \mid X)$) and that characterizes the containment of $\mathscr{L}(\mathcal{S} \mid X)$ in $\mathscr{L}(\mathcal{T} \mid Y)$ for every component $Y$ of $\mathcal{T}$ (i.e., $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y)$ if and only if $C_X \in \mathscr{L}(\mathcal{T} \mid Y)$). We say that a context $C$ is *cyclic* for a component $X$ if there is a state $q \in X$ such that $q \in \delta(q, C)$.

LEMMA 6.9. *For all $X \in \mathsf{SCC}(\mathcal{S})$, there is a cyclic context $C_X \in \mathscr{L}(\mathcal{S} \mid X)$ such that for all $Y \in \mathsf{SCC}(\mathcal{T})$,*

$$\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y) \qquad \textit{if and only if} \qquad C_X \in \mathscr{L}(\mathcal{T} \mid Y).$$

PROOF. Let $X$ be a component of $\mathcal{S}$, and let $Y_1, \ldots, Y_m \in \mathsf{SCC}(\mathcal{T})$ be all components of $\mathcal{T}$. We construct the cyclic context $C_X$ by exploiting an induction over the number $m$ of components of $\mathcal{T}$—that is, we prove that for every $0 \leq i \leq m$, there is a cyclic context $C_i \in \mathscr{L}(\mathcal{S} \mid X)$ such that

$$\forall 1 \leq j \leq i. \qquad \mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y_j) \qquad \text{if and only if} \qquad C_i \in \mathscr{L}(\mathcal{T} \mid Y_j). \qquad (\star)$$

Clearly, the statement of the lemma follows from $(\star)$ when we let $C_X = C_m$.

The base case $i = 0$ holds vacuously for $C_i = \bullet$, so we focus on the inductive step. Suppose that we defined a context $C_i$ that satisfies $(\star)$ for some index $0 \leq i < m$. To construct a context $C_{i+1}$ that satisfies $(\star)$, we need to distinguish two cases, depending on whether $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y_{i+1})$ or not.

If $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y_{i+1})$, then we define $C_{i+1} = C_i$ and observe that $(\star)$ holds trivially for $i + 1$.

Otherwise, if $\mathscr{L}(\mathcal{S} \mid X) \not\subseteq \mathscr{L}(\mathcal{T} \mid Y_{i+1})$, we let $C$ be a context in $\mathscr{L}(\mathcal{S} \mid X) \setminus \mathscr{L}(\mathcal{T} \mid Y_{i+1})$. Since $C_i$ is cyclic and $C \in \mathscr{L}(\mathcal{S} \mid X)$, there exist some states $p, q, r \in X$ such that $r \in \delta(r, C_i)$ and $q \in \delta(p, C)$. Let $C'$ and $C''$ be some other contexts in $\mathscr{L}(\mathcal{S} \mid X)$ such that $r \in \delta(q, C')$ and $p \in \delta(r, C'')$ (such contexts exist since $p, q, r$ are states within the same strongly connected component of $\mathcal{S}$). We then define

$$C_{i+1} \;=\; C_i \circ C' \circ C \circ C''.$$

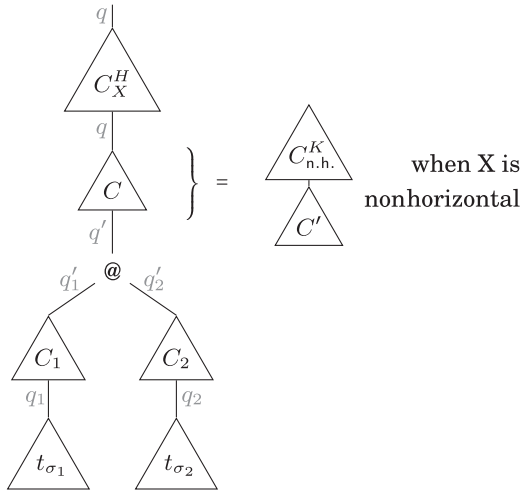



Fig. 12. Construction of the witness tree.

We claim that $C_{i+1}$ is a cyclic context in $\mathscr{L}(\mathcal{S} \mid X)$. Indeed, we have the following runs in the source automaton $\mathcal{S}$:

$$r \xrightarrow{C''} p \xrightarrow{C} q \xrightarrow{C'} r \xrightarrow{C_i} r.$$

It is also easy to see that $C_{i+1} \notin \mathscr{L}(\mathcal{T} \mid Y_{i+1})$. Indeed, if $C_{i+1} \in \mathscr{L}(\mathcal{T} \mid Y_{i+1})$, then there would exist some states $p', q' \in Y_{i+1}$ such that $q' \in \delta'(p', C)$, which would contradict the fact that $C \notin \mathscr{L}(\mathcal{T} \mid Y_{i+1})$.

We have just constructed a cyclic context $C_{i+1} \in \mathscr{L}(\mathcal{S} \mid X)$ such that $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y_{i+1})$ if and only if $C_{i+1} \in \mathscr{L}(\mathcal{T} \mid Y_{i+1})$. To conclude the proof, we recall from the inductive hypothesis that for all $1 \leq j \leq i$, $\mathscr{L}(\mathcal{S} \mid X) \nsubseteq \mathscr{L}(\mathcal{T} \mid Y_j)$ implies that $C_i \notin \mathscr{L}(\mathcal{T} \mid Y_j)$, and hence since $C_{i+1}$ contains an occurrence of $C_i$, $C_{i+1} \notin \mathscr{L}(\mathcal{T} \mid Y_j)$. Symmetrically, $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y_j)$ implies that $C_{i+1} = C_i \in \mathscr{L}(\mathcal{T} \mid Y_j)$. All together, this shows that for all $1 \leq j \leq i+1$, $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y_j)$ if and only if $C_{i+1} \in \mathscr{L}(\mathcal{T} \mid Y_j)$. □

For the rest of this section, we fix for each component $X$ of $\mathcal{S}$ a context $C_X$ that satisfies Lemma 6.9, which we call the *fingerprint context*. We also fix an arbitrary primitive synopsis tree $\sigma$ of $\mathcal{S}$.

Next we construct the so-called witness tree $t_\sigma$ by exploiting a structural induction on the primitive synopsis tree $\sigma$. In doing so, we will guarantee that there exists a run of $\mathcal{S}$ on $t_\sigma$ that assigns to the root of $t_\sigma$ some state that belongs to the same component that labels the root of $\sigma$, namely $\delta(t_\sigma) \cap \sigma(\varepsilon) \neq \varnothing$. We omit the construction for the base case, where $\sigma$ is a singleton, as it can be easily derived from what follows.We assume that $X$ is the component at the root of $\sigma$ and that $\sigma_1$ and $\sigma_2$ are the (nonempty) left and right subtrees of $\sigma$. Suppose that $t_{\sigma_1}$ and $t_{\sigma_2}$ are the recursively defined witness trees for $\sigma_1$ and $\sigma_2$. Moreover, choose arbitrarily some $q_1$ (respectively, $q_2$) in the nonempty set $\delta(t_{\sigma_1}) \cap \sigma_1(\varepsilon)$ (respectively, $\delta(t_{\sigma_2}) \cap \sigma_2(\varepsilon)$). The witness tree $t_\sigma$ for $\sigma$ is obtained from the following series of transformations (we suggest that the reader refers to Figure 12 for a graphical representation):

(1) The first transformation merges the trees $t_{\sigma_1}$ and $t_{\sigma_2}$ into a single tree that induces a run of $\mathcal{S}$ ending in the component $X$. We know from the definition of primitive synopsis tree that $\sigma$ respects the transition function of $\mathcal{S}$. In particular, this means that there exist some states $q' \in X$, $q_1' \in \sigma_1(\varepsilon)$, and $q_2' \in \sigma_2(\varepsilon)$ such that $q' \in \delta(q_1', q_2')$.

Moreover, since $q_1$ and $q_1'$ (respectively, $q_2$ and $q_2'$) belong to the same component at the root of $\sigma_1$ (respectively, $\sigma_2$), there exist some contexts $C_1 \in \mathscr{L}(\mathcal{S} \mid \sigma_1(\varepsilon))$ and $C_2 \in \mathscr{L}(\mathcal{S} \mid \sigma_2(\varepsilon))$ such that $q_1' \in \delta(q_1, C_1)$ and $q_2' \in \delta(q_2, C_2)$. This allows us to construct the tree

$$(C_1 \circ t_{\sigma_1}) @ (C_2 \circ t_{\sigma_2})$$

that induces a run of $\mathcal{S}$ ending in state $q'$ within the component $X$.

(2) The second transformation extends the tree obtained in the previous step in such a way that one can later append repetitions of the fingerprint context $C_X$. This is done by identifying a "recurrent" state $q$ such that $q \in \delta(q, C_X)$ (this state exists since $C_X$ is cyclic) and then connecting it to the state $q'$ via a suitable context $C \in \mathscr{L}(\mathcal{S} \mid X)$ such that $q \in \delta(q', C)$ (note that $q$ and $q'$ belong to the same component $X$). The resulting tree is of the following form:

$$C \circ ((C_1 \circ t_{\sigma_1}) @ (C_2 \circ t_{\sigma_2})).$$

To avoid that an editing of the witness tree $t_\sigma$ could modify the ancestorship of $C_X$ with the nodes of the two subtrees $t_{\sigma_1}$ and $t_{\sigma_2}$, we further assume that if $X$ is a *nonhorizontal* component, then the context $C$ that is used for connecting $q$ to $q'$ is of the form $C_{\text{n.h.}}^K \circ C'$, where $K = \mathsf{dist}(\mathscr{L}(\mathcal{S}), \mathscr{L}(\mathcal{T}))$, $C', C_{\text{n.h.}} \in \mathscr{L}(\mathcal{S} \mid X)$, $C_{\text{n.h.}}$ is some cyclic *nonhorizontal* context, and $C_{\text{n.h.}}^K$ is the $K$-fold repetition of $C_{\text{n.h.}}$ (recall that the ancestorship of nonhorizontal contexts is preserved by the editing operations).

(3) The last transformation adds a sufficiently large repetition of the fingerprint context $C_X$. For this, we define $H = m \cdot (2K + 1)$, where $K = \mathsf{dist}(\mathscr{L}(\mathcal{S}), \mathscr{L}(\mathcal{T}))$ and $m$ is the number of components of $\mathcal{T}$. We then attach to the tree so far constructed the $H$-fold repetition $C_X^H$ of the fingerprint context $C_X$, finally obtaining the desired witness tree:

$$t_\sigma = C_X^H \circ C \circ ((C_1 \circ t_{\sigma_1}) @ (C_2 \circ t_{\sigma_2})).$$

We observe that, thanks to the preceding constructions, the automaton $\mathcal{S}$ admits a run on $t_\sigma$ that ends in state $q \in X$. This shows that the invariant $\delta(t_\sigma) \cap X \neq \varnothing$ is satisfied. We also remark that it may happen that $\delta(t_\sigma) \cap F = \varnothing$, and hence $t_\sigma \notin \mathscr{L}(\mathcal{S})$. Technically speaking, this could violate the claim that one can repair $t_\sigma$ into $\mathscr{L}(\mathcal{T})$ with at most $K$ edits. However, from the assumption that $\mathcal{S}$ is trimmed, it follows that there is a context $C_F$ such that $\delta(q, C_F) \cap F \neq \varnothing$, and hence one can always prolong $t_\sigma$ to obtain a tree inside the language $\mathscr{L}(\mathcal{S})$. From now on, we assume for the sake of simplicity that $t_\sigma \in \mathscr{L}(\mathcal{S})$.

*Building the covering from a repaired witness tree*. We now turn toward extracting a covering of $\sigma$ from a repair of $t_\sigma$. We fix, once and for all, the tree $t_\sigma'$ in the target language $\mathscr{L}(\mathcal{T})$ that is obtained by repairing $t_\sigma$ with at most $K$ edits.

We recall that the witness tree $t_\sigma$ contains $H = m \cdot (2K + 1)$ copies of the fingerprint context $C_X$, for each node $x$ in $\sigma$, where $X = \sigma(x)$. As a consequence, the repaired tree $t_\sigma'$ must contain an $m$-fold repetition $C_X^m$ of each fingerprint context $C_X$. In the following, we will look at the occurrences of these fingerprint contexts inside the repaired tree $t_\sigma'$ and compare their post-order and ancestor relationships with those for the analogous occurrences in $t_\sigma$. We need some preliminary definitions.

Given a context $C$ and a node $x$ of a tree $t$, we say that $C$ *occurs at node* $x$ if there exist a context $C'$ and a tree $t''$ such that (i) $t$ can be written as $C' \circ C \circ t''$ and (ii) $x$ is the node from where the subtree $C \circ t''$ of $t$ hangs. We denote an occurrence of a context $C$ at a node $x$ of a tree $t$ by the pair $(C, x)$. Furthermore, we say that two occurrences $(C, x)$ and $(C', x')$ of two contexts inside the same tree $t$ are *nonoverlapping* (respectively, in *post-order relation*, *ancestor relation*) if $\{x\} \cdot \mathsf{nodes}(C) \cap \{x'\} \cdot \mathsf{nodes}(C') = \varnothing$ (respectively, if $x \preccurlyeq_t^{\mathsf{post}} x'$, if $x \preccurlyeq_t^{\mathsf{anc}} x'$).

The following lemma shows that the occurrences of the contexts $C_X^m$ inside $t_\sigma$ are in the same post-order relation as some corresponding occurrences inside $t'_\sigma$, and similarly for the ancestor relation when $X$ is a nonhorizontal component.

LEMMA 6.10. *One can find a mapping $f$ from the nontrivial nodes $x$ of the primitive synopsis tree $\sigma$ to the nodes $f(x)$ of the repaired witness tree $t'_\sigma$ such that*

—*the context $C_X^m$, where $X = \sigma(x)$, occurs at node $f(x)$ in $t'_\sigma$, for all nontrivial nodes $x$ of $\sigma$;*
—*all occurrences $(C_X^m, f(x))$, with $X = \sigma(x)$ and $x$ nontrivial node of $\sigma$, are pairwise nonoverlapping;*
—*$x \preccurlyeq_\sigma^{\mathsf{post}} y$ if and only if $f(x) \preccurlyeq_{t'_\sigma}^{\mathsf{post}} f(y)$ for all nontrivial nodes $x, y$ of $\sigma$; and*
—*$x \preccurlyeq_\sigma^{\mathsf{anc}} y$ if and only if $f(x) \preccurlyeq_{t'_\sigma}^{\mathsf{anc}} f(y)$ for all nontrivial nodes $x, y$ of $\sigma$, with $\sigma(x)$ the nonhorizontal component.*

PROOF. We begin by establishing a property that concerns the occurrences of contexts in a tree that has been edited. Intuitively, the following claim implies that if $t'$ is a tree obtained from $t$ by applying at most $K$ edit operations and $t$ contains an occurrence of the $2K+1$-fold repetition of a context $C$, then $t''$ contains at least one occurrence of the same context $C$.

CLAIM 1. *Let $t$ be a curried tree, and let $t''$ be the curried tree obtained from $t$ after a deletion or an insertion of a single node. If $t$ contains at least $n$ nonoverlapping occurrences of the same context $C$, then $t''$ contains at least $n-2$ nonoverlapping occurrences of $C$.*

PROOF. We prove the claim for the deletion operation only, as the arguments for the case of an insertion are similar. Let $x$ be the node that is deleted from $t$. As mentioned in Section 5.3, there is a unique way to represent the deletion of $x$ using composition of trees and contexts, namely we can write

$$t = C'' \circ (t'''@(C' \circ a)),$$

where $a$ is the label of node $x$ and $C'$ is the *horizontal* context that represents the forest of subtrees under $x$. The result of the deletion of node $x$ gives the curried tree

$$t'' = C'' \circ (C' \circ t''').$$

Note that the deletion operation, performed on curry encodings, removes exactly two nodes: the $a$-labeled leaf that corresponds to $x$ and the @-labeled node $y$ that connects $t''$ to $C' \circ a$. All other nodes are preserved (but possibly rearranged) by this transformation. In particular, this means that if $(C, z)$ is an occurrence of the context $C$ in $t$ that does not overlap with the $a$-labeled node $x$ nor with the @-labeled node $y$, then $C$ occurs in either $C''$, $C'$, or $t'''$. This shows that $C$ occurs at least once in $t''$. In general, suppose that there are $n$ nonoverlapping occurrences of $C$ in $t$. Since the deletion operation affects only two nodes, $x$ and $y$, we have that, in the worst-case, all but two of these occurrences of $C$ can be found in $t''$, and hence $t''$ contains at least $n-2$ occurrences of $C$. Finally, it is easy to see that the deletion operation preserves the property of occurrences of being nonoverlapping. □

We continue now with the proof of Lemma 6.10. We consider a nontrivial node $x$ of the primitive synopsis tree $\sigma$ and let $X = \sigma(x)$ be the associated component. By definition of $t_\sigma$, we know that the context $C_X^H$ occurs in $t_\sigma$. We also recall that $H = m \cdot (2K+1)$. In particular, $t_\sigma$ contains $(2K+1)$ occurrences of the context $C_X^m$. As $t'_\sigma$ is obtained from $t_\sigma$ by applying at most $K$ edit operations to it, we know from the preceding claim that

the context $C_X^m$ occurs at least once in $t'_\sigma$. We denote by $f(x)$ some node of $t'_\sigma$ where $C_X^m$ occurs. We have just proved the first part of the lemma.

For second part, let $x$ and $y$ be two distinct nontrivial nodes of $\sigma$, and let $X = \sigma(x)$ and $Y = \sigma(y)$. Furthermore, let $f(x)$ and $f(y)$ be the nodes in $t'_\sigma$ where the contexts $C_X^m$ and $C_Y^m$ occurs. Note that the occurrences of these two contexts in $t_\sigma$ are nonoverlapping. Since deletion and insertion operations preserve the property of context occurrences of being nonoverlapping, we have that $(C_X^m, f(x))$ and $(C_Y^m, f(y))$ are also nonoverlapping occurrences in $t'_\sigma$.

Now suppose that $x \preccurlyeq_\sigma^{\mathsf{post}} y$. Let $x'$ and $y'$ be the nodes in $t_\sigma$ that carry the corresponding occurrences of the contexts $C_X^H$ and $C_Y^H$, respectively. It is routine to check, by exploiting the recursive definition of $t_\sigma$, that $x' \preccurlyeq_{t_\sigma}^{\mathsf{post}} y'$. In addition, we know that edit operations preserve the post-order relationships between nodes. As discussed previously, at least one occurrence of $C_X^m$ (respectively, $C_Y^m$) inside $C_X^H$ (respectively, $C_Y^H$) is not affected by the edit operations that transform $t_\sigma$ into $t'_\sigma$. This means that the corresponding occurrences $(C_X^m, f(x))$ and $(C_Y^m, f(y))$ in $t'_\sigma$ are in the same post-order relationship as $x$ and $y$, namely $f(x) \preccurlyeq_{t'_\sigma}^{\mathsf{post}} f(y)$. The converse implication follows from the fact that $\preccurlyeq_\sigma^{\mathsf{post}}$ is a total order (hence, it is sufficient to swap the roles of $x$ and $y$ above).

We finally check the last condition. Suppose that $X = \sigma(x)$ is a nonhorizontal component and that $x \preccurlyeq_\sigma^{\mathsf{anc}} y$. As before, let $x'$ and $y'$ be the nodes in $t_\sigma$ that carry the occurrences of $C_X^H$ and $C_Y^H$, respectively. Thanks to the construction of $t_\sigma$, we have $x' \preccurlyeq_\sigma^{\mathsf{anc}} y'$. Moreover, recall that during the construction of $t_\sigma$, we inserted $K$ copies of the nonhorizontal context $C_{\text{n.h.}}$ immediately below $C_X^H$ (and thus above $C_Y^H$). This implies that the path in $t_\sigma$ that connects the node $x'$ to its descendant $y'$ visits at least $K$ right edges. If we now look at the unranked tree $\mathsf{ext}^{-1}(t_\sigma)$ encoded by $t_\sigma$, we observe that there is a bijection between the right edges in $t_\sigma$ and the vertical edges in $\mathsf{ext}^{-1}(t_\sigma)$, and this bijection preserves the ancestor order. This means that the two portions of the unranked tree $\mathsf{ext}^{-1}(t_\sigma)$ that are encoded by $C_X^H$ and $C_Y^H$, namely $\mathsf{ext}^{-1}(C_X^H)$ and $\mathsf{ext}^{-1}(C_Y^H)$, are separated by at least $K$ vertical edges. Finally, each deletion or insertion operation performed on $\mathsf{ext}^{-1}(t_\sigma)$ can only bring two nodes closer by one level at a time. This means that after at most $K$ edit operations, the resulting curried tree $t'_\sigma$ contains the occurrence of $C_X^m$ at node $f(x)$ is still above the occurrence of $C_Y^m$ at node $f(y)$. We have just proved that $x \preccurlyeq_\sigma^{\mathsf{anc}} y$ implies $f(x) \preccurlyeq_{t'_\sigma}^{\mathsf{anc}} f(y)$. The converse implication follows by symmetric arguments. □

It now remains to show how to extract a basic synopsis tree $\tau$ that covers $\sigma$ from the tree $t'_\sigma$. Recall that $t'_\sigma \in \mathscr{L}(\mathcal{T})$, and let $\rho$ be an accepting run of $\mathcal{T}$ on $t'_\sigma$. Further, let $f$ be the mapping from nontrivial nodes of $\sigma$ to nodes of $t'_\sigma$, as defined in Lemma 6.10. Consider a nontrivial node $x$ of $\sigma$, and let $X$ be its label. We know that the context $C_X^m$ occurs at node $f(x)$ in $t'_\sigma$. Let $y$ be the node of the fingerprint context $C_X$ that is labeled with the placeholder symbol $\bullet$. Clearly, $C_X$ occurs in its $m$-fold iteration $C_X^m$ at positions $\varepsilon$, $y$, $y \cdot y, \ldots, y^{m-1}$. Analogous (nonoverlapping) occurrences exist in $t'_\sigma$, namely at positions $y_{x,0} = f(x), y_{x,1} = f(x) \cdot y, y_{x,2} = f(x) \cdot y \cdot y, \ldots, y_{x,m-1} = f(x) \cdot y^{m-1}$. For convenience, let $y_{x,m} = f(x) \cdot y^m$.

Next we consider the states that occur at the $m + 1$ nodes $y_{x,0}, y_{x,1}, \ldots, y_{x,m}$ of the run $\rho$ on $t'_\sigma$. By the pigeonhole principle, we know that two among these states, say $\rho(y_{x,i})$ and $\rho(y_{x,j})$ for some $0 \leq i < j \leq m$, belong to the same component $Y$ of $\mathcal{T}$. In fact, from the definition of strongly connected component, we can even assume that $j = i + 1$, from which we immediately obtain $C_X \in \mathscr{L}(\mathcal{T} \mid Y)$. It is now time to exploit

the property of the fingerprint context $C_X$, which is shown in Lemma 6.9. In particular, from the fact that $C_X \in \mathscr{L}(\mathcal{T} \mid Y)$, we derive that $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y)$.

We have just shown that it is possible to find a mapping from any nontrivial node $x$ of $\sigma$ to a node $y_x$ $(= y_{x,i})$ in $t'_\sigma$ such that $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y)$, where $X = \sigma(x)$ and $Y$ is the component of the state $\rho(y_x)$. Thanks to Lemma 6.10, we can also claim that for all nontrivial nodes $x, x'$ in $\sigma$,

—$x \neq x'$ implies that $y_x \neq y_{x'}$,
—$x \preccurlyeq^{\mathsf{post}}_{\sigma} x'$ if and only if $y_x \preccurlyeq^{\mathsf{post}}_{t'_\sigma} y_{x'}$, and
—$x \preccurlyeq^{\mathsf{anc}}_{\sigma} x'$ if and only if $y_x \preccurlyeq^{\mathsf{anc}}_{t'_\sigma} y_{x'}$, provided that the component $\sigma(x)$ is nonhorizontal.

Toward a conclusion, we now define the basic synopsis tree $\tau$ that covers $\sigma$. The domain of $\tau$ coincides with the domain of $t'_\sigma$ (i.e., $\mathsf{nodes}(\tau) = \mathsf{nodes}(t'_\sigma)$). The labeling function of $\tau$ maps every trivial node $x$ of $\tau$ to the trivial component $\varepsilon$ and every nontrivial node $x$ to the component that contains the state $\rho(y_x)$ associated with the corresponding node $y_x$ in $t'_\sigma$. It is easy to see that $\tau$ satisfies the properties of a basic synopsis tree (in particular, it respects the transitions of $\mathcal{T}$ because its labeling is essentially the lifting of a valid run $\rho$ of $\mathcal{T}$). It remains to define the mapping $\lambda$ that witnesses the coverability of $\sigma$ by $\tau$: for this, we simply let $\lambda(x) = y_x$ for every nontrivial node $x$ of $\sigma$. The fact that $\lambda$ satisfies Definition 5.6 follows easily from the properties described by the preceding three items (i.e., the fact that $\lambda$ is injective follows from the first item). This proves that every primitive synopsis tree of $\mathcal{S}$ is covered by a basic synopsis tree of $\mathcal{T}$.

## 7. COMPLEXITY ANALYSIS

In this section, we investigate the complexity of deciding whether a regular tree language $S$ is bounded repairable into a regular tree language $T$, and we assume that $S$ and $T$ are represented by automata $\mathcal{S}$ and $\mathcal{T}$, respectively. One can propose a straightforward decision procedure following the characterization of bounded repairability with synopsis trees (Theorem 5.7): for every primitive synopsis tree of $\mathcal{S}$, it suffices to guess a covering basic synopsis tree of $\mathcal{T}$. We also recall from Remark 5.3 that the size of a primitive synopsis tree of $\mathcal{S}$ is bounded by a function exponential in the number of strongly connected components of $\mathcal{S}$. Hence, by Lemma 6.8, an analogous bound holds for the size of basis synopsis trees of $\mathcal{T}$. It is also easy to see that testing the covering of a primitive synopsis tree by a basic synopsis tree can be performed efficiently in the size of the synopsis trees. These observations show that deciding bounded repairability for languages represented by tree automata is in ExpSpace.

We show, however, that a more efficient procedure exists: rather than inspecting individual elements of $\mathsf{PST}(\mathcal{S})$ and verifying that they are covered by elements of $\mathsf{BST}(\mathcal{T})$, it checks inclusion of the sets of *normalized* synopsis trees. More precisely, we first relabel synopsis trees in $\mathsf{BST}(\mathcal{T})$ with compatible connected components of $\mathcal{S}$, and as a result, we deal with synopsis trees labeled with elements of $\mathsf{SCC}(\mathcal{S})$. Next, we define a *serialization* of a synopsis tree, a string representation of the synopsis tree, and show that serialization of a synopsis tree is the same as the serialization of the normal form of the synopsis tree. Naturally, this reduces bounded repairability to the inclusion of serializations of $\mathsf{PST}(\mathcal{S})$ and $\mathsf{BST}(\mathcal{T})$, respectively. Testing this inclusion is not trivial, because although both sets $\mathsf{PST}(\mathcal{S})$ and $\mathsf{BST}(\mathcal{T})$ can be captured with tree automata, the serializations versions are string languages that need not be regular. We show, however, that serializations of flattened versions of $\mathsf{PST}(\mathcal{S})$ and $\mathsf{BST}(\mathcal{T})$ can be captured with context-free grammars and, moreover, that the context-free grammar for $\mathsf{PST}(\mathcal{S})$ is nonrecursive. We then use existing results on testing inclusion of nonrecursive context-free grammar in another (possibly) recursive context-free grammar and obtain a coNExp upper bound. Finally, we show that the coNExp upper bound is

tight even if we restrict the tree languages provided as input to those definable with deterministic nonrecursive DTDs.

### 7.1. Upper Bound

We show that the complexity of testing bounded repairability between two regular tree languages represented with tree automata is in coNExp. For the remainder of this section, we fix automata $\mathcal{S}$ and $\mathcal{T}$ that recognize the source tree language and the target tree language, respectively.

We begin by recalling the notion of embedding from Section 6.1. Given a synopsis tree $\theta$ of $\mathcal{S}$ and a synopsis tree $\tau$ of $\mathcal{T}$, we say that $\theta$ *is embedded into* $\tau$, denoted $\theta \leftharpoondown \tau$, if $\theta$ and $\tau$ have the same domain (i.e., $\mathsf{nodes}(\theta) = \mathsf{nodes}(\tau)$), and that $\theta$ is covered by $\tau$ via the identity function (i.e., $\mathscr{L}(\mathcal{S} \mid \theta(x)) \subseteq \mathscr{L}(\mathcal{T} \mid \tau(x))$ for all nodes $x$). We define the set

$$\mathsf{Emb}_{\mathcal{S}}(\tau) = \{\theta \text{ synopsis tree of } \mathcal{S} : \theta \leftharpoondown \tau\}$$

of all synopsis trees of $\mathcal{S}$ that are embedded into $\tau$, and we extend the notation to any set $S$ of synopsis trees by letting $\mathsf{Emb}_{\mathcal{S}}(S) = \bigcup_{\tau \in S} \mathsf{Emb}_{\mathcal{S}}(\tau)$.

Now we introduce a variant of the notion of serialization for synopsis trees, and we show that this can be used as an alternative representation of the normal form that we introduced in Section 6. Such a serialization takes a synopsis tree $\theta$ and produces a well-nested word $\hat{\theta}$ over the alphabet $\mathsf{tags}(\mathcal{S})$ that consists of opening tags of the form $\langle X \rangle$ and closing tags of the form $\langle /X \rangle$, with $X \in \mathsf{SCC}(\mathcal{S})$. It is important to remark that the serialization $\hat{\theta}$ does not represent the specific tree $\theta$, but rather the class of synopsis trees that have the same normal form as $\theta$. Formally, we define the *serialization* $\hat{\theta}$ *of a synopsis tree* $\theta$ *of* $\mathcal{S}$ recursively as follows:

$$\hat{\theta} = \begin{cases} \hat{\theta}_1 \cdot \hat{\theta}_2 & \text{if } \theta = X(\theta_1, \theta_2) \text{ and } X \text{ is a trivial component,} \\ \hat{\theta}_1 \cdot \hat{\theta}_2 \cdot \langle X \rangle \cdot \langle /X \rangle & \text{if } \theta = X(\theta_1, \theta_2) \text{ and } X \text{ is a nontrivial horizontal component,} \\ \langle X \rangle \cdot \hat{\theta}_1 \cdot \hat{\theta}_2 \cdot \langle /X \rangle & \text{if } \theta = X(\theta_1, \theta_2) \text{ and } X \text{ is a nonhorizontal component.} \end{cases}$$

Note that the trivial components disappear in the serialization $\hat{\theta}$ of a synopsis tree $\theta$. As usual, we extend serializations to sets of synopsis trees by letting $\hat{U} = \{\hat{\theta} : \theta \in U\}$.

It is easy to see that serializations are unaffected by the editing operations on synopsis trees that are used to attain the normal form.

Lemma 7.1. *Given two synopsis trees* $\theta$ *and* $\zeta$ *of* $\mathcal{S}$ *such that* $\theta \to_{\mathsf{op}}^{*} \zeta$, *we have* $\hat{\theta} = \hat{\zeta}$. *In particular, we have* $\hat{\theta} = \widehat{\theta^*}$, *where* $\theta^*$ *is the normal form of* $\theta$.

Proof. By induction, it suffices to show that applying a single editing operation does not change the serialization. A quick inspection of the definitions of the editing operations of promotion, demotion, and reduction (see Figure 11) shows that the claim holds trivially. □

The preceding lemma, together with the results proven in Section 6.1, implies the following.

Corollary 7.2. *The language* $\mathscr{L}(\mathcal{S})$ *is bounded repairable into the language* $\mathscr{L}(\mathcal{T})$ *if and only if* $\hat{U} \subseteq \hat{V}$, *where* $U = \mathsf{PST}(\mathcal{S})$ *and* $V = \mathsf{Emb}_{\mathcal{S}}(\mathsf{BST}(\mathcal{T}))$.

Proof. To prove the left-to-right implication, suppose that $\mathscr{L}(\mathcal{S})$ is bounded repairable into $\mathscr{L}(\mathcal{T})$, and consider a primitive synopsis tree $\sigma \in U$. We know from Theorem 5.7 that $\sigma$ is covered by some basic synopsis tree $\tau$ of $\mathcal{T}$. Moreover, by Lemma 6.2, there exists a synopsis tree $\theta$ of $\mathcal{S}$ that "interpolates" $\sigma$ and $\tau$, namely such that $\sigma$ is

strongly covered by $\theta$ (denoted $\sigma \hookrightarrow\!\!\!\!\!\rightarrow \theta$) and $\theta$ embedded into $\tau$ (denoted $\theta \leftrightharpoons \tau$). In particular, we have that $\theta$ belongs to the set $V = \mathsf{Emb}_{\mathcal{S}}(\tau)$. Moreover, since $\sigma \hookrightarrow\!\!\!\!\!\rightarrow \theta$, we know from Lemma 6.5 that the normal forms of $\sigma$ and $\theta$ coincide, and hence by Lemma 7.1 we have $\hat{\sigma} = \hat{\theta}$. We conclude that $\hat{\sigma}$ belongs to $\hat{V}$.

The proof of the converse direction is symmetric, namely we assume that $\hat{U} \subseteq \hat{V}$, we consider a primitive synopsis tree $\sigma$ of $\mathcal{S}$, and we prove that $\sigma$ is covered by some basic synopsis tree of $\mathcal{T}$. Indeed, since $\sigma \in U$ and $\hat{U} \subseteq \hat{V}$, we have $\hat{\sigma} \in \hat{V}$. This means that there exist a synopsis tree $\theta$ of $\mathcal{S}$ and a basic synopsis tree $\tau$ of $\mathcal{T}$ such that $\theta \leftrightharpoons \tau$ and $\hat{\theta} = \hat{\sigma}$. In particular, Lemma 7.1 implies that $\theta$ and $\sigma$ have the same normal form, and hence they strongly cover each other, from which $\sigma \hookrightarrow\!\!\!\!\!\rightarrow \theta \leftrightharpoons \tau$. We conclude that $\sigma$ is covered by $\tau$, and hence by Theorem 5.7, $\mathscr{L}(\mathcal{S})$ is bounded repairable into $\mathscr{L}(\mathcal{T})$. □

We conclude the section by showing how to effectively test the inclusion from Corollary 7.2. For this, we introduce some context-free grammars that capture the languages $\hat{U}$ and $\hat{V}$, where $U = \mathsf{PST}(\mathcal{S})$ and $V = \mathsf{Emb}_{\mathcal{S}}(\mathsf{BST}(\mathcal{T}))$. We can define these grammars on the basis of the components of $\mathcal{S}$ and $\mathcal{T}$ and the transitions of $\mathcal{S}$ and $\mathcal{T}$ lifted to these components. More precisely, the grammar $G_{\mathcal{S}}$ that defines the language $\hat{U}$ uses nonterminals $X, X_1, X_2, \ldots$ that correspond to components of $\mathcal{S}$ and rules of the forms

$$
\begin{aligned}
X &::= \langle X\rangle\ \langle /X\rangle && \\
X &::= X_1\ X_2 && \text{if } X \text{ is a trivial component,} \\
X &::= X_1\ X_2\ \langle X\rangle\ \langle /X\rangle && \text{if } X \text{ is a nontrivial horizontal component,} \\
X &::= \langle X\rangle\ X_1\ X_2\ \langle /X\rangle && \text{if } X \text{ is a nonhorizontal component,}
\end{aligned}
$$

where $X_1 \neq X \neq X_2$, $q \in \delta(q_1, q_2)$ for some $q \in X$, $q_1 \in X_1$, $q_2 \in X$, and $\delta$ is the transition function of $\mathcal{S}$.

The grammar $G_{\mathcal{S},\mathcal{T}}$ that defines the language $\hat{V}$ uses the same nonterminals $X, X_1, X_2 \in \mathsf{SCC}(\mathcal{S})$ and the same rules as earlier, but instead of enforcing $X_1 \neq X \neq X_2$ and $q \in \delta(q_1, q_2)$ for some $q \in X$, $q_1 \in X_1$, $q_2 \in X$, it requires that there exist some components $Y, Y_1, Y_2$ of the target automaton $\mathcal{T}$ such that (i) $\mathscr{L}(\mathcal{S} \mid X) \subseteq \mathscr{L}(\mathcal{T} \mid Y)$, (ii) $\mathscr{L}(\mathcal{S} \mid X_1) \subseteq \mathscr{L}(\mathcal{T} \mid Y_1)$, (iii) $\mathscr{L}(\mathcal{S} \mid X_2) \subseteq \mathscr{L}(\mathcal{T} \mid Y_2)$, and (iv) $q \in \gamma(q_1, q_2)$ for some $q \in Y$, $q_1 \in Y_1$, and $q_2 \in Y_2$, where $\gamma$ is the transition function of $\mathcal{T}$.

Although testing the inclusion of two generic context-free languages is known to be undecidable [Hopcroft and Ullman 1979], here we can exploit the fact that the grammar $G_{\mathcal{S}}$ is nonrecursive to decide the inclusion $\mathscr{L}(G_{\mathcal{S}}) \subseteq \mathscr{L}(G_{\mathcal{S},\mathcal{T}})$. Indeed, a nonrecursive grammar defines a finite language of words whose lengths are uniformly bounded by an exponent in the size of the grammar. Consequently, a nondeterministic Turing machine can guess a word $w$ of length exponential in the size of $G_{\mathcal{S}}$ and decide the noncontainment $\mathscr{L}(G_{\mathcal{S}}) \not\subseteq \mathscr{L}(G_{\mathcal{S},\mathcal{T}})$ by checking that $w \in \mathscr{L}(G_{\mathcal{S}})$ and $w \notin \mathscr{L}(G_{\mathcal{S},\mathcal{T}})$. It is also known that the membership problem of context-free languages can be solved in polynomial time [Younger 1967; Earley 1970]. The only subtlety here is that although the grammars $G_{\mathcal{S}}$ and $G_{\mathcal{S},\mathcal{T}}$ are of polynomial size with respect to $\mathcal{S}$ and $\mathcal{T}$, this reduction takes exponential time in the size of $\mathcal{S}$ and $\mathcal{T}$: indeed, the the definition of $G_{\mathcal{S},\mathcal{T}}$ requires checking some containment relationships between the languages recognized by the components of $\mathcal{S}$ and $\mathcal{T}$. The latter problem, however, is EXP-complete [Seidl 1990] and hence dominated by the time that is required to guess the word $w$. We can thus claim the following complexity upper bound to the bounded repair problem.

THEOREM 7.3. *The bounded repair problem between languages represented by stepwise tree automata is in* coNEXP.

### 7.2. Lower Bound

Here we show that the complexity bound established in Theorem 7.3 is tight. More precisely, we prove a matching coNEXP lower bound for the bounded repair problem, which remarkably holds even for tree languages represented by nonrecursive deterministic DTDs.

We recall the results in Champavère et al. [2009], particularly Proposition 4 and Theorem 5, which show that any deterministic DTD can be transformed, in polynomial time, into an equivalent deterministic stepwise automaton. This means that the complexity lower bound for the bounded repair problem of languages represented by (nonrecursive) deterministic DTDs can be immediately transferred to languages represented by deterministic stepwise automata.

We also recall the folklore PSPACE upper bound for the containment problem of nondeterministic DTDs: given two DTDs $D$ and $D'$, one can decide whether the language defined by $D$ is contained in the language defined by $D'$ by first removing the useless rules and then checking that for all letters $a$ in the alphabet of $D$, the regular language associated with $a$ in $D$ is contained in the regular language associated with $a$ in $D'$. This upper bound result is tight due to the PSPACE-hardness of containment of regular expressions [Stockmeyer and Meyer 1973]. We finally observe that the complexity of the containment problem lowers to P as soon as deterministic DTDs are considered. Interestingly, the situation is completely different for the complexity of the bounded repair problem.

THEOREM 7.4. *The bounded repair problem between languages represented by nonrecursive deterministic DTDs is* coNEXP*-hard.*

PROOF. The proof is by a reduction from the problem of tiling a square grid of exponential size [Boas 1997]. An instance of the latter problem is given by a tuple $I = (n, S, H, V, s_\bot, s_\top)$, where $n$ is a natural number encoded in unary and representing the width $2^n$ of the square grid, $S$ is a finite set of tiles, $H, V \subseteq S \times S$ are the set of vertical and horizontal constraints, and $s_\bot$ and $s_\top$ are the tiles that should mark the lower left and upper right corners. A tiling is a function $f$ mapping pairs $(i, j) \in \{1, \ldots, 2^n\} \times \{1, \ldots, 2^n\}$ to tiles $f(i, j) \in S$. We say that a tiling $f$ satisfies the constraints of $I$ if the following conditions are satisfied:

(1) $f(1, 1) = s_\bot$ and $f(2^n, 2^n) = s_\top$,
(2) $(f(i, j-1), f(i, j)) \in H$ for all $1 \leq i \leq 2^n$ and all $1 < j \leq 2^n$, and
(3) $(f(i-1, j), f(i, j)) \in V$ for all $1 < i \leq 2^n$ and all $1 \leq j \leq 2^n$.

The exponential tiling problem is the problem of deciding whether there exists a tiling $f$ that satisfies all constraints in a given instance $I$. This problem is known to be NEXP-complete [Boas 1997].

Now we fix an instance $I = (n, S, H, V, s_\bot, s_\top)$ of the exponential tiling problem, and we construct some regular languages $S, T$ of unranked trees such that $S$ is bounded repairable into $T$ if and only if there is *no tiling* satisfying the constraints of $I$. The basic idea is to let the source language $S$ contain encodings of the possible tilings and the target language $T$ contain modified encodings that expose violations of some constraint of $I$. The intended relation between $S$ and $T$ can be phrased as follows: if every tree in $S$ can be transformed into a tree in $T$ with a small (i.e., uniformly bounded) amount of edits, then every tiling of the exponential grid violates some constraint in $I$ and vice versa. To forbid a repair strategy to modify the encoded tiling with a bounded amount of edits, we will allow some redundancy in our encodings. For convenience, we first describe the languages $S$ and $T$ as if they were given by means of stepwise automata of polynomial size with respect to the instance $I$. Toward the end of the proof, we will

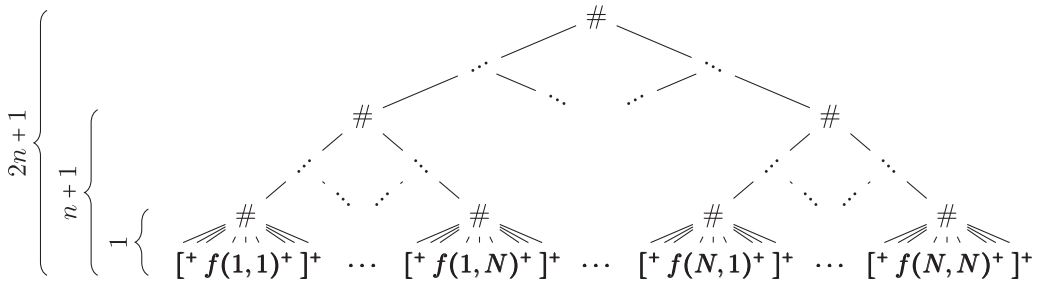



Fig. 13. Redundant encoding of a tiling by an unranked tree.

show how to modify the constructions to get languages representable by nonrecursive deterministic DTDs of polynomial size.

*Source language*. We begin by describing the trees in the language $S$. For the sake of the brevity, we let $N = 2^n$ be the width of the grid to be tiled and consider a generic tiling $f : \{1, \ldots, N\} \times \{1, \ldots, N\} \to S$. A tree that encodes the tiling $f$ is labeled over an alphabet consisting of tiles in $S$, separator symbols [and], and a dummy symbol #. Each cell $(i, j)$ in the grid is encoded by a series of consecutive leaves that spell out a word of the form $[[\ldots[f(i,j)\ f(i,j)\ \ldots\ f(i,j)]\ldots]]$, where each symbol $f(i, j)$ occurs at least once and the square brackets are not necessarily well parenthesized. The repetitions of the symbols $f(i, j)$ are used to ensure robustness to any repair strategy of bounded cost. From now on, such repetitions will be simply represented by a superscript $+$. The above word encoding a cell $(i, j)$ is called a *cell block*. Cell blocks are then juxtaposed to form the frontier of a tree, following the left-to-right bottom-to-top order of the corresponding cells in the grid:

$$\underbrace{[^+ f(1,1)^+ ]^+ \ldots [^+ f(1,N)^+ ]^+}_{\text{row } 1}\ \underbrace{[^+ f(2,1)^+ ]^+ \ldots [^+ f(2,N)^+ ]^+}_{\text{row } 2}\ \ldots\ \underbrace{[^+ f(N,1)^+ ]^+ \ldots [^+ f(N,N)^+ ]^+}_{\text{row } N}$$

Finally, #-labeled internal nodes are introduced to guarantee that the frontier is well formed, namely it contains exactly $N$ rows, each one consisting of $N$ cell blocks. This can be done by enforcing the existence of $2n + 1$ levels above the frontier and by requiring that each internal node at level $\ell = 0 \ldots 2n - 1$ has exactly two children, and each internal node at level $2n$ has a cell block as childhood Figure 13).

The source language $S$ is defined as the set of all tree-shaped encodings of tilings $f$ that satisfy the first constraint of $I$, namely those tilings $f$ such that $f(1, 1) = s_\perp$ and $f(N, N) = s_\top$. The language $S$ is clearly regular. Furthermore, it is not difficult to construct a stepwise automaton $\mathcal{S}$ that recognizes $S$ and has size polynomial in $n$ and $|S|$, and hence also in $|I|$ (we omit the formal definition of such an automaton).

*Target language*. We now turn to the target language $T$, which intuitively contains encodings of $S$ modified in a suitable way so as to expose violations of horizontal or vertical constraints, which can then be checked by an automaton of small size.

We begin by analyzing the simpler case of a tiling $f$ that violates a horizontal constraint, say between two tiles $f(i, j - 1)$ and $f(i, j)$. Observe that in the frontier of every tree of $S$ that encodes $f$, the violating tiles are represented by two consecutive cell blocks of the form $[^+ f(i, j-1)^+ ]^+$ and $[^+ f(i, j)^+ ]^+$. It is then convenient to expose the violation at the least common ancestor of these two cell blocks, which must occur at some level $\ell \in \{n, \ldots, 2n - 1\}$. For example, this can be done by relabeling the least common ancestor with the pair $(f(i, j - 1), f(i, j))$ $(\notin H)$. In this case, the modified encoding looks like the unranked tree in Figure 14 (for the sake of clarity, we highlighted the cell blocks corresponding to the tiles that violate the horizontal constraint).

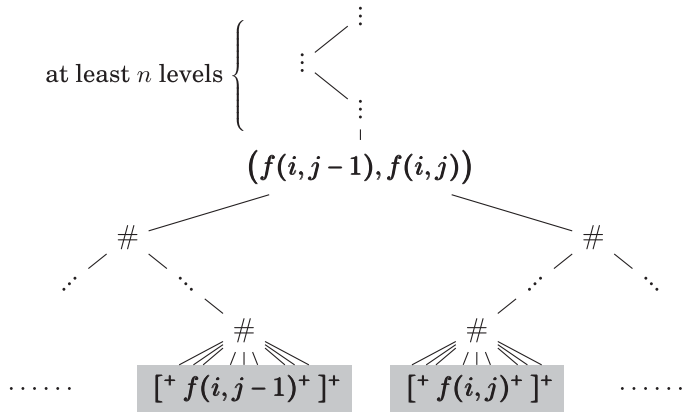



Fig. 14. Exposure of a violation of a horizontal constraint.

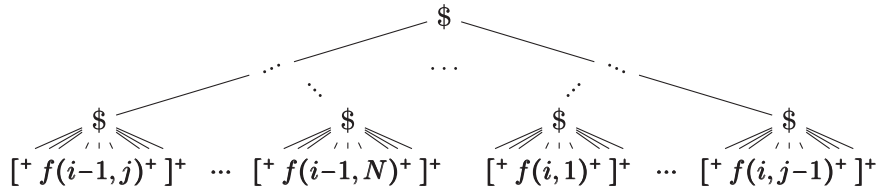



Fig. 15. Factor of a frontier delimited by vertically adjacent tiles.

We denote by $T_H$ the language of all trees that can be obtained by relabeling a node of some tree in $S$ as described earlier. We observe that the language $T_H$ is regular, and furthermore, one can construct a stepwise automaton that recognizes $T_H$ and has size polynomial in $|I|$.

We now deal with the case of a tiling $t$ that violates a vertical constraint, say between tiles $f(i-1, j)$ and $f(i, j)$. The basic idea here is to "hide" under a new subtree the factor of the frontier that starts with the corresponding occurrence of the cell block $[^+ f(i-1, j)^+ ]^+$ and ends just before the occurrence of the cell block $[^+ f(i, j)^+ ]^+$. Note that this factor contains exactly $N$ cell blocks, so it can be hidden under a complete binary tree of height $n$, such as the one depicted in Figure 15.

Similarly, the remaining part of the frontier consists of $N-1$ sequences, each one containing $N$ cell blocks, so this shape can be enforced using an almost complete binary tree of height $2n$, where exactly one node at level $n$ (e.g., the right-most one) is a leaf. Putting all together, the modified encoding for the tiling $f$ has the shape depicted in Figure 16 (as before, we highlighted the cell-blocks corresponding to the tiles that violate the vertical constraint).

Accordingly, we define the language $T_V$ of all unranked trees of the preceding form for all possible choices of $1 < i \leq N$ and $1 \leq j \leq N$ such that $(f(i-1, j), f(i, j)) \notin V$. Note that the latter condition can be checked by a small automaton that compares the highlighted cell blocks in the figure. In particular, the language $T_V$ is recognized by a stepwise automaton of size polynomial in $|I|$.

We can finally construct the target language $T$ as the union of $T_H$ and $T_V$, and recall that this is also recognized by an automaton of polynomial size in $|I|$.

*Reduction*. Now we need to argue that $S$ is bounded repairable into $T$ if and only if every tiling of the exponential grid violates some constraint of $I$. We begin with the easier direction, which assumes that every tiling violates some constraint of $I$. Consider

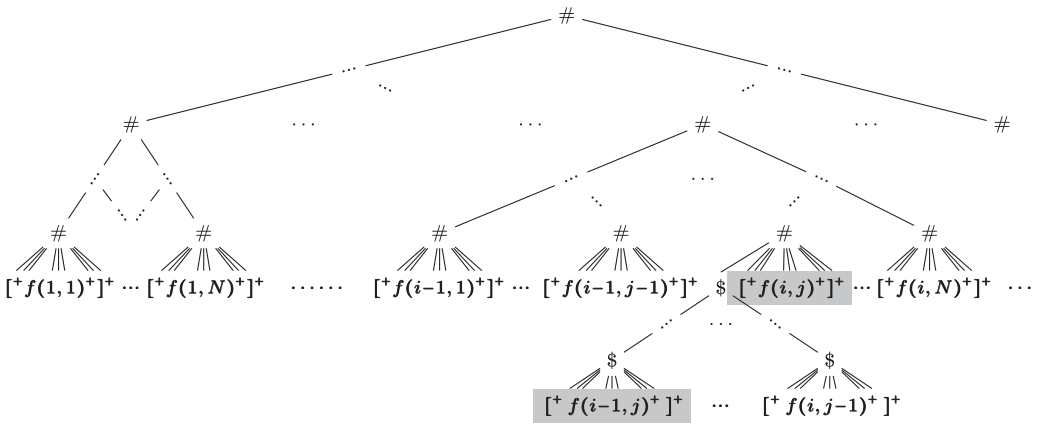



Fig. 16. Exposure of a violation of a vertical constraint.

a generic tree $t \in S$ that encodes a tiling $f$, and let $t' \in T$ be the modified encoding that exposes a violation of a horizontal or vertical constraint, as described earlier. We observe that the frontiers of $t$ and $t'$ spell out the same sequence of cell blocks. In particular, $t'$ can be obtained from $t$ by deleting all internal nodes and by inserting new internal nodes. Since the number of internal nodes in $t$ and $t'$ is uniformly bounded by a constant (roughly $\mathcal{O}(2^{2n})$), we know that $S$ is bounded repairable into $T$.

As for the other direction, suppose that there is a tiling $f$ that satisfies all constraints of $I$. We fix an arbitrarily large number $K$ and prove that some tree $t \in S$ requires at least $K$ edits to be transformed into a tree of $T$. The tree $t$ is nothing but the encoding of the tiling $f$, where each symbol in a cell block is repeated $K$ times; more precisely, $t$ is the unranked tree of Figure 13 where every superscript $+$ is replaced with $K$. We observe that every transformation of $t$ consisting of less than $K$ edits preserves at least one occurrence of each symbol in the frontier, and it also preserves the post-order relationships between these occurrences. Furthermore, note that occurrences of symbols [and] ensure that every transformation do not change or mix the order of the cell block. This means that every such transformation produces a tree whose frontier contains a subsequence that encodes the same tiling $f$ as $t$. Consider now a generic tree $t' \in T$. We observe that the frontier of $t'$ contains exactly $N^2$ cell blocks, so it encodes a tiling $f'$ of the exponential grid. Moreover, by definition of $T$, the tiling $f'$ must violate some constraint of $I$. We thus conclude that $f$ and $f'$ must be different tilings, and hence $t'$ cannot be obtained as an editing of $t$ with cost less than $K$. The preceding argument holds for any arbitrarily large number $K$, so this proves that $S$ is not bounded repairable into $T$.

*From automata to DTDs*. It remains to show how to modify the languages $S$ and $T$ in such a way that they can be succinctly described by nonrecursive deterministic DTDs. The general idea is to annotate the internal nodes of the trees in $S$ and $T$ with enough information so as to ease a deterministic top-down processing. First of all, we need to annotate the internal nodes of all trees of $S$ and $T$ with their levels: this is possible thanks to the fact that the considered trees have height at most $3n+2$. In addition, we mark the left-most and right-most paths of the trees of $S$ with special labels, say 1 and 2, respectively (the marking at the root is irrelevant): this makes it possible to check, by means of a DTD, that the first and last cell blocks are of the form $[^+ s_\perp^+ ]^+$ and $[^+ s_\top^+ ]^+$. As for the trees in $T$, the crux is to ease the certification of a violation of a horizontal or vertical constraint. To do so, we can promote the information about the violating

tiles up to the root. More precisely, on the trees depicted in Figure 14 and Figure 16, we consider the access path to the first highlighted cell block and annotate all internal nodes along this path with the corresponding tile; in a similar way, we add a second annotation for the access path to the second highlighted cell block. For example, the parent of the cell block $[^+ f(i,j)^+]^+$ of the tree of Figure 16 will be labeled with the tuple $(\#, 2n, f(i-1,j), f(i,j))$, where $2n$ is the level of that node, and $f(i-1,j)$ and $f(i,j)$ indicate the tiles corresponding to the first and second highlighted cell blocks.

The additional information on the labeling of the trees of $S$ and $T$ makes it easy to describe these languages by means of nonrecursive deterministic DTDs of size polynomial in $|I|$. Finally, because only internal nodes are affected by the new annotation, the same arguments for the proof of the reduction can be used here. □

Combining Theorems 7.3 and 7.4, we obtain that the bounded repair problem for tree languages represented by all standard specifications [Martens et al. 2006]—that is, stepwise tree automata, deterministic stepwise tree automata, XML schema, DTDs, nonrecursive deterministic DTDs—is coNExp-complete.

### 7.3. Simpler Instances

To find subcases of the bounded repair problem with a lower complexity, we consider a specialization of the problem where the alphabet $\Sigma$ of the source language is fixed. We show that in this case, the problem is PSpace-complete for languages represented by nondeterministic DTDs and coNP-complete for languages represented by deterministic DTDs.

Let us first discuss the complexity upper bounds. Suppose that $D$ is a DTD defining a source language over the fixed alphabet $\Sigma$. A close inspection to the translation from DTDs to stepwise automata [Champavère et al. 2009] discloses the following crucial property (see the Appendix for the proof).

Lemma 7.5. *Given a nondeterministic (respectively, deterministic) DTD $D$ that defines a source language $S$ over an alphabet $\Sigma$, one can compute in polynomial time a nondeterministic (respectively, deterministic) stepwise automaton $\mathcal{S} = (\Sigma, Q, \delta, \delta_0, F)$ that recognizes $S$ and whose state space can be partitioned into $k \leq 2|\Sigma|$ subsets $Q_1, \ldots, Q_k$ such that*

*—every component of $\mathcal{S}$ is contained in some set $Q_i$, and*
*—for all states $q_1, q_2, q \in Q$, if $q \in \delta(q_1, q_2)$ and $q_2$ and $q$ are in different components, then $q_2 \in Q_i$ and $q \in Q_j$ for some $1 \leq i < j \leq k$.*

For example, the automaton $\mathcal{S}$ described in Example 3.1 is a deterministic stepwise automaton whose state space can be partitioned into nine sets that satisfy the first part of the claim: $Q_1 = \{p_0^a\}$, $Q_2 = \{p_1^a\}$, $Q_3 = \{p_0^b\}$, $Q_4 = \{p_1^b\}$, $Q_5 = \{p_0^c\}$, $Q_6 = \{p_0^d\}$, $Q_7 = \{p_1^d\}$, $Q_8 = \{p_0^r, p_1^r\}$, $Q_9 = \{p_2^r\}$.

Lemma 7.5 implies that any path in the transition graph of $\mathcal{S}$ (i.e., see the left-hand side graph of Figure 7) traverses at most $2|\Sigma| - 1$ *vertical* edges that connect pairs of states in different components. As a consequence, any primitive synopsis tree of $\mathcal{S}$ has size at most $|Q|^{2|\Sigma|}$—that is, polynomial in the size of $\mathcal{S}$ when $\Sigma$ is fixed.

Putting together Lemma 7.5, Corollary 7.2, and Theorem 5.7, one obtains a PSpace (respectively, coNP) algorithm that decides whether $\mathsf{cost}(S, T) < \infty$, where $S$ and $T$ are languages defined by nondeterministic (respectively, deterministic) DTDs and $S$ is over a fixed alphabet $\Sigma$. The algorithm has the same structure of the algorithm sketched before Theorem 7.3. Namely, it translates the input DTDs into equivalent stepwise automata $\mathcal{S}$ and $\mathcal{T}$, then it translates $\mathcal{S}$ and $\mathcal{T}$ into the grammars $G_S$ and $G_{S,T}$, and finally it checks whether $\mathscr{L}(G_S) \subseteq \mathscr{L}(G_{S,T})$. As stated previously, the last step of the

algorithm can be done in coNP by universally guessing a word $w$ of polynomial size from $\mathscr{L}(G_{\mathcal{S}})$ and checking whether $w \in \mathscr{L}(G_{\mathcal{S},\mathcal{T}})$ in polynomial time. Note that the translation of $\mathcal{S}$ and $\mathcal{T}$ into $G_{\mathcal{S}}$ and $G_{\mathcal{S},\mathcal{T}}$ takes polynomial space for nondeterministic DTDs and polynomial time for deterministic DTDs. The blow-up of the complexity for the former is because one has to check language containment between regular languages, which can be done with polynomial space.

PROPOSITION 7.6. *The bounded repair problem between a source language represented by a nondeterministic (respectively, deterministic) DTD over a fixed alphabet and a target language represented by a nondeterministic (respectively, deterministic) DTD over an arbitrary alphabet is in* PSPACE *(respectively, in* coNP*).*

Finally, we show that even strong restrictions, including fixing both alphabets, cannot get us below PSPACE in the nondeterministic case. Indeed, one can easily reduce the containment problem between regular expressions to a bounded repair problem between languages defined by *nonrecursive nondeterministic* DTDs, thus showing that the latter problem is PSPACE-hard. To see this, consider two regular expressions $E_1$ and $E_2$. Let # be a fresh symbol, and let $E_1^{\#}$ and $E_2^{\#}$ be the expressions obtained from $E_1$ and $E_2$, respectively, by substituting every occurrence of a symbol $a$ with the expression $a^*\,\#$. Let $r$ be another fresh letter reserved for the roots of the trees. Clearly, the language defined by $E_1$ is contained in the language defined by $E_2$ if and only if the DTD $r \to E_1^{\#}$ is bounded repairable into the DTD $r \to E_2^{\#}$ (one direction is trivial and the other is easily shown by contraposition). As the latter DTDs are nonrecursive (they define trees of height two), this shows that the bounded repair problem between nonrecursive nondeterministic DTDs is PSPACE-hard, and this holds even when the alphabets are fixed.

We can also provide a coNP lower bound for the analogous problem when the languages are represented by nonrecursive *deterministic* DTDs over fixed alphabets. This lower bound follows easily from a reduction from the validity problem for propositional formulas in disjunctive form. A similar reduction was given in Benedikt et al. [2013] for languages of words recognized by deterministic finite automata. The additional complication here is that we have to fix the source and the target alphabets; however, the reduction is still possible by encoding the valuation of each variable with a block of nodes labeled over a binary alphabet.

PROPOSITION 7.7. *The bounded repair problem between languages represented by nonrecursive nondeterministic (respectively, deterministic) DTDs with both source and target alphabets fixed is* PSPACE*-hard (respectively,* coNP*-hard).*

## 8. THE UNIVERSAL CASE

In this section, we consider the so-called universal case of the bounded repairability problem, namely a variant of the problem where the source language is assumed universal (i.e., equal to $\mathcal{T}_\Sigma$) and the target language is represented by a stepwise automaton $\mathcal{T}$.

We recall the assumption that any stepwise automaton $\mathcal{T}$ is trimmed (i.e., every state of $\mathcal{T}$ appears in some accepting run of $\mathcal{A}$ on some input tree). Under this assumption, we say that an automaton $\mathcal{T}$ is *complete over* $\Sigma$ if for every tree $t \in \mathcal{T}_\Sigma$ there is a (possibly nonaccepting) run of $\mathcal{T}$ on $t$.

Here we also make use of deterministic visibly pushdown transducers [Raskin and Servais 2008; Alur and Madhusudan 2009] as suitable devices that transform unranked trees in a streaming fashion. These devices receive the serialized version of an unranked tree and output the serialized version of another unranked tree. By a slight abuse of notation, we identify unranked trees with their serializations.

The following result gives equivalent conditions for bounded repairability in the universal case.

PROPOSITION 8.1. *Given an alphabet $\Sigma$ and an automaton $\mathcal{T} = (\Delta, Q, \delta, \delta_0, F)$, the following conditions are equivalent:*

—*$\mathcal{T}_\Sigma$ is bounded repairable into $\mathcal{L}(\mathcal{T})$,*
—*$\mathcal{T}$ is complete over $\Sigma$, and*
—*there exist $k \in \mathbb{N}$ and a deterministic visibly pushdown transducer that receives any unranked tree $t$ over $\Sigma$ and outputs an unranked tree $t'$ such that $\mathsf{dist}(t, t') \leq k$ and $\mathsf{ext}(t') \in \mathcal{L}(\mathcal{T})$.*

PROOF. Here we only prove that the second item implies the third one (the other two directions are explained in the Appendix). Suppose that $\mathcal{T} = (\Delta, Q, \delta, \delta_0, F)$ is a (trimmed) stepwise automaton that is complete over $\Sigma$. It is not difficult to show that from the fact that $\mathcal{T}$ is complete over $\Sigma$, it follows that $\mathcal{T}_\Sigma$ is bounded repairable into $\mathcal{L}(\mathcal{T})$. The interesting result is that when we identify unranked trees with their serializations, the repair strategy of $\mathcal{T}_\Sigma$ into $\mathcal{L}(\mathcal{T})$ can be implemented by a deterministic visibly pushdown transducer. More specifically, the deterministic visibly pushdown transducer outputs, at the very first step and independently of the input, a fixed prefix of a serialized unranked tree (this represents a portion of the repaired tree); then it copies the input $t$ as a continuation of the prefix formerly constructed, mimicking at the same time the computation of the stepwise automaton $\mathcal{T}$ on $\mathsf{ext}(t)$; and finally, the transducer terminates by outputting a suitable suffix in such a way that the corresponding repaired tree belongs to the language $\mathsf{ext}^{-1}(\mathcal{L}(\mathcal{T}))$. The difficult part of this proof is to show that there is a single prefix that, no matter how it is prolonged, can be completed into a serialized tree that belongs to the language $\mathsf{ext}^{-1}(\mathcal{L}(\mathcal{T}))$. Namely, to complete the proof, we need to show the following claim.

CLAIM 2. *There are a symbol $a \in \Sigma$, a state $p \in Q$, and a sequence of unranked trees $u_1, \ldots, u_n$ over $\Sigma$ such that for every unranked tree $t$ over $\Sigma$, there is a sequence of unranked trees $v_1, \ldots, v_m$ over $\Sigma$ satisfying $p \in \delta(\mathsf{ext}(a(u_1, \ldots, u_n, t, v_1, \ldots, v_m)))$.*

We prove the claim by contraposition. Suppose that ($\star$) for every symbol $a \in \Sigma$, every state $p \in Q$, and every sequence $u_1, \ldots, u_n$ of trees, there exists a tree $t$ such that for every sequence of trees $v_1, \ldots, v_m$, $p \notin \delta(\mathsf{ext}(a(u_1, \ldots, u_n, t, v_1, \ldots, v_m)))$. We fix an arbitrary symbol $a \in \Sigma$ and an enumeration $p_1, \ldots, p_N$ of all states in $Q$. Then, by applying the hypothesis ($\star$) to the symbol $a$, to each state $p \in \{p_1, \ldots, p_N\}$, and to increasing sequences of trees $u_1, \ldots, u_n$, we construct a tree $t'$ over $\Sigma$ on which $\mathcal{T}$ has no valid run (this would imply that $\mathcal{T}$ is not complete over $\Sigma$). First, we let $p = p_1$ and $n = 0$, and we obtain from ($\star$) that there is a tree $t_1$ such that $p_1 \notin \delta(\mathsf{ext}(a(t_1, v_1, \ldots, v_m)))$ for all sequences of trees $v_1, \ldots, v_m$. Similarly, if we let $p = p_2$, $n = 1$, and $u_1 = t_1$, we know from ($\star$) that there is a tree $t_2$ such that $p_2 \notin \delta(\mathsf{ext}(a(t_1, t_2, v_1, \ldots, v_m)))$ for all sequences of trees $v_1, \ldots, v_m$. By applying a simple inductive argument, we can construct a sequence of trees $t_1, \ldots, t_N$ such that for every index $1 \leq i \leq N$, $p_i \notin \delta(\mathsf{ext}(a(t_1, \ldots, t_N)))$. Since $p_1, \ldots, p_N$ are all and only the states of $\mathcal{T}$, we derive that $\mathcal{T}$ has no valid run on $\mathsf{ext}(a(t_1, \ldots, t_N))$. This shows that $\mathcal{T}$ is not complete over $\Sigma$. □

From the preceding characterization, one can derive a polynomial-time algorithm that decides whether $\mathcal{T}_\Sigma$ is bounded repairable into $\mathcal{L}(\mathcal{T})$ when $\mathcal{T}$ is given by a *deterministic* stepwise automaton. For this, it is sufficient to turn $\mathcal{T}$ into a trimmed deterministic automaton $\mathcal{T}' = (\Sigma, Q', \delta', \delta'_0, F')$ over $\Sigma$ and then check that (i) for every symbol $a \in \Sigma$, $\delta'_0(a) \neq \varnothing$ and (ii) for every pair of states $q_1, q_2 \in Q'$, $\delta'(q_1, q_2) \neq \varnothing$. When the target language is represented by a *nondeterministic* stepwise automaton $\mathcal{T}$, the

complexity increases to EXP: one can simply determinize $\mathcal{T}$ and then use the decision procedure for the deterministic case.

As one could expect, the above complexity bounds (i.e., P for deterministic stepwise automata and EXP for nondeterministic stepwise automata) are tight. The hardness proofs can be derived from reductions of the emptiness and universality problems, respectively, on the corresponding classes of automata (see the Appendix).

PROPOSITION 8.2. *The bounded repair problem in the universal case when the target language is represented by a nondeterministic (respectively, deterministic) stepwise automaton is* EXP*-complete (respectively,* P*-complete).*

## 9. CONCLUSIONS

In this article, we have investigated the bounded repairability problem for regular tree languages. We have provided an effective characterization of bounded repairability and characterized the complexity of testing whether a given source language $S$ is bounded repairable with respect to a given target language $T$. The characterization can be used with several different formalisms for representing the tree languages: tree automata, XML schemas, and DTDs, as well as their nonrecursive and deterministic restrictions. Although generally the problem is coNEXP-complete, its complexity is considerably reduced for DTDs over fixed alphabets. In the latter case, the problem becomes coNP-complete or PSPACE-complete, depending on whether the DTDs are deterministic or not. Finally, we have also considered the variant of the problem when the source language is set to be universal. In this case, we have shown that the problems is EXP-complete in general and becomes tractable (P-complete, in fact) when a deterministic bottom-up automaton is used.

Several directions for future work can be envisioned. Bounded repairability is essentially a generalization of inclusion between tree languages modulo a bounded number of editing operations. One could attempt to further generalize it by allowing several editing operations that are bounded by a ratio of the size of the input tree. In Benedikt et al. [2014], it is shown how such a generalized notion of repairability can be computed for regular string languages. It would be interesting to see if the employed methods can be adapted to the setting of regular tree languages. Another direction is bounded repairability in the streaming setting: not only the pair of source and target languages need to be bounded repairable but also the repair must be executable by a transducer, namely a machine with a possibly infinite state space that makes one pass over the serialization of the input tree while producing a serialization of the output tree. Proposition 8.1 shows that in the unrestricted case, visibly pushdown transducers are expressive enough to implement bounded-cost streaming repairs whenever these exist. In the general case, however, visibly pushdown transducers may be too limiting [Bourhis et al. 2013].

## ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

## REFERENCES

Foto Afrati and Phokion Kolaitis. 2009. Repair checking in inconsistent databases: Algorithms and complexity. In *Proceedings of the 12th International Conference on Database Theory (ICDT'09)*. ACM, New York, NY, 31–41.

Alfred Aho and Thomas Peterson. 1972. A minimum distance error-correcting parser for context-free languages. *SIAM Journal on Computing* 1, 4, 305–312.

Rajeev Alur and Parthasarathy Madhusudan. 2009. Adding nesting structure to words. *Journal of the ACM* 56, 3, 16.

Shunichi Amano, Leonid Libkin, and Filip Murlak. 2009. XML schema mappings. In *Proceedings of the 28th Symposium on Principles of Database Systems*. 33–42.

Timos Antonopoulos, Floris Geerts, Wim Martens, and Frank Neven. 2013. Generating, sampling and counting subclasses of regular tree languages. *Theory of Computing Systems* 52, 3, 542–585.

Marcelo Arenas, Leopoldo Bertossi, and Jan Chomicki. 1999. Consistent query answers in inconsistent databases. In *Proceedings of the 18th ACM SIGMOD Symposium on Principles of Database Systems (PODS'99)*. ACM, New York, NY, 68–79.

Marcelo Arenas and Leonid Libkin. 2008. XML data exchange: Consistency and query answering. *Journal of the ACM* 55, 2, Article No. 7.

Charles Babbage. 1864. *Passages from the Life of a Philosopher*. Longman, Green, Longman, Roberts, & Green, London, England.

Michael Benedikt, Gabriele Puppis, and Cristian Riveros. 2013. Bounded repairability of word languages. *Journal of Computer and System Sciences* 79, 8, 1302–1321.

Michael Benedikt, Gabriele Puppis, and Cristian Riveros. 2014. The per-character cost of repairing word languages. *Theoretical Computer Science* 539, 38–67.

Leopoldo Bertossi. 2011. Database repairing and consistent query answering. *Synthesis Lectures on Data Management* 3, 5, 1–121.

Philip Bille. 2005. A survey on tree edit distance and related problems. *Theoretical Computer Science* 337, 1, 217–239.

Peter Van Emde Boas. 1997. The convenience of tilings. In *Complexity, Logic, and Recursion Theory*. Marcel Dekker, 331–363.

Mikołaj Bojańczyk, Leszek A. Kołodziejczyk, and Filip Murlak. 2011. Solutions in XML data exchange. In *Proceedings of the 14th International Conference on Database Theory*. ACM, New York, NY, 102–113.

Utsav Boobna and Michel de Rougemont. 2004. Correctors for XML data. In *Database and XML Technologies*. Lecture Notes in Computer Science, Vol. 3186. Springer, 97–111.

Pierre Bourhis, Gabriele Puppis, and Cristian Riveros. 2013. Which DTDs are streaming bounded repairable? In *Proceedings of the 16th International Conference on Database Theory (ICDT'13)*. ACM, New York, NY, 57–68.

Anne Brüggemann-Klein and Derick Wood. 1998. One-unambiguous regular languages. *Information and Computation* 142, 2, 182–206.

Julien Carme, Joachim Niehren, and Marc Tommasi. 2004. Querying unranked trees with stepwise tree automata. In *Rewriting Techniques and Applications*. Springer, 105–118.

Jérôme Champavère, Rémi Gilleron, Aurélien Lemay, and Joachim Niehren. 2009. Efficient inclusion checking for deterministic tree automata and XML schemas. *Information and Computation* 207, 11, 1181–1208.

Shan Chen, Dan Hong, and Vincent Y. Shen. 2005. An experimental study on validation problems with existing HTML Webpages. In *Proceedings of the 2005 International Conference on Internet Computing (ICOMP'05)*. 373.

Dario Colazzo, Giorgio Ghelli, Luca Pardini, and Carlo Sartiani. 2013. Almost-linear inclusion for XML regular expression types. *ACM Transactions on Database Systems* 38, 3, 15.

Hubert Comon, Max Dauchet, Rémi Gilleron, Christof Löding, Florent Jacquemard, Denis Lugiez, Sophie Tison, and Marc Tommasi. 2007. Tree Automata Techniques and Applications. Retrieved June 2, 2016, from http://www.grappa.univ-lille3.fr/tata.

G. Cormode and S. Muthukrishnan. 2007. The string edit distance matching problem with moves. *ACM Transactions on Algorithms* 3, 1, 2:1–2:19. DOI:http://dx.doi.org/10.1145/1186810.1186812

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the ACM* 13, 2, 94–102.

David Fallside and Priscilla Walmsley. 2004. *XML Schema Part 0: Primer Second Edition*. W3C Recommendation.

Wenfei Fan and Philip Bohannon. 2008. Information preserving XML schema embedding. *ACM Transactions on Database Systems* 33, 1, 4.

Sergio Flesca, Filippo Furfaro, Sergio Greco, and Ester Zumpano. 2005. Querying and repairing inconsistent XML data. In *Web Information Systems Engineering*. Lecture Notes in Computer Science, Vol. 3806. Springer, 175–188.

Gösta Grahne and Alex Thomo. 2004. Query answering and containment for regular path queries under distortions. In *Foundations of Information and Knowledge Systems*. Springer, 98–115.

Steven Grijzenhout and Maarten Marx. 2013. The quality of the XML Web. *Web Semantics: Science, Services and Agents on the World Wide Web* 19, 59–68.

John Hopcroft and Jeffrey Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Flip Korn, Barna Saha, Divesh Srivastava, and Shanshan Ying. 2013. On repairing structural problems in semi-structured data. *Proceedings of the VLDB Endowment* 6, 9, 601–612.

William Lidwell, Kritina Holden, and Jill Butler. 2010. *Universal Principles of Design, Revised and Updated: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions*. Rockport Publishers.

Wim Martens, Frank Neven, and Thomas Schwentick. 2009. Complexity of decision problems for XML schemas and chain regular expressions. *SIAM Journal on Computing* 39, 4, 1486–1530.

Wim Martens, Frank Neven, Thomas Schwentick, and Geert Jan Bex. 2006. Expressiveness and complexity of XML schema. *ACM Transactions on Database Systems* 31, 3, 770–813.

Wim Martens and Joachim Niehren. 2007. On the minimization of XML schemas and tree automata for unranked trees. *Journal of Computer and System Sciences* 73, 4, 550–583.

Ejike Ofuonye, Patricia Beatty, Scott Dick, and James Miller. 2010. Prevalence and classification of Web page defects. *Online Information Review* 34, 1, 160–174.

Gabriele Puppis, Cristian Riveros, and Slawek Staworko. 2012. Bounded repairability for regular tree languages. In *Proceedings of the 15th International Conference on Database Theory (ICDT'12)*. Berlin, Germany, 155–168.

Erhard Rahm and Philip A. Bernstein. 2001. A survey of approaches to automatic schema matching. *VLDB Journal* 10, 4, 334–350.

Jean-François Raskin and Frédéric Servais. 2008. Visibly pushdown transducers. In *Automata, Languages and Programming*. Lecture Notes in Computer Science, Vol. 5126. Springer, 386–397.

Thomas Schwentick. 2007. Automata for XML—a survey. *Journal of Computer and System Sciences* 73, 3, 289–315.

Luc Segoufin and Cristina Sirangelo. 2007. Constant-memory validation of streaming XML documents against DTDs. In *Proceedings of the 11th International Conference on Database Theory (ICDT'07)*. 299–313.

Luc Segoufin and Victor Vianu. 2002. Validating streaming XML documents. In *Proceedings of the 21st ACM SIGMOD Symposium on Principles of Database Systems (PODS'02)*. ACM, New York, NY, 53–64.

Helmut Seidl. 1990. Deciding equivalence of finite tree automata. *SIAM Journal on Computing* 19, 3, 424–437.

Slawomir Staworko and Jan Chomicki. 2006. Validity-sensitive querying of XML databases. In *Current Trends in Database Technology (EDBT Workshops)*. Lecture Notes in Computer Science, Vol. 4254. Springer, 164–177.

Slawomir Staworko, Emmanuel Filiot, and Jan Chomicki. 2008. Querying regular sets of XML documents. In *Proceedings of the International Workshop on Logic in Databases (LiD'08)*.

Larry Stockmeyer and Albert Meyer. 1973. Word problems requiring exponential time (preliminary report). In *Proceedings of the 15th Annual ACM Symposium on Theory of Computing (STOC'73)*. ACM, New York, NY, 1–9.

Nobutaka Suzuki. 2005. Finding an optimum edit script between an XML document and a DTD. In *Proceedings of the 2005 ACM Symposium on Applied Computing*. ACM, New York, NY, 647–653.

Kuo-Chung Tai. 1979. The tree-to-tree correction problem. *Journal of the ACM* 26, 3, 422–433.

Robert Wagner. 1974. Order-n correction for regular languages. *Communications of the ACM* 17, 5, 265–268.

Daniel H. Younger. 1967. Recognition and parsing of context-free languages in time n^3. *Information and Control* 10, 2, 189–208.

Received November 2014; revised October 2015; accepted March 2016