Optimizing Schema Languages for XML: Numerical Constraints and Interleaving

OPTIMIZING SCHEMA LANGUAGES FOR XML: NUMERICAL CONSTRAINTS AND INTERLEAVING[†]

WOUTER GELADE[‡¶], WIM MARTENS[§], AND FRANK NEVEN[‡]

Abstract. The presence of a schema offers many advantages in processing, translating, querying, and storage of XML data. Basic decision problems like equivalence, inclusion, and non-emptiness of intersection of schemas form the basic building blocks for schema optimization and integration, and algorithms for static analysis of transformations. It is thereby paramount to establish the exact complexity of these problems. Most common schema languages for XML can be adequately modeled by some kind of grammar with regular expressions at right-hand sides. In this paper, we observe that apart from the usual regular operators of union, concatenation and Kleene-star, schema languages also allow numerical occurrence constraints and interleaving operators. Although the expressiveness of these operators remain within the regular languages, their presence or absence has significant impact on the complexity of the basic decision problems. We present a complete overview of the complexity of the basic decision problems for DTDs, XSDs and Relax NG with regular expressions incorporating numerical occurrence constraints and interleaving. We also discuss chain regular expressions and the complexity of the schema simplification problem incorporating the new operators.

Key words. XML schema languages, complexity, optimization, regular expressions

AMS subject classifications. 68P15, 68Q45, 68Q17

1. Introduction. XML is the lingua franca for data exchange on the Internet [1]. Within applications or communities, XML data is usually not arbitrary but adheres to some structure imposed by a schema. The presence of such a schema not only provides users with a global view on the anatomy of the data, but far more importantly, it enables automation and optimization of standard tasks like (*i*) searching, integration, and processing of XML data (cf., e.g., [12, 23, 26, 44]); and, (*ii*) static analysis of transformations (cf., e.g., [2, 17, 27, 34]). Decision problems like equivalence, inclusion and non-emptiness of intersection of schemas, hereafter referred to as *the basic decision problems*, constitute essential building blocks in solutions for the just mentioned optimization and static analysis problems. Additionally, the basic decision problems are fundamental for schema minimization (cf., e.g., [10, 30]). Because of their widespread applicability, it is therefore important to establish the exact complexity of the basic decision problems for the various XML schema languages.

The most common schema languages for XML are DTD, XML Schema [40], and Relax NG [9] and can be modeled by grammar formalisms [33]. In particular, DTDs correspond to context-free grammars with regular expressions (REs) at right-hand sides, while Relax NG is abstracted by extended DTDs (EDTDs) [35] or equivalently, unranked tree automata [6], defining the regular unranked tree languages. While XML Schema is usually abstracted by unranked tree automata as well, recent results indicate that XSDs correspond to a strict subclass of the regular tree languages and are much closer to DTDs than to tree automata [29]. In fact, they can be abstracted by single-type EDTDs. As detailed in [28], the relationship between schema formalisms

[†]A preliminary version of this work was presented at the 11th International Conference on Database Theory, Barcelona, Spain, 2007.

[‡]Hasselt University and Transnational University of Limburg, School for Information Technology, firstname.lastname@uhasselt.be

[§]Universität Dortmund, Department of Computer Science, wim.martens@udo.edu

[¶]Research Assistant of the Fund for Scientific Research - Flanders (Belgium).

shop	$\rightarrow$	regular* & discount-box*
regular	$\rightarrow$	cd
discount-box	$\rightarrow$	cd$^{[10,12]}$ price
cd	$\rightarrow$	artist & title & price

FIG. 1.1. *A sample schema using the numerical occurrence and interleave operators. The schema defines a shop that sells CDs and offers a special price for boxes of 10–12 CDs.*

and grammars provides direct upper and lower bounds for the complexity of the basic decision problems.

A closer inspection of the various schema specifications reveals that the above abstractions in terms of grammars with regular expressions is too coarse. Indeed, in addition to the conventional regular expression operators like concatenation, union, and Kleene-star, the XML Schema and the Relax NG specification allow two other operators as well:

(1) Both the XML Schema and the Relax NG specification allow a certain form of unordered concatenation: the `ALL` and the `interleave` operator, respectively. This operator is actually the resurrection of the &-operator from SGML DTDs that was excluded from the definition of XML DTDs. Although there are restrictions on the use of `ALL` and `interleave`, we consider the operator in its unrestricted form. We refer by RE(&) to such regular expressions with the interleaving operator.

(2) The XML Schema specification allows to express numerical occurrence constraints which define the minimal and maximal number of times a regular construct can be repeated. We refer by RE(#) to such regular expressions with numerical occurrence constraints.

We illustrate these additional operators in Figure 1.1. Their formal definition is given in Section 2. Although the new operators can be expressed by the conventional regular operators, they cannot do so succinctly [14], which has severe implications on the complexity of the basic decision problems.

The goal of this paper is to study the complexity of the basic decision problems for DTDs, XSDs, and Relax NG with regular expressions extended with interleaving and numerical occurrence constraints. The latter class of regular expressions is denoted by RE(#, &). As observed in Section 5, the complexity of inclusion and equivalence of RE(#, &) expressions (and subclasses thereof) carries over to DTDs and single-type EDTDs. We therefore first establish the complexity of the basic decision problems for RE(#, &) expressions and frequently occurring subclasses. These results are summarized in Table 1.1 and Table 4.1. Of independent interest, we introduce NFA(#, &)s, an extension of NFAs with counter and split/merge states for dealing with numerical occurrence constraints and interleaving operators. Finally, we revisit the simplification problem introduced in [29] for schemas with RE(#, &) expressions. This problem is defined as follows: given an extended DTD, can it be rewritten into an equivalent DTD or a single-type EDTD?

In this paper, we do not consider deterministic or one-unambiguous regular expressions which form a strict subclass of the regular expressions [7]. The reason is two-fold. First of all, one-unambiguity is a highly debatable constraint (cf., e.g., pg 98 of [42] and [25, 39]) which is only required for DTDs and XML Schema, not for Relax NG. Actually, the only direct advantage of one-unambiguity is that it gives rise to PTIME algorithms for some of the basic decision problems for standard regular ex-

	INCLUSION	EQUIVALENCE	INTERSECTION
RE	PSPACE ([41])	PSPACE ([41])	PSPACE ([24])
RE(&)	EXPSPACE ([31])	EXPSPACE ([31])	**PSPACE**
RE(#)	EXPSPACE ([32])	EXPSPACE ([32])	**PSPACE**
RE(#, &)	**EXPSPACE**	**EXPSPACE**	**PSPACE**
NFA(#), NFA(&), and NFA(#, &)	**EXPSPACE**	**EXPSPACE**	**PSPACE**
DTDs with RE	PSPACE ([41])	PSPACE ([41])	PSPACE ([24])
DTDs with RE(#), RE(&), or RE(#, &)	**EXPSPACE**	**EXPSPACE**	**PSPACE**
single-type EDTDs with RE	PSPACE ([28])	PSPACE ([28])	EXPTIME ([28])
single-type EDTDs with RE(#), RE(&), or RE(#, &)	**EXPSPACE**	**EXPSPACE**	**EXPTIME**
EDTDs with RE	EXPTIME ([37])	EXPTIME ([37])	EXPTIME ([38])
EDTDs with RE(#), RE(&), or RE(#, &)	**EXPSPACE**	**EXPSPACE**	**EXPTIME**

TABLE 1.1

Overview of new and known complexity results. All results are completeness results. The new results are printed in bold.

pressions. The latter does not hold anymore for RE(#, &) expressions rendering the notion even less attractive. Indeed, already intersection for one-unambiguous regular expressions is PSPACE-hard [28] and inclusion for one-unambiguous RE(#) expressions is coNP-hard [19]. A second reason is that, in contrast to conventional regular expressions, one-unambiguity is not yet fully understood for regular expressions with numerical occurrence constraints and interleaving operators. Some initial results are provided by Bruggemann-Klein [5], and Kilpeläinen and Tuhkanen [22] who give algorithms for deciding one-unambiguity of RE(&) and RE(#) expressions, respectively. However, the results of Bruggemann-Klein are on the SGML interleaving operator, which is not the same as the Relax NG interleaving operator considered here. Furthermore, no study investigating the properties of these one-unambiguous languages has been undertaken. Such a study, although definitely relevant, is outside the scope of this paper.

Outline. In Section 2, we provide the necessary definitions. In Section 3, we define NFA(#, &). In Section 4 and Section 5, we establish the complexity of the basic decision problems for regular expressions and schema languages, respectively. We discuss simplification in Section 6. We conclude in Section 7.

2. Definitions.

2.1. Regular Expressions with Counting and Interleaving. For the rest of the paper, Σ always denotes a finite alphabet. A Σ-*symbol* (or simply symbol) is an element of Σ, and a Σ-*string* (or simply string) is a finite sequence $w = a_1 \cdots a_n$ of Σ-symbols. We define the length of w, denoted by $|w|$, to be n. We denote the empty string by ε. The set of *positions of* w is $\{1, \ldots, n\}$ and the *symbol of* w *at position* i is a_i. By $w_1 \cdot w_2$ we denote the *concatenation* of two strings w_1 and

w_2. For readability, we usually denote the concatenation of w_1 and w_2 by w_1w_2. The set of all strings is denoted by Σ^*. A *string language* is a subset of Σ^*. For two string languages $L, L' \subseteq \Sigma^*$, we define their concatenation $L \cdot L'$ to be the set $\{ww' \mid w \in L, w' \in L'\}$. We abbreviate $L \cdot L \cdots L$ (i times) by L^i. By $w_1 \,\&\, w_2$ we denote the set of strings that is obtained by *interleaving* or *shuffling* w_1 and w_2 in every possible way. That is, for $w, w_1, w_2 \in \Sigma^*$ and $a, b \in \Sigma$, $w \,\&\, \varepsilon = \varepsilon \,\&\, w = \{w\}$, and $a \cdot w_1 \,\&\, b \cdot w_2 = (\{a\} \cdot (w_1 \,\&\, b \cdot w_2)) \cup (\{b\} \cdot (a \cdot w_1 \,\&\, w_2))$. Here, $\cdot$ has precedence over &. The operator & is then extended to languages in the canonical way.

The set of *regular expressions* over Σ, denoted by RE, is defined in the usual way: ε, and every Σ-symbol is a regular expression; and when r and s are regular expressions, then rs, $r + s$, and r^* are also regular expressions. By RE(#, &) we denote RE extended with two new operators: *interleaving* and *numerical occurrence constraints*. That is, when r and s are RE(#, &) expressions then so are $r \,\&\, s$ and $r^{[k,\ell]}$ for $k, \ell \in \mathbb{N}$ with $k \leq \ell$ and $\ell > 0$. By RE(#) and RE(&), we denote RE extended only with counting and interleaving, respectively. Notice that we disallow $\emptyset$ as it does not occur in practical schema languages.

The language defined by a regular expression r, denoted by $L(r)$, is inductively defined as follows: $L(\varepsilon) = \{\varepsilon\}$; $L(a) = \{a\}$; $L(rs) = L(r) \cdot L(s)$; $L(r+s) = L(r) \cup L(s)$; $L(r^*) = \{\varepsilon\} \cup \bigcup_{i=1}^{\infty} L(r)^i$, $L(r^{[k,\ell]}) = \bigcup_{i=k}^{\ell} L(r)^i$; and, $L(r \,\&\, s) = L(r) \,\&\, L(s)$. The *size* of a regular expression r over Σ, denoted by $|r|$, is the number of Σ-symbols and operators occurring in r plus the sizes of the binary representations of the integers. By $r?$ and r^+, we abbreviate the expression $r + \varepsilon$ and rr^*, respectively. We assume familiarity with finite automata such as nondeterministic finite automata (NFAs) and deterministic finite automata (DFAs) [16].

2.2. Schema Languages for XML. The set of *unranked Σ-trees*, denoted by $\mathcal{T}_\Sigma$, is the smallest set of strings over Σ and the parenthesis symbols "(" and ")" such that, for $a \in \Sigma$ and $w \in (\mathcal{T}_\Sigma)^*$, $a(w)$ is in $\mathcal{T}_\Sigma$. So, a tree is either ε (empty) or is of the form $a(t_1 \cdots t_n)$ where each t_i is a tree. In the tree $a(t_1 \cdots t_n)$, the subtrees $t_1, \ldots, t_n$ are attached to the root labeled a. We write a rather than $a()$. Notice that there is no a priori bound on the number of children of a node in a Σ-tree; such trees are therefore *unranked*. For every $t \in \mathcal{T}_\Sigma$, the *set of nodes* of t, denoted by $\mathrm{Dom}(t)$, is the set defined as follows: (*i*) if $t = \varepsilon$, then $\mathrm{Dom}(t) = \emptyset$; and (*ii*) if $t = a(t_1 \cdots t_n)$, where each $t_i \in \mathcal{T}_\Sigma$, then $\mathrm{Dom}(t) = \{\varepsilon\} \cup \bigcup_{i=1}^{n} \{iu \mid u \in \mathrm{Dom}(t_i)\}$. In the sequel, whenever we say tree, we always mean Σ-tree. A *tree language* is a set of trees.

We make use of the following definitions to abstract from the commonly used schema languages:

Definition 2.1. *Let $\mathcal{R}$ be a class of regular expressions over Σ.*

1. *A* DTD($\mathcal{R}$) *over Σ is a tuple (Σ, d, s_d) where d is a function that maps Σ-symbols to elements of $\mathcal{R}$ and $s_d \in \Sigma$ is the start symbol. For convenience of notation, we denote (Σ, d, s_d) by d and leave the start symbol s_d implicit whenever this cannot give rise to confusion.*
   A tree t satisfies *d if* (i) *$lab^t(\varepsilon) = s_d$ and,* (ii) *for every $u \in Dom(t)$ with n children, $lab^t(u1) \cdots lab^t(un) \in L(d(lab^t(u)))$. By $L(d)$ we denote the set of trees satisfying d.*
2. *An* extended DTD *(EDTD($\mathcal{R}$)) over Σ is a 5-tuple $D = (\Sigma, \Sigma', d, s, \mu)$, where Σ' is an alphabet of* types*, (Σ', d, s) is a DTD($\mathcal{R}$) over Σ', and μ is a mapping from Σ' to Σ.*
   A tree t then satisfies *an extended DTD if $t = \mu(t')$ for some $t' \in L(d)$. Here we abuse notation and let μ also denote its extension to define a homomor-*

phism on trees. Again, we denote by $L(D)$ the set of trees satisfying D. For ease of exposition, we always take $\Sigma' = \{a^i \mid 1 \leq i \leq k_a, a \in \Sigma, i \in \mathbb{N}\}$ for some natural numbers k_a, and we set $\mu(a^i) = a$.

3. *A* single-type EDTD *(EDTD$^{st}(\mathcal{R})$) over Σ is an EDTD($\mathcal{R}$) $D = (\Sigma, \Sigma', d, s, \mu)$ with the property that for every $a \in \Sigma'$, in the regular expression $d(a)$ no two types b^i and b^j with $i \neq j$ occur.*

We denote by EDTD, EDTD(#), EDTD(&), and EDTD(#,&), the classes EDTD (RE), EDTD(RE(#)), EDTD(RE(&)), and EDTD(RE(#, &)), respectively. The same notation is used for EDTD$^{\text{st}}$s and DTDs.

For clarity, we sometimes write $a \to r$ rather than $d(a) = r$ in examples and proofs. Following this notation, a simple example of an EDTD is the following:

shop^1	$\to$	$(\text{cd}^1 + \text{cd}^2)^*\text{cd}^2(\text{cd}^1 + \text{cd}^2)^*$	title^1	$\to$	ε
cd^1	$\to$	$\text{title}^1\ \text{price}^1$	price^1	$\to$	ε
cd^2	$\to$	$\text{title}^1\ \text{price}^1\ \text{discount}^1$	discount^1	$\to$	ε

Here, cd^1 defines ordinary CDs, while cd^2 defines CDs on sale. The rule for shop^1 specifies that there should be at least one CD on sale. Notice that the above EDTD is not a single-type EDTD as cd^1 and cd^2 occur in the same rule.

As explained in [33, 29], EDTDs and single-type EDTDs correspond to Relax NG and XML Schema, respectively.

2.3. Decision Problems. The following problems are fundamental to this paper.

DEFINITION 2.2. *Let $\mathcal{M}$ be a class of regular expressions, string automata, or extended DTDs. We define the following problems:*

- INCLUSION *for $\mathcal{M}$: Given two elements $e, e' \in \mathcal{M}$, is $L(e) \subseteq L(e')$?*
- EQUIVALENCE *for $\mathcal{M}$: Given two elements $e, e' \in \mathcal{M}$, is $L(e) = L(e')$?*
- INTERSECTION *for $\mathcal{M}$: Given an arbitrary number of elements $e_1, \ldots, e_n \in \mathcal{M}$, is $\bigcap_{i=1}^n L(e_i) \neq \emptyset$?*
- MEMBERSHIP *for $\mathcal{M}$: Given an element $e \in \mathcal{M}$ and a string or a tree f, is $f \in L(e)$?*

We recall the known results concerning the complexity of REs and EDTDs.

THEOREM 2.3.

(1) INCLUSION, EQUIVALENCE, *and* INTERSECTION *for REs are* PSPACE*-complete [24, 41].*

(2) INCLUSION *and* EQUIVALENCE *for RE(&) and RE(#) are* EXPSPACE*-complete [31, 32].*

(3) INCLUSION *and* EQUIVALENCE *for EDTDst are* PSPACE*-complete [28];* INTERSECTION *for EDTDst is* EXPTIME*-complete [28].*

(4) INCLUSION, EQUIVALENCE, *and* INTERSECTION *for EDTDs are* EXPTIME*-complete [37, 38].*

(5) MEMBERSHIP *for RE(&) is* NP*-complete [31].*

(6) MEMBERSHIP *for RE(#) is in* PTIME *[20].*

2.4. Relating decision problems for regular expressions to DTDs and single-type EDTDs. In [28] it was shown for any subclass of the REs that the complexity of INCLUSION and EQUIVALENCE is the same as the complexity of the corresponding problem for DTDs and single-type EDTDs. The same holds for INTERSECTION and DTDs. The proofs of these theorems carry over literally to RE(#, &).

We call a complexity class $\mathcal{C}$ *closed under positive reductions* if the following holds for every $O \in \mathcal{C}$. Let L' be accepted by a deterministic polynomial-time Turing machine M with oracle O (denoted $L' = L(M^O)$). Let M further have the property that $L(M^A) \subseteq L(M^B)$ whenever $L(A) \subseteq L(B)$. Then L' is also in $\mathcal{C}$. For a more precise definition of this notion we refer the reader to [15]. For our purposes, it is sufficient that important complexity classes like PTIME, NP, coNP, PSPACE, and EXPSPACE have this property, and that every such class contains PTIME.

PROPOSITION 2.4 ([28]). *Let $\mathcal{R}$ be a subclass of RE(#, &) and let $\mathcal{C}$ be a complexity class closed under positive reductions. Then the following are equivalent:*
(a) INCLUSION *for $\mathcal{R}$ expressions is in $\mathcal{C}$.*
(b) INCLUSION *for DTD($\mathcal{R}$) is in $\mathcal{C}$.*
(c) INCLUSION *for EDTDst($\mathcal{R}$) is in $\mathcal{C}$.*
The corresponding statement holds for EQUIVALENCE.

The previous proposition can be generalized to INTERSECTION of DTDs as well.

PROPOSITION 2.5 ([28]). *Let $\mathcal{R}$ be a subclass of RE(#, &) and let $\mathcal{C}$ be a complexity class which is closed under positive reductions. Then the following are equivalent:*
(a) INTERSECTION *for $\mathcal{R}$ expressions is in $\mathcal{C}$.*
(b) INTERSECTION *for DTD($\mathcal{R}$) is in $\mathcal{C}$.*

The above proposition does not hold for single-type EDTDs. Indeed, there is a class of regular expressions $\mathcal{R}'$ for which INTERSECTION is NP-complete while INTERSECTION for EDTD$^{\text{st}}$($\mathcal{R}'$) is EXPTIME-complete [28].

3. Automata for Occurrence Constraints and Interleaving. We introduce the automaton model NFA(#, &). In brief, an NFA(#, &) is an NFA with two additional features: *(i)* split and merge transitions to handle interleaving; and, *(ii)* counting states and transitions to deal with numerical occurrence constraints. The idea of split and merge transitions stems from Jędrzejowicz and Szepietowski [18]. Their automata are more general as they can express shuffle-closure which is not regular. Counting states are also used in the counter automata of Kilpeläinen and Tuhkanen [21], and Reuter [36] although these counter automata operate quite differently from NFA(#)s. Zilio and Lugiez [11] also proposed an automaton model that incorporates counting and interleaving by means of Presburger formulas. None of the cited papers consider the complexity of the basic decision problems of their model. We will use NFA(#, &)s to obtain complexity upper bounds in Sections 4 and 5.

For readability, we denote $\Sigma \cup \{\varepsilon\}$ by Σ_ε. We then define an NFA(#, &) as follows.

DEFINITION 3.1. *An NFA(#, &) is a 5-tuple $A = (Q, \Sigma, s, f, \delta)$ where*

- *Q is a finite set of states. To every $q \in Q$, we associate a lower bound $\min(q) \in \mathbb{N}$ and an upper bound $\max(q) \in \mathbb{N}$.*
- *$s, f \in Q$ are the start and final states, respectively.*
- *δ is the transition relation and is a subset of the union of the following sets:*

(1)	$Q \times \Sigma_\varepsilon \times Q$	*ordinary transition (resets the counter)*
(2)	$Q \times \{store\} \times Q$	*transition that does not reset the counter*
(3)	$Q \times \{split\} \times Q \times Q$	*split transition*
(4)	$Q \times Q \times \{merge\} \times Q$	*merge transition*

Let $\max(A) = \max\{\max(q) \mid q \in Q\}$ be the largest upper bound occurring in A. A *configuration* γ is a pair (P, α) where $P \subseteq Q$ is a set of states and $\alpha : Q \to \{0, \ldots, \max(A)\}$ is the value function mapping states to the value of their counter. For a state $q \in Q$, we denote by α_q the value function mapping q to 1 and every other state to 0. The initial configuration γ_s is $(\{s\}, \alpha_s)$. The final configuration γ_f is

$(\{f\}, \alpha_f)$. When α is a value function then $\alpha[q = 0]$ and $\alpha[q^{++}]$ denote the functions obtained from α by setting the value of q to 0 and incrementing the value of q by 1, respectively, while leaving all other values unchanged.

We now define the transition relation between configurations. Intuitively, the value of the state at which the automaton arrives is always incremented by one. When exiting a state, the state's counter is always reset to zero, except when we exit through a *counting transition*, in which case the counter remains the same. In addition, exiting a state through a non-counting transition is only allowed when the value of the counter lies between the allowed minimum and maximum. The latter, hence, ensures that the occurrence constraints are satisfied. *Split* and *merge transitions* start and close a parallel composition.

A configuration $\gamma' = (P', \alpha')$ *immediately follows* a configuration $\gamma = (P, \alpha)$ by reading $\sigma \in \Sigma_\varepsilon$, denoted $\gamma \rightarrow_{A,\sigma} \gamma'$, when one of the following conditions hold:

1. **(ordinary transition)** there is a $q \in P$ and $(q, \sigma, q') \in \delta$ such that $\min(q) \leq \alpha(q) \leq \max(q)$, $P' = (P - \{q\}) \cup \{q'\}$, and $\alpha' = \alpha[q = 0][q'^{++}]$. That is, A is in state q and moves to state q' by reading σ (note that σ can be ε). The latter is only allowed when the counter value of q is between the lower and upper bound. The state q is replaced in P by q'. The counter of q is reset to zero and the counter of q' is incremented by one.
2. **(counting transition)** there is a $q \in P$ and $(q, \text{store}, q') \in \delta$ such that $\alpha(q) < \max(q)$, $P' = (P - \{q\}) \cup \{q'\}$, and $\alpha' = \alpha[q'^{++}]$. That is, A is in state q and moves to state q' by reading ε when the counter of q has not reached its maximal value yet. The state q is replaced in P by q'. The counter of q is not reset but remains the same. The counter of q' is incremented by one.
3. **(split transition)** there is a $q \in P$ and $(q, \text{split}, q_1, q_2) \in \delta$ such that $\min(q) \leq \alpha(q) \leq \max(q)$, $P' = (P - \{q\}) \cup \{q_1, q_2\}$, and $\alpha' = \alpha[q = 0][{q_1}^{++}][{q_2}^{++}]$. That is, A is in state q and splits into states q_1 and q_2 by reading ε when the counter value of q is between the lower and upper bound. The state q in P is replaced by (split into) q_1 and q_2. The counter of q is reset to zero, and the counters of q_1 and q_2 are incremented by one.
4. **(merge transition)** there are $q_1, q_2 \in P$ and $(q_1, q_2, \text{merge}, q) \in \delta$ such that, for each $j = 1, 2$, $\min(q_j) \leq \alpha(q_j) \leq \max(q_j)$, $P' = (P - \{q_1, q_2\}) \cup \{q\}$, and $\alpha' = \alpha[q_1 = 0][q_2 = 0][q^{++}]$. That is, A is in states q_1 and q_2 and moves to state q by reading ε when the respective counter values of q_1 and q_2 are between the lower and upper bounds. The states q_1 and q_2 in P are replaced by (merged into) q, the counters of q_1 and q_2 are reset to zero, and the counter of q is incremented by one.

For a string w and two configurations γ, γ', we denote by $\gamma \Rightarrow_{A,w} \gamma'$ when there is a sequence of configurations $\gamma \rightarrow_{A,\sigma_1} \cdots \rightarrow_{A,\sigma_n} \gamma'$ such that $w = \sigma_1 \cdots \sigma_n$. The latter sequence is called a *run* when γ is the initial configuration γ_s. A string w is *accepted* by A iff $\gamma_s \Rightarrow_{A,w} \gamma_f$ with γ_f the final configuration. We usually denote $\Rightarrow_{A,w}$ simply by $\Rightarrow_w$ when A is clear from the context. We denote by $L(A)$ the set of strings accepted by A. The size of A, denoted by $|A|$, is $|Q| + |\delta| + \Sigma_{q \in Q} \log(\max(q))$. So, each $\max(q)$ is represented in binary.

An example of an NFA(#, &) defining $dvd^{[10,12]}$ & $cd^{[10,12]}$ is shown in Figure 3.1.

An NFA(#) is an NFA(#, &) without split and merge transitions. An NFA(&) is an NFA(#, &) without counting transitions. An NFA is an NFA(#) without counting transitions.

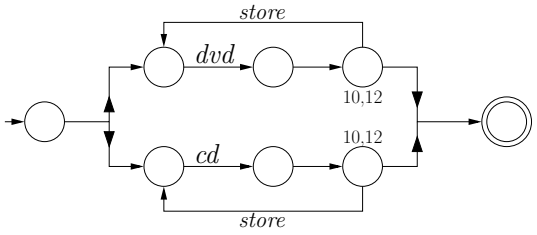

FIG. 3.1. *An NFA(#, &) for the language* $dvd^{[10,12]} \& cd^{[10,12]}$. *For readability, we only displayed the alphabet symbol on non-epsilon transitions and counters for states* q *where* $\min(q)$ *and* $\max(q)$ *are different from one. The arrows from the initial state and to the final state are split and merge transitions, respectively. The arrows labeled* store *represent counting transitions.*

Clearly NFA(#, &) accept all regular languages. The next theorem shows the complexity of translating between RE(#, &) and NFA(#, &), and NFA(#, &) and NFA.

THEOREM 3.2.

(1) Given an RE(#, &) expression r*, an equivalent NFA(#, &) can be constructed in time linear in the size of* r.

(2) Given an NFA(#, &) A*, an equivalent NFA can be constructed in time exponential in the size of* A.

Proof. (1) We prove the theorem by induction on the structure of RE(#, &)-expressions. For every r we define a corresponding NFA(#, &) $A(r) = (Q_r, \Sigma, s_r, f_r, \delta_r)$ such that $L(r) = L(A(r))$.

For r of the form ε, a, $r_1 \cdot r_2$, $r_1 + r_2$ and r_1^* these are the usual RE to NFA with ε-transition constructions as displayed in text books such as [16].

We perform the following steps for the numerical occurrence and interleaving operator which are graphically illustrated in Figure 3.2. The construction for the interleaving operator comes from [18].

(i) If $r = (r_1)^{[k,\ell]}$ and $A(r_1) = (Q_1, \Sigma, s_1, f_1, \delta_1)$, then
- $Q_r := Q_{r_1} \uplus \{s_r, f_r, q_r\}$;
- $\min(s_r) = \max(s_r) = \min(f_r) = \max(f_r) = 1$, $\min(q_r) = k$, and $\max(q_r) = \ell$;
- if $k \neq 0$ then $\delta_r := \delta_{r_1} \uplus \{(s_r, \varepsilon, s_{r_1}), (f_{r_1}, \varepsilon, q_r), (q_r, \text{store}, s_{r_1}), (q_r, \varepsilon, f_r)\}$; and,
- if $k = 0$ then $\delta_r := \delta_{r_1} \uplus \{(s_r, \varepsilon, s_{r_1}), (f_{r_1}, \varepsilon, q_r), (q_r, \text{store}, s_{r_1}), (q_r, \varepsilon, f_r), (s_r, \varepsilon, f_r)\}$.

(ii) If $r = r_1 \,\&\, r_2$, $A(r_1) = (Q_{r_1}, \Sigma, s_{r_1}, f_{r_1}, \delta_{r_1})$ and $A(r_2) = (Q_{r_2}, \Sigma, s_{r_2}, f_{r_2}, \delta_{r_2})$, then
- $Q_r := Q_{r_1} \uplus Q_{r_2} \uplus \{s_r, f_r\}$;
- $\min(s_r) = \max(s_r) = \min(f_r) = \max(f_r) = 1$;
- $\delta_r := \delta_{r_1} \uplus \delta_{r_2} \uplus \{(s_r, \text{split}, s_{r_1}, s_{r_2}), (f_{r_1}, f_{r_2}, \text{merge}, f_r)\}$.

Notice that in each step of the construction, a constant number of states are added to the automaton. Moreover, the constructed counters are linear in the size of r. It follows that the size of $A(r)$ is linear in the size of r. The correctness of the construction can easily be proved by induction on the structure of r. □

We next turn to the complexity of the basic decision problems for NFA(#, &).

THEOREM 3.3.

(1) EQUIVALENCE *and* INCLUSION *for NFA(#, &) are* EXPSPACE*-complete;*

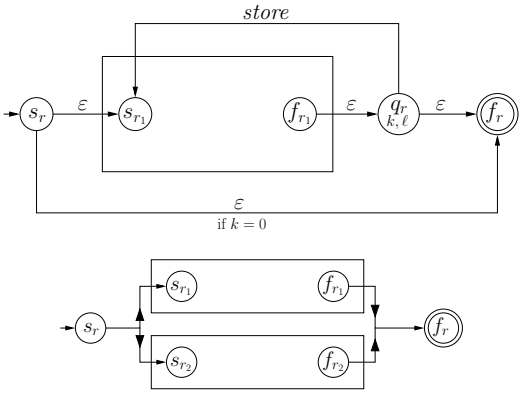

FIG. 3.2. *From RE(#, &) to NFA(#, &).*

(2) INTERSECTION *for NFA(#, &) is* PSPACE*-complete; and,*

(3) MEMBERSHIP *for NFA(#) is* NP*-hard,* MEMBERSHIP *for NFA(&) and NFA(#,&) are* PSPACE*-complete.*

Proof. (1) EXPSPACE-hardness follows from Theorem 3.2(1) and the EXPSPACE-hardness of EQUIVALENCE for RE(&) [31]. Membership in EXPSPACE follows from Theorem 3.2(2) and the fact that INCLUSION for NFAs is in PSPACE [41].

(2) The lower bound follows from [24]. We show that the problem is in PSPACE. For $j \in \{1, \ldots, n\}$, let $A_j = (Q_j, \Sigma, s_j, f_j, \delta_j)$ be an NFA(#, &). The algorithm proceeds by guessing a Σ-string w such that $w \in \bigcap_{j=1}^{n} L(A_j)$. Instead of guessing w at once, we guess it symbol by symbol and keep for each A_j one current configuration γ_j on the tape. More precisely, at each time instant, the tape contains for each A_j a configuration $\gamma_j = (P_j, \alpha_j)$ such that $\gamma_{s_j} \Rightarrow_{A_j, w_i} (P_j, \alpha_j)$, where $w_i = a_1 \cdots a_i$ is the prefix of w guessed up to now. The algorithm accepts when each γ_j is a final configuration. Formally, the algorithm operates as follows.

1. Set $\gamma_j = (\{s_j\}, \alpha_{s_j})$ for $j \in \{1, \ldots, n\}$;
2. While not every γ_j is a final configuration
   (*i*) Guess an $a \in \Sigma$.
   (*ii*) Non-deterministically replace each γ_j by a (P'_j, α'_j) such that $(P_j, \alpha_j) \Rightarrow_{A_j, a} (P'_j, \alpha'_j)$.

As the algorithm only uses space polynomial in the size of the NFA(#, &) and step (2,*ii*) can be done PSPACE, the overall algorithm operates in PSPACE.

(3) The MEMBERSHIP problem for NFA(#, &)s is easily seen to be in PSPACE by an on-the-fly implementation of the construction in Theorem 3.2(2). Indeed, as a configuration of an NFA(#, &) $A = (Q, \Sigma, s, f, \delta)$ has size at most $|Q| + |Q| \cdot \log(\max(A))$, we can store a configuration using only polynomial space.

We show that the MEMBERSHIP problem for NFA(#)s is NP-hard by a reduction from a modification of INTEGER KNAPSACK. We define this problem as follows. Given a set of natural numbers $W = \{w_1, \ldots, w_k\}$ and two integers m and n, all in binary notation, the problem asks whether there exists a mapping $\tau : W \to \mathbb{N}$ such that $m \leq \sum_{w \in W} \tau(w) \times w \leq n$. The latter mapping is called a solution. This problem is known to be NP-complete [13].

We construct an NFA(#) $A = (Q, \Sigma, s, f, \delta)$ such that $L(A) = \{\varepsilon\}$ if W, m, n has a solution, and $L(A) = \emptyset$ otherwise.

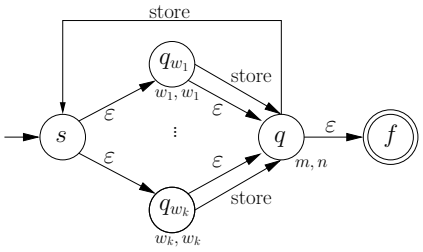

FIG. 3.3. NP-*hardness of* MEMBERSHIP *for NFA(#).*

The state set Q consists of the start and final states s and f, a state q_{w_i} for each weight w_i, and a state q. Intuitively, a successful computation of A loops at least m and at most n times through state q. In each iteration, A also visits one of the states q_{w_i}. Using numerical occurrence constraints, we can ensure that a computation accepts if and only if it passes at least m and at most n times through q and a multiple of w_i times through each q_{w_i}. Hence, an accepting computation exists if and only if there is a mapping τ such that $m \leq \sum_{w \in W} \tau(w) \times w \leq n$.

Formally, the transitions of A are the following:

- $(s, \varepsilon, q_{w_i})$ for each $i \in \{1, \ldots, k\}$;
- $(q_{w_i}, \text{store}, q)$ for each $i \in \{1, \ldots, k\}$;
- $(q_{w_i}, \varepsilon, q)$ for each $i \in \{1, \ldots, k\}$;
- (q, store, s); and,
- (q, ε, f).

We set $\min(s) = \max(s) = \min(f) = \max(f) = 1$, $\min(q) = m$, $max(q) = n$ and $\min(q_{w_i}) = \max(q_{w_i}) = w_i$ for each q_{w_i}. The automaton is graphically illustrated in Figure 3.3.

Finally, we show that MEMBERSHIP for NFA(&)s is PSPACE-hard. Before giving the proof, we describe some n-ary merge and split transitions which can be rewritten in function of the regular binary split and merge transitions.

1. $(q_1, q_2, \text{merge-split}, q'_1, q'_2)$: States q_1 and q_2 are read, and immediately split into states q'_1 and q'_2.
2. $(q_1, q_2, q_3, \text{merge-split}, q'_1, q'_2, q'_3)$: States q_1, q_2 and q_3 are read, and immediately split into states q'_1, q'_2 and q'_3.
3. $(q_1, \text{split}, q'_1, \ldots, q'_n)$: State q_1 is read, and is immediately split into states $q'_1, \ldots, q'_n$.
4. $(q_1, \ldots, q_n, \text{merge}, q'_1)$: States $q_1, \ldots, q_n$ are read, and are merged into state q'_1.

Transitions of type 1 (resp. 2) can be rewritten using 2 (resp. 4) *regular* transitions, and 1 (resp. 3) new auxiliary states. Transitions of type 3 and 4 can be rewritten using $(n-1)$ regular transitions and $(n-1)$ new auxiliary states. For example, the transition $(q_1, q_2, \text{merge-split}, q'_1, q'_2)$ is equal to the transitions $(q_1, q_2, \text{merge}, q_h)$, and $(q_h, \text{split}, q'_1, q'_2)$, where q_h is a new auxiliary state.

To show that MEMBERSHIP for NFA(&)s is PSPACE-hard, we reduce from CORRIDOR TILING. A *tiling instance* is a tuple $T = (X, H, V, \overline{b}, \overline{t}, n)$ where X is a finite set of tiles, $H, V \subseteq X \times X$ are the horizontal and vertical constraints, n is an integer in unary notation, and $\overline{b}, \overline{t}$ are n-tuples of tiles ($\overline{b}$ and $\overline{t}$ stand for *bottom row* and *top row*, respectively).

A *correct corridor tiling for* T is a mapping $\lambda : \{1, \ldots, m\} \times \{1, \ldots, n\} \to X$ for

some $m \in \mathbb{N}$ such that the following constraints are satisfied:

- the bottom row is $\overline{b}$: $\overline{b} = (\lambda(1,1), \ldots, \lambda(1,n))$;
- the top row is $\overline{t}$: $\overline{t} = (\lambda(m,1), \ldots, \lambda(m,n))$;
- all vertical constraints are satisfied: $\forall i < m$, $\forall j \leq n$, $(\lambda(i,j), \lambda(i+1,j)) \in V$; and,
- all horizontal constraints are satisfied: $\forall i \leq m$, $\forall j < n$, $(\lambda(i,j), \lambda(i,j+1)) \in H$.

The CORRIDOR TILING problem asks, given a tiling instance, whether there exists a correct corridor tiling. The latter problem is PSPACE-complete [43].

Given a tiling instance $T = (X, H, V, \overline{b}, \overline{t}, n)$, we construct an NFA(&) A over the empty alphabet ($\Sigma = \emptyset$) which accepts ε iff there exists a correct corridor tiling for T.

The automaton constructs the tiling row by row. Therefore, A must at any time reflect the current row in its state set. (recall that an NFA(&) can be in more than one state at once) To do this, A contains for every tile x, a set of states $x^1, \ldots, x^n$, where n is the length of each row. If A is in state x^i, this means that the ith tile of the current row is x. For example, if $\overline{b} = x_1x_3x_1$, and $\overline{t} = x_2x_2x_1$, then the initial state set is $\{x_1^1, x_3^2, x_1^3\}$, and A can accept when the state set is $\{x_2^1, x_2^2, x_1^3\}$.

It remains to define how A can transform the current row ("state set"), into a state set which describes a valid row on top of the current row. This transformation proceeds on a tile by tile basis and begins with the first tile, say x_i, in the current row which is represented by x_i^1 in the state set. Now, for every tile x_j, for which $(x_i, x_j) \in V$, we allow x_i^1 to be replaced by x_j^1, since x_j can be the first tile of the row on top of the current row. For the second tile of the next row, we have to replace the second tile of the current row, say x_k, by a new tile, say x_ℓ, such that the vertical constraints between x_k and x_ℓ are satisfied and such that the horizontal constraints between x_ℓ and the tile we just placed at the first position of the first row, x_j, are satisfied as well.

The automaton proceeds in this manner for the remainder of the row. For this, the automaton needs to know at any time at which position a tile must be updated. Therefore, an extra set of states $p_1, \ldots, p_n$ is created, where the state p_i says that the tile at position i has to be updated. So, the state set always consists of one state p_i, and a number of states which represent the current and next row. Here, the states up to position i already represent the tiles of the next row, the states from position i still represent the current row, and i is the next position to be updated.

We can now formally construct an NFA(&) $A = (Q, \Sigma, s, f, \delta)$ which accepts ε iff there exists a correct corridor tiling for a tiling instance $T = (X, H, V, \overline{b}, \overline{t}, n)$ as follows:

- $Q = \{x^j \mid x \in X, 1 \leq j \leq n\} \cup \{p_i \mid 1 \leq i \leq n\} \cup \{s, f\}$
- $\Sigma = \emptyset$
- δ is the union of the following transitions
  - $(s, \text{split}, p_1, \overline{b}_1^1, \ldots, \overline{b}_n^n)$: From the initial state the automaton immediately goes to the states which represent the bottom row.
  - $(p_1, \overline{t}_1^1, \ldots, \overline{t}_n^n, \text{merge}, f)$: When the state set represents a full row (the automaton is in state p_1), and the current row is the accepting row, all states are merged to the accepting state.
  - $\forall x_i, x_j \in X, (x_j, x_i) \in V$: $(p_1, x_j^1, \text{merge-split}, p_2, x_i^1)$: When the first tile has to be updated, the automaton only has to check the vertical constraints with the first tile of the previous row.

– $\forall x_i, x_j, x_k \in X, m \in \mathbb{N}, 2 \leq m \leq n, (x_k, x_i) \in V, (x_j, x_i) \in H$:
$(p_m, x_k^m, x_j^{m-1}, \text{merge-split}, p_{(m \mod n)+1}, x_i^m, x_j^{m-1})$: When a tile at the mth ($m \neq 1$) position has to be updated, the automaton has to check the vertical constraint with the mth tile at the previous row, and the horizontal constraint with the $(m-1)$th tile of the new row.

Clearly, if there exists a correct corridor tiling for T, there exists a run of A accepting ε. Conversely, the construction of our automaton, in which the updates are always determined by the position p_i, and the horizontal and vertical constraints, assures that when there is an accepting run of A on ε, this run simulates a correct corridor tiling for T. □

4. Complexity of Regular Expressions. Before we turn to schemas, we first deal with the complexity of regular expressions and frequently used subclasses as these are directly related to the complexities of DTDs and single-type EDTDs.

Mayer and Stockmeyer [31] and Meyer and Stockmeyer [32] already established the EXPSPACE-completeness of INCLUSION and EQUIVALENCE for RE(&) and RE(#), respectively. From Theorem 3.2(1) and Theorem 3.3(1) it then directly follows that allowing both operators does not increase the complexity. It further follows from Theorem 3.2(1) and Theorem 3.3(2) that INTERSECTION for RE(#, &) is in PSPACE. We stress that the latter results could also have been obtained without making use of NFA(#, &)s but by translating RE(#, &)s directly to NFAs. However, in the case of INTERSECTION such a construction should be done in an on-the-fly fashion to not go beyond PSPACE. Although such an approach certainly is possible, we prefer the shorter and more elegant construction using NFA(#, &)s.

THEOREM 4.1.

1. EQUIVALENCE *and* INCLUSION *for RE(#, &) are in* EXPSPACE*; and*

2. INTERSECTION *for RE(#, &) is* PSPACE*-complete.*

Proof. (1) Follows directly from Theorem 3.2(1) and Theorem 3.3(1).

(2) The upper bound follows directly from Theorem 3.2(1) and Theorem 3.3(2). The lower bound is already known for ordinary regular expressions. □

Bex et al. [4] established that the far majority of regular expressions occurring in practical DTDs and XSDs are of a very restricted form as defined next. The class of *chain regular expressions (CHAREs)* are those REs consisting of a sequence of factors $f_1 \cdots f_n$ where every factor is an expression of the form $(a_1 + \cdots + a_n)$, $(a_1 + \cdots + a_n)?$, $(a_1 + \cdots + a_n)^+$, or, $(a_1 + \cdots + a_n)^*$, where $n \geq 1$ and every a_i is an alphabet symbol. For instance, the expression $a(b + c)^* d^+ (e + f)?$ is a CHARE, while $(ab + c)^*$ and $(a^* + b?)^*$ are not.[1]

We introduce some additional notation to define subclasses and extensions of CHAREs. By CHARE(#) we denote the class of CHAREs where also factors of the form $(a_1 + \cdots + a_n)^{[k,\ell]}$ are allowed. For the following fragments, we list the admissible types of factors. Here, a, $a?$, a^* denote the factors $(a_1 + \cdots + a_n)$, $(a_1 + \cdots + a_n)?$, and $(a_1 + \cdots + a_n)^*$, respectively, with $n = 1$, while $a\#$ denotes $a^{[k,\ell]}$, and $a\#^{>0}$ denotes $a^{[k,\ell]}$ with $k > 0$.

Table 4.1 lists the new and the relevant known results. We first show that adding numerical occurrence constraints to CHAREs increases the complexity of INCLUSION by one exponential. We reduce from EXP-CORRIDOR TILING.

THEOREM 4.2. INCLUSION *for CHARE(#) is* EXPSPACE*-complete.*

Proof. The EXPSPACE upper bound already follows from Theorem 4.1(1).

[1] We disregard here the additional restriction used in [3] that every symbol can occur only once.

	INCLUSION	EQUIVALENCE	INTERSECTION
CHARE	PSPACE [28]	in PSPACE [41]	PSPACE [28]
CHARE(#)	**EXPSPACE**	**in EXPSPACE**	**PSPACE**
CHARE($a, a?$)	coNP [28]	in PTIME [28]	NP [28]
CHARE(a, a^*)	coNP [28]	in PTIME [28]	NP [28]
CHARE($a, a?, a\#$)	**coNP**	**in PTIME**	**NP**
CHARE($a, a\#^{>0}$)	**in PTIME**	**in PTIME**	**in PTIME**

TABLE 4.1

Overview of new and known complexity results concerning Chain Regular Expressions. All results are completeness results, unless otherwise mentioned. The new results are printed in bold.

The proof for the EXPSPACE lower bound is similar to the proof for PSPACE-hardness of INCLUSION for CHAREs in [28]. The main difference is that the numerical occurrence operator allows to compare tiles over a distance exponential in the size of the tiling instance.

The proof is a reduction from EXP-CORRIDOR TILING. A *tiling instance* is a tuple $T = (X, H, V, x_\bot, x_\top, n)$ where X is a finite set of tiles, $H, V \subseteq X \times X$ are the horizontal and vertical constraints, $x_\bot, x_\top \in X$, and n is a natural number in unary notation. A *correct exponential corridor tiling for* T is a mapping $\lambda : \{1, \ldots, m\} \times \{1, \ldots, 2^n\} \to X$ for some $m \in \mathbb{N}$ such that the following constraints are satisfied:

- the first tile of the first row is $x_\bot$: $\lambda(1, 1) = x_\bot$;
- the first tile of the m-th row is $x_\top$: $\lambda(m, 1) = x_\top$;
- all vertical constraints are satisfied: $\forall i < m, \forall j \leq 2^n, (\lambda(i, j), \lambda(i+1, j)) \in V$; and,
- all horizontal constraints are satisfied: $\forall i \leq m, \forall j < 2^n, (\lambda(i, j), \lambda(i, j+1)) \in H$.

The EXP-CORRIDOR TILING problem asks, given a tiling instance, whether there exists a correct exponential corridor tiling. The latter problem is easily shown to be EXPSPACE-complete [43].

We proceed with the reduction from EXP-CORRIDOR TILING. Thereto, let $T = (X, H, V, x_\bot, x_\top, n)$ be a tiling instance. Without loss of generality, we assume that $n \geq 2$. We construct two CHARE(#) expressions r_1 and r_2 such that

$L(r_1) \subseteq L(r_2)$ if and only if

there exists no correct exponential corridor tiling for T.

As EXPSPACE is closed under complement, the EXPSPACE-hardness of INCLUSION for CHARE(#) follows.

Set $\Sigma = X \uplus \{\$, \triangle\}$. For ease of exposition, we denote $X \cup \{\triangle\}$ by $X_\triangle$ and $X \cup \{\triangle, \$\}$ by $X_{\triangle,\$}$. We encode candidates for a correct tiling by a string in which the rows are separated by the symbol $\triangle$, that is, by strings of the form

$$\triangle R_0 \triangle R_1 \triangle \cdots \triangle R_m \triangle, \qquad (\dagger)$$

in which each R_i represents a row, that is, belongs to X^{2^n}. Moreover, R_0 is the bottom row and R_m is the top row. The following regular expressions detect strings of this form which do not encode a correct tiling for T:

- $X_\triangle^* \triangle X^{[0,2^n-1]} \triangle X_\triangle^*$. This expression detects rows that are too short, that is, contain less than 2^n symbols.

- $X^*_\triangle \triangle X^{[2^n+1,2^n+1]} X^* \triangle X^*_\triangle$. This expression detects rows that are too long, that is, contain more than 2^n symbols.
- $X^*_\triangle x_1 x_2 X^*_\triangle$, for every $x_1, x_2 \in X$, $(x_1, x_2) \notin H$. These expressions detect all violations of horizontal constraints.
- $X^*_\triangle x_1 X^{[2^n,2^n]}_\triangle x_2 X^*_\triangle$, for every $x_1, x_2 \in X$, $(x_1, x_2) \notin V$. These expressions detect all violations of vertical constraints.

Let $e_1, \ldots, e_k$ be an enumeration of the above expressions. Notice that $k = \mathcal{O}(|X|^2)$. It is straightforward that a string w in (†) does not match $\bigcup_{i=1}^k e_i$ if and only if w encodes a correct tiling.

Let $e = e_1 \cdots e_k$. Because of leading and trailing $X^*_\triangle$ expressions, $L(e) \subseteq L(e_i)$, for every $i \in \{1, \ldots, k\}$. We are now ready to define r_1 and r_2:

$$r_1 = \overbrace{\$e\$e\$\cdots\$e\$}^{k \text{ times } e} \triangle x_\perp X^{[2^n-1,2^n-1]} \triangle X^*_\triangle \triangle x_\top X^{[2^n-1,2^n-1]} \triangle \overbrace{\$e\$e\$\cdots\$e\$}^{k \text{ times } e}$$

$$r_2 = \$X^*_{\triangle,\$}\$e_1\$e_2\$\cdots\$e_k\$X^*_{\triangle,\$}\$$$

Notice that both r_1 and r_2 are in CHARE(#) and can be constructed in polynomial time. It remains to show that $L(r_1) \subseteq L(r_2)$ if and only if there is no correct tiling for T.

We first show the implication from left to right. Thereto, let $L(r_1) \subseteq L(r_2)$. Let uwu' be an arbitrary string in $L(r_1)$ such that $u, u' \in L(\$e\$e\$\cdots\$e\$)$ and $w \in \triangle x_\perp X^{[2^n-1,2^n-1]} \triangle X^*_\triangle \triangle x_\top X^{[2^n-1,2^n-1]} \triangle$. By assumption, $uwu' \in L(r_2)$.

Notice that uwu' contains $2k+2$ times the symbol "\$". Moreover, the first and the last "\$" of uwu' is always matched onto the first and last "\$" of r_2. This means that $k+1$ consecutive \$-symbols of the remaining $2k$ \$-symbols in uwu' must be matched onto the \$-symbols in $\$e_1\$e_2\$\cdots\$e_k\$$. Hence, w is matched onto some e_i. So, w does not encode a correct tiling. As the subexpression $\triangle x_\perp X^{[2^n-1,2^n-1]} \triangle X^*_\triangle \triangle x_\top X^{[2^n-1,2^n-1]} \triangle$ of r_1 defines all candidate tilings, the system T has no solution.

To show the implication from right to left, assume that there is a string $uwu' \in L(r_1)$ that is not in r_2, where $u, u' \in L(\$e\$e\$\cdots\$e\$)$. Then $w \notin \bigcup_{i=1}^k L(e_i)$ and, hence, w encodes a correct tiling. □

Adding numerical occurrence constraints to the fragment CHARE($a, a?$) keeps EQUIVALENCE in PTIME, INTERSECTION in NP and INCLUSION in coNP.

THEOREM 4.3.

(1) EQUIVALENCE *for CHARE($a, a?, a\#$) is in* PTIME.

(2) INCLUSION *for CHARE($a, a?, a\#$) is co*NP*-complete.*[2]

(3) INTERSECTION *for CHARE($a, a?, a\#$) is* NP*-complete.*

Proof. (1) It is shown in [28] that two CHARE($a, a?$) expressions are equivalent if and only if they have the same *sequence normal form* (which is defined below). As $a^{[k,\ell]}$ is equivalent to $a^k(a?)^{\ell-k}$, it follows that two CHARE($a, a?, a\#$) expressions are equivalent if and only if they have the same sequence normal form. It remains to argue that the sequence normal form of CHARE($a, a?, a\#$) expressions can be computed in polynomial time. To this end, let $r = f_1 \cdots f_n$ be a CHARE($a, a?, a\#$) expression with factors $f_1, \ldots, f_n$. The *sequence normal form* is then obtained in the following way. First, we replace every factor of the form

- a by $a[1, 1]$;
- $a?$ by $a[0, 1]$; and,
- $a^{[k,\ell]}$ by $a[k, \ell]$,

[2] In the previous version of this article presented at ICDT'07 the complexity was wrongly attributed to lie between PSPACE and EXPSPACE.

where a is an alphabet symbol. We call a the *base symbol* of the factor $a[i,j]$. Then, we replace successive subexpressions $a[i_1,j_1]$ and $a[i_2,j_2]$ carrying the same base symbol a by $a[i_1+i_2, j_1+j_2]$ until no further replacements can be made anymore. For instance, the sequence normal form of $aa?a^{[2,5]}a?bb?b?b^{[1,7]}$ is $a[3,8]b[2,10]$. Obviously, the above algorithm to compute the sequence normal form of CHARE$(a, a?, a\#)$ expressions can be implemented in polynomial time. It can then be tested in linear time whether two sequence normal forms are the same.

(2) coNP-hardness is immediate since INCLUSION is already coNP-complete for CHARE $(a, a?)$ expressions [28].

We show that the problem remains in coNP. To this end, we represent strings w by their sequence normal form as discussed above, where we take each string w as the regular expression defining w. We call such strings *compressed*. Let r_1 and r_2 be two CHARE$(a, a?, a\#)$s. We can assume that they are in sequence normal form.

To show that $L(r_1) \not\subseteq L(r_2)$, we guess a compressed string w of polynomial size for which $w \in L(r_1)$, but $w \notin L(r_2)$. We guess $w \in L(r_1)$ in the following manner. We iterate from left to right over the factors of r_1. For each factor $a[k,\ell]$ we guess an h such that $k \leq h \leq \ell$, and add a^h to the compressed string w. This algorithm gives a compressed string of polynomial size which is defined by r_1. Furthermore, this algorithm is capable of guessing every possible string defined by r_1. It is however possible that in the compressed string there are two consecutive *elements* a^i, a^j with the same *base symbol* a. If this is the case we merge these elements to a^{i+j} which gives a proper compressed string.

The following lemma shows that testing $w \notin L(r_2)$ can be done in PTIME.

LEMMA 4.4. *Given a compressed string w and an expression r in sequence normal form, deciding whether $w \in L(r)$ is in* PTIME.

Proof. Let $w = a_1^{p_1} \cdots a_n^{p_n}$, and $r = b_1[k_1,\ell_1] \cdots b_m[k_m,\ell_m]$. Denote $b_i[k_i,\ell_i]$ by f_i. For every position i of w $(0 < i \leq n)$, we define C_i as a set of factors $b[k,\ell]$ of r. Formally, $f_j \in C_i$ when $a_1^{p_1} \cdots a_{i-1}^{p_{i-1}} \in L(f_1 \cdots f_{j-1})$ and $a_i = b_j$. We compute the C_i as follows.

- C_1 is the set of all $b_j[k_j,\ell_j]$ such that $a_1 = b_j$, and $\forall h < j, k_h = 0$. These are all the factors of r which can match the first symbol of w.
- Then, for all $i \in \{2,\ldots,n\}$, we compute C_i from C_{i-1}. In particular, $f_h = b_h[k_h,\ell_h] \in C_i$ when there is a $f_j = b_j[k_j,\ell_j] \in C_{i-1}$ such that $a_{i-1}^{p_{i-1}} \in f_j \cdots f_{h-1}$ and $a_i = b_h$. That is, the following conditions should hold:
  - $j < h$: f_h occurs after f_j in r.
  - $b_h = a_i$: f_h can match the first symbol of $a_i^{p_i}$.
  - $\forall e \in \{j,\ldots,h-1\}$, if $b_e \neq a_{i-1}$ then $k_e = 0$: in between factors f_j and f_h it is possible to match only symbols a_{i-1}.
  - Let $\min = \sum_{e\in\{j,\ldots,h-1\}, b_e = a_{i-1}} k_e$ and $\max = \sum_{e\in\{j,\ldots,h-1\}, b_e = a_{i-1}} \ell_e$. Then $\min \leq p_{i-1} \leq \max$. That is, p_{i-1} symbols a_{i-1} should be matched from f_j to f_{h-1}.

Then, $w \in L(r)$ iff there is an $f_j \in C_n$ such that $a_n^{p_n} \in L(f_j \cdots f_n)$. As the latter test and the computation of $C_1,\ldots,C_n$ can be done in PTIME the lemma follows. ∎

(3) NP-hardness is immediate since INTERSECTION is already NP-complete for CHARE $(a, a?)$ expressions [28].

We show that the problem remains in NP. As in the proof of Theorem 4.3(2) we represent a string w as a compressed string. Let $r_1,\ldots,r_n$ be CHARE$(a, a?, a\#)$ expressions.

LEMMA 4.5. *If* $\bigcap_{i=1}^{n} L(r_i) \neq \emptyset$*, then there exists a string* $w = a_1^{p_1} \cdots a_m^{p_m} \in \bigcap_{i=1}^{n} L(r_i)$ *such that* $m \leq \min\{|r_i| \mid i \in \{1, \ldots, n\}\}$ *and, for each* $i \in \{1, \ldots, n\}$*,* j_i *is not larger than the largest integer occurring in* $r_1, \ldots, r_n$*.*

Proof. Suppose that there exists a string $w = a_1^{p_1} \cdots a_m^{p_m} \in \bigcap_{i=1}^{n} L(r_i)$, with $a_i \neq a_{i+1}$ for every $i \in \{1, \ldots, m-1\}$. Since w is matched by every expression $r_1, \ldots, r_n$, and since a factor of a CHARE$(a, a?, a\#)$ expression can never match a strict superstring of $a_i^{p_i}$ for $i \in \{1, \ldots, n\}$, we have that $m \leq \min\{|r_i| \mid i \in \{1, \ldots, n\}\}$.

Furthermore, since w is matched by every expression $r_1, \ldots, r_n$, no j_i can be larger than the largest integer occurring in $r_1, \ldots, r_n$. ∎

The NP algorithm then consists of guessing a compressed string w of polynomial size and verifying whether $w \in \bigcap_{i=1}^{n} L(r_i)$. If we represent $r_1, \ldots, r_n$ by their sequence normal form, this verification step can be done in polynomial time by Lemma 4.4. ∎

Finally, we exhibit a tractable subclass with numerical occurrence constraints:

THEOREM 4.6. INCLUSION, EQUIVALENCE, *and* INTERSECTION *for CHARE(*a, $a\#^{>0}$*) are in* PTIME.

Proof. The upper bound for EQUIVALENCE is immediate from Theorem 4.3(2).

For INCLUSION, let r_1 and r_2 be two CHARE$(a, a\#^{>0})$s in sequence normal form. (as defined in the proof of Theorem 4.3) Let $r_1 = a_1[k_1, \ell_1] \cdots a_n[k_n, \ell_n]$ and $r_2 = a'_1[k'_1, \ell'_1] \cdots a'_{n'}[k'_{n'}, \ell'_{n'}]$. Notice that every number k_i and k'_j is greater than zero. We claim that $L(r_1) \subseteq L(r_2)$ if and only if $n = n'$ and for every $i \in \{1, \ldots, n\}$, $a_i = a'_i$, $k_i \geq k'_i$, and $\ell_i \leq \ell'_i$.

Indeed, if $n \neq n'$, or if there exists an i such that $a_i \neq a'_i$ or $k_i < k'_i$, then $a_1^{k_1} \cdots a_n^{k_n} \in L(r_1) - L(r_2)$. If there exists an i such that $\ell_i > \ell'_i$, then $a_1^{\ell_1} \cdots a_n^{\ell_n} \in L(r_1) - L(r_2)$. Conversely, it is immediate that every string in $L(r_1)$ is also in $L(r_2)$. It is straightforward to test these conditions in linear time.

For INTERSECTION, let, for every $i \in \{1, \ldots, n\}$, $r_i = a_{i,1}[k_{i,1}, \ell_{i,1}] \cdots a_{i,m_i}[k_{i,m_i}, \ell_{i,m_i}]$ be a CHARE$(a, a\#^{>0})$ in sequence normal form. Notice that every number $k_{i,j}$ is greater than zero. We claim that $\bigcap_{i=1}^{n} L(r_i) \neq \emptyset$ if and only if

(i) $m_1 = m_2 = \cdots = m_n$;
(ii) for every $i, j \in \{1, \ldots, n\}$ and $x \in \{1, \ldots, m_1\}$, $a_{i,x} = a_{j,x}$; and,
(iii) for every $x \in \{1, \ldots, m_1\}$, $\max\{k_{i,x} \mid 1 \leq i \leq n\} \leq \min\{\ell_{i,x} \mid 1 \leq i \leq n\}$.

Indeed, if the above conditions hold, we have that $a_{1,1}^{K_1} \cdots a_{1,m_1}^{K_{m_1}}$ is in $\bigcap_{i=1}^{n} L(r_i)$, where $K_x = \max\{k_{i,x} \mid 1 \leq i \leq n\}$ for every $x \in \{1, \ldots, m_1\}$. If $m_i \neq m_j$ for some $i, j \in \{1, \ldots, n\}$, then the intersection between r_i and r_j is empty. So assume that condition (i) holds. If $a_{i,x} \neq a_{j,x}$ for some $i, j \in \{1, \ldots, n\}$ and $x \in \{1, \ldots, m_1\}$, then we also have that the intersection between r_i and r_j is empty. Finally, if condition (iii) does not hold, take i, j, and x such that $k_{i,x} = \max\{k_{i,x} \mid 1 \leq i \leq n\}$ and $\ell_{j,x} = \min\{\ell_{i,x} \mid 1 \leq i \leq n\}$. Then the intersection between r_i and r_j is empty.

Finally, testing conditions (i)–(iii) can be done in linear time. ∎

5. Complexity of Schemas.

5.1. DTDs and Single-Type EDTDs. By Proposition 2.4 the results on the EQUIVALENCE and INCLUSION problem of the previous section carry over to DTDs and single-type EDTDs. For the INTERSECTION problem, the results only carry over to DTDs (Proposition 2.5). The only remaining problem is INTERSECTION for single-type EDTDs with counting and interleaving. However, INTERSECTION for EDTD$^{\text{st}}$(RE) is EXPTIME-hard and in the next section we will see that even for EDTD(#,&) INTERSECTION remains in EXPTIME. It immediately follows that INTERSECTION for

EDTDst(#), EDTDst(&), and EDTDst(#,&) is also EXPTIME-complete.

5.2. Extended DTDs. We next consider the complexity of the basic decision problems for EDTDs with numerical occurrence constraints and interleaving. As the basic decision problems are EXPTIME-complete for EDTD(RE), the straightforward approach of translating every RE(#, &) expression into an NFA and then applying the standard algorithms gives rise to a double exponential time complexity. By using NFA(#, &), we can do better: EXPSPACE for INCLUSION and EQUIVALENCE, and, more surprisingly, EXPTIME for INTERSECTION.

THEOREM 5.1.

(1) EQUIVALENCE *and* INCLUSION *for EDTD(#,&) are in* EXPSPACE*;*

(2) EQUIVALENCE *and* INCLUSION *for EDTD(#) and EDTD(&) are* EXPSPACE*-hard; and,*

(3) INTERSECTION *for EDTD(#,&) is* EXPTIME*-complete.*

Proof. (1) We show that INCLUSION is in EXPSPACE. The upper bound for EQUIVALENCE then immediately follows.

First, we introduce some notation. For an EDTD $D = (\Sigma, \Sigma', d, s, \mu)$, we will denote elements of Σ', i.e., types, by τ. We denote by (D, τ) the EDTD D with start symbol τ. We define the *depth* of a tree t, denoted by $\text{depth}(t)$, as follows: if $t = \varepsilon$, then $\text{depth}(t) = 0$; and if $t = \sigma(t_1 \cdots t_n)$, then $\text{depth}(t) = \max\{\text{depth}(t_i) \mid i \in \{1, \ldots, n\}\} + 1$.

Suppose that we have two EDTDs $D_1 = (\Sigma, \Sigma'_1, d_1, s_1, \mu_1)$ and $D_2 = (\Sigma, \Sigma'_2, d_2, s_2, \mu_2)$. We provide an EXPSPACE algorithm that decides whether $L(D_1) \not\subseteq L(D_2)$. As EXPSPACE is closed under complement, the theorem follows. The algorithm computes a set E of pairs $(C_1, C_2) \in 2^{\Sigma'_1} \times 2^{\Sigma'_2}$ where $(C_1, C_2) \in E$ iff there exists a tree t such that $C_j = \{\tau \in \Sigma'_j \mid t \in L((D_j, \tau))\}$ for each $j = 1, 2$. That is, every C_j is the set of types that can be assigned by D_j to the root of t. Or when viewing D_j as a tree automaton, C_j is the set of states that can be assigned to the root in a run on t. Therefore, we say that t *is a witness for* (C_1, C_2). Notice that $t \in L(D_1)$ (resp., $t \in L(D_2)$) if $s_1 \in C_1$ (resp. $s_2 \in C_2$). Hence, $L(D_1) \not\subseteq L(D_2)$ iff there exists a pair $(C_1, C_2) \in E$ with $s_1 \in C_1$ and $s_2 \notin C_2$.

We compute the set E in a bottom-up manner as follows:

1. Initially, set $E_1 := \{(C_1, C_2) \mid \exists a \in \Sigma, \tau_1 \in \Sigma'_1, \tau_2 \in \Sigma'_2$ such that $\mu_1(\tau_1) = \mu_2(\tau_2) = a$, and for $i = 1, 2$, $C_i = \{\tau \in \Sigma'_i \mid \varepsilon \in d_i(\tau) \wedge \mu_i(\tau) = a\}\}$.
2. For every $k > 1$, E_k is the union of E_{k-1} and the pairs (C_1, C_2) for which there are $a \in \Sigma$, $n \in \mathbb{N}$ and a string $(C_{1,1}, C_{2,1}) \cdots (C_{1,n}, C_{2,n})$ in E^*_{k-1} such that

$$C_j = \{\tau \in \Sigma'_j \mid \mu_j(\tau) = a, \exists b_{j,1} \in C_{j,1}, \ldots, b_{j,n} \in C_{j,n} \text{ with } b_{j,1} \cdots b_{j,n} \in d_j(\tau)\}, \text{ for each } j = 1, 2.$$

Let $E := E_\ell$ for $\ell = 2^{|\Sigma'_1|} \cdot 2^{|\Sigma'_2|}$. The algorithm then accepts when there is a pair $(C_1, C_2) \in E$ with $s_1 \in C_1$ and $s_2 \notin C_2$ and rejects otherwise.

We argue that the algorithm is correct. As $E_k \subseteq E_{k+1}$, for every k, it follows that $E_\ell = E_{\ell+1}$. Hence, the algorithm computes the largest set of pairs. The following lemma then shows that the algorithm decides whether $L(D_1) \not\subseteq L(D_2)$. The lemma can be proved by induction on k.

LEMMA 5.2. *For every $k \geq 1$, $(C_1, C_2) \in E_k$ if and only if there exists a witness tree for (C_1, C_2) of depth at most k.*

It remains to show that the algorithm can be carried out using exponential space. Step (1) reduces to a linear number of tests $\varepsilon \in L(r)$, for some RE(#, &) expressions r which is in PTIME by [20]. Step (3) can be carried out in exponential time, since the size of E is exponential in the input. For step (2), it suffices to argue that, when E_{k-1} is known, it is decidable in EXPSPACE whether a pair (C_1, C_2) is in E_k. As there are only an exponential number of such possible pairs, the result follows. To this end, we need to verify that there exists a string $W = (C_{1,1}, C_{2,1}) \cdots (C_{1,n}, C_{2,n})$ in E_{k-1}^* such that for each $j = 1, 2$,

(A) for every $\tau \in C_j$, there exist $b_{j,1} \in C_{j,1}, \ldots, b_{j,n} \in C_{j,n}$ with $b_{j,1} \cdots b_{j,n} \in d_j(\tau)$; and,

(B) for every $\tau \in \Sigma'_j \setminus C_j$, there do *not* exist $b_{j,1} \in C_{j,1}, \ldots, b_{j,n} \in C_{j,n}$with $b_{j,1} \cdots b_{j,n} \in d_j(\tau)$.

Assume that $\Sigma'_1 \cap \Sigma'_2 = \emptyset$. Let, for each $j = 1, 2$ and $\tau \in \Sigma'_j$, $N(\tau)$ be the NFA(#, &) accepting $d_j(\tau)$. Intuitively, we guess the string W one symbol at a time and compute the set of reachable configurations Γ_τ for each $N(\tau)$.

Initially, Γ_τ is the singleton set containing the initial configuration of $N(\tau)$. Suppose that we have guessed a prefix $(C_{1,1}, C_{2,1}) \cdots (C_{1,m-1}, C_{2,m-1})$ of W and that we guess a new symbol $(C_{1,m}, C_{2,m})$. Then, we compute the set $\Gamma'_\tau = \{\gamma' \mid \exists b \in C_{j,m}, \gamma \in \Gamma_\tau \text{ such that } \gamma \Rightarrow_{N(\tau),b} \gamma'\}$ and set Γ_τ to Γ'_τ. Each set Γ'_τ can be computed in exponential space from Γ_τ. We accept (C_1, C_2) when for every $\tau \in \Sigma'_j$, $\tau \in C_j$ iff Γ_τ contains an accepting configuration.

(2) It is shown by Mayer and Stockmeyer [31] and Meyer and Stockmeyer [32] that EQUIVALENCE and INCLUSION are EXPSPACE-hard for RE(&)s and RE(#), respectively. Hence, EQUIVALENCE and INCLUSION are also EXPSPACE-hard for EDTD(&) and EDTD(#).

(3) The lower bound follows from [38]. We argue that the problem is in EXPTIME. Thereto, let, for each $i \in \{1, \ldots, n\}$, $D_i = (\Sigma, \Sigma'_i, d_i, s_i, \mu_i)$ be an EDTD(#,&). We assume w.l.o.g. that the sets Σ'_i are pairwise disjoint. We also assume that the start type s_i never appears at the right-hand side of a rule. Finally, we assume that no derivation tree consists of only the root. For each type $\tau \in \Sigma'_i$, let $N(\tau)$ denote an NFA(#, &) for $d_i(\tau)$. According to Theorem 3.2, $N(\tau)$ can be computed from $d_i(\tau)$ in polynomial time. We provide an alternating polynomial space algorithm that guesses a tree t and accepts if $t \in L(D_1) \cap \cdots \cap L(D_n)$. As APSPACE = EXPTIME [8], this shows the theorem.

We guess t node by node in a top-down manner. For every guessed node v, the following information is written on the tape of the TM: for every $i \in \{1, \ldots, n\}$, the triple $c_i = (\tau_v^i, \tau_p^i, \gamma^i)$ where τ_v^i is the type assigned to v by grammar D_i, τ_p^i is the type of the parent assigned by D_i, and γ^i is the current configuration $N(\tau_p^i)$ is in after reading the string formed by the left siblings of v. In the following, we say that $\tau \in \Sigma'_i$ is an a-type when $\mu_i(\tau) = a$.

The algorithm proceeds as follows:

1. As for each grammar the types of the roots are given, we start by guessing the first child of the root. That is, we guess an $a \in \Sigma$, and for each $i \in \{1, \ldots, n\}$, we guess an a-type τ^i and write the triple $c_i = (\tau^i, s_i, \gamma_s^i)$ on the tape where γ_s^i is the start configuration of $N(s_i)$.
2. For $i \in \{1, \ldots, n\}$, let $c_i = (\tau^i, \tau_p^i, \gamma^i)$ be the triples on the tape. The algorithm now universally splits into two parallel branches as follows:
   (a) **Downward extension**: When for every i, $\varepsilon \in d_i(\tau^i)$ then the current node can be a leaf node and the branch accepts. Otherwise, guess an

$a \in \Sigma$ and for each i, guess an a-type θ^i. Replace every c_i by the triple $(\theta^i, \tau^i, \gamma^i_s)$ and proceed to step (2). Here, γ^i_s is the start configuration of $N(\tau^i)$.

(b) **Extension to the right**: For every $i \in \{1, \ldots, n\}$, compute a configuration γ'^i for which $\gamma^i \Rightarrow_{N(\tau^i_p),\tau^i} \gamma'^i$. When every γ'^i is a final configuration, then we do not need to extend to the right anymore and the algorithm accepts. Otherwise, guess an $a \in \Sigma$ and for each i, guess an a-type θ^i. Replace every c_i by the triple $(\theta^i, \tau^i, \gamma'^i)$ and proceed to step (2).

We argue that the algorithm is correct. If the algorithm accepts, we have guessed a tree t and, for every $i \in \{1, \ldots, n\}$, a tree t'_i with $\mu_i(t'_i) = t$ and $t'_i \in L(d_i)$. Therefore, $t \in \bigcap_{i=1}^n L(D_i)$. For the other direction, suppose that there exists a tree $t \in \bigcap_{i=1}^n L(D_i)$ and t is minimal in the sense that no subtree t_0 of t is in $\bigcap_{i=1}^n L(D_i)$. Then, there is a run of the above algorithm that guesses t and guesses trees t'_i with $\mu_i(t'_i) = t$. The tree t must be minimal since the algorithm stops extending the tree as soon as possible.

The algorithm obviously uses only polynomial space. □

6. Simplification. The simplification problem is defined as follows: Given an EDTD, check whether it has an equivalent EDTD of a restricted type, i.e., an equivalent DTD or single-type EDTD. In [29], this problem was shown to be EXPTIME-complete for EDTDs with standard regular expressions. We revisit this problem in the context of RE(#, &).

We need a bit of terminology. Let t be a tree and v be a node. By anc-str$^t(v)$ we denote the string formed by the labels on the path from the root to v, i.e., $\mathrm{lab}^t(\varepsilon)\mathrm{lab}^t(i_1)\mathrm{lab}^t(i_1 i_2)\cdots\mathrm{lab}^t(i_1 i_2 \cdots i_k)$ where $v = i_1 i_2 \cdots i_k$.

We say that a tree language L is *closed under ancestor-guarded subtree exchange* if the following holds. Whenever for two trees $t_1, t_2 \in L$ with nodes $u_1 \in \mathrm{Dom}(t_1)$ and $u_2 \in \mathrm{Dom}(t_2)$, anc-str$^{t_1}(u_1) =$ anc-str$^{t_2}(u_2)$ implies $t_1[u_1 \leftarrow \mathrm{subtree}^{t_2}(u_2)] \in L$. Here, $t_1[u_1 \leftarrow \mathrm{subtree}^{t_2}(u_2)]$ denotes the tree obtained from t_1 by replacing its subtree rooted at u_1 by the subtree rooted at u_2 in t_2.

We recall the following theorem from [29]:

THEOREM 6.1 (Theorem 7.1 in [29]). *Let L be a tree language defined by an EDTD. Then the following conditions are equivalent.*
(a) L is definable by a single-type EDTD.
(b) L is closed under ancestor-guarded subtree exchange.

We are now ready for the following theorem.

THEOREM 6.2. Given an EDTD(#,&), deciding whether it is equivalent to an EDTD$^{\mathrm{st}}$(#,&) or DTD(#,&) is EXPSPACE-complete.

Proof. We first show that the problem is hard for EXPSPACE. We use a reduction from EQUIVALENCE of RE(#), which is EXPSPACE-complete [32].

Let r_1, r_2 be RE(#) expressions over Σ and let b and s be two symbols not occurring in Σ. By definition $L(r_j) \neq \emptyset$, for $j = 1, 2$. Define $D = (\Sigma \cup \{b, s\}, \Sigma \cup \{s, b^1, b^2\}, d, s, \mu)$ as the EDTD with the following rules:

$$\begin{array}{lcl} s & \rightarrow & b^1 b^2 \\ b^1 & \rightarrow & r_1 \\ b^2 & \rightarrow & r_2, \end{array}$$

where for every $\tau \in \Sigma \cup \{s\}$, $\mu(\tau) = \tau$, and $\mu(b^1) = \mu(b^2) = b$. We claim that D is equivalent to a single-type DTD or a DTD iff $L(r_1) = L(r_2)$. Clearly, if r_1 is

equivalent to r_2, then D is equivalent to the DTD (and therefore also to a single-type EDTD)

$$\begin{array}{lcl} s & \to & bb \\ b & \to & r_1. \end{array}$$

Conversely, suppose that there exists an EDTD$^{\text{st}}$ which defines the language $L(D)$. Towards a contradiction, assume that r_1 is not equivalent to r_2. So, there exists a string w_1 such that $w_1 \in L(r_1)$ and $w_1 \notin L(r_2)$, or $w_1 \notin L(r_1)$ and $w_1 \in L(r_2)$. We only consider the first case, the second is identical. Now, let w_2 be a string in $L(r_2)$ and consider the tree $t = s(b(w_1)b(w_2))$. Clearly, t is in $L(D)$. However, the tree $t' = s(b(w_2)b(w_1))$ obtained from t by switching its left and right subtree is not in $L(D)$. According to Theorem 6.1, every tree language defined by a single-type EDTD is closed under such an exchange of subtrees. So, this means that $L(D)$ cannot be defined by an EDTD$^{\text{st}}$, which leads to the desired contradiction.

We now proceed with the upper bounds. The following algorithms are along the same lines as the EXPTIME algorithms in [29] for the simplification problem without numerical occurrence or interleaving operators. We first give an EXPSPACE algorithm which decides whether an EDTD is equivalent to a EDTD$^{\text{st}}$. Let $D = (\Sigma, \Sigma', d, s, \mu)$ be an EDTD. Intuitively, we compute a EDTD$^{\text{st}}$ $D_0 = (\Sigma, \Sigma'_0, d_0, s, \mu_0)$ which is the closure of D under the single-type property. The EDTD$^{\text{st}}$ D_0 has the following properties:

(a) Σ'_0 is in general exponentially larger than Σ';
(b) the RE$(\#, \&)$ expressions in the definition of d_0 are only polynomially larger than the RE$(\#, \&)$ expressions in the definition of d;
(c) $L(D) \subseteq L(D_0)$; and,
(d) $L(D_0) = L(D) \Leftrightarrow D$ is equivalent to a EDTD$^{\text{st}}$.

Hence, we have that D is equivalent to a EDTD$^{\text{st}}$ if and only if $L(D_0) \subseteq L(D)$.

We first show how D_0 can be constructed. We can assume w.l.o.g. that, for each type $a^i \in \Sigma'$, there exists a tree $t' \in L(d)$ such that a^i is a label in t'. Indeed, every *useless* type can be removed from D in a simple preprocessing step. Then, for a string $w \in \Sigma^*$ and $a \in \Sigma$ let types(wa) be the set of all types $a^i \in \Sigma'$, for which there is a tree t and a tree $t' \in L(d)$ with $\mu(t') = t$, and a node v in t such that anc-str$^t(v) = wa$ and the type of v in t' is a^i. We show how to compute types(wa) in exponential time. To this end, we enumerate all sets types(w). Let $s = c^1$. Initially, set $W := \{c\}$, Types$(c) := \{c^1\}$ and $R := \{\{c^1\}\}$. Repeat the following until W becomes empty:

(1) Remove a string wa from W.
(2) For every $b \in \Sigma$, let Types(wab) contain all b^i for which there exists an a^j in Types(wa) and a string in $d(a^j)$ containing b^i. If Types(wab) is not empty and not already in R, then add it to R and add wab to W.

Since we add every set only once to R, the algorithm runs in time exponential in the size of D. Moreover, we have that Types$(w) = \mathit{types}(w)$ for every w, and that $R = \Sigma'_0$.

For each $a \in \Sigma$, let types(D, a) be the set of all nonempty sets types(wa), with $w \in \Sigma^*$. Clearly, each types(D, a) is finite. We next define $D_0 = (\Sigma, \Sigma'_0, d_0, s, \mu_0)$. Its set of types is $\Sigma'_0 := \bigcup_{a \in \Sigma} \text{types}(D, a)$. Note that $s \in \Sigma'_0$. For every $\tau \in \text{types}(D, a)$, set $\mu_0(\tau) = a$. In d_0, the right-hand side of the rule for each types(wa) is the disjunction of all $d(a^i)$ for $a^i \in \text{types}(wa)$, with each b^j in $d(a^i)$ replaced by types(wab).

We show that properties (a)–(d) hold. Since $\Sigma'_0 \subseteq 2^{\Sigma'}$, we immediately have that (a) holds. The RE$(\#, \&)$ expressions that we constructed in D_0 are unions of a

linear number of RE(#, &) expressions in D, but have types in $2^{\Sigma'}$ rather than in Σ'. Hence, the size of the RE(#, &) expressions in D_0 is at most quadratic in the size of D. Finally, we note that it has been shown in Theorem 7.1 in [29] that (c) and (d) also hold.

It remains to argue that it can be decided in EXPSPACE that $L(D_0) \subseteq L(D)$. A direct application of the EXPSPACE algorithm in Theorem 5.1(1) leads to a 2EXPSPACE algorithm to test whether $L(D_0) \subseteq L(D)$, due to the computation of C_1. Indeed, the algorithm remembers, given the EDTDs $D_0 = (\Sigma, \Sigma'_0, d_0, s_0, \mu_0)$ and $D = (\Sigma, \Sigma', d, s, \mu)$, all possible pairs (C_1, C_2) such that there exists a tree t with $C_1 = \{\tau \in \Sigma'_0 \mid t \in L((D_0, \tau))\}$ and $C_2 = \{\tau \in \Sigma' \mid t \in L((D, \tau))\}$. It then accepts if there exists such a pair (C_1, C_2) with $s_0 \in C_1$ and $s \notin C_2$. However, when we use non-determinism, notice that it is not necessary to compute the entire set C_1. Indeed, as we only test whether there *exist* elements in C_1 in the entire course of the algorithm, we can adapt the algorithm to compute pairs (c_1, C_2), where c_1 is an element of C_1, rather than the entire set. Since NEXPSPACE = EXPSPACE, we can use this adaption to test whether $L(D_0) \subseteq L(D)$ in EXPSPACE.

Finally, we give the algorithm which decides whether an EDTD $D = (\Sigma, \Sigma', d, s, \mu)$ is equivalent to a DTD. We compute a DTD (Σ, d_0, s_d) which is equivalent to D iff $L(D)$ is definable by a DTD. Thereto, let for each $a^i \in \Sigma'$, $r_{a,i}$ be the expression obtained from $d(a^i)$ by replacing each symbol b^j in $d(a^i)$ by b. For every $a \in \Sigma$, define $d_0(a) = \bigcup_{a^i \in \Sigma'} r_{a,i}$. Again, it is shown in [29] that $L(D) = L(d_0)$ iff $L(D)$ is definable by a DTD. By Theorem 5.1(1) and since d_0 is of size polynomial in the size of D, this can be tested in EXPSPACE. □

7. Conclusion. The present work gives an overview of the complexity of the basic decision problems for abstractions of several schema languages including numerical occurrence constraints and interleaving. W.r.t. INTERSECTION the complexity remains the same, while for INCLUSION and EQUIVALENCE the complexity increases by one exponential for DTDs and single-type EDTDs, and goes from EXPTIME to EXPSPACE for EDTDs. The results w.r.t. CHAREs also follow this pattern. We further showed that the complexity of simplification increases to EXPSPACE.

We emphasize that this is a theoretical study delineating the worst case complexity boundaries for the basic decision problems. Although these complexities must be studied, we note that the regular expressions used in the hardness proofs do not correspond at all to those employed in practice. Further, w.r.t. XSDs, our abstraction is not fully adequate as we do not consider the one-unambiguity (or unique particle attribution) constraint. However, it is doubtful that this constraint is the right one to get tractable complexities for the basic decision problems. Indeed, already intersection for unambiguous regular expressions is PSPACE-hard [28] and inclusion for one-unambiguous RE(#) expressions is coNP-hard [19]. It would therefore be desirable to find robust subclasses for which the basic decision problems are in PTIME.

REFERENCES

[1] S. Abiteboul, P. Buneman, and D. Suciu, *Data on the Web : From Relations to Semistructured Data and XML*, Morgan Kaufmann, 1999.

[2] M. Benedikt, W. Fan, and F. Geerts, *Xpath satisfiability in the presence of DTDs*, J. ACM, 55 (2008).

[3] G.J. Bex, F. Neven, T. Schwentick, and K. Tuyls, *Inference of concise DTDs from XML data*, in International Conference on Very Large Data Bases (VLDB), 2006, pp. 115–126.

[4] G.J. Bex, F. Neven, and J. Van den Bussche, *DTDs versus XML schema: A practical study*, in The World Wide Web and Databases (WebDB), 2004, pp. 79–84.
[5] A. Brüggemann-Klein, *Unambiguity of extended regular expressions in SGML document grammars*, in Algorithms, First Annual European Symposium (ESA), 1993, pp. 73–84.
[6] A. Brüggemann-Klein, M. Murata, and D. Wood, *Regular tree and regular hedge languages over unranked alphabets: Version 1, april 3, 2001*, Tech. Report HKUST-TCSC-2001-0, The Hongkong University of Science and Technology, 2001.
[7] A. Brüggemann-Klein and D. Wood, *One-unambiguous regular languages*, Inform. and Comput., 142 (1998), pp. 182–206.
[8] A. K. Chandra, D. Kozen, and L. J. Stockmeyer, *Alternation.*, J. ACM, 28 (1981), pp. 114–133.
[9] J. Clark and M. Murata, *RELAX NG Specification*, OASIS, December 2001.
[10] J. Cristau, C. Löding, and W. Thomas, *Deterministic automata on unranked trees.*, in Fundamentals of Computation Theory (FCT), M. Liskiewicz and R. Reischuk, eds., vol. 3623 of Lecture Notes in Computer Science, Springer, 2005, pp. 68–79.
[11] S. Dal-Zilio and D. Lugiez, *XML schema, tree logic and sheaves automata.*, in Rewriting Techniques and Applications (RTA), Robert Nieuwenhuis, ed., vol. 2706 of Lecture Notes in Computer Science, Springer, 2003, pp. 246–263.
[12] A. Deutsch, M. F. Fernandez, and D. Suciu, *Storing Semistructured Data with STORED.*, in ACM SIGMOD International Conference on Management of Data, 1999, pp. 431–442.
[13] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, 1979.
[14] W. Gelade, *Succinctness of regular expressions with interleaving, intersection and counting*, in Mathematical Foundations of Computer Science (MFCS), 2008, pp. 363–374.
[15] L. Hemaspaandra and M. Ogihara, *Complexity Theory Companion*, Springer, 2002.
[16] J.E. Hopcroft, R. Motwani, and J.D. Ullman and, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, third ed., 2007.
[17] H. Hosoya and B. C. Pierce, *Xduce: A statically typed XML processing language.*, ACM Transactions on Internet Technologies, 3 (2003), pp. 117–148.
[18] J. Jędrzejowicz and A. Szepietowski, *Shuffle languages are in P*, Theoret. Comput. Sci., 250 (2001), pp. 31–53.
[19] P. Kilpeläinen, *Inclusion of unambiguous #REs is NP-hard.* Unpublished note, University of Kuopio, Finland, May 2004.
[20] P. Kilpeläinen and R. Tuhkanen, *Regular expressions with numerical occurrence indicators — preliminary results.*, in Symposium on Programming Languages and Software Tools (SPLST), 2003, pp. 163–173.
[21] ———, *Towards efficient implementation of XML schema content models.*, in DOCENG 2004, ACM, 2004, pp. 239–241.
[22] ———, *One-unambiguity of regular expressions with numeric occurrence indicators.*, Inform. and Comput., 205 (2007), pp. 890–916.
[23] C. Koch, S. Scherzinger, N. Schweikardt, and B. Stegmaier, *Schema-based scheduling of event processors and buffer minimization for queries on structured data streams*, in International Conference on Very Large Data Bases (VLDB), 2004, pp. 228–239.
[24] D. Kozen, *Lower bounds for natural proof systems*, in Foundations of Computer Science (FOCS), IEEE, 1977, pp. 254–266.
[25] M. Mani, *Keeping chess alive — Do we need 1-unambiguous content models?*, in Extreme Markup Languages, Montreal, Canada, 2001.
[26] I. Manolescu, D. Florescu, and D. Kossmann, *Answering XML Queries on Heterogeneous Data Sources*, in International Conference on Very Large Data Bases (VLDB), 2001, pp. 241–250.
[27] W. Martens and F. Neven, *Frontiers of tractability for typechecking simple XML transformations.*, J. Comput. System Sci., 73 (2007), pp. 362–390.
[28] W. Martens, F. Neven, and T. Schwentick, *Complexity of decision problems for simple regular expressions.*, in Mathematical Foundations of Computer Science (MFCS), Jirí Fiala, Václav Koubek, and Jan Kratochvíl, eds., vol. 3153 of Lecture Notes in Computer Science, Springer, 2004, pp. 889–900.
[29] W. Martens, F. Neven, T. Schwentick, and G. J. Bex, *Expressiveness and complexity of XML schema.*, ACM Transactions on Database Systems, 31 (2006), pp. 770–813.
[30] W. Martens and J. Niehren, *On the minimization of XML schemas and tree automata for unranked trees.*, J. Comput. System Sci., 73 (2007), pp. 550–583.
[31] A. J. Mayer and L. J. Stockmeyer, *Word problems — this time with interleaving*, Inform. and Comput., 115 (1994), pp. 293–311.

[32] A. R. Meyer and L. J. Stockmeyer, *The equivalence problem for regular expressions with squaring requires exponential space*, in Foundations of Computer Science (FOCS), IEEE, 1972, pp. 125–129.
[33] M. Murata, D. Lee, M. Mani, and K. Kawaguchi, *Taxonomy of XML schema languages using formal language theory*, ACM Transactions on Internet Techologies, 5 (2005), pp. 1–45.
[34] F. Neven and T. Schwentick, *On the complexity of XPath containment in the presence of disjunction, DTDs, and variables.*, Log. Methods Comput. Sci., 2 (2006).
[35] Y. Papakonstantinou and V. Vianu, *DTD inference for views of XML data*, in ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), New York, 2000, ACM Press, pp. 35–46.
[36] F. Reuter, *An enhanced W3C XML Schema-based language binding for object oriented programming languages.* Manuscript, 2006.
[37] H. Seidl, *Deciding equivalence of finite tree automata*, SIAM J. Comput., 19 (1990), pp. 424–437.
[38] ———, *Haskell overloading is DEXPTIME-complete*, Inf. Process. Lett., 52 (1994), pp. 57–60.
[39] C.M. Sperberg-McQueen, *XML Schema 1.0: A language for document grammars*, in XML 2003, 2003.
[40] C.M. Sperberg-McQueen and H. Thompson, *XML Schema*. http://www.w3.org/XML/Schema, 2005.
[41] L.J. Stockmeyer and A.R. Meyer, *Word problems requiring exponential time: Preliminary report*, in ACM Symposium on Theory of Computing (STOC), ACM Press, 1973, pp. 1–9.
[42] E. van der Vlist, *XML Schema*, O'Reilly, 2002.
[43] P. van Emde Boas, *The convenience of tilings*, in Complexity, Logic and Recursion Theory, vol. 187 of Lect. Notes Pure Appl. Math., 1997, pp. 331–363.
[44] G. Wang, M. Liu, J. X. Yu, B. Sun, G. Yu, J. Lv, and H. Lu, *Effective schema-based XML query optimization techniques*, in International Database Engineering and Applications Symposium (IDEAS), 2003, pp. 230–235.