ON THE COMPLEXITY OF VERIFYING CONSISTENCY OF XML SPECIFICATIONS

MARCELO ARENAS†, WENFEI FAN‡, AND LEONID LIBKIN§

Abstract. XML specifications often consist of a type definition (typically, a DTD) and a set of integrity constraints. It has been shown previously that such specifications can be inconsistent, and thus it is often desirable to check consistency at compile-time. It is known [16] that for general keys and foreign keys, and DTDs, the consistency problem is undecidable; however, it becomes NP-complete when all keys are one-attribute (unary), and tractable, if no foreign keys are used.

In this paper, we consider a variety of previously studied constraints for XML data, and investigate the complexity of the consistency problem. Our main conclusion is that in the presence of foreign key constraints, compile-time verification of consistency is infeasible. We look at absolute constraints that hold in the entire document, and relative constraints that only hold in a part of the document. For absolute constraints, we prove decidability and establish complexity bounds for primary multi-attribute keys and unary foreign keys, and study unary constraints that involve regular expressions. For relative constraints, we prove that even for unary constraints, the consistency problem is undecidable. We also show that results continue to hold for extended DTDs, a more expressive typing mechanism for XML.

1. Introduction. XML data, just like relational and object-oriented data, can be specified in a certain data definition language. While the exact details of XML data definition languages are still being worked out, it is clear that all of them would contain a form of document description, as well as integrity constraints. Constraints are naturally introduced when one considers transformations between XML and relational databases [10, 12, 18, 19, 23, 30, 31], as well as integrating several XML documents [2, 3, 4, 15].

Document descriptions usually come in the form of DTDs (Document Type Definition), and constraints are typically natural analogs of the most common relational integrity constraints: keys and foreign keys. Indeed, a large number of proposals (e.g., [35, 38, 36, 5]) support specifications for keys and foreign keys.

We investigate XML specifications with DTDs and keys and foreign keys. We study the consistency, or satisfiability, of such specifications: given a DTD and a set of constraints, whether there are XML documents conforming to the DTD and satisfying the constraints. In other words, we want to validate XML documents statically, at compile-time. Invalid XML specifications are likely to be more common than invalid specifications of other kinds of data, due to the rather complex interaction of DTDs and constraints. Furthermore, many specifications are not written at once, but rather in stages: as new requirements are discovered, they are added to the constraints, and thus it is quite possible that at some point they may be contradictory.

An alternative to the static validation would be a dynamic approach: simply attempt to validate a document with respect to a DTD and a set of constraints. This, however, would not tell us whether repeated failures are due to a bad specification, or problems with the documents.

The consistency analysis for XML specifications is not nearly as easy as for relational data (any set of keys and foreign keys can be declared on a set of relational attributes). Indeed, [16] showed that for DTDs and arbitrary keys and foreign keys, the consistency problem is undecidable. Furthermore, under the restriction that all keys and foreign keys are unary (single-attribute), the problem is NP-complete.

These results only revealed the tip of the iceberg, as many other flavors of XML constraints exist, and are likely to be added to future standards for XML such as XML Schema [38]. One of our goals is to study such constraints. In particular, we concentrate on constraints with regular expressions, and relative constraints that only hold in a part of the document. We now give examples of new kinds of constraints considered here, and explain their consistency problem.

Constraints with regular expressions. As XML data is hierarchically structured, one is often interested in constraints specified by regular expressions. For example, consider an XML document (represented as a node-labeled tree) in Figure 1.1, which conforms to the following DTD for schools:

```xml
<!ELEMENT r (students, courses, faculty, labs)>
<!ELEMENT students (student*)>
<!ELEMENT courses (cs340, cs108, cs434)>
```

†An extended abstract was presented at the 21st ACM Symposium on Principles of Database Systems (PODS 2002).
‡Pontificia Universidad Católica de Chile, Santiago, Chile (marenas@ing.puc.cl)
§School of Informatics, University of Edinburgh, Edinburgh EH8 9LE, UK (wenfei@inf.ed.ac.uk)
§School of Informatics, University of Edinburgh, Edinburgh EH8 9LE, UK (libkin@inf.ed.ac.uk)
Here we omit the descriptions of elements whose type is string (PCDATA). Assume that each record element has an attribute @id, each takenBy has an attribute @sid (for student id), and each acc has an attribute @num. One may impose the following constraints over the DTD of that document:

\[
\begin{align*}
& r.*.(\text{student} \cup \text{prof}).\text{record}.@id \rightarrow r.*.(\text{student} \cup \text{prof}).\text{record}, \\
& r.*.\text{cs434}.\text{takenBy}.@sid \subseteq_{\text{FK}} r.*.\text{student}.\text{record}.@id, \\
& r.*.\text{dbLab}.\text{acc}.@num \subseteq_{\text{FK}} r.*.\text{cs434}.\text{takenBy}.@sid. \\
\end{align*}
\]

Here ‘*’ is a wildcard that matches any label (tag) and ‘.’ is its Kleene closure that matches any path. The first constraint says that @id is a key for all records of students and professors. The other constraints specify foreign keys, which assert that cs434 can only be taken by students, and only students who are taking cs434 can have an account in the database lab. Recall that a foreign key also imposes a key constraint on the target elements, e.g., the last foreign key above also says that @sid is a key for students taking cs434.

Clearly, there is an XML tree satisfying both the DTD and the constraints. As was mentioned earlier, specifications are rarely written at once. Now suppose a new requirement is discovered: all faculty members must have a dbLab account. Consequently, one adds a new foreign key:

\[
r.\text{faculty}.\text{prof}.\text{record}.@id \subseteq_{\text{FK}} r.*.\text{dbLab}.\text{acc}.@num. \\
\]

However, this addition makes the whole specification inconsistent. This is because previous constraints postulate that dbLab users are students taking cs434, and no professor can be a student since @id is a key for both students and professors, while the new foreign key insists upon professors also being dbLab users and the DTD enforces at least one professor to be present in the document. Thus no XML document both conforms to the DTD and satisfies all the constraints.

The consistency problem for regular expression constraints is at least as hard as for constraints specified for element types with simple attributes: NP-hard in the unary case and undecidable in general [16]. We use results from [1, 16, 27] to show that in the unary case, the problem remains decidable, but the lower bound becomes PSPACE.

Relative integrity constraints. Many types of constraints are specified for an entire document. A different kind of constraints, called relative, was proposed in [5] – those constraints only hold in a part of a document. As an example, consider an XML document that for each country lists its administrative subdivisions (e.g., into provinces or states), as well as capitals of provinces. A DTD is given below and an XML document conforming to it is depicted in Figure 1.2.
Each country has a nonempty sequence of provinces and a nonempty sequence of province capitals, and for each province we specify its capital and perhaps other cities. Each country and province has an attribute @name, and each capital has an attribute @inProvince.

Now suppose we want to define keys for countries and provinces. One can state that country @name is a key for country elements. It is also tempting to say that @name is a key for province, but this may not be the case. The example in Figure 1.2 clearly shows that; which Limburg one is interested in probably depends on whether one’s interests are in database theory, or in the history of the European Union. To overcome this problem, we define @name to be a key for province relative to a country; indeed, it is extremely unlikely that two provinces of the same country would have the same name. Thus, our constraints are:

\[
\text{country.}@name \rightarrow \text{country}, \\
\text{country(province.}@name \rightarrow \text{province),} \\
\text{country(}^*.\text{capital.}@inProvince \rightarrow ^*.\text{capital),} \\
\text{country(}^*.\text{capital.}@inProvince \subseteq_{FK} ^*.\text{province.}@name).}
\]

The first constraint is like those we have encountered before: it is an absolute key, which applies to the entire document. The rest are relative constraints which are specified for sub-documents rooted at country elements. They assert that for each country, @name is a key of all province descendants of the country element and @inProvince is a key of all capital descendants of the country element and it is a foreign key referring to @name of province elements in the same sub-document. The foreign keys assure that for each capital element in a country element (sub-document), its @inProvince attribute refers to a province in the same country (recall that capital elements immediately below country also denote province capitals). Note that these constraints are somewhat related to the notion of keys for weak entities in relational databases (cf. [33]). In contrast to regular expression constraints given earlier, these constraints are defined for element types, e.g., the first constraint is a key for all country elements in the entire document, and the second constraint is a (relative) key for all capital elements in a sub-document rooted at a country node.

To illustrate the interaction between constraints and DTDs, observe that the above specification – which might look reasonable at first – is actually inconsistent! To see this, let T be a tree that satisfies the specification. The constraints say that for any sub-document rooted at a country c, the number of its capital elements is at most the number of province elements among c’s descendants. The DTD says that for each province element in a country element (sub-document), its @inProvince attribute refers to a province in the same country (recall that capital elements immediately below country also denote province capitals). Thus, the number of capital descendants of c is larger than the number of province descendants of c, which contradicts the previous bound. Hence, the specification is inconsistent. We note that one can make the specification consistent by replacing country(\(^*.\text{capital.}@inProvince \rightarrow ^*.\text{capital}) with two keys:
country(capital.@inProvince → capital) and country(province.capital.@inProvince → province.capital),
which allow capital.@inProvince and province.capital.@inProvince to share the same value.

Relative constraints appear to be quite useful for capturing information about XML documents that
cannot possibly be specified by absolute constraints. It turns out, however, that the consistency problem is
much harder for them: it is undecidable even for single-attribute keys and foreign keys.

Decidable restrictions. Since expensive lower bounds, and even undecidability, were established for most
versions of the consistency problem, we would like to see some interesting tractable, or decidable, restrictions.
In case of absolute constraints, the results of [16] consider either single attributes or multi-attribute sets for
both keys and foreign keys, and thus say nothing about the intermediate case in which only keys are allowed
to be multi-attribute. This class of constraints is rather common and arises when relational data is translated
into XML. While often identifiers are used as single-attribute keys, other sets of attributes can form a key
as well (e.g., via SQL unique declaration) and those typically contain more than one attribute. We show
that the consistency problem for this class of constraints, when every key is primary (i.e., at most one key
is defined for each element type), remains decidable.

The main conclusion of this paper is that while many proposals such as XML Schema [38] and XML
Data [36] support the facilities provided by the DTDs as well as integrity constraints, and while it is possible
to write inconsistent specifications, checking consistency at compile-time appears to be infeasible, even for fairly small specifications.

Related work. Consistency was studied for other data models, such as object-oriented and extended
relational (e.g., with support for cardinality constraints), see [8, 9, 22].

A number of specifications for XML keys and foreign keys have been proposed, e.g., XML Schema [38],
XML-Data [36]. A recent proposal [5] introduced relative constraints. To the best of our knowledge, consistency
of XML constraints in the presence of schema specifications was only investigated in [16]. However,
[16] did not consider relative constraints, constraints defined with regular expressions and the class of multi-
attribute keys and unary foreign keys. Other constraints for semi-structured data, different from those
considered here, were studied in, e.g., [1, 6, 17]. The latter also studies the consistency problem; the special
form of constraints used there makes it possible to encode consistency as an instance of conjunctive query
containment. Application of constraints in data transformations was studied in [23, 12]; usefulness of keys
and foreign keys in query optimization has also been recognized [13, 14].

Organization. Section 2 defines DTDs, absolute keys and foreign keys for XML. Section 3 studies the
class of absolute multi-attribute keys and unary foreign keys, and the class of regular expression constraints
which is an extension of absolute constraints with regular path expressions. Section 4 defines and investigates
relative keys and foreign keys. Section 5 provides lower and upper bounds for the consistency problem for
extended DTDs, a slight extension of DTDs which captures unranked tree automata, and several different
classes of keys and foreign keys. Section 6 summarizes the main results of the paper.

2. Notations.

2.1. DTDs, XML Trees and Paths. Assume that we have the following disjoint sets: $El$ of element
names, $Att$ of attribute names, $S$ of possible values of attributes and raw text, and $Vert$ of node identifiers.
All attribute names start with the symbol @, and these are the only ones starting with this symbol. We let $S$
be a reserved symbol not in any of those sets.

We formalize the notion of DTDs as follows (cf. [35, 7, 25, 16]).

**Definition 2.1.** A DTD (Document Type Definition) is defined to be
$D = (E, A, P, R, r)$, where:

- $E \subseteq El$ is a finite set of element types;
- $A \subseteq Att$ is a finite set of attributes;
- $P$ is a mapping from $E$ to element type definitions: Given $\tau \in E$, $P(\tau) = S$ or $P(\tau)$ is a regular
  expression $\alpha$ defined as follows:

  $$\alpha ::= \epsilon \mid \tau' \mid \alpha | \alpha \mid \alpha \mid \alpha^*$$

  where $S$ denotes the string type, $\tau' \in E$, $\epsilon$ is the empty word, and $\mid$, ” and “*” denote union,
  concatenation, and the Kleene closure, respectively;
- $R$ is a mapping from $E$ to the powerset of $A$. If $\exists \alpha \in R(\tau)$, we say that $\exists \alpha$ is defined for $\tau$.


The parent-child edge relation on documents w.r.t. a DTD, along the same lines as XQuery [39], XML Schema [38] and DOM [34].

We normally denote element types by $\tau$, and assume that $R(r) = \emptyset$ and $r$ does not appear in $P(\tau)$ for any $\tau \in E$. We also assume that each $\tau$ in $E \setminus \{r\}$ is connected to $r$, i.e., either $\tau$ appears in $P(r)$, or it appears in $P(\tau')$ for some $\tau'$ that is connected to $r$. In this paper we also use the following shorthands for regular expressions: $\alpha^+$ for $(\alpha^+)$, and $\alpha?$ for $(\epsilon|\alpha)$. Finally, notice that mixed content is not allowed in XML trees; for every $\tau \in E$, $P(\tau)$ is either $\emptyset$ or a regular expression over $E$.

**Example 2.2.** Let us consider the DTD $D$ given in Section 1 for storing information about countries and their administrative subdivisions. In our formalism, $D$ can be represented as $(E, A, P, R, r)$, where $E = \{db, country, province, capital, city\}$, $A = \{@name, @inProvince\}$, $r = db$ and $P, R$ are as follows:

$$
\begin{align*}
P(db) &= country^+ \quad R(db) = \emptyset \\
P(country) &= (province^+, capital^+) \quad R(country) = \{@name\} \\
P(province) &= (capital, city^+) \quad R(province) = \{@name\} \\
P(capital) &= S \quad R(capital) = \{@inProvince\} \\
P(city) &= S \quad R(city) = \emptyset
\end{align*}
$$

An XML document is typically modeled as a node-labeled tree. Below we describe valid XML documents w.r.t. a DTD, along the same lines as XQuery [39], XML Schema [38] and DOM [34].

**Definition 2.3.** Let $D = (E, A, P, R, r)$ be a DTD. An XML tree $T$ conforming to $D$, written $T \models D$, is defined to be $(V, lab, ele, att, root)$, where

- $V \subseteq Vert$ is a finite set of nodes;
- $lab : V \to E$; if $lab(v) = \tau$ ($v \in V$), $\tau$ is said to be the element type of $v$;
- $ele : V \to S \cup V^*$, where $V^*$ is the set of all the finite sequences of values from $V$, such that for every $v \in V$, if $P(lab(v)) = S$, then $ele(v) = [s]$, where $s \in S$, otherwise $ele(v) = [v_1, \ldots, v_n]$, and the string $lab(v_1) \cdots lab(v_n)$ is in the regular language defined by $P(lab(v))$.
- $att$ is a partial function from $V \times A$ to $S$ such that for any $v \in V$ and $@l \in A$, $att(v, @l)$ is defined iff $@l \in R(lab(v))$.
- $root$ is the root of $T$: $root \in V$ and $lab(root) = r$.

The parent-child edge relation on $V$, $\{(v_1, v_2) \mid v_2$ occurs in $ele(v_1)\}$, is required to form a rooted tree.

In an XML tree $T$, for each $v \in V$, there is a unique path of parent-child edges from the root to $v$, and each node has at most one incoming edge. The root is a unique node labeled with $r$. If a node $x$ is labeled $\tau$ in $E$, then function $ele$ defines the children of $x$ and function $att$ defines the attributes of $x$. The children of $x$ are ordered and their labels observe the regular expression $P(\tau)$. In contrast, their attributes are unordered and are identified by their labels (names).

**Example 2.4.** Figure 2.1 (a) shows an XML document storing information about provinces in Canada and conforming to the DTD shown in Example 2.2. Figure 2.1 (b) shows an XML tree $T = (V, lab, ele, att, v_0)$ representing this document. In this tree, $V = \{v_i \mid i \in [0, 6]\}$ and $lab$ is defined as:

$$
\begin{align*}
lab(v_0) &= db \\
lab(v_1) &= country \\
lab(v_2) &= province \\
lab(v_3) &= capital \\
lab(v_4) &= province \\
lab(v_5) &= capital
\end{align*}
$$

Furthermore, function $ele$ is defined as:

![An XML document represented as a tree.](image)
Finally, function $\text{att}$ is defined as:

- $\text{att}(v_1, \text{name}) = \text{Canada}$
- $\text{att}(v_1, \text{inProvince}) = \text{Ontario}$
- $\text{att}(v_2, \text{name}) = \text{Alberta}$
- $\text{att}(v_2, \text{inProvince}) = \text{Alberta}$
- $\text{att}(v_3, \text{name}) = \text{Manitoba}$
- $\text{att}(v_3, \text{inProvince}) = \text{Manitoba}$

Our model is simpler than the models of XQuery and XML Schema as DTDs support only one basic type (PCDATA or string) and do not have complex type constructs. Unlike the data model of XQuery, we do not consider nodes representing namespaces, processing instructions and references. These simplifications do not affect the lower bounds, however.

We also use the following notations. Referring to an XML tree $T$, if $x$ is a $\tau$-element in $T$ and $@l$ is an attribute in $R(\tau)$, then $x.@l$ denotes the $@l$-attribute value of $x$, i.e., $x.@l = \text{att}(x, @l)$. If $X$ is a list $[@l_1, \ldots, @l_n]$ of attributes in $R(\tau)$, then $x[X] = [x.@l_1, \ldots, x.@l_n]$. For any element type $\tau' \in E$, $\text{ext}(\tau)$ denotes the set of all the $\tau'$-elements in $T$. For any $\tau \in R(\tau)$, $\text{values}(\tau, \omega)$ denotes $\{x.\omega \mid x \in \text{ext}(\tau)\}$, the set of all the $@\omega$-attribute values of $\tau$-nodes. We write $|S|$ for the cardinality of a set $S$.

Given a DTD $D$ and a set $\Sigma$ of constraints, we also use $|D|$ and $|\Sigma|$ to denote their sizes, respectively.

Given a DTD $D = (E, A, P, R, r)$ and element types $\tau, \tau' \in E$, a string $\tau_1.\tau_2.\cdots.\tau_n$ over $E$ is a path in $D$ from $\tau$ to $\tau'$ if $\tau_1 = \tau$, $\tau_n = \tau'$ and for each $i \in [2, n]$, $\tau_i$ is a symbol in the alphabet of $P(\tau_{i-1})$. Moreover, $\text{paths}(D) = \{ p \mid \text{there is } \tau \in E \text{ such that } p \text{ is a path in } D \text{ from } r \text{ to } \tau \}$. We say that a DTD is non-recursive if $\text{paths}(D)$ is finite, and recursive otherwise. We also say that $D$ is a no-star DTD if the Kleene star does not occur in any regular expression $P(\tau)$ (note that this is a stronger restriction than being $*$-free: a regular expression without the Kleene star yields a finite language, while the language of a $*$-free regular expression may still be infinite as it allows boolean operators including complement).

2.2. Keys and Foreign Keys. We consider two forms of constraints for XML: absolute constraints that hold on the entire document, denoted by $\mathcal{AC}$; and relative constraints that hold on certain sub-documents, denoted by $\mathcal{RC}$. Below we define absolute keys and foreign keys, and we shall define relative constraints in Section 4. The constraints given in Section 1 are instances of absolute constraints and relative constraints.

Regular expression constraints. To capture the hierarchical nature of XML data, absolute constraints, in their general form, are defined on a collection of elements identified by a regular path expression. It is common to find path expressions in specification and query languages for XML (e.g., XML Schema [38], XQuery [39], XSL [40]). We define a regular (path) expression over a set of element types $E$ as follows:

$$\beta ::= \epsilon \mid \tau \mid \beta.\beta \mid \beta \cup \beta \mid \beta^*,$$

where $\epsilon$ denotes the empty word, $\tau$ is an element type in $E$ and $\cdot$, $\cup$ and $\ast$ denote concatenation, union and Kleene closure, respectively. A regular expression defines a language over the alphabet $E$, which will be denoted by $\beta$ as well. Given a DTD $D = (E, A, P, R, r)$ and a regular expression $\beta$ over $E$, we say that $\beta$ is a regular (path) expression over $D$ if $\beta$ is of the form $r.\beta'$ where $\beta'$ does not include $r$. In this section, we use $\cdot$ as a shorthand for $E \setminus \{r\}$.

Recall that a path in a DTD is a list of $E$ symbols, that is, a string in $E^*$. Given an XML tree $T = (V, \text{lab}, \text{ele}, \text{att}, \text{root})$, a pair of nodes $x, y$ in $T$ with $y$ a descendant of $x$ and a path $w = \tau_1.\cdots.\tau_n$ over $E$, we say that $w$ is a path from $x$ to $y$ if there exists a sequence of nodes $v_1, \ldots, v_n$ in $T$ such that (1) $v_1 = x$ and $v_n = y$, (2) $v_{i+1}$ is a child of $v_i$ in $T$, for every $i \in [1, n-1]$, and (3) $\text{lab}(v_i) = \tau_i$, for every $i \in [1, n]$. Any pair of nodes $x, y$ in an XML tree $T$ with $y$ a descendant of $x$ uniquely determines the path, denoted by $\rho(x, y)$, from $x$ to $y$. We say that $y$ is reachable from $x$ by following a regular expression $\beta$ over $D$, denoted by $T \models \beta(x, y)$, if $\rho(x, y) \in \beta$. For any fixed $T$, let $\text{nodes}(\beta)$ stand for the set of nodes reachable from the root by following the regular expression $\beta$: $\text{nodes}(\beta) = \{ y \mid T \models \beta(\text{root}, y) \}$. Note that for any element type $\tau \in E \setminus \{r\}$, $\text{nodes}(\tau, \cdot, \tau) = \text{ext}(\tau)$.

We now define unary XML keys and foreign keys with regular path expressions. Let DTD $D = (E, A, P, R, r)$:

- A key over $D$ is an expression $\varphi$ of the form $\beta.\tau[X] \rightarrow \beta.\tau$, where
  - $\tau \in E$;
  - $X$ is a nonempty set of attributes in $R(\tau)$; and
- $\beta$ is a regular expression over $D$.

For any XML tree $T$ that conforms to $D$, the tree $T$ satisfies $\varphi$, denoted by $T \models \varphi$, if

$$\forall x, y \in \text{nodes}(\beta, \tau). (x[X] = y[X] \rightarrow x = y).$$

- A foreign key over $D$ is an expression $\varphi$ of the form $\beta_1.\tau_1[X] \subseteq_{FK} \beta_2.\tau_2[Y]$, where
  - $\tau_i \in E$ for $i = 1, 2$;
  - $\beta$ is a regular expression over $D$, for $i = 1, 2$; and
  - $X, Y$ are nonempty lists of attributes in $R(\tau_1), R(\tau_2)$ of the same length.

Here $T \models \varphi$ if $T \models \beta_2.\tau_2[Y] \rightarrow \beta_2.\tau_2$, and

$$\forall x \in \text{nodes}(\beta_1, \tau_1) \exists y \in \text{nodes}(\beta_2, \tau_2) (x[X] = y[Y]).$$

We use two notions of equality to define keys: value equality is assumed when comparing attributes, and node identity is used when comparing elements. We shall use the same symbol ‘$=$’ for both, as it will never lead to ambiguity.

The above constraints are generally referred to as multi-attribute regular expression constraints as they may be defined with multiple attributes. A regular expression key (foreign key) is said to be unary if it is defined in terms of a single attribute; that is, $|X|=1$ ($|Y|=1$) in the above definition. In that case, we write $\beta.\tau.\emptyset l \rightarrow \beta.\tau$ for regular expression unary keys, and $\beta_1.\tau_1.\emptyset l_1 \subseteq_{FK} \beta_2.\tau_2.\emptyset l_2$ for regular expression unary foreign keys.

From [16], we immediately obtain that the consistency problem for regular expression constraints is undecidable. Thus, in this paper we only study the consistency problem for unary constraints defined with regular expressions. We denote this class of constraints by $\mathcal{AC}_{\text{reg}}^{FK}$, where subscripts $K$ and $FK$ stand for keys and foreign keys, respectively. For example, the constraints over the school DTD that we have seen in Section 1 are instances of $\mathcal{AC}_{\text{reg}}^{FK}$.

Constraints associated with element types. A class of absolute keys and foreign keys, denoted by $\mathcal{AC}_{FK}^{\ast}$ (we shall explain the notation shortly), has been studied in [16]. It is a special case of regular-expression constraints and is defined for element types as follows. An $\mathcal{AC}_{FK}^{\ast}$-constraint $\varphi$ over a DTD $D = (E, A, P, R, r)$ has one of the following forms:

- Key. $\tau[X] \rightarrow \tau$, where $\tau \in E$ and $X$ is a nonempty set of attributes in $R(\tau)$. An XML tree $T$ satisfies $\varphi$, denoted by $T \models \varphi$, if

$$\forall x, y \in \text{ext}(\tau). (x[X] = y[X] \rightarrow x = y).$$

- Foreign key. $\tau_1[X] \subseteq_{FK} \tau_2[Y]$, where $\tau_1, \tau_2 \in E$ and $X, Y$ are nonempty lists of attributes in $R(\tau_1), R(\tau_2)$ of the same length. It is satisfied by a tree $T$ if $T \models \tau_2[Y] \rightarrow \tau_2$, and in addition

$$\forall x \in \text{ext}(\tau_1) \exists y \in \text{ext}(\tau_2) (x[X] = y[Y]).$$

That is, $\tau[X] \rightarrow \tau$ says that the $X$-attribute values of a $\tau$-element uniquely identify the element in $\text{ext}(\tau)$. Furthermore, $\tau_1[X] \subseteq_{FK} \tau_2[Y]$ says that the list of $X$-attribute values of every $\tau_1$-node in $T$ must match the list of $Y$-attribute values of some $\tau_2$-node in $T$ and the $Y$-attribute values of a $\tau_2$-element uniquely identify the element in $\text{ext}(\tau_2)$.

Note that an $\mathcal{AC}_{FK}^{\ast}$-constraint can be readily expressed as a regular-expression constraint, by using $r^\ast.\tau$ for $\tau$. As for the case of regular expression constraints, an $\mathcal{AC}_{FK}^{\ast}$-constraint is generally referred to as a multi-attribute constraint as it may be defined with multiple attributes. An $\mathcal{AC}_{FK}^{\ast}$-constraint is said to be unary if it is defined in terms of a single attribute; that is, $|X|=1$ ($|Y|=1$) in the above definition. In that case, we write $\tau.\emptyset l \rightarrow \tau$ for unary keys, and $\tau_1.\emptyset l_1 \subseteq_{FK} \tau_2.\emptyset l_2$ for unary foreign keys. As in relational databases, we also consider primary keys for each element type, at most one key can be defined.

We shall use the following notations for subclasses of $\mathcal{AC}_{FK}^{\ast}$: subscripts $K$ and $FK$ denote keys and foreign keys, respectively. When the primary key restriction is imposed, we use subscript $PK$ instead of $K$. The superscript ‘$\ast$’ denotes multi-attribute, and ‘$1$’ means unary. When both superscripts are left out,
Meaning

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{AC}^*_{K,FK}$</td>
<td>multi-attribute keys and foreign keys</td>
</tr>
<tr>
<td>$\mathcal{AC}^1_{PK,FK}$</td>
<td>multi-attribute primary keys, unary foreign keys</td>
</tr>
<tr>
<td>$\mathcal{AC}_K,FK$</td>
<td>unary keys and foreign keys</td>
</tr>
<tr>
<td>$\mathcal{AC}_{PK,FK}$</td>
<td>primary unary keys and unary foreign keys</td>
</tr>
<tr>
<td>$\mathcal{AC}^{reg}_{K,FK}$</td>
<td>regular expression unary keys and foreign keys</td>
</tr>
</tbody>
</table>

Fig. 2.2. Notation summary.

we mean that both keys and foreign keys are unary. We shall be dealing with the following subclasses of $\mathcal{AC}^*_{K,FK}$: $\mathcal{AC}^*_{K,FK}$ denotes the class of multi-attribute keys and unary foreign keys; $\mathcal{AC}^1_{PK,FK}$ is the class of primary multi-attribute keys and unary foreign keys; $\mathcal{AC}_K,FK$ is the class of unary keys and unary foreign keys; and $\mathcal{AC}_{PK,FK}$ is the class of primary unary keys and unary foreign keys. We note that since a key is part of a foreign key, the restriction of $\mathcal{AC}^*_{K,FK}$ to unary keys and multi-attributes foreign keys ($\mathcal{AC}^{1*}_{K,FK}$) does not make sense.

For easy reference, in Figure 2.2 we summarize our notation for absolute constraints.

### 2.3. The Consistency Problem

We are interested in the consistency, or satisfiability problem for XML constraints considered together with DTDs: that is, whether a given set of constraints and a DTD are satisfiable by an XML tree. Formally, for a class $\mathcal{C}$ of integrity constraints we define the XML specification consistency problem $\text{SAT}(\mathcal{C})$ as follows:

**PROBLEM:** $\text{SAT}(\mathcal{C})$

**INPUT:** A DTD $D$, a finite set $\Sigma$ of $C$-constraints.

**QUESTION:** Is there an XML tree $T$ such that $T \models D$ and $T \models \Sigma$?

It is known [16] that $\text{SAT}(\mathcal{AC}^*_{K,FK})$ is undecidable, but $\text{SAT}(\mathcal{AC}^1_{K,FK})$ and $\text{SAT}(\mathcal{AC}_{PK,FK})$ are NP-complete. Nothing was known however about $\text{SAT}(\mathcal{AC}^{1*}_{K,FK})$, where only keys are allowed to be multi-attribute, or about $\text{SAT}(\mathcal{AC}^{reg}_{K,FK})$, where regular expressions are used to define unary keys and foreign keys. These problems will be studied in Section 3.

In what follows, we write $T \models (D, \Sigma)$ instead of $T \models D$ and $T \models \Sigma$.

### 3. Absolute Integrity Constraints

In this section, we establish the decidability and lower bounds for $\text{SAT}(\mathcal{AC}^1_{PK,FK})$ and $\text{SAT}(\mathcal{AC}^{reg}_{K,FK})$, the consistency problems for absolute primary multi-attribute keys and unary foreign keys, and for regular-expression unary keys and unary foreign keys.

#### 3.1. Consistency of Multi-attribute Keys

We know that $\text{SAT}(\mathcal{AC}_K,FK)$, the consistency problem for unary absolute keys and foreign keys, is NP-complete [16]. In contrast, $\text{SAT}(\mathcal{AC}^{1*}_{K,FK})$ is undecidable [16]. This leaves a large gap: namely, $\text{SAT}(\mathcal{AC}^{1*}_{K,FK})$, where only keys are allowed to be multi-attribute.

The reason for the undecidability of $\text{SAT}(\mathcal{AC}^*_{K,FK})$ is that the implication problem for functional and inclusion dependencies can be reduced to it [16]. However, this implication problem is known to be decidable – in fact, in cubic time – for single-attribute inclusion dependencies [11], thus giving us hope to get decidability for multi-attribute keys and unary foreign keys.

The problem we resolve here is $\text{SAT}(\mathcal{AC}^1_{PK,FK})$: the consistency problem for primary multi-attribute keys and unary foreign keys. Recall that a set $\Sigma$ of $\mathcal{AC}^1_{PK,FK}$-constraints is said to be primary if for each element type $\tau$, there is at most one key in $\Sigma$ defined for $\tau$-elements (including key dependencies defined by foreign key constraints). Even dealing with this version of $\text{SAT}(\mathcal{AC}^1_{PK,FK})$ one encounters considerable difficulties: with a rather involved proof, we manage to show that this problem is equivalent to a certain decidable version of Diophantine equations problem whose exact complexity has been an open problem for a while [21]:

**PROBLEM:** PDE (Prequadratic Diophantine Equations)

**INPUT:** An integer $n \times m$ matrix $A$, a vector $\vec{b} \in \mathbb{Z}^n$, and a set $E \subseteq \{1, \ldots, m\}$.

**QUESTION:** Is there a vector $\vec{x} \in \mathbb{N}^m$ such that $A\vec{x} \leq \vec{b}$ and $x_i \leq x_j \cdot x_k$ for all $(i,j,k) \in E$?
Note that for $E = \emptyset$, this is exactly the integer linear programming problem [27]. Thus, PDE can be thought of as integer linear programming extended with inequalities of the form $x \leq y - z$ among variables. It is therefore NP-hard, and [21] proved an NEXPTIME upper bound for PDE. The exact complexity of the problem remains unknown.

Recall that two problems $P_1$ and $P_2$ are polynomially equivalent if there are PTIME reductions from $P_1$ to $P_2$ and from $P_2$ to $P_1$. We now show the following.

**Theorem 3.1.** SAT($\mathcal{AC}_{PK,FK}^*$) and PDE are polynomially equivalent. □

**Proof.** The proof consists of two PTIME reductions, one for each direction.

a) A reduction from SAT($\mathcal{AC}_{PK,FK}^*$) to PDE. We first define a class of simplified DTDs called narrow DTDs, and we explain how to reduce the consistency problem for $\mathcal{AC}_{PK,FK}^*$-constraints over arbitrary DTDs to that over narrow DTDs. Then we show how to encode the consistency problem for narrow DTDs and $\mathcal{AC}_{PK,FK}^*$-constraints by a prequadratic Diophantine system.

We start by explaining the process of narrowing the DTDs. Intuitively, we replace long “horizontal” regular expressions in $P(\tau)$ by shorter ones. Formally, consider a DTD $D = (E, A, P, R, r)$. $D$ is basically an extended regular grammar (cf. [7, 25]); for each $\tau \in E$, $P(\tau)$ is a regular expression in the context), where

$$\tau \rightarrow \tau_1, \tau_2 \quad \tau \rightarrow \tau_1 \mid \tau_2 \quad \tau \rightarrow \tau_1^* \quad \tau \rightarrow \tau' \quad \tau \rightarrow S \quad \tau \rightarrow \epsilon$$

where $\tau, \tau_1, \tau_2$ are element types in $E \cup N$, $\tau' \in E$, $S$ is the string type and $\epsilon$ denotes the empty word. More specifically, we conduct the following “narrowing” process on the production rule $\tau \rightarrow \alpha$:

- If $\alpha = (\tau_1, \tau_2)$, then we introduce two new element types $\tau_1, \tau_2$ and replace $\tau \rightarrow \alpha$ with a new rule $\tau \rightarrow \tau_1, \tau_2$. We proceed to process $\tau_1 \rightarrow \alpha_1$ and $\tau_2 \rightarrow \alpha_2$ in the same way.
- If $\alpha = (\tau_1, \tau_2)$, then we introduce two new element types $\tau_1, \tau_2$ and replace $\tau \rightarrow \alpha$ with a new rule $\tau \rightarrow \tau_1 \mid \tau_2$. We proceed to process $\tau_1 \rightarrow \alpha_1$ and $\tau_2 \rightarrow \alpha_2$ in the same way.
- If $\alpha = \alpha_1^*$, then we introduce a new element type $\tau_1$ and replace $\tau \rightarrow \alpha$ with $\tau \rightarrow \tau_1^*$. We proceed to process $\tau_1 \rightarrow \alpha_1$ in the same way.
- If $\alpha$ is one of $\tau' \in E$, $S$ or $\epsilon$, then the rule for $\tau$ remains unchanged.

We refer to the set of new element types introduced when processing $\tau \rightarrow P(\tau)$ as $N_\tau$ and the set of production rules generated/revised as $P_\tau$. Observe that $N_\tau \cap E = \emptyset$ for any $\tau \in E$. We define a new DTD $D_N = (E_N, A, P_N, R_N, r)$, referred to as the narrowed DTD of $D$ (or just a narrow DTD if $D$ is clear from the context), where

- $E_N = E \cup \bigcup_{\tau \in E} N_\tau$, i.e., all element types of $E$ and new element types introduced in the narrowing process;
- $P_N = \bigcup_{\tau \in E} P_\tau$, i.e., production rules generated/revised in the narrowing process;
- $R_N(\tau) = R(\tau)$ for each $\tau \in E$, and $R_N(\tau) = \emptyset$ for each $\tau \in E_N \setminus E$.

Note that the root element type $r$ and the set $A$ of attributes remain unchanged. Moreover, elements of any type in $E_N \setminus E$ do not have any attribute. The only kind of $P_N$ production rules whose right-hand side contains element type of $E$ are of the form $\tau \rightarrow \tau'$, where $\tau' \in E$. It is easy to see that $D_N$ is computable in polynomial time.

Obviously, any set $\Sigma$ of $\mathcal{AC}_{PK,FK}^*$-constraints over $D$ is also a set of $\mathcal{AC}_{PK,FK}^*$-constraints over the narrow DTD $D_N$ of $D$. The next lemma establishes the connection between $D$ and $D_N$, which allows us to consider only narrow DTDs from now on.

**Lemma 3.2.** Let $D$ be a DTD, $D_N$ the narrowed DTD of $D$ and $\Sigma$ a set of $\mathcal{AC}_{PK,FK}^*$-constraints over $D$. Then there exists an XML tree $T_1$ such that $T_1 \models (D, \Sigma)$ iff there exists an XML tree $T_2$ such that $T_2 \models (D_N, \Sigma)$.

**Proof.** Given an element type $\tau$ and a sequence of attributes $\#l_1, \ldots, \#l_n \in R(\tau)$, define values($\tau[\#l_1, \ldots, \#l_n]$) as $\{x[\#l_1, \ldots, \#l_n] \mid x \in ext(\tau)\}$.

To prove the lemma, it suffices to show the following:
Claim: Given any XML tree $T_1 \models D$ one can construct an XML tree $T_2$ by modifying $T_1$ such that $T_2 \models D_N$, and vice versa. Furthermore, for every element type $\tau$ in $D$ and $\langle \alpha_1, \ldots, \alpha_n \rangle \in R(\tau)$, $|\text{ext}(\tau)|$ in $T_2$ equals $|\text{ext}(\tau)|$ in $T_1$, and $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_2$ equals $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_1$.

For the claim holds, we can show the lemma as follows. Assume that there exists an XML tree $T_1$ such that $T_1 \models D$ and $T_2 \models \Sigma$. By the claim, there is $T_2$ such that $T_2 \models D_N$. Suppose, by contradiction, there is $\varphi \in \Sigma$ such that $T_2 \not\models \varphi$. (1) If $\varphi$ is a key $\tau[\alpha_1, \ldots, \alpha_n] \rightarrow \tau$, then there exist two distinct nodes $x, y \in \text{ext}(\tau)$ in $T_2$ such that $\langle \alpha_i \rangle = y, \langle \alpha_i \rangle$ for every $i \in [1, n]$. In other words, $|\text{values}(\tau[\alpha_1, \ldots, \alpha_n])| < |\text{ext}(\tau)|$ in $T_2$. Since $T_1 \models \varphi$, it must be the case that $|\text{values}(\tau[\alpha_1, \ldots, \alpha_n])| = |\text{ext}(\tau)|$ in $T_1$ because the tuple $(x.\langle \alpha_1 \rangle, \ldots, x.\langle \alpha_n \rangle)$ of each $x \in \text{ext}(\tau)$ uniquely identifies $x$ among $\text{ext}(\tau)$. This contradicts the claim that $|\text{ext}(\tau)|$ in $T_2$ equals $|\text{ext}(\tau)|$ in $T_1$ and $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_2$ equals $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_1$.

(2) If $\varphi$ is a unary foreign key: $\tau_1[\alpha_1] \subseteq_{FK} \tau_2[\alpha_2]$, then either $T_2 \not\models \tau_2[\alpha_2] \rightarrow \tau_2$ or there is $x \in \text{ext}(\tau_1)$ in $T_2$ such that for all $y \in \text{ext}(\tau_2)$ in $T_2$, $x.\langle \alpha_1 \rangle \not= y.\langle \alpha_2 \rangle$. In the first case, we reach a contradiction as in (1). In the second case, we have $x.\langle \alpha_1 \rangle \not\in \text{values}(\tau_2[\alpha_2])$ in $T_2$. By the claim, $x.\langle \alpha_1 \rangle \in \text{values}(\tau_1[\alpha_1])$ in $T_1$. Since $T_1 \models \varphi$, $x.\langle \alpha_1 \rangle \in \text{values}(\tau_2[\alpha_2])$ in $T_1$. Again by the claim, we have $x.\langle \alpha_1 \rangle \in \text{values}(\tau_2[\alpha_2])$ in $T_2$, which contradicts the assumption. The proof for the other direction is similar.

We next verify the claim. Given an XML tree $T_1 = (V_1, lab_1, ele_1, att, root)$ such that $T_1 \models D$, we construct an XML tree $T_2$ by modifying $T_1$ such that $T_2 \models D_N$. Consider a $\tau$-element $v$ in $T_1$. Let $ele_1(v) = \langle v_1, \ldots, v_n \rangle$ and $w = lab_1(v_1) \ldots lab_1(v_n)$. Recall $N_\tau$ and $P_\tau$, the set of nonterminals and the set of production rules generated when narrowing $\tau \rightarrow P(\tau)$. Let $Q_\tau$ be the set of E symbols that appear in $P_\tau$ plus $S$. We can view $G = (Q_\tau, N_\tau \cup \{\tau\}, P_\tau)$ as an extended context free grammar, where $Q_\tau$ is the set of terminals, $N_\tau \cup \{\tau\}$ the set of nonterminals, $P_\tau$ the set of production rules and $\tau$ the start symbol. Since $T_1 \models D$, we have $w \in P(\tau)$. By a straightforward induction on the structure of $P_N(\tau)$ it can be verified that $w$ is in the language defined by $G$. Thus there is a parse tree $T(w)$ w.r.t. the grammar $G$ for $w$, and $w$ is the frontier (the list of leaves from left to right) of $T(w)$. Without loss of generality, assume that the root of $T(w)$ is $v$, and the leaves are $v_1, \ldots, v_n$. Observe that the internal nodes of $T(w)$ are labeled with element types in $N_\tau$ except that the root $v$ is labeled $\tau$. Intuitively, we construct $T_2$ by replacing each element $v$ in $T_1$ by such a parse tree. More specifically, let $T_2 = (V_2, lab_2, ele_2, att, root)$. Here $V_2$ consists of nodes in $V_1$ and the internal nodes introduced in the parse trees. For each $x \in V_2$, let $lab_2(x) = lab_1(x)$ if $x \in V_1$, and otherwise let $lab_2(x)$ be the node label of $x$ in the parse tree where $x$ belongs. Note that nodes in $V_2 \setminus V_1$ are elements of some type in $E_N \setminus E$. For every $x \in V_1$, let $ele_2(x)$ be the list of its children in the parse tree having $x$ as root. For every $x \in V_2 \setminus V_1$, let $ele_2(x)$ be the list of its children in the parse tree of an element in $V_1$ that contains $x$. Note that $att$ and $root$ remain unchanged. By the construction of $T_2$ it can be verified that $T_2 \models D_N$; and moreover, for every element type $x \in D$ and $\langle \alpha_1, \ldots, \alpha_n \rangle \in R(\tau)$, $|\text{ext}(\tau)|$ in $T_2$ equals $|\text{ext}(\tau)|$ in $T_1$ and $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_2$ equals $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_1$ because, among other things, (1) none of the new nodes, i.e., nodes in $V_2 \setminus V_1$, is labeled with an $E$-type; (2) no new attributes are defined; and (3) attribute function $att$ is unchanged.

Conversely, assume that there is $T_2 = (V_2, lab_2, ele_2, att, root)$ such that $T_2 \models D_N$. We construct an XML tree $T_1$ by modifying $T_2$ such that $T_1 \models D$. For every node $v \in V_2$ with $lab_2(v) = \tau$ and $\tau \in E_N \setminus E$, we substitute $v$ in $ele_2(v')$ by the children of $v$, where $v'$ is the parent of $v$. In addition, we remove $v$ from $V_2$, $lab_2(v)$ from $lab_1(v)$, and $ele_2(v)$ from $ele_1(v)$. Observe that by the definition of $D_N$, no attributes are defined for elements of any type in $E_N \setminus E$. We repeat the process until there is no node labeled with element type in $E_N \setminus E$. Now let $T_1 = (V_1, lab_1, ele_1, att, root)$, where $V_1$, $lab_1$ and $ele_1$ are $V_2$, $lab_2$ and $ele_2$ at the end of the process, respectively. Notice that $att$ and $root$ remain unchanged. By the definition of $T_1$ it can be verified that $T_1 \models D$; and in addition, for every element type $\tau$ in $D$ and $\langle \alpha_1, \ldots, \alpha_n \rangle \in R(\tau)$, $|\text{ext}(\tau)|$ in $T_2$ equals $|\text{ext}(\tau)|$ in $T_1$ and $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_2$ equals $\text{values}(\tau[\alpha_1, \ldots, \alpha_n])$ in $T_1$ because, among other things, none of the nodes removed is labeled with a type of $E$ and the attribute function $att$ is unchanged.

By Lemma 3.2, in the rest of this proof we consider only narrow DTDs. Next we show how to encode $\mathcal{A}_{PK,FK}^{\tau}$-constraints by a prequadratic Diophantine system. Let $D = (E, A, P, R, \tau)$ be a narrow DTD and $\Sigma$ be a set of $\mathcal{A}_{PK,FK}^{\tau}$-constraints, i.e., primary $\mathcal{A}_{PK,FK}^{\tau}$-constraints. We encode $\Sigma$ with a set $C_\Sigma$ of

\[1\text{If } \tau \in P(\tau), \text{ i.e., if } \tau \text{ is recursively defined, we need to rename } \tau \text{ in } Q_\tau \text{ to ensure that } Q_\tau \text{ and } N_\tau \cup \{\tau\} \text{ are disjoint. It is straightforward to handle that case.}\]
integer constraints, referred to as the cardinality constraints determined by $\Sigma$. For every $\varphi \in \Sigma$,

- if $\varphi$ is a key constraint $\tau[\alpha_1, \ldots, \alpha_k] \rightarrow \tau$, then $C_\Sigma$ contains $|\text{ext}(\tau)| \leq |\text{values}(\tau[\alpha_1])| \cdot \ldots \cdot |\text{values}(\tau[\alpha_k])|$;
- if $\varphi$ is a unary foreign key $\tau_1[\alpha_1] \subseteq\text{FK}_2 \tau_2[\alpha_2]$, then $C_\Sigma$ contains $|\text{values}(\tau_1[\alpha_1])| \leq |\text{values}(\tau_2[\alpha_2])|$ and $|\text{ext}(\tau_2)| \leq |\text{values}(\tau_2[\alpha_2])|$;
- furthermore, for any $\tau \in E$, if $R(\tau) = \emptyset$, then $0 \leq |\text{ext}(\tau)|$ is in $C_\Sigma$. Otherwise, for every $\alpha \in R(\tau)$, $|\text{values}(\tau[\alpha])| \leq |\text{ext}(\tau)|$ and $0 \leq |\text{values}(\tau[\alpha])|$ are in $C_\Sigma$.

Observe that for a unary key $\tau[\alpha] \rightarrow \tau$ we have both $|\text{values}(\tau[\alpha])| \leq |\text{ext}(\tau)|$ and $|\text{ext}(\tau)| \leq |\text{values}(\tau[\alpha])|$ in $C_\Sigma$. Thus $C_\Sigma$ assures $|\text{ext}(\tau)| = |\text{values}(\tau[\alpha])|$.

We write $T \models C_\Sigma$ if $T$ satisfies all the constraints of $C_\Sigma$, and we write $T \models (D, C_\Sigma)$ if $T$ conforms to a narrow DTD $D$ and satisfies $C_\Sigma$. Note that $C_\Sigma$ is equivalent (in fact, can be converted in polynomial time) to a prequadratic Diophantine system since $x \leq x_1 \ldots x_4$ can be written as constraints of the form $x \leq y \cdot z$ by introducing $k-2$ fresh variables, e.g., $x \leq x_1 \cdot x_2 \cdot x_3 \cdot x_4$ is equivalent to $x \leq x_1 \cdot z_1 \cdot z_2 \cdot x_2$ and $z_2 \leq x_3 \cdot x_4$ (in the sense that the former is satisfiable iff the latter is). Thus, without loss of generality, assume that $C_\Sigma$ consists of linear and prequadratic integer constraints only. It should be noted that $C_\Sigma$ can be computed in time polynomial in the size of $\Sigma$ and $D$. The lemma below shows that $C_\Sigma$ characterizes the consistency of $\Sigma$ if keys in $\Sigma$ are primary.

**Lemma 3.3.** Let $D$ be a narrow DTD and $\Sigma$ a set of $\mathcal{AC}_\text{FK}^*$-constraints over $D$. Then every XML tree conforming to $D$ and satisfying $\Sigma$ also satisfies $C_\Sigma$. In addition, if there exists an XML tree $T_2$ such that $T_2 \models (D, C_\Sigma)$, then there exists an XML tree $T_1$ such that $T_1 \models (D, \Sigma)$.

**Proof.** It is easy to see that for every XML tree $T_1$ that satisfies $\Sigma$, it must be the case that $T_1 \models C_\Sigma$.

Conversely, we show that if there exists an XML tree $T_2 = (V, \text{lab}, \text{ele}, \text{att}_2, \text{root})$ such that $T_2 \models (D, C_\Sigma)$, then we can construct an XML tree $T_1 = (V, \text{lab}, \text{ele}, \text{att}_1, \text{root})$ such that $T_1 \models (D, \Sigma)$. We construct $T_1$ from $T_2$ by modifying the function $\text{att}_2$ while leaving $V$, lab, ele and root unchanged. More specifically, let $S = \{\tau.[\alpha] | \tau \in E, \alpha \in R(\tau)\}$. To define the new function, denoted by $\text{att}_1$, we first associate a set of string values with each $\tau[\alpha]$ in $S$. Let $N$ be the maximum cardinality of $|\text{values}(\tau[\alpha])|$ in $T_2$, i.e., $N \geq |\text{values}(\tau[\alpha])|$ in $T_2$ for all $\tau[\alpha] \in S$. Let $V_\Sigma = \{a_i | i \in [1,N]\}$ be a set of distinct string values. For each $\tau[\alpha] \in S$, let $V_{\tau[\alpha]} = \{a_i | i \in [1,|\text{values}(\tau[\alpha])|]\}$, and for each $\tau \in \text{ext}(\tau)$, let $\text{att}_1(x[\alpha])$ be a string value in $V_{\tau[\alpha]}$ such that in $T_1$, $\text{values}(\tau[\alpha]) = V_{\tau[\alpha]}$. The value $\text{att}_1(x[\alpha])$ can be selected in such a way that for each key $\varphi = \tau[\alpha_1, \ldots, \alpha_k] \rightarrow \tau$ in $\Sigma$, $x[\alpha_1, \ldots, \alpha_k]$ is a distinct list of string values from $V_{\tau[\alpha_1]} \times \ldots \times V_{\tau[\alpha_k]}$. This is possible because by the definition of $T_1$, (1) $|\text{ext}(\tau)|$ in $T_1$ equals $|\text{ext}(\tau)|$ in $T_2$; (2) $|\text{values}(\tau[\alpha])|$ in $T_1$ equals $|\text{values}(\tau[\alpha])|$ in $T_2$; (3) $T_2 \models C_\Sigma$ and $|\text{ext}(\tau)| \leq |\text{values}(\tau[\alpha])|$ for every $\tau \in E$; and (4) since $\varphi$ is the only key defined for $\tau$-elements, when we populate attributes $\alpha_1, \ldots, \alpha_k$ of $x$, we only need to select the value of $\text{att}_1(x[\alpha])$ from $V_{\tau[\alpha]}$ such that $x[\alpha_1, \ldots, \alpha_k]$ is distinct, without worrying about whether the population may hamper “other keys” defined on $x$ (note that in the absence of the primary key assumption, the populations of different keys may interact with each other and as a result, the simply population strategy given above may no longer work; this is why we assume primary keys). It should be noted that it may be the case that $V_{\tau_1[\alpha_1]} \subseteq V_{\tau_2[\alpha_2]}$, even if $\Sigma$ does not imply $\tau_1[\alpha_1] \subseteq\text{FK}_1 \tau_2[\alpha_2]$. This does not lose generality as we do not intend to capture negation of foreign keys. We next show that $T_1$ is indeed what we want.

It is easy to verify that $T_1 \models D$ given the construction of $T_1$ from $T_2$ and the assumption that $T_2 \models D$. To show that $T_1 \models \Sigma$, we consider $\varphi \in \Sigma$ in the following cases. (1) If $\varphi$ is a key $\tau[\alpha_1, \ldots, \alpha_k] \rightarrow \tau$, it is immediate from the definition of $T_1$ that $T_1 \models \varphi$ since for any $x \in \text{ext}(\tau)$, $x[\alpha_1, \ldots, \alpha_k]$ is a distinct list of string values from $V_{\tau[\alpha_1]} \times \ldots \times V_{\tau[\alpha_k]}$. (2) If $\varphi$ is a unary foreign key $\tau_1[\alpha_1] \subseteq\text{FK}_2 \tau_2[\alpha_2]$, then $T_2 \models |\text{values}(\tau_1[\alpha_1])| \leq |\text{values}(\tau_2[\alpha_2])|$ in $T_2 \models C_\Sigma$. By the definition of $\text{att}_1$, for $i = 1, 2$, $V_{\tau[\alpha_i]} = \{a_i | i \in [1,|\text{values}(\tau[\alpha_i])|]\}$ and hence $T_1$, $|\text{values}(\tau_1[\alpha_1])| \leq |\text{values}(\tau_2[\alpha_2])|$ in $T_1$. Furthermore, given that $|\text{ext}(\tau_2)| \leq |\text{values}(\tau_2[\alpha_2])|$ and $|\text{values}(\tau_2[\alpha_2])| \leq |\text{ext}(\tau)|$ are both in $C_\Sigma$, $T_2 \models C_\Sigma$, $|\text{ext}(\tau_2)|$ in $T_2$ is equal to $|\text{ext}(\tau_2)|$ in $T_1$ and $|\text{values}(\tau_2[\alpha_2])|$ in $T_2$ is equal to $|\text{values}(\tau_2[\alpha_2])|$ in $T_1$, we conclude that $|\text{ext}(\tau_2)|$ equal to $|\text{values}(\tau_2[\alpha_2])|$ in $T_1$ and, hence, $T_1 \models \tau_2[\alpha_2] \rightarrow \tau_2$ since each $x \in \text{ext}(\tau_2)$ in $T_1$ has a distinct $\alpha_2$-attribute value and thus the value of its $\alpha_2$-attribute uniquely identifies $x$ among nodes in $\text{ext}(\tau_2)$. Therefore, $T_1 \models \varphi$ and, thus, $T_1 \models (D, \Sigma)$. This concludes the proof of the lemma.

The above lemma takes care of coding the constraints; the next step is to code DTDs. For that, we use the technique developed in [16]: for each narrow DTD $D$, one can compute in polynomial time in the size
of $D$ a set $\Psi_D$ of linear inequalities on nonnegative integers, referred to as the set of cardinality constraints determined by $D$, which includes $|\text{ext}(\tau)|$ as a variable for each element type $\tau$ in $D$, but it does not have $|\text{values}(\tau, \@l)|$ as a variable for any attribute $\@l$ of $\tau$. More specifically, for each symbol $\tau \in E \cup \{\Sigma\}$, $|\text{ext}(\tau)|$ is treated as a distinct variable, which keeps track of the number of all $\tau$ elements in an XML tree $T$ conforming to $D$. In addition, for each occurrence of $\tau$ in the definition $P(\tau')$ of some element type $\tau'$, we also create distinct variables as follows: if $P(\tau') = \tau_1$ for $\tau_1 \in E \cup \{\Sigma\}$, then we create a distinct variable $x^1_{\tau_1,\tau'}$; if $P(\tau') = (\tau_1, \tau_2)$ or $P(\tau') = (\tau_1|\tau_2)$, then we create two distinct variables $x^1_{\tau_1,\tau'}$ and $x^2_{\tau_1,\tau'}$. Intuitively, for $i \in [1, 2]$, $x^i_{\tau_1,\tau'}$ keeps track of the number of $\tau_1$ subelements at position $i$ under all $\tau'$ elements in $T$. Let $X_\tau$ be the set of all variables of the form $x^i_{\tau_1,\tau'}$. Using these variables, for each $\tau \in E$, we define a set $\psi_\tau$ of linear integer constraints that characterizes $P(\tau)$ quantitatively, as follows:

- If $P(\tau) = \tau_1$ for $\tau_1 \in E \cup \{\Sigma\}$, then $\psi_\tau$ includes $|\text{ext}(\tau)| = x^1_{\tau_1,\tau}$. Referring to an XML tree $T$ that conforms to $D$, this assures that each $\tau$ element has a unique $\tau_1$ subelement.
- If $P(\tau') = (\tau_1, \tau_2)$, then $\psi_\tau$ includes $|\text{ext}(\tau)| = x^1_{\tau_1,\tau'}$ and $|\text{ext}(\tau)| = x^2_{\tau_1,\tau'}$. These assure that each $\tau$ element in $T$ must have a unique $\tau_1$ subelement and a unique $\tau_2$ subelement.
- If $P(\tau') = (\tau_1|\tau_2)$, then $\psi_\tau$ includes $|\text{ext}(\tau)| = x^1_{\tau_1,\tau'} + x^2_{\tau_1,\tau'}$. These assure that each $\tau$ element in $T$ must have either a $\tau_1$ subelement or a $\tau_2$ subelement, and thus the sum of the numbers of these $\tau_1$ and $\tau_2$ subelements equals the number of $\tau$ elements in $T$.

The set $\Psi_D$ of cardinality constraints determined by DTD $D$ consists of the following:

- $|\text{ext}(\tau)| = 1$; i.e., there is a unique root in any XML tree valid w.r.t. $D$;
- constraints of $\psi_\tau$ for each $\tau \in E$; these assure that $P(\tau)$ is satisfied;
- $|\text{ext}(\tau)| = \sum_{x^i_{\tau_1,\tau'} \in X_\tau} x^i_{\tau_1,\tau'}$ for each $\tau \in (E \setminus \{\tau\}) \cup \{\Sigma\}$; this indicates that the set $\text{ext}(\tau)$ includes all $\tau$ elements no matter where they occur in an XML tree;
- $x \geq 0$ for any variable $x$ used above; i.e., the number of elements (subelements) is nonnegative.

It has been shown [16] that $\Psi_D$ has a nonnegative integer solution if and only if there exists an XML tree $T$ conforming to $D$ such that the cardinality of $\text{ext}(\tau)$ in $T$ equals the value of the variable $|\text{ext}(\tau)|$ in the solution for each element type $\tau$ in $D$.

We now combine this coding with the coding for $\mathcal{AC}_{PK,FK}^{n+1}$-constraints. Given a narrow DTD $D$ and a set $\Sigma$ of $\mathcal{AC}_{PK,FK}^{n+1}$-constraints over $D$, we define the set of cardinality constraints determined by $D$ and $\Sigma$ to be

$$\Psi(D, \Sigma) = \Psi_D \cup C_\Sigma \cup \{|\text{ext}(\tau)| > 0 \rightarrow (|\text{values}(\tau, \@l)| > 0) \mid \tau \in E, \@l \in R(\tau)\},$$

where $C_\Sigma$ is the set of cardinality constraints determined by $\Sigma$, $\Psi_D$ is the set of cardinality constraints determined by $D$, and constraints $|\text{ext}(\tau)| > 0 \rightarrow (|\text{values}(\tau, \@l)| > 0)$ are to ensure that every $\tau$-element has an $\@l$-attribute (note that $|\text{values}(\tau, \@l)| \leq |\text{ext}(\tau)|$ is already in $C_\Sigma$). Constraints in $\Psi(D, \Sigma)$ are either linear integer constraints, or inequalities of the form $x \leq y \cdot z$, which come from $C_\Sigma$, or constraints of the form $x > 0 \rightarrow y > 0$. Note that if we leave out constraints of the form $x > 0 \rightarrow y > 0$, $\Psi(D, \Sigma)$ is a prequadratic Diophantine system. Also note that $\Psi(D, \Sigma)$ can be computed in polynomial time in the size of $D$ and $\Sigma$.

We say that $\Psi(D, \Sigma)$ is consistent if and only if $\Psi(D, \Sigma)$ admits a nonnegative integer solution. That is, there is a nonnegative integer assignment to the variables in $\Psi(D, \Sigma)$ such that all the constraints in $\Psi(D, \Sigma)$ are satisfied.

**Lemma 3.4.** Let $D$ be a narrow DTD and $\Sigma$ a set of $\mathcal{AC}_{PK,FK}^{n+1}$-constraints over $D$. Then $\Psi(D, \Sigma)$ is consistent if and only if there is an XML tree $T$ such that $T \models (D, \Sigma)$.

**Proof.** Suppose that there exists an XML tree $T$ such that $T \models (D, \Sigma)$. Then there is a nonnegative integer solution to $\Psi_D$ such that for each element type $\tau$ in $D$, the value of the variable $|\text{ext}(\tau)|$ equals the number of $\tau$-elements in $T$ [16]. By Lemma 3.3 and $T \models \Sigma$, we have $T \models C_\Sigma$. We extend the solution of $\Psi_D$ to be one to $\Psi(D, \Sigma)$ by letting the variable $|\text{values}(\tau, \@l)|$ equal the number of distinct $\@l$-attribute values of all $\tau$-elements in $T$, for each element type $\tau$ and attribute $\@l$ of $\tau$ in $D$. Since $T \models C_\Sigma$, this extended assignment satisfies all the constraints in $C_\Sigma$. In addition, if $|\text{ext}(\tau)| > 0$ then $|\text{values}(\tau, \@l)| > 0$ since every $\tau$-element in $T$ has an $\@l$-attribute. Hence the assignment is indeed a nonnegative solution to $\Psi(D, \Sigma)$ and, therefore, $\Psi(D, \Sigma)$ is consistent.
Conversely, suppose that $\Psi(D, \Sigma)$ admits a nonnegative integer solution. Then there exists an XML tree $T$ such that $T \models D$ and moreover, for each element type $\tau$ in $D$, the cardinality of $\text{ext}(\tau)$ in $T$ equals the value of the variable $|\text{ext}(\tau)|$ in the solution [16]. We construct a new tree $T''$ from $T$ by modifying the definition of the function $\text{att}$ such that in $T''$, for each element type $\tau$ and attribute $\text{att}$ of $\tau$, the number of distinct $\text{att}$-attribute values of all $\tau$-elements equals the value of the variable $|\text{values}(\tau, \text{att})|$ in the solution. This is possible since $|\text{values}(\tau, \text{att})| \leq |\text{ext}(\tau)|$ and $(|\text{ext}(\tau)| > 0) \rightarrow (|\text{values}(\tau, \text{att})| > 0)$ are in $\Psi(D, \Sigma)$. The assignment is also a solution to $C_S$. Thus $T'' \models D$ and $T' \models C_S$. Hence by Lemma 3.3, there exists an XML tree $T''$ such that $T'' \models (D, \Sigma)$. This concludes the proof of the lemma.

We now conclude the proof of reduction from SAT($\mathcal{A}^*_P, \mathcal{F}_K$) to PDE. By Lemma 3.2, given an arbitrary DTD $D$ and a set $\Sigma$ of $\mathcal{A}^*_P, \mathcal{F}_K$-constraints, one can compute a narrow DTD $D_N$ such that $(D, \Sigma)$ is consistent iff $(D_N, \Sigma)$ is consistent. By Lemma 3.4, $(D_N, \Sigma)$ is consistent iff $\Psi(D_N, \Sigma)$ has a nonnegative integer solution. Such a solution requires $|\text{values}(\tau, \text{att})| > 0$ if $|\text{ext}(\tau)| > 0$. To ensure this, let $\Phi(D_N, \Sigma)$ be a system that includes all linear integer constraints and prequadratic constraints in $\Psi(D_N, \Sigma)$ and moreover, $|\text{ext}(\tau)| \leq |\text{values}(\tau, \text{att})| \cdot |\text{ext}(\tau)|$ for each $(|\text{ext}(\tau)| > 0) \rightarrow (|\text{values}(\tau, \text{att})| > 0)$ in $\Psi(D_N, \Sigma)$. Now $\Phi(D_N, \Sigma)$ is a prequadratic Diophantine system. In addition, $\Psi(D_N, \Sigma)$ has a nonnegative integer solution iff $\Phi(D_N, \Sigma)$ has a nonnegative integer solution. To see this, observe that for any nonnegative integer assignment to $|\text{ext}(\tau)|$ and $|\text{values}(\tau, \text{att})|$, $(|\text{ext}(\tau)| > 0) \rightarrow (|\text{values}(\tau, \text{att})| > 0)$ iff $|\text{ext}(\tau)| \leq |\text{values}(\tau, \text{att})| \cdot |\text{ext}(\tau)|$. Thus, $(D, \Sigma)$ is consistent iff $\Psi(D_N, \Sigma)$ has a nonnegative integer solution. Note that $D_N$ can be computed in polynomial time in the size of $D$, $\Psi(D_N, \Sigma)$ can be computed in polynomial time in the size of $D_N$ and $\Sigma$, and $\Phi(D_N, \Sigma)$ can be computed in polynomial time in the size of $\Psi(D_N, \Sigma)$. Hence, it takes polynomial time to compute $\Phi(D_N, \Sigma)$ from $D$ and $\Sigma$. Therefore, there is a PTIME reduction from SAT($\mathcal{A}^*_P, \mathcal{F}_K$) to PDE.

b) A reduction from PDE to SAT($\mathcal{A}^*_P, \mathcal{F}_K$). We now move to the other direction. Given an instance of PDE, i.e., a system $S$ consisting of a set $S_L$ of linear equations/inequalities on integers and a set $S_P$ of prequadratic constraints of the form $x \leq y \cdot z$, we define a DTD $D$ and a set $\Sigma$ of $\mathcal{A}^*_P, \mathcal{F}_K$-constraints such that $S$ has a nonnegative solution iff there is an XML tree $T$ satisfying $\Sigma$ and conforming to $D$. We use $X = \{x_i \mid i \in [1, n]\}$ to denote the set of all the variables in $S$. Assume that $S_L = \{c_j \mid j \in [1, m]\}$ and $c_j$ is of the form: $a_1^j x_1 + \cdots + a_n^j x_n + c_j \leq b_1^j x_1 + \cdots + b_n^j x_n + d_j$, where $a_i^j$ (i.e., $i \in [1, n]$), $b_i^j$ (i.e., $i \in [1, n]$), $c_j$ and $d_j$ are nonnegative integers. Also, assume that $S_P = \{p_j \mid j \in [1, l]\}$, where $p_j$ is a prequadratic equation of the form $x \leq y \cdot z$. Then we define DTD $D = (E, A, P, R, r)$ as follows:

1. For each variable $x_i$, we define an element type $X_i$. In addition, for each $p_j \in S_P$ of the form $x_i \leq x_j \cdot x_k$, we define an element type $U_i^j$. For each linear constraint $c_j$, we define distinct element types $E_j$, $A_1^j, \ldots, A_n^j$, $C_j$, $F_j$, $B_1^j, \ldots, B_n^j$, $D_j$. We use $r$ to denote the root element type. That is,

$$E = \{r\} \cup \{X_i \mid i \in [1, n]\} \cup$$

$$\{E_j, A_1^j, \ldots, A_n^j, C_j, F_j, B_1^j, \ldots, B_n^j, D_j \mid j \in [1, m]\} \cup \{U_i^j \mid p_s = x_i \leq x_j \cdot x_k \in S_P\}.$$

Intuitively, referring to an XML tree conforming to $D$, we use $|\text{ext}(X_i)|$ to code the value of the variable $x_i$ in $S$. For every equation $c_j$, we use $|\text{ext}(A_i^j)|$, $|\text{ext}(A_n^j)|$, $|\text{ext}(C_j)|$ to code the values of constants $a_i^j$, $a_n^j$, $c_j$; $|\text{ext}(E_j)|$ to code the value of the expression $a_1^j x_1 + \cdots + a_n^j x_n + c_j$; $|\text{ext}(B_i^j)|$, $\ldots$, $|\text{ext}(B_n^j)|$, $|\text{ext}(D_j)|$ to code the values of constants $b_1^j$, $\ldots$, $b_n^j$, $d_j$; and $|\text{ext}(F_j)|$ to code the value of the expression $b_1^j x_1 + \cdots + b_n^j x_n + d_j$. Furthermore, for each prequadratic equation $p_s = x_i \leq x_j \cdot x_k$ in $S_P$, we create a distinct copy $U_i^j$ of $X_i$. The reason to use $U_i^j$ instead of $X_i$ is to ensure that the set $\Sigma$ of $\mathcal{A}^*_P, \mathcal{F}_K$-constraints defined below is primary.

2. $A = \{@c, @d, @e\}$. Intuitively, we shall define $@e$ as a key and use $@c$ and $@d$ to code prequadratic constraint of the form $x \leq y \cdot z$.

---

2For example, we represent equation $-3x + 5y \leq -7$ as $0x + 5y + 7 \leq 3x + 0y + 0$. 13
(3) We define production rules as follows. For the root of the DTD:

$$P(r) = (X_1, U_1^{a_1,1}, \ldots, U_1^{a_l,n})^*, \ldots, (X_n, U_n^{a_1,1}, \ldots, U_n^{a_l,n})^*,$$

where \( \{s_{i,j} \mid i \in [1,n] \} \) is the set of indexes \( \{ s \mid p_s = x_i \leq x_j, x_k \in S_P \} \). Furthermore, for every \( i \in [1,n] \) and every \( j \in [1,m] \):

- \( P(A_i) = E_j \)
- \( P(C_j) = E_j \)
- \( P(B_i^l) = F_j \)
- \( P(D_j) = F_j \)
- \( P(X_i) = A_{i,1}^{a_1}, \ldots, A_{i,m}^{a_l}, B_{i,1}^1, \ldots, B_{i,m}^1, \ldots, B_{i,m}^{b_l} \).

Finally, for every \( i \in [1,n] \) and every \( s \in [1,l] \) such that \( p_s = x_i \leq x_j, x_k \in S_P \), \( P(U_i^s) = c \).

(4) We define the attribute function \( R \) as follows: for every \( j \in [1,m] \), \( R(E_j) = R(F_j) = \{ @e \} \). In addition, for every \( i \in [1,n] \), \( R(X_i) = \{ @e \} \), and for every \( s \in [1,l] \) such that \( p_s = x_i \leq x_j, x_k \in S_P \), \( R(U_i^s) = \{ @e, @d \} \).

For example, Figure 3.1 (a) shows an XML tree conforming to the DTD constructed from the set of equations \( S_L = \{ 2x_1 \leq x_2 + 4 \} \) and \( S_P = \{ x_1 \leq x_2 \cdot x_3 \} \). We note that this tree codes solution \( x_1 = 1, x_2 = 2, x_3 = 1 \) for this system of equations.

Given DTD \( D \), we define a set \( \Sigma \) of AC\(^*\)\(_{\text{PK,FK}}\)-constraints over \( D \). For each \( j \in [1,m] \), \( \Sigma \) includes keys \( E_j \cup @e \rightarrow E_i, F_j \cup @e \rightarrow F_i \) and foreign key \( E_j \cup @e \subseteq_{FK} F_j \cup @e \). Furthermore, for every \( i, j, k \in [1,n] \) and \( s \in [1,l] \) such that \( p_s = x_i \leq x_j \cdot x_k \in S_P \), \( \Sigma \) includes the following constraints:

- \( U_i^s \cup @e, @d \rightarrow U_i^s \)
- \( U_i^s \cup @e \subseteq_{FK} X_j \cup @e \)
- \( U_i^s \cup @d \subseteq_{FK} X_k \cup @e \).

Clearly, the set \( \Sigma \) is primary, i.e., for any element type \( \tau \) there is at most one key defined. In fact, we use copies \( U_i^s \) of \( X_i \) just to ensure that \( \Sigma \) is primary.

We next show that the encoding is indeed a reduction from PDE to SAT\((\text{AC}_{\text{PK,FK}})^*\). Suppose that \( S \) has a nonnegative solution. Then we construct an XML tree conforming to \( D \) as shown in Figure 3.1 (a).

![Diagram](a)

For every \( i \in [1,n] \) and every \( x_i \)-element \( x \) in \( T \), we let \( x \cup @e \) be a distinct value such that in \( T \), \( \text{values}(X_i \cup @e) = \text{ext}(X_i) \). For every \( j \in [1,m] \) and every \( E_j \)-element \( x \) in \( T \), we let \( x \cup @e \) be a distinct value such that in \( T \), \( \text{values}(E_j \cup @e) = \text{ext}(E_j) \). Likewise, we assign values to the \( @e \)-attribute of the nodes in \( \text{ext}(F_j) \) in such a way that \( \text{values}(F_j \cup @e) = \text{ext}(F_j) \) in \( T \). Finally, for every \( i, j, k \in [1,n] \) and \( s \in [1,l] \) such that \( p_s = x_i \leq x_j \cdot x_k \in S_P \) and for every node \( x \) in \( T \) of type \( U_i^s \), we
let $x[@c, @d]$ be a distinct list of string values from $values(X_j, @e) \times values(X_k, @e)$. This is possible since $x_i \leq x_j \cdot x_k \in S_P$ and by the definition of $T$, $|\mathit{ext}(U_i^c)| = |\mathit{ext}(X_i)| = x_i$, $|values(X_j, @e)| = |\mathit{ext}(X_j)| = x_j$ and $|values(X_k, @e)| = |\mathit{ext}(X_k)| = x_k$. Since $T$ codes a solution of $S$, it is straightforward to prove that $T \models C_S$, the set of cardinality constraints determined by $\Sigma$. Thus, by Lemma 3.3 we conclude that there exists an XML tree $T'$ such that $T' \models (D, \Sigma)$ and, hence, $(D, \Sigma)$ is consistent. Conversely, suppose that there exists an XML tree $T$ such that $T \models (D, \Sigma)$. We construct a solution of $S$ by letting variable $x_i$ equal $|\mathit{ext}(X_i)|$ in $T$. By the definitions of $D$ and $\Sigma$, it is easy to verify that this is indeed a nonnegative integer solution for $S$. In particular, each $p_k = x_i \leq x_j \cdot x_k$ in $S_P$ holds because $T \models (D, \Sigma)$ and, thus, $|\mathit{ext}(X_i)| = |\mathit{ext}(U_i^c)| \leq |values(U_i^c, @e)| \cdot |values(U_i^c, @d)| \leq |values(X_j, @e)| \cdot |values(X_k, @e)| \leq |\mathit{ext}(X_j)| \cdot |\mathit{ext}(X_k)|$.

Theorem 3.1 it follows immediately:

It is known that the linear integer programming problem is NP-hard and can be solved in $\text{NEXPSPACE}$. Thus from $\text{PDE}$ and, thus, it is NP-hard and can be solved in $\text{NEXPSPACE}$. 2

2

lower bound, we encode the quantified boolean formula problem (QBF) as an instance of $\text{SAT}(\mathcal{A}^{*}_{\text{DR}}, \text{FK})$.

Assume that $a = \sum_{i=0}^{k} a_i \cdot 2^i$, where each $a_i (i \in [0, k−1])$ is either 0 or 1 and $a_k = 1$, that is, the binary representation of a is $a_k a_{k−1} \cdots a_0$. To code $a$ in a DTD we include element types $A, Y_0, \ldots, Y_k$ and we define $P$ on these elements as follows:

$$P(Y_i) = \begin{cases} \epsilon & i = 0 \\ Y_{i−1}, Y_{i−1} & \text{Otherwise} \end{cases}$$

and $P(A) = Y_0, \ldots, Y_k$, where $i_1 > \cdots > i_\ell \geq 0$ and $\{i_1, \ldots, i_\ell\}$ is the set of indexes $\{j \in [0, k] | a_j = 1\}$. We note that the size of this set of rules is polynomial in the size of $a$. Furthermore, if an XML tree $T$ conforms to this DTD, then $|\mathit{ext}(Y_0)| = a$ in $T$. For example, if $a = 5$, then $P(A) = Y_2, Y_0, P(Y_2) = Y_1, Y_1, P(Y_1) = Y_0, Y_0$ and $P(Y_0) = \epsilon$ and an XML tree conforming to these rules is of the form shown in Figure 3.1 (b).

Thus, by using this coding in our original reduction of PDE to $\text{SAT}(\mathcal{A}^{*}_{\text{DR}}, \text{FK})$ we can show that there is a PTIME reduction from PDE to $\text{SAT}(\mathcal{A}^{*}_{\text{DR}}, \text{FK})$. This completes the proof of Theorem 3.1.

It is known that the linear integer programming problem is NP-hard and PDE is in $\text{NEXPSPACE}$. Thus from Theorem 3.1 it follows immediately:

Corollary 3.5. $\text{SAT}(\mathcal{A}^{*}_{\text{DR}}, \text{FK})$ is NP-hard, and can be solved in $\text{NEXPSPACE}$. □

Obviously we cannot obtain the exact complexity of $\text{SAT}(\mathcal{A}^{*}_{\text{DR}}, \text{FK})$ without resolving the corresponding question for PDE, which appears to be quite hard [21]. The result of Theorem 3.1 can be generalized to disjoint $\mathcal{A}^{*}_{\text{DR}, \text{FK}}$-constraints: that is, a set $\Sigma$ of $\mathcal{A}^{*}_{\text{DR}, \text{FK}}$-constraints in which for every element type $\tau$ and every two distinct keys $\tau[X] \to \tau$ and $\tau[Y] \to \tau$ in $\Sigma$ (including key dependencies defined by foreign key constraints), $X \cap Y = \emptyset$. The proof of Theorem 3.1 applies almost verbatim to show the following.

Corollary 3.6. The restriction of $\text{SAT}(\mathcal{A}^{*}_{\text{DR}, \text{FK}})$ to disjoint constraints is polynomially equivalent to PDE and, thus, it is NP-hard and can be solved in $\text{NEXPSPACE}$. □

3.2. Consistency of Regular Expression Constraints. Specifications of $\mathcal{A}^{*}_{\text{DR}, \text{FK}}$-constraints are associated with element types. We next consider $\mathcal{A}^{*}_{\text{DR}, \text{FK}}$, the class of unary keys and foreign keys defined in terms of regular path expressions. For $\text{SAT}(\mathcal{A}^{*}_{\text{DR}, \text{FK}})$, we are able to establish both an upper and a lower bound. The lower bound already indicates that the problem is perhaps infeasible in practice, even for very simple DTDs. Finding the precise complexity of the problem remains open, and does not appear to be easy. In fact, even the current proof of the upper bound is quite involved, and relies on combining the techniques from [16] for coding DTDs and constraints with integer linear inequalities, and from [1] for reasoning about constraints given by regular expressions by using the product automaton for all the expressions involved in the constraints.

Theorem 3.7.

a) $\text{SAT}(\mathcal{A}^{*}_{\text{DR}, \text{FK}})$ can be solved in 2-NEXPSPACE.

b) $\text{SAT}(\mathcal{A}^{*}_{\text{DR}, \text{FK}})$ is PSPACE-hard, even for non-recursive no-star DTDs. □

Proof. We reduce $\text{SAT}(\mathcal{A}^{*}_{\text{DR}, \text{FK}})$ to the existence of solution of an (almost) instance of linear integer programming, which happens to be of double-exponential size; hence the 2-NEXPSPACE bound. For the lower bound, we encode the quantified boolean formula problem (QBF) as an instance of $\text{SAT}(\mathcal{A}^{*}_{\text{DR}, \text{FK}})$.
A in \( \Sigma \), we include in \( \Sigma \), we add

used in constraints. For that, we adopt the technique of \cite{1}, and tag all the variables in the coding of DTDs have the information about the path leading to the node, and its relationship with the regular expressions used in constraints. For that, we adopt the technique of \cite{1}, and tag all the variables in the coding of DTDs with states of a certain automaton (the product automaton for all the automata corresponding to the regular expressions used in constraints).

Now it is time to fill in all the details. First, we need some additional notation. For every regular expression \( \beta \) and every attribute \( \forall \), write \( \text{values}(\beta, \forall) \) to denote the set \( \{y, \exists \mid y \in \text{nodes}(\beta) \text{ and } y \in \forall \} \). Observe that for any \( \tau \in E \setminus \{r\} \), and \( \forall \in R(\tau) \), \( \text{values}(\tau \ast \tau, \forall) \) corresponds to our original definition of \( \text{values}(\tau, \forall) \).

We say that a DTD \( D \) is one-attribute if \( D \) contains only one attribute and no element type \( \tau \) such that \( P(\tau) = \emptyset \). We start by showing that SAT\( (\mathcal{AC}_{\text{FK}}^{\text{reg}}) \) can be reduced to the consistency problem for regular expression constraints over one-attribute DTDs. Let \( D = (E, A, P, R, \tau) \) be a DTD and \( \Sigma \) a set of \( \mathcal{AC}_{\text{FK}}^{\text{reg}} \)-constraints over \( D \). First, define DTD \( D_U = (E_U, A_U, P_U, R_U, \tau) \) as follows. For every \( \tau \in E \) and \( \forall \in R(\tau) \), assume that \( \tau_0 \forall \) is a fresh element type symbol. Then define \( E_U \) as \( E \cup \{\tau_0 \mid \tau \in \exists \text{ and } \forall \in R(\tau) \} \) and \( A_U = \{\forall \} \), where \( \forall \) is a fresh attribute symbol. Furthermore, define functions \( P_U \) and \( R_U \) as:

- For every \( \tau \in \exists \) such that \( P(\tau) = \emptyset \), if \( R(\tau) = \{\forall_1, \ldots, \forall_n\} \), where \( n \geq 0 \), then \( P_U(\tau) = \tau_{\forall_1}, \ldots, \tau_{\forall_n} \) and \( R_U(\tau) = \emptyset \).
- For every \( \tau \in \exists \) such that \( P(\tau) \) is a regular expression over \( E \), if \( R(\tau) = \{\forall_1, \ldots, \forall_n\} \), where \( n \geq 0 \), then \( P_U(\tau) = \tau_{\forall_1}, \ldots, \tau_{\forall_n} \) and \( R_U(\tau) = \emptyset \).
- For every \( \tau \in \exists \) and \( \forall \in R(\tau) \), then \( P_U(\tau) = \emptyset \).

We note that if \( P(\tau) = \emptyset \) and \( R(\tau) = \emptyset \), then \( P_U(\tau) = \emptyset \).

Second, define the set \( \Sigma_U \) of \( \mathcal{AC}_{\text{FK}}^{\text{reg}} \)-constraints over \( D_U \) as follows. For every key constraint \( \beta.\forall \rightarrow \beta.\tau \) in \( \Sigma \), we include \( \beta.\tau.\forall \rightarrow \beta.\tau.\tau_0 \) in \( \Sigma_U \), and for every foreign key constraint \( \beta.\tau.\forall \subseteq \tau \beta'.\forall' \) in \( \Sigma \), we add \( \beta.\tau.\forall \subseteq \mathcal{FK} \beta'.\tau'.\forall' \) to \( \Sigma_U \).

**Lemma 3.8.** Let \( D \) be a DTD, \( \Sigma \) be a set of \( \mathcal{AC}_{\text{FK}}^{\text{reg}} \)-constraints over \( D \), and \( D_U, \Sigma_U \) be as defined above. Then there exists an XML tree \( T_1 \) such that \( T_1 \models (D, \Sigma) \) if and only if there exists an XML tree \( T_2 \) such that \( T_2 \models (D_U, \Sigma_U) \).

**Proof.** (\( \Rightarrow \)) Let \( T_1 = (V_1, \text{lab}_1, \text{ele}_1, \text{att}_1, \text{root}) \) be an XML tree such that \( T_1 \models (D, \Sigma) \). We define an XML tree \( T_2 \) from \( T_1 \) such that \( T_2 \models (D_U, \Sigma_U) \). More specifically, \( T_2 = (V_2, \text{lab}_2, \text{ele}_2, \text{att}_2, \text{root}) \), where \( V_2, \text{lab}_2, \text{ele}_2 \) and \( \text{att}_2 \) are defined as follows. Let \( v \) be a node in \( T_1 \) such that \( \text{lab}_1(v) = \tau \in E \) and \( R(\tau) = \{\forall_1, \ldots, \forall_n\} \). Then \( V_2 \) contains node \( v \) and fresh nodes \( v_{\forall_1}, \ldots, v_{\forall_n} \) such that \( \text{lab}_2(v_\forall) = \tau \) and \( \text{lab}_2(v_{\forall_i}) = \tau_{\forall_i} \), for every \( i \in [1, k] \). Furthermore, if \( \text{ele}_1(v) = [s] \), where \( s \in \Sigma \), then \( \text{ele}_2(v) = [v_{\forall_1}, \ldots, v_{\forall_n}] \). Otherwise, \( \text{ele}_1(v) = [v_1, \ldots, v_n] \), where \( n \geq 0 \) and each \( v_i \) is an element node, and \( \text{ele}_2(v) = [v_1, \ldots, v_n, v_{\forall_1}, \ldots, v_{\forall_n}] \). Finally, \( \text{att}_2(v, \forall_e) \) is not defined and \( \text{att}_2(v_{\forall_1}, \forall_e) = \text{att}_1(v, \forall_1) \), for every \( i \in [1, k] \).

Next we show that \( T_2 \models (D_U, \Sigma_U) \).

By the definition of \( D_U \) and given that \( T_1 \models D \), it is easy to see that \( T_2 \models D_U \). Assume that \( T_2 \not\models \Sigma_U \). Then there exists \( \varphi \in \Sigma_U \) such that \( T_2 \not\models \varphi \). (1) If \( \varphi \) is a key \( \beta.\tau.\forall \rightarrow \beta.\tau.\tau_0 \) in \( T_2 \) such that \( \text{att}_2(v_1, \forall_e) = \text{att}_2(v_2, \forall_e) \). Let \( u_1, u_2 \) be the parents of \( v_1, v_2 \) in \( T_2 \), respectively. By the definition of \( D_U \) and that \( v_1 \neq v_2 \), we have that \( u_1 \neq u_2 \). Thus, by the definition of \( T_2 \), \( u_1 \) and \( u_2 \) are nodes in \( T_1 \) such that \( u_1, u_2 \in \text{nodes}(\beta, \tau) \) and \( \text{att}_1(u_1, \forall) = \text{att}_1(u_2, \forall) = \text{att}_2(v_1, \forall) \). Therefore, \( T_1 \not\models \beta.\tau.\forall \rightarrow \beta.\tau \), which contradicts the assumption that \( T_1 \models \Sigma \). (2) If \( \varphi \) is a foreign key \( \beta.\tau.\forall \subseteq \mathcal{FK} \beta'.\tau'.\forall' \), then either \( T_2 \not\models \beta'.\tau'.\forall' \rightarrow \beta'.\tau'.\tau_0' \) or there exists \( v \in \text{nodes}(\beta.\tau.\forall) \) such that \( \text{att}_2(v, \forall_e) \not\in \text{values}(\beta'.\tau'.\forall', \forall_e) \) in \( T_2 \). In the former case, we reach a contradiction as in (1). In the latter case, assume that \( u \) is the parent of \( v \) in \( T_2 \). By the definition of \( T_2 \), we have that \( u \) is a
node in $T_1$ such that $u \in \text{nodes}(\beta.\tau.)$ and $\text{att}_1(u, \ell l) = \text{att}_2(v, \ell e)$. Thus, given that $\text{values}(\beta'.\tau'.\tau'_{\ell e}\ell p, \ell e)$ in $T_2$ is equal to $\text{values}(\beta'.\tau'.\ell l')$ in $T_1$, we conclude that $\text{att}_1(u, \ell l) \notin \text{values}(\beta'.\tau'.\ell l')$ in $T_1$. Therefore, $T_1 \notin \beta.\tau.\ell l \subseteq \text{FK} \beta'.\tau'.\ell l'$, which contradicts the assumption that $T_1 \models \Sigma$.

$(\Leftarrow)$ Let $T_2 = (V_2, \text{lab}_2, \text{ele}_2, \text{att}_2, \text{root})$ be an XML tree such that $T_2 \models (D_U, \Sigma_U)$. We define an XML tree $T_1$ from $T_2$ such that $T_1 \models (D, \Sigma)$. More specifically, $T_1 = (V_1, \text{lab}_1, \text{ele}_1, \text{att}_1, \text{root})$, where $V_1$, $\text{lab}_1$, $\text{ele}_1$ and $\text{att}_1$ are defined as follows. Let $v$ be a node in $T_2$ such that $\text{lab}_2(v) = \tau$, $\tau \in E$ and $R(\tau) = \{v_{11}, \ldots, v_{1k}\}$. Then $V_1$ also contains node $v$ with $\text{lab}_1(v) = \tau$. Furthermore, if $P(\tau) = S$, then $\text{ele}_2(v) = [v_{a1}, \ldots, v_{a1}]$, and we define $\text{ele}_1(v)$ as $[s]$, where $s$ is an arbitrary string in $S$, and we define $\text{att}_1(v, l_i) = \text{att}_2(v_{a1}, \ell e)$, for every $i \in [1,k]$. Otherwise, if $P(\tau)$ is a regular expression over $E$ and $\text{ele}_2(v) = [v_{11}, \ldots, v_{a1}, v_{a1}, \ldots, v_{a1}]$, where $\text{lab}(v_i) = E (i \in [1,n])$ and $\text{lab}(v_{a1}) = \tau_{a1}$, ($j \in [1,k]$), and we define $\text{ele}_1(v)$ as $[v_{11}, \ldots, v_{a1}]$ and $\text{att}_1(v, l_i) = \text{att}_2(v_{a1}, \ell e)$, for every $i \in [1,k]$. Next we show that $T_1 \models (D, \Sigma)$.

By the definition of $D_U$ and given that $T_2 \models D_U$, it is easy to see that $T_1 \models D$. Assume that $T_1 \notin \Sigma$. Then there exists $\varphi \in \Sigma$ such that $T_1 \models \varphi$. (1) If $\varphi$ is a key $\beta.\tau.\ell l \rightarrow \beta.\tau$, then there exists distinct $u_1, u_2 \in \text{nodes}(\beta.\tau.)$ in $T_1$ such that $\text{att}_1(u_1, \ell l) = \text{att}_1(u_2, \ell l)$. By the definition of $T_1$, $u_1$ and $u_2$ are also in $\text{nodes}(\beta.\tau.)$ in $T_2$. Let $v_1, v_2$ be the children of $u_1, u_2$ in $T_2$ of type $\tau_{a1}, \tau_{a1}$, respectively. Given that $u_1 \neq u_2$, we have that $v_1 \neq v_2$. Thus, by the definition of $T_1$, $v_1$ and $v_2$ are nodes in $T_2$ such that $\varphi(v_1, \ell e) = \varphi(v_2, \ell e), \text{for every } i \in [1,k]$. Therefore, $T_2 \models \beta.\tau.\tau_{a1}
\text{For if the claim holds, we can show the lemma as follows. Assume that there exists an XML tree $T_1$ such that $T_1 \models (D, \Sigma)$. By the claim, there is $T_2$ such that $T_2 \models D_N$. Suppose, by contradiction, there is $\varphi \in \Sigma_N$ such that $T_2 \notin \varphi$. (1) If $\varphi$ is a key $f(\beta.\tau.\ell l) \rightarrow f(\beta.\tau)$, then there exist two distinct nodes $x, y \in \text{nodes}(f(\beta.\tau.))$ in $T_2$ such that $x.\ell l = y.\ell l$. In other words, $|\text{values}(f(\beta.\tau.\ell l))| < |\text{nodes}(f(\beta.\tau.))|$ in $T_2$. Since $T_1 \models \varphi$, it must be that $|\text{values}(\varphi.\ell l)| = |\text{nodes}(\varphi.\ell l)|$ in $T_1$ because the value $x.\ell l$ of
each $x \in \text{nodes}(\beta, \tau)$ uniquely identifies $x$ among $\text{nodes}(\beta, \tau)$. This contradicts the claim that $|\text{nodes}(f(\beta, \tau))|$ in $T_2$ equals $|\text{nodes}(f(\beta, \tau))|$ in $T_1$ and values($f(\beta, \tau, \text{nil})$) in $T_2$ equals values($\beta, \tau, \text{nil}$) in $T_1$. (2) If $\varphi$ is a foreign key: $f(\beta_1) \cdot \tau_1, \text{nil} \subseteq \text{FK}_{f(\beta_2)}$, then either $T_2 \not\models f(\beta_2, \tau_2, \text{nil}) \rightarrow f(\beta_2, \tau_2)$ or there is $x \in \text{nodes}(f(\beta_1), \tau_1)$ such that for all $y \in \text{nodes}(f(\beta_2), \tau_2)$ in $T_2$, $y, \text{nil} \neq y, \text{nil}$. In the first case, we reach a contradiction as in (1). In the second case, we have $x, \text{nil} \not\in \text{values}(f(\beta_2), \tau_2, \text{nil})$ in $T_2$. By the claim, $x, \text{nil} \in \text{values}(\beta_1, \tau_1, \text{nil})$ in $T_1$. Since $T_1 \models \varphi$, $x, \text{nil} \in \text{values}(\beta_1, \tau_1, \text{nil})$ in $T_1$. Again by the claim, we have $x, \text{nil} \in \text{values}(f(\beta_2), \tau_2, \text{nil})$ in $T_2$, which contradicts the assumption. The proof for the other direction is similar.

We next verify the claim. Given an XML tree $T_1 = (V_1, lab_1, ele_1, att, root)$ such that $T_1 \models D$, we construct an XML tree $T_2$ by modifying $T_1$ such that $T_2 \models D_N$. Consider a $\tau$-element $v$ in $T_1$. Let $e_1(v) = [v_1, \ldots, v_n]$ and $w = lab_2(v_1) \ldots lab_2(v_n)$. Recall $N_{\tau}$ and $P_{\tau}$, the set of nonterminals and the set of production rules generated when narrowing $\tau \rightarrow P(\tau)$ (see proof of Theorem 3.1). Let $Q_\tau$ be the set of E symbols that appear in $P_{\tau}$. We can view $G = (Q_\tau, N_\tau \cup \{\tau\}, P_{\tau}, \tau)$ as an extended context-free grammar, where $Q_\tau$ is the set of terminals, $N_\tau \cup \{\tau\}$ the set of nonterminals, $P_{\tau}$ the set of production rules and $\tau$ the start symbol. Since $T_1 \models D$, we have $w \in P(\tau)$. By a straightforward induction on the structure of $P_N(\tau)$ it can be verified that $w$ is in the language defined by $G$. Thus there is a parse tree $T(w)$ w.r.t. the grammar $G$ for $w$, and $w$ is the frontier (the list of leaves from left to right) of $T(w)$. Without loss of generality, assume that the root of $T(w)$ is $v$, and the leaves are $v_1, \ldots, v_n$. Observe that the internal nodes of $T(w)$ are labeled with element types in $N_{\tau}$ except that the root $v$ is labeled $\tau$. Intuitively, we construct $T_2$ by replacing each element $v$ in $T_1$ by such a parse tree. More specifically, let $T_2 = (V_2, lab_2, ele_2, att, root)$. Here $V_2$ consists of nodes in $V_1$ and the internal nodes introduced in the parse trees. For each $x \in V_2$, let $lab_2(x) = lab_2(x)$ if $x \in V_1$, and otherwise let $lab_2(x)$ be the node label of $x$ in the parse tree where $x$ belongs. Note that nodes in $V_2 \setminus V_1$ are elements of some type in $E_N \setminus E$. For every $x \in V_2$, let $ele_2(x)$ be the list of its children in the parse tree having $x$ as root. For every $x \in V_2 \setminus V_1$, let $ele_2(x)$ be the list of its children in the parse tree of an element in $V_1$ that contains $x$. Note that $att$ remains unchanged. By the construction of $T_2$ it can be verified that $T_2 \models D_N$; and moreover, for every regular expression $\beta \cdot \tau$ over $D$ and $\text{nil} \in R(\tau)$, $|\text{nodes}(f(\beta, \tau))|$ in $T_2$ equals $|\text{nodes}(f(\beta, \tau))|$ in $T_1$ and values($f(\beta, \tau, \text{nil})$) in $T_2$ equals values($\beta, \tau, \text{nil}$) in $T_1$ because, among other things, (1) if a string $r, \tau_1, \ldots, \tau_n, \tau$ over $E$ is in $\beta, \tau$, then for every sequence of strings $w_0, \ldots, w_n$ in $(E_N \setminus E)^*$, $w_0, \tau_1, w_1, \ldots, \tau_n, w_n$ is in $f(\beta, \tau)$. (2) if a string $r, w_0, \tau_1, w_1, \ldots, \tau_n, w_n, \tau$ is in $f(\beta, \tau)$, where $r_1, \ldots, \tau_n, \tau$ are element types in $E$ and $w_0, \ldots, w_n$ are strings in $(E_N \setminus E)^*$, then $r, \tau_1, \ldots, \tau_n, \tau$ is in $\beta; \tau$. (3) none of the new nodes, i.e., nodes in $V_2 \setminus V_1$, is labeled with an $E$-type; (4) no new attributes are defined; and (5) the ancestor-descendant relation on $T_1$-elements is not changed in $T_2$.

Conversely, assume that there is $T_2 = (V_2, lab_2, ele_2, att, root)$ such that $T_2 \models D_N$. We construct an XML tree $T_1$ by modifying $T_2$ such that $T_1 \models D$. For any node $v \in V_2$ with $lab(v) = \tau$ and $\tau \in E_N \setminus E$, we replace $v$ in $ele_2(v')$ by the children of $v$, where $v'$ is the parent of $v$. In addition, we remove $v$ from $V_2$, $lab_2(v)$ from $lab_2$, and $ele_2(v)$ from $ele_2$. Observe that by the definition of $D_N$, no attributes are defined for elements of any type in $E_N \setminus E$. We repeat the process until there is no node labeled with element type in $E_N \setminus E$. Now let $T_1 = (V_1, lab_1, ele_1, att, root)$, where $V_1$, $lab_1$, and $ele_1$ are $V_2$, $lab_2$, and $ele_2$ at the end of the process, respectively. Observe that $att$ and $root$ remain unchanged. By the definition of $T_1$ it can be verified that $T_1 \models D$; and in addition, for any regular expression $\beta \cdot \tau$ over $D$ and $\text{nil} \in R(\tau)$, $|\text{nodes}(f(\beta, \tau))|$ in $T_1$ equals $|\text{nodes}(f(\beta, \tau))|$ in $T_2$, and values($\beta, \tau, \text{nil}$) in $T_1$ equals values($f(\beta, \tau, \text{nil})$) in $T_2$, because of (1) and (2) above and, among other things, the fact that none of the nodes removed is labeled with a type of $E$ and the attribute function $att$ is unchanged.

We now move to encoding of DTDs, more specifically, narrow one-attribute DTDs. Let $D = (E, \{\text{nil}\}, P, R, r)$ be a narrow one-attribute DTD and $\Sigma$ a set of $AC^\text{reg}_{K, \text{FK}}$-constraints over $D$. We encode $D$ with a system $\Psi^D_\Sigma$ of integer constraints such that there exists an XML tree conforming to $D$ if $\Psi^D_\Sigma$ admits a nonnegative solution. The coding is developed w.r.t. $\Sigma$. More specifically, assume that $\beta_1, \tau_1, \text{nil}, \ldots, \beta_k, \tau_k, \text{nil}$ is an enumeration of all regular expressions and attributes that appear in $\Sigma$ and $\Theta$ be the set of functions $\theta : \{1, \ldots, k\} \rightarrow \{0, 1\}$ which are not identically 0. For every $\theta \in \Theta$, define a regular

---

Footnote: As in the proof of Lemma 3.2, if $\tau$ is in $P(\tau)$, then we need to rename $\tau$ in $Q_\tau$ to ensure that $Q_\tau$ and $N_\tau \cup \{\tau\}$ are disjoint. It is straightforward to handle that case.
expression:

$$r_\theta = \left( \bigcap_{i : \theta(i) = 1} \beta_i.\tau_i \right) \cap \left( \bigcap_{j : \theta(j) = 0} \overline{\beta_j.\tau_j} \right),$$

(3.1)

where $\overline{\beta_j.\tau_j}$ is the complement $\beta_j.\tau_j$. We allow intersection and complement operators only in regular expressions $r_\theta$. We note that for every $i \in [1, k]$: 4

$$\beta_i.\tau_i = \bigcup_{\theta : \theta(i) = 1} r_\theta.$$

Then to capture the interaction between $D$ and constraints of $\Sigma$, the system $\Psi^D_\Sigma$ has a variable $|nodes(\beta_i.\tau_i)|$, for every $i \in [1, k]$, and $|nodes(r_\theta)|$, for every $\theta \in \Theta$. In other words, $\Psi^D_\Sigma$ specifies the dependencies imposed by $D$ on the number of elements reachable by following $\beta_i.\tau_i$ ($i \in [1, k]$) and $r_\theta$ ($\theta \in \Theta$).

To capture $\beta_i.\tau_i$ ($i \in [1, k]$) and $r_\theta$ ($\theta \in \Theta$) in $\Psi^D_\Sigma$, consider, for each regular expression $\beta_i.\tau_i$ ($i \in [1, k]$), a deterministic automaton that recognizes that expression. Let $M$ be the deterministic automaton equivalent to the product of all these automata. We refer to $M$ as the DFA for $\Sigma$. Let $s_M$ be the start state of $M$ and $\delta$ be its transition function. Given an XML tree $T$ conforming to $D$, for each node $x \in T$ we define state$(x)$ as $s$, if there is a simple path $\rho$ over $D$ such that $T \models \rho(root, x)$ and $s = \delta(s_M, \rho)$. The connection between $M$ and $T$ w.r.t. $\beta_i.\tau_i$ ($i \in [1, k]$) is described by the following lemma:

**Lemma 3.10.** Let $D$ be a narrow one-attribute DTD, $\Sigma$ a set of $\mathcal{AC}^{\text{eq}}_K$, constraints over $D$, $M$ the DFA for $\Sigma$ and $\beta_i.\tau_i$ a regular expression in $\Sigma$. Then for every XML tree $T$ conforming to $D$ and every $\tau$-element $x$ in $T$, $x \in nodes(\beta_i.\tau_i)$ in $T$ iff state$(x)$ contains some final state $f_{\beta_i.\tau_i}$ of the automaton for $\beta_i.\tau_i$.

In other words, $nodes(\beta_i.\tau_i)$ in $T$ consists of all $\tau$-elements $x$ such that state$(x)$ (which is a tuple of states of automata corresponding to regular expressions in $\Sigma$) contains some final state $f_{\beta_i.\tau_i}$ of the automaton for $\beta_i.\tau_i$. A similar idea was exploited in [1].

**Proof.** Since $T$ is a tree, there exists a unique simple path $\rho$ over $D$ such that $T \models \rho.\tau_i(root, x).$ Thus $x \in nodes(\beta_i.\tau_i)$ in $T$ iff $\rho.\tau_i \in \beta_i.\tau_i$. If $x \in nodes(\beta_i.\tau_i)$ in $T$, then $\rho.\tau_i \in \beta_i.\tau_i$ and, therefore, there must be a final state $f_{\beta_i.\tau_i}$ in the automaton for $\beta_i.\tau_i$ and a state $s$ in $M$ such that $s = \delta(s_M, \rho.\tau_i)$ and $s$ contains $f_{\beta_i.\tau_i}$. Thus state$(x) = s$ contains some final state $f_{\beta_i.\tau_i}$ of the automaton for $\beta_i.\tau_i$. Conversely, if state$(x)$ contains a final state $f_{\beta_i.\tau_i}$ in the automaton for $\beta_i.\tau_i$, then $\rho.\tau_i \in \beta_i.\tau_i$ since $s = \delta(s_M, \rho.\tau_i)$. Therefore, $x \in nodes(\beta_i.\tau_i)$ in $T$.

We next define a system $\Psi^\Sigma$ of integer constraints. The variables used in the constraints of $\Psi^D_\Sigma$ are as follows. Let $\tau \in E$ be an element type and $s = \delta(s_M, \rho.\tau)$ for some simple path $\rho.\tau \in E^\ast$. For each such pair we create a distinct variable $x^{\tau}_s$. Intuitively, in an XML tree $T$ conforming to $D$, we use $x^{\tau}_s$ to keep track of the number of $\tau$-elements with state $s$. Furthermore, define $Y^\Sigma_\tau$ as the set of pairs $(\tau', s')$ such that $\tau' \in E$, $s' = \delta(s_M, \rho.\tau')$ for some simple path $\rho.\tau' \in E^\ast$, $\tau$ is mentioned in $P(\tau')$ and $s = \delta(s', \tau')$. For each such pair $(\tau', s')$, we create a variable $x^{\tau',\tau'}_s$. Intuitively, in an XML tree $T$ conforming to $D$, we use $x^{\tau',\tau'}_s$ to keep track of the number of $\tau$-elements with state $s$ that are children of a node of type $\tau'$ with state $s'$. There are exponentially many variables (in the size of $D$ and $\Sigma$) in total since $M$ is a DFA. Using these, we define an integer constraint to specify $\tau \rightarrow P(\tau)$ at state $s$ as follows. Let us use $\Psi^\Sigma_\tau$ to denote the set of integer constraints defined for $\tau$ at $s$.

- If $P(\tau) = \tau_1$, then $\Psi^\Sigma_\tau$ includes $x^{\tau_1}_s$ where $s_1 = \delta(s, \tau_1)$.
- If $P(\tau) = (\tau_1, \tau_2)$, then $\Psi^\Sigma_\tau$ includes $x^{\tau_1}_s = x^{\tau_1}_{s_1}$ and $x^{\tau_2}_s = x^{\tau_2}_{s_2}$, where $s_i = \delta(s, \tau_i)$ for $i = 1, 2$.

Referring to the XML tree $T$, these assure that each $\tau$-element in $T$ must have a $\tau_1$-subelement and a $\tau_2$-subelement.

- If $P(\tau) = (\tau_1 | \tau_2)$, then $\Psi^\Sigma_\tau$ includes $x^{\tau}_s = x^{\tau_1}_{s_1} + x^{\tau_2}_{s_2}$, where $s_i = \delta(s, \tau_i)$ for $i = 1, 2$. This assures that each $\tau$-element in $T$ must have either a $\tau_1$-subelement or a $\tau_2$-subelement, and thus the sum of the number of these $\tau_1$-subelements and the number of $\tau_2$-subelements equals the number of $\tau$-elements.

4Recall that the regular language defined by a regular expression $\beta$ is denoted by $\beta$ as well.
• If $P(\tau) = \tau_i^*$, then $\Psi_{\tau}^*$ includes $(x_{\tau_1}^{s_1}; \ldots; x_{\tau_k}^{s_k} > 0) \rightarrow (x_{\tau_1}^s > 0)$, where $s_1 = \delta(s, \tau_1)$.

In addition, $\Psi_{\tau}^*$ includes $x_{\tau}^s = \sum_{(\tau', s') \in Y_\tau^x} x_{\tau'}^{s'}$.

Recall that $\beta_1, \tau_1, \ldots, \beta_k, \tau_k, \Theta$ is an enumeration of all regular expressions and attributes that appear in $\Sigma$. That $\Theta$ is the set of functions $\theta: \{1, \ldots, k\} \rightarrow \{0, 1\}$ which are not identically 0 and that for each such function $\theta$, $\tau_0$ is a regular expression defined as in (3.1). For each $i \in [1, k]$, we define $F_{\beta_i, \tau_i}$ as the set of states $s = (s_1, \ldots, s_k)$ of the DFA for $\Sigma$ such that $s_i$ is a final state of the DFA for $\beta_i, \tau_i$. Notice that by Lemma 3.10, for every XML tree $T$ conforming to $D$ and every node $x$ of $T$, $x \in \text{nodes}(\beta_i, \tau_i)$ in $T$ if and only if $\text{state}(x) \in F_{\beta_i, \tau_i}$. Furthermore, for each $\theta \in \Theta$, we define $F_\theta$ as the set of states $s = (s_1, \ldots, s_k)$ of the DFA for $\Sigma$ such that for every $i \in [1, k]$, $s_i$ is a final state of the DFA for $\beta_i, \tau_i$ if and only if $\theta(i) = 1$. Notice that by Lemma 3.10, for every XML tree $T$ conforming to $D$ and every node $x$ of $T$, $x \in \text{nodes}(\tau_0)$ in $T$ if and only if $\text{state}(x) \in F_{\beta_i, \tau_i}$.

By our restriction on regular expressions regarding element type $r$, there is a unique variable $x_r^s$ associated with $r$, where $s = \delta(s_M, r)$. We write $x_r$ for $x_r^1$. Then we define the set of cardinality constraints determined by $\text{DTD}$ $D$ w.r.t. a set $\Sigma$ of $\mathcal{AC}_{\text{reg}}^{\text{K,FK}}$-constraints over $D$, denoted by $\Psi_D^\Sigma$, as follows:

- For each $\tau \in E$ and each state $s$ given above, $\Psi_D^\Sigma$ contains all the constraints in $\Psi_\tau^*$. 
- $\Psi_D^\Sigma \tau$ contains constraint $x_r = 1$; i.e., there is a unique root in each XML tree conforming to $D$.
- For every $i \in [1, k]$, $\Psi_D^\Sigma$ contains constraint $|\text{nodes}(\beta_i, \tau_i)| = \sum_{s \in F_{\beta_i, \tau_i}} x_s^i$.
- For every $\theta \in \Theta$ such that $\tau_0 \neq \emptyset$, $\Psi_D^\Sigma$ contains constraint $|\text{nodes}(\tau_0)| = \sum_{s \in F_\emptyset} x_{\tau_0}^s = 0$.

Note that $\Psi_D^\Sigma$ can be computed in EXPTIME in the size of $D$ and $\Sigma$. We say that $\Psi_D^\Sigma$ is consistent if it has a nonnegative solution. We next show that $\Psi_D^\Sigma$ indeed characterizes narrow one-attribute $\text{DTD}$ $D$.

**Lemma 3.11.** Let $D$ be a narrow one-attribute $\text{DTD}$, $\Sigma$ a set of $\mathcal{AC}_{\text{reg}}^{\text{K,FK}}$-constraints over $D$ and $\Psi_D^\Sigma$ the set of cardinality constraints determined by $D$ w.r.t. $\Sigma$. Then $\Psi_D^\Sigma$ is consistent if and only if there is an XML tree $T$ such that $T \models D$. In addition, for every $i \in [1, k]$ and $\theta \in \Theta$, $|\text{nodes}(\beta_i, \tau_i)|$ and $|\text{nodes}(\tau_0)|$ in $T$ equal the value of variables $|\text{nodes}(\beta_i, \tau_i)|$ and $|\text{nodes}(\tau_0)|$ given by the solution to $\Psi_D^\Sigma$.

**Proof.** First, assume that there is an XML tree $T = (V, \text{lab}, \text{ele}, \text{att}, \text{root})$ conforming to $D$. We define a nonnegative solution of $\Psi_D^\Sigma$ as follows. For each variable $x_{\tau_1}^{s_1} \ldots x_{\tau_k}^{s_k}$ in $\Psi_D^\Sigma$, let its value be the number of $\tau$-elements $x$ in $T$ such that $x$ is a child of a node $y$ of type $\tau'$ with $\text{state}(x) = s$ and $\text{state}(y) = s'$. Furthermore, let $x_r$ be 1 and for every variable $x_r^s$ in $\Psi_D^\Sigma$, let $x_r^s$ be the sum of the variables $x_{\tau'}^{s'}$, where $(\tau', s') \in Y_\tau^x$. Finally, for every $i \in [1, k]$ and every $\theta \in \Theta$, let $|\text{nodes}(\beta_i, \tau_i)|$ and $|\text{nodes}(\tau_0)|$ be $\sum_{s \in F_{\beta_i, \tau_i}} x_s^i$ and $\sum_{s \in F_\emptyset} x_{\tau_0}^s$, respectively. This defines a nonnegative assignment since $T$ is finite. It can be verified that the assignment is a solution of $\Psi_D^\Sigma$. Indeed, it satisfies the constraint $x_r = 1$ and constraints of the form $x_{\tau_1}^{s_1} \ldots x_{\tau_k}^{s_k}$, $|\text{nodes}(\beta_i, \tau_i)| = \sum_{s \in F_{\beta_i, \tau_i}} x_s^i$, and $|\text{nodes}(\tau_0)| = \sum_{s \in F_\emptyset} x_{\tau_0}^s$ by the definition of the assignment. Moreover, one can verify that it also satisfies the constraints of each $\Psi_\tau^*$, by considering the four different cases corresponding to the four different types of regular expressions in $D$. In particular, it satisfies constraints of the form $(x_{\tau_1}^{s_1}; \ldots; x_{\tau_k}^{s_k} > 0) \rightarrow (x_{\tau_1}^s > 0)$ for each $\tau \rightarrow \tau_1^*$ in $P$, since if $x_{\tau_1}^s > 0$, then there exists a $\tau_1$-node in $T$ having as its parent a $\tau$-node $y$ with $\text{state}(y) = s$. Thus, $x_r^s > 0$ by the definition of the assignment. Therefore, $\Psi_D^\Sigma$ is consistent. Moreover, by Lemma 3.10, for every $i \in [1, k]$ and $\theta \in \Theta$, the values of variables $|\text{nodes}(\beta_i, \tau_i)|$ and $|\text{nodes}(\tau_0)|$ in the solution are indeed $|\text{nodes}(\beta_i, \tau_i)|$ and $|\text{nodes}(\tau_0)|$ in $T$.

Conversely, assume that $\Psi_D^\Sigma$ admits a nonnegative solution. We show that there exists an XML tree $T = (V, \text{lab}, \text{ele}, \text{att}, \text{root})$ such that $T \models D$. To do so, for each element type $\tau$ and state $s$ for $\tau$, we create $x_\tau^s$ many distinct $\tau$-elements. Let $\text{ext}(\tau)$ denote the set of all $\tau$-elements created above and

$$V = \bigcup_{\tau \in E} \text{ext}(\tau).$$
Then function \( \text{lab} \) is defined as \( \text{lab}(v) = \tau \) if \( v \in \text{ext}(\tau) \), and function \( \text{att} \) is defined as follows:

\[
\text{att}(v, @l) = \begin{cases} 
\text{empty string} & \text{if } @l \in R(\text{lab}(v)) \\
\text{undefined} & \text{otherwise}
\end{cases}
\]

It is easy to verify that these functions are well defined. Let \( \text{root} \) be the node labeled \( r \), which is unique since \( x_r = 1 \) is in \( \Psi^E_D \). Finally, to define function \( \text{ele} \), we do the following. For each \( x_{\tau,\tau'} \) in \( \Psi^E_D \), we choose \( x_{\tau,\tau'} \) many distinct vertices labeled \( \tau \) and mark them with \( x_{\tau,\tau'} \). Note that every \( \tau \)-element in \( \Psi^E_D \) can be marked once and only once. Starting at \( \text{root} \), for each \( \tau \)-element \( x \) marked with \( x_{\tau,\tau'} \), for some \((\tau', s') \in Y^s_s \), consider \( P(\tau) \) and constraints of \( \Psi^E_D \). If \( P(\tau) = \tau_1 \in E \), then we choose a distinct \( \tau_1 \)-element \( y \) marked with \( x_{\tau_1,\tau} \) and let \( \text{ele}(x) = [y] \), where \( x_{\tau_1,\tau} = x_{\tau_1,\tau} \) is in \( \Psi^E_D \). If \( P(\tau) = (\tau_1, \tau_2) \), then we choose a \( \tau_1 \)-element \( y_1 \) marked with \( x_{\tau_1,\tau} \) and a \( \tau_2 \)-element \( y_2 \) marked with \( x_{\tau_2,\tau} \) and let \( \text{ele}(x) = [y_1, y_2] \), where \( x_{\tau_1,\tau} = x_{\tau_1,\tau} \) and \( x_{\tau_2,\tau} \) are in \( \Psi^E_D \). If \( P(\tau) = (\tau_1, \tau_2) \), then we choose an element \( y \) marked with \( x_{\tau_1,\tau} \) or \( x_{\tau_2,\tau} \) and let \( \text{ele}(x) = [y] \), where \( x_{\tau_1,\tau} = x_{\tau_1,\tau} + x_{\tau_2,\tau} \) is in \( \Psi^E_D \). If \( P(\tau) = \tau_1 \), then we choose a list \([y_1, \ldots, y_n] \) (\( n \geq 0 \)) of \( \tau_1 \)-elements marked with \( x_{\tau_1,\tau} \) and let \( \text{ele}(x) = [y_1, \ldots, y_n] \), where \( x_{\tau_1,\tau} > 0 \) \( \rightarrow \) \( x_{\tau_1,\tau} > 0 \) is in \( \Psi^E_D \). By the constraints in \( \Psi^E_D \), each element of \( V \) can be chosen once and only once. One can verify that \( T \) defined in this way is indeed an XML tree and \( T \models D \). Hence, there exists an XML tree conforming to \( D \).

Finally, to see that for every \( i \in [1, k] \) and \( \theta \in \Theta \), \( |\text{nodes}(\beta_\theta, \tau_i)\| \) and \( |\text{nodes}(r_\theta)\| \) in \( T \) equals the values of variables \( |\text{nodes}(\beta_\theta, \tau_i)\| \) and \( |\text{nodes}(r_\theta)\| \) in the solution, respectively, it suffices to show, by Lemma 3.10, that for each node \( x \) in \( T \), if \( x \) is marked with \( x_{\tau,\tau'} \) in the construction, then \( \text{state}(x) = s \). Since \( T \) is a tree, there is a unique simple path \( \rho \in E^* \) such that \( T \models \rho(\text{root}, x) \). We show the claim by induction on the length \( |\rho| \) of \( \rho \). If \( |\rho| = 1 \), i.e., \( \rho = r \), then \( x \) is the root and obviously, \( \text{state}(x) = \delta(s_M, r) \). Assume the claim for \( \rho \) and we show that the claim holds for \( r \tau \). Let \( y \) be the \( r \tau \)-element in \( T \) such that \( T \models \rho(\text{root}, y) \) and \( y \) is the parent of \( x \). Suppose that \( y \) is marked with \( x_{\tau,\tau'} \) in the construction. By the induction hypothesis \( \text{state}(y) = s' \). It is easy to see \( \text{state}(x) = \delta(s', \tau) \). By the definition of \( \Psi^s \), we have that \( s \) is precisely the state \( \delta(s', \tau) \). Thus \( \text{state}(x) = s \). This proves the claim and thus the lemma.

We now move to encoding \( \mathcal{AC}^{\text{reg}}_{K, FK} \)-constraints in terms of integer constraints. Let \( D \) be a DTD \((E, \{@l\}, P, R, r)\) and \( \Sigma \) a set of \( \mathcal{AC}^{\text{reg}}_{K, FK} \)-constraints over \( D \). By Lemmas 3.8 and 3.9, we assume, without loss of generality, that \( D \) is a narrow one-attribute DTD. To encode \( \Sigma \), let \( \beta_\theta, \tau_1, @l_1, \ldots, \beta_\theta, \tau_k, @l_k \) be an enumeration of all regular expressions and attributes that appear in \( \Sigma \), and for every function \( \theta : \{1, \ldots, k\} \rightarrow \{0, 1\} \) which is not identically 0, let regular expression \( r_\theta \) be defined as in (3.1). Then for every nonempty \( \Omega \subseteq \Theta \), we introduce a new variables \( z_\Omega \). In any XML tree conforming to \( D \), the intended interpretations of \( z_\Omega \) is the cardinality of

\[
\left( \bigcap_{\theta : \theta \in \Omega} \text{values}(r_\theta, @l_\theta) \right) \setminus \left( \bigcup_{\theta : \theta \in \Omega \setminus \Theta} \text{values}(r_\theta, @l_\theta) \right).
\]

Note that the number of new variables is double-exponential in the number of regular expression in \( \Sigma \). Using these variables, we define the set of the cardinality constraints determined by \( \Sigma \), denoted by \( C\Sigma \), which consists of the following:

\footnote{We assume that \( \text{root} \) is marked with \( x_r \), where \( s = \delta(s_M, r) \) and \( s_M \) is the initial state of the DFA for \( \Sigma \).}

21
\[ \sum_{\Omega: \theta \in \Omega} z_\Omega = |\text{values}(r_\theta, \emptyset)| \quad \text{for every } \theta \in \Theta, \]

\[ \sum_{\Omega: \Omega \cap \{\theta|\theta(i)=1\} \neq \emptyset} z_\Omega = |\text{values}(\beta_i, \tau_i, \emptyset)| \quad \text{for every } i \in [1, k], \]

\[ |\text{values}(\beta_i, \tau_i, \emptyset)| = |\text{nodes}(\beta_i, \tau_i)| \quad \text{for every } \beta_i, \tau_i, \emptyset \rightarrow \beta_i, \tau_i \text{ in } \Sigma, \]

\[ |\text{values}(\beta_i, \tau_i, \emptyset)| = |\text{nodes}(\beta_i, \tau_i)| \quad \text{for every } \beta_i, \tau_i, \emptyset \subseteq FK \beta_j, \tau_j, \emptyset \text{ in } \Sigma, \]

\[ \sum_{\Omega: \Omega \cap \{\theta|\theta(i)=1\} \neq \emptyset, \Omega \cap \{\theta'|\theta'(j)=1\} = \emptyset} z_\Omega = 0 \quad \text{for every } \beta_i, \tau_i, \emptyset \subseteq FK \beta_j, \tau_j, \emptyset \text{ in } \Sigma, \]

\[ |\text{values}(\beta_i, \tau_i, \emptyset)| \leq |\text{nodes}(\beta_i, \tau_i)| \quad \text{for every } i \in [1, k], \]

\[ |\text{values}(r_\theta, \emptyset)| \leq |\text{nodes}(r_\theta)| \quad \text{for every } \theta \in \Theta. \]

Note that the size of \( C_{\Sigma} \) is double-exponential in the size of \( \Sigma \).

We now combine the encodings for constraints and the DTDs, and present a system \( \Psi(D, \Sigma) \) of linear integer constraints for a DTD \( D \) and a set \( \Sigma \) of \( \mathcal{AC}^\text{reg}_{FK} \)-constraints. Assuming that \( D \) and \( \Sigma \) are as above, the set \( \Psi(D, \Sigma) \), called the \textit{set of cardinality constraints determined by } \( D \) and \( \Sigma \), is defined to be:

\[
\Psi_D^\Sigma \cup C_{\Sigma} \cup \{(|\text{nodes}(\beta_i, \tau_i)| > 0) \rightarrow (|\text{values}(\beta_i, \tau_i, \emptyset)| > 0) \mid i \in [1, k]\} \cup
\{(|\text{nodes}(r_\theta)| > 0) \rightarrow (|\text{values}(r_\theta, \emptyset)| > 0) \mid \theta \in \Theta\},
\]

where \( C_{\Sigma} \) is the set of cardinality constraints determined by \( \Sigma \), and \( \Psi_D^\Sigma \) is the set of cardinality constraints determined by \( D \) w.r.t. \( \Sigma \). The system \( \Psi(D, \Sigma) \) is said to be consistent if it has a nonnegative solution that satisfies all of its constraints. Observe that \( \Psi(D, \Sigma) \) can be partitioned into two sets: \( \Psi(D, \Sigma) = \Psi^D(D, \Sigma) \cup \Psi^\Sigma(D, \Sigma) \), where \( \Psi^D(D, \Sigma) \) consists of linear integer constraints, and \( \Psi^\Sigma(D, \Sigma) \) consists of constraints of the form \((x > 0 \rightarrow y > 0)\). Also note that the size of \( \Psi(D, \Sigma) \) is double-exponential in the size of \( D \) and \( \Sigma \).

We next show that \( \Psi(D, \Sigma) \) indeed characterizes the consistency of \( D \) and \( \Sigma \).

**Lemma 3.12.** Let \( D \) be a narrow one-attribute DTD, \( \Sigma \) a finite set of \( \mathcal{AC}^\text{reg}_{FK} \)-constraints over \( D \) and \( \Psi(D, \Sigma) \) the set of cardinality constraints determined by \( D \) and \( \Sigma \). Then \( \Psi(D, \Sigma) \) is consistent if and only if there is an XML tree \( T \) such that \( T \models (D, \Sigma) \).

**Proof.** Suppose that there exists an XML tree \( T \) such that \( T \models (D, \Sigma) \). Then by Lemma 3.11, there exists a nonnegative solution for \( \Psi^\Sigma_D \) such that for every \( i \in [1, k] \) and \( \theta \in \Theta \), the values of variables \( |\text{nodes}(r_\theta)| \) and \( |\text{nodes}(\beta_i, \tau_i)| \) in this solution coincide with \( |\text{values}(r_\theta)| \) and \( |\text{values}(\beta_i, \tau_i)| \) in \( T \), respectively. From this solution, it is easy to generate a solution to \( \Psi(D, \Sigma) \) by assigning to variable \( |\text{values}(r_\theta, \emptyset)| \) the size of \( |\text{values}(r_\theta, \emptyset)| \) in \( T \), for every \( \theta \in \Theta \), assigning to variable \( |\text{nodes}(\beta_i, \tau_i, \emptyset)| \) the size of \( |\text{values}(\beta_i, \tau_i, \emptyset)| \) in \( T \), for every \( i \in [1, k] \), and then assigning to each variable \( z_\Omega \) the cardinality of set (3.2) above. It is straightforward to verify that this assignment is a solution to \( \Psi(D, \Sigma) \).

Conversely, suppose that \( \Psi(D, \Sigma) \) has an integer solution. We show that there is an XML tree \( T \) such that \( T \models (D, \Sigma) \). By Lemma 3.11, given an integer solution to \( \Psi(D, \Sigma) \), we can construct an XML tree \( T' = (V, \text{lab}, \text{ele}, \text{att}', \text{root}) \) such that \( T' \models D \). Moreover, for every \( i \in [1, k] \), there are exactly \( n_{\beta_i, \tau_i} \) elements in \( T' \) reachable by following \( \beta_i, \tau_i \), where \( n_{\beta_i, \tau_i} \) is the value of the variable \( |\text{nodes}(\beta_i, \tau_i)| \) in \( \Psi(D, \Sigma) \), and for every \( \theta \in \Theta \), there are exactly \( n_{\theta} \) values in \( T' \) reachable by following \( r_\theta \), where \( n_{\theta} \) is the value of the variable \( |\text{nodes}(r_\theta)| \) in \( \Psi(D, \Sigma) \). We modify the definition of the function \( \text{att}' \), while leaving \( \text{lab}, \text{ele} \) and \( \text{root} \) unchanged, to generate a tree \( T = (V, \text{lab}, \text{ele}, \text{att}, \text{root}) \) such that \( T \models (D, \Sigma) \). More specifically, we modify \( \text{att}'(v, \emptyset) \) if \( v \) is in \( \text{nodes}(\beta, \tau) \) for some regular expression \( \beta, \tau \) mentioned in \( \Sigma \), and leave \( \text{att}'(v, \emptyset) \) unchanged otherwise. To do this, for each variable \( z_\Omega \) we create a set \( s_\Omega \) of distinct string values such that \( \{s_\Omega\} = z_\Omega \) and \( s_\Omega \cap s_{\Omega'} = \emptyset \) if \( \Omega \neq \Omega' \). Then for every \( \Omega \subseteq \Theta \), we let \( |\text{values}(r_\theta, \emptyset)| \) in \( T \) to contain \( s_\Omega \) if and only if \( \theta \in \Omega \). This is possible because (1) \( \sum_{\Omega: \theta \in \Omega} z_\Omega = |\text{values}(r_\theta, \emptyset)| \) is in \( C_{\Sigma} \), for every \( \theta \in \Theta \); (2) \( \sum_{\Omega: \Omega \cap \{\theta|\theta(i)=1\} \neq \emptyset} z_\Omega = |\text{values}(\beta_i, \tau_i, \emptyset)| \) is in \( C_{\Sigma} \), for every \( i \in [1, k] \); (3) if \( r_\theta = \emptyset \), then
\[\text{nodes}(r_\theta) = 0\] for every \(\theta \in \Theta; (4) \text{values}(\beta_i, \tau_i, \bar{a}l) \leq \text{nodes}(\beta_i, \tau_i)\] is in \(C_\Sigma\), for every \(i \in [1, k]; (5) \text{values}(r_\theta, \bar{a}l) \leq \text{nodes}(r_\theta)\] is in \(C_\Sigma\), for every \(\theta \in \Theta;\) and (6) \text{nodes}(\beta)\] in \(T‘\) equals \text{nodes}(\beta)\] in \(T‘\), for every regular expression \(\beta\) over \(D\).

We next show that \(T\] has the desired properties. It is easy to verify \(T \models D\] given the construction of \(T\] from \(T‘\) and the assumption \(T‘ \models D\). By the definition of \(T\], we have that for every \(i \in [1, k]\) and \(\theta \in \Theta, \text{nodes}(\beta_i, \tau_i)\], \(\text{values}(\beta_i, \tau_i, \bar{a}l)\], \(\text{values}(r_\theta)\] and \(\text{values}(r_\theta, \bar{a}l)\] in \(T‘\) equal the value of variables \(\text{nodes}(\beta_i, \tau_i)\], \(\text{values}(\beta_i, \tau_i, \bar{a}l)\], \(\text{nodes}(r_\theta)\] and \(\text{values}(r_\theta, \bar{a}l)\] given by the solution to \(\Psi(D, \Sigma)\). We use this property to show that \(T \models \Sigma\). Let \(\varphi\] be a constraint in \(\Sigma\). (1) If \(\varphi\] is a key \(\beta_i, \tau_i, \bar{a}l \rightarrow \beta_j, \tau_l\] it is immediate from the definition of \(T\] that \(T \models \varphi\] since \(\text{values}(\beta_i, \tau_i, \bar{a}l)\] = \(\text{values}(\beta_j, \tau_l)\] is a constraint in \(C_\Sigma\) and, hence, \(\text{values}(\beta_i, \tau_i, \bar{a}l)\] = \(\text{values}(\beta_j, \tau_l)\] in \(T\). That is, each \(x \in \text{values}(\beta_i, \tau_i)\] in \(T\) has a distinct \(\bar{a}l\)-attribute value and thus the value of its \(\bar{a}l\)-attribute uniquely identifies \(x\) among nodes in \(\text{nodes}(\beta_i, \tau_i)\]. (2) If \(\varphi\] is \(\beta_i, \tau_i, \bar{a}l \subseteq_{FK} \beta_j, \tau_j, \bar{a}l\], it is easy to see that in \(T\):

\[\text{values}(\beta_i, \tau_i, \bar{a}l) \setminus \text{values}(\beta_j, \tau_j, \bar{a}l) = \bigcup_{\Omega : \Omega \cap \{\theta | \theta(i) = 1\} \neq \emptyset, \Omega \cap \{\theta’ | \theta’(j) = 1\} = \emptyset} s_\Omega,\]

Since \(s_\Omega \cap s_{\Omega’} = \emptyset\] if \(\Omega \neq \Omega’\],

\[\text{values}(\beta_i, \tau_i, \bar{a}l) \setminus \text{values}(\beta_j, \tau_j, \bar{a}l) = \sum_{\Omega : \Omega \cap \{\theta | \theta(i) = 1\} \neq \emptyset, \Omega \cap \{\theta’ | \theta’(j) = 1\} = \emptyset} z_\Omega.\]

Thus, given that

\[\sum_{\Omega : \Omega \cap \{\theta | \theta(i) = 1\} \neq \emptyset, \Omega \cap \{\theta’ | \theta’(j) = 1\} = \emptyset} z_\Omega = 0\]

is in \(C_\Sigma\) (since \(\beta_i, \tau_i, \bar{a}l \subseteq_{FK} \beta_j, \tau_j, \bar{a}l \in \Sigma\)), we have \(\text{values}(\beta_i, \tau_i, \bar{a}l) \setminus \text{values}(\beta_j, \tau_j, \bar{a}l) = 0\] in \(T\), that is, \(\text{values}(\beta_i, \tau_i, \bar{a}l) \subseteq \text{values}(\beta_j, \tau_j, \bar{a}l)\] in \(T\). Furthermore, \(T \models \beta_i, \tau_i, \bar{a}l \rightarrow \beta_j, \tau_j\] since \(\text{values}(\beta_j, \tau_l)\] is a constraint in \(C_\Sigma\). Thus \(T \models \varphi\). This concludes the proof of the lemma.

We need another lemma for a mild generalization of linear integer constraints.

**Lemma 3.13.** Given a system \(Ax \leq \bar{b}\) of linear integer constraints together with conditions of the form \((x_i > 0) \rightarrow (x_j > 0)\), where \(A\) is an \(n \times m\) matrix on integers, \(\bar{b}\] is an \(n\)-vector on integers and \(1 \leq i, j \leq m\), the problem of determining whether the system admits a nonnegative integer solution is in \(\text{NP}\).

**Proof.** Let \(c_1, \ldots, c_p\] enumerate the conditions of the form \((x > 0) \rightarrow (y > 0)\), \(c_k\] being \((x_k^1 > 0) \rightarrow (x_k^2 > 0)\). Consider \(2^p\) instances \(I_j\) of integer programming obtained by adding, for each \(k \leq p\), either \(x_k^1 = 0\) or \(x_k^2 > 0\) to \(Ax \leq \bar{b}\). Clearly, the original system of constraints has a solution iff some \(I_j\] has a solution. By [27], \(I_j\] has a solution iff it has a solution whose size is polynomial in \(A, \bar{b}\] and \(p\). Hence, to check if the original system of constraints has a solution, it suffices to guess a system \(I_j\] and then guess a polynomial size solution for it; thus, the problem is in \(\text{NP}\).

We now conclude the proof of the first part of the theorem. By Lemma 3.8, given an arbitrary DTD \(D\] and a set \(\Sigma\) of \(\mathcal{A}_{\Sigma}^\mathcal{FK}\)-constraints over \(D\), it is possible to compute a one-attribute DTD \(D’\) and a set \(\Sigma’\) of \(\mathcal{A}_{\Sigma}^\mathcal{FK}\)-constraints over \(D’\) such that \((D, \Sigma)\) is consistent iff \((D’, \Sigma’)\) is consistent. By Lemma 3.9, one can compute a narrow one-attribute DTD \(D_N\) and a set \(\Sigma_N\) of \(\mathcal{A}_{\Sigma}^\mathcal{FK}\)-constraints over \(D_N\) such that \((D’, \Sigma’)\) is consistent iff \((D_N’, \Sigma_N’)\) is consistent. By Lemma 3.12, \((D_N’, \Sigma_N’)\) is consistent iff \(\Psi(D_N’, \Sigma_N’)\) has a nonnegative integer solution. Thus, \((D, \Sigma)\) is consistent iff \(\Psi(D_N’, \Sigma_N’)\) has a nonnegative integer solution. Note that \((D’, \Sigma’)\) can be computed in polynomial time on \(|D| + |\Sigma|\), \((D_N’, \Sigma_N’)\) can be computed in polynomial time on \(|D’| + |\Sigma’|\), and \(\Psi(D_N’, \Sigma_N’)\) can be computed in double-exponential time on \(|D_N’| + |\Sigma_N’|\). Thus, by Lemma 3.13, one can check in \(2\text{-NEXPTIME}\) whether there exists an XML tree \(T\) such that \(T \models (D, \Sigma)\).

**Proof of b)** We establish the PSPACE-hardness by reduction from the QBF-CNF problem. An instance of QBF-CNF is a quantified boolean formula in prenex conjunctive normal form. The problem is to determine whether this formula is valid. QBF-CNF is known to be PSPACE-complete [20, 28].

Let \(\theta\) be a formula of the form

\[\neg q_1 x_1 \cdots \neg q_m x_m \psi,\] (3.3)
where each $Q_i \in \{\forall, \exists\}$ (1 ≤ $i$ ≤ m) and $\psi$ is a propositional formula in conjunctive normal form, say $C_1 \land \cdots \land C_n$, that mentions variables $x_1, \ldots, x_m$. We construct a DTD $D_\theta$ and a set $\Sigma_\theta$ of $AC_{CK, FK}$-constraint such that $\theta$ is valid if and only if there is an XML tree conforming to $D_\theta$ and satisfying $\Sigma_\theta$.

We construct a DTD $D_\theta = (E, A, P, R, r)$ as follows. $E = \{r, C\} \cup \bigcup_{i=1}^m \{x_i, \bar{x}_i, N_{x_i}, P_{x_i}\}$, $A = \{\emptyset\}$ and $P$ is defined by considering the quantifiers of $\theta$. We use the set of constraint $\Sigma_\theta$ of $AC_{FK}$-constraint such that $\theta$ is valid if and only if there is an XML tree conforming to $D_\theta$ and satisfying $\Sigma_\theta$.

In general, for each 1 ≤ $i$ ≤ m − 1, we consider quantifier $Q_{i+1}$ to define $P(N_{x_{i+1}})$ and $P(P_{x_{i+1}})$:

$$P(N_{x_{i+1}}) = P(P_{x_{i+1}}) = \begin{cases} N_{x_{i+1}} | P_{x_{i+1}} & Q_{i+1} = \exists \\ N_{x_{i+1}} | P_{x_{i+1}} & Q_{i+1} = \forall \end{cases}$$

We represent formula $\psi$ as a regular expression. Given a clause $C_j = \bigvee_{i=1}^n y_i \lor \forall_j z_i \land \exists_j w_i (j \in [1,n])$, $tr(C_j)$ is defined to be the regular expression $y_1 \lor \cdots \lor y_n \lor z_1 \cdots \lor z_q$. We define $P$ on element types $N_{x_m}$ and $P_{x_m}$ as $P(N_{x_m}) = P(P_{x_m}) = tr(C_1), \ldots, tr(C_n)$. For the remaining elements of $E$, we define $P$ as $\epsilon$. We define function $R$ as follows:

$$R(r) = R(P_{x_m}) = R(N_{x_1}) = \emptyset \quad 1 \leq i \leq m$$

$$R(C) = R(x_1) = R(\bar{x}_1) = \{\emptyset\} \quad 1 \leq i \leq m.$$ 

Finally, $\Sigma_\theta$ contains the following foreign keys:

$$r \cdot N_{x_1} \cdot x_i \cdot \emptyset \subseteq FK \quad r.C.C \cdot \emptyset,$$  

$$r \cdot P_{x_2} \cdot x_i \cdot \emptyset \subseteq FK \quad r.C.C \cdot \emptyset, \quad i \in [1,m].$$

For instance, for the formula $\forall x_1 \exists x_2 \forall x_3 (x_1 \lor x_2 \lor \neg x_3)$, an XML tree conforming to $D$ is shown in Figure 3.2. In this tree, a node of type $N_{x_1}$ represents a negative value (0) for the variable $x_1$, and a node of type $P_{x_2}$ represents a positive value (1) for this variable. Thus, given that the root has two children of types $N_{x_2}$ and $P_{x_3}$, the values 0 and 1 are assigned to $x_1$ (representing the quantifier $\forall x_1$). Nodes of type $N_{x_2}$ have one child of type either $N_{x_3}$ or $P_{x_3}$, and, therefore, either 0 or 1 is assigned to $x_2$ (representing the quantifier $\exists x_2$). The same holds for nodes of type $P_{x_3}$. The fourth level of the tree represents the quantifier $\forall x_3$. Note that in any XML tree $T$ conforming to $D$, there is no node in $T$ reachable by following the path $r.C.C$.

In Figure 3.2, every path from the root $r$ to a node of type either $N_{x_3}$ or $P_{x_3}$ represents a truth assignment for the variables $x_1$, $x_2$, $x_3$. For example, the path from the root to the node $u$ represents the truth assignment $\sigma_u$: $\sigma_u(x_1) = 0$, $\sigma_u(x_2) = 1$ and $\sigma_u(x_3) = 0$. To verify that all these assignments satisfy the formula $x_1 \lor x_2 \lor \neg x_3$, we use the set of constraint $\Sigma_\theta$.

Next we prove that $\theta$, defined in (3.3), is valid if and only if there is an XML tree $T$ conforming to $D_\theta$ and satisfying $\Sigma_\theta$. We show only the “if” direction. The “only if” direction is similar.

Suppose that there is an XML tree $T$ such that $T \models (D_\theta, \Sigma_\theta)$. To prove that $\theta$ is valid, it suffices to prove that each path from the root $r$ to a node of type either $N_{x_m}$ or $P_{x_m}$ represents a truth assignment
satisfying ψ. Let p be one of these paths and let v be the node of type either \( N_{xm} \) or \( P_{xm} \) reachable from the root by following p. We define the truth assignment \( σ_p \) as follows:

\[
σ_p(x_i) = \begin{cases} 
1 & \text{p contains a node of type } P_{x_i} \\
0 & \text{Otherwise.}
\end{cases}
\]

We have to prove that \( σ_p(C_i) = 1 \) for each \( i \in [1, n] \). Given that \( T \vDash D_σ \), v has as child a node \( v' \) whose type is in \( tr(C_i) \). If the type of \( v' \) is \( x_j \), then given that \( T \vDash r.C.C. x_j. @l \subseteq FK \) r.C.C. @l and that there exists no node in \( T \) reachable by following the path \( r.C.C. \), \( p \) contains a node of type \( P_{x_j} \), and, therefore, \( σ_p(x_j) = 1 \) since \( σ_p(x_j) = 1 \). If the type of \( v' \) is \( x_j \), then given that \( T \vDash r.C.C. x_j. @l \subseteq FK \) r.C.C. @l, \( p \) contains a node of type \( P_{x_j} \) and it does not contain a node of type \( P_{x_j} \) and, therefore, \( σ_p(C_i) = 1 \) since \( σ_p(¬x_j) = 1 \). Thus, we conclude that \( θ \) is valid. This concludes the proof of part b) of the theorem.

4. Relative integrity constraints. Since XML documents are hierarchically structured, one may be interested in the entire document as well as in its sub-documents. The latter gives rise to relative integrity constraints [5], that only hold on certain sub-documents. Below we define relative keys and foreign keys. Recall that we use \( RC \) to denote various classes of such constraints. We use the notation \( x \prec y \) when \( x \) and \( y \) are two nodes in an XML tree and \( y \) is a descendant of \( x \).

We first define unary relative keys and foreign keys associated with element types. Let \( D = (E, A, P, R, τ) \) be a DTD. A relative key is an expression \( ϕ \) of the form \( τ(1.l → τ_1) \), where \( @l \in R(τ_1) \). It says that relative to each node \( x \) of element type \( τ \), \( @l \) is a key for all the \( τ_1 \)-nodes that are descendants of \( x \). That is, if a tree \( T \) conforms to \( D \), then \( T \vDash ϕ \) if

\[
∀x ∈ ext(τ) ∀y, z ∈ ext(τ_1) ((x ∼ y) ∧ (x ∼ z) ∧ (y.@[l] = z.@[l]) → y = z).
\]

A relative foreign key is an expression \( ϕ \) of the form \( τ(τ_1.@[l] ⊆ FK τ_2.@[l_2]) \), where \( @l_1 \in R(τ_1) \), \( i = 1, 2 \). It indicates that for each \( x \) in \( ext(τ) \), \( @[l_1] \) is a foreign key of descendants of \( x \) of type \( τ_1 \) that references a key \( @[l_2] \) of \( τ_2 \)-descendants of \( x \). That is, \( T \vDash ϕ \) if \( T \vDash τ(τ_2.@[l_2] → τ_2) \) and \( T \) satisfies

\[
∀x ∈ ext(τ) ∀y_1 ∈ ext(τ_1) ((x ∼ y_1) ⇒ ∃y_2 ∈ ext(τ_2) ((x ∼ y_2) ∧ y_1.@[l_1] = y_2.@[l_2])).
\]

Here \( τ \) is called the context type of \( ϕ \). Note that absolute constraints are a special case of relative constraints when \( τ = r \); i.e., \( r(τ.@[l] → τ) \) is the usual absolute key. Thus, the consistency problem for multi-attribute relative constraints is undecidable [16], and hence we only consider unary relative constraints here.

Following the notations for \( AC \), we use \( RC_{FK, FK} \) to denote the class of all unary relative keys and foreign keys defined for element types; \( RC_{FK, FK} \) means the primary key restriction. For example, the constraints given in Section 1 over the country/province/capital DTD can be expressed in \( RC_{FK, FK} \) as follows:

\[
\begin{align*}
&\text{country.}@\text{name} \rightarrow \text{country}, \\
&\text{country(province.}@\text{name} \rightarrow \text{province}), \\
&\text{country(capital.}@\text{inProvince} \rightarrow \text{capital}), \\
&\text{country(capital.}@\text{inProvince} \subseteq FK \text{ province.}@\text{name}).
\end{align*}
\]

A more general form of unary relative constraints is defined in terms of regular path expressions, along the same lines as \( AC_{FK, FK} \). For example, the constraints given in Section 1 over the country/province/capital DTD are instances of this general form of relative constraints. Since \( RC_{FK, FK} \) constraints are a special case of the general regular-expression relative constraints (by substituting \( \tau \) for \( τ \)), the lower bound for \( SAT(\mathcal{RC}_{FK, FK}) \) carries over to the consistency problem for unary relative constraints defined in terms of regular path expressions.

Recall that \( SAT(\mathcal{AC}_{FK, FK}) \), the consistency problem for absolute unary constraints, is NP-complete. One would be tempted to think that \( SAT(\mathcal{RC}_{FK, FK}) \), the consistency problems for relative unary constraints, is decidable as well. We next show, however, that there is an enormous difference between unary absolute constraints and unary relative constraints: while clearly \( SAT(\mathcal{RC}_{FK, FK}) \) is r.e., it turns out that one cannot lower this bound.
Theorem 4.1. SAT(\(R_C_{K,FK}\)) is undecidable.

Proof. We establish the undecidability of the consistency problem for unary relative keys and foreign keys by reduction from the Hilbert’s 10th problem [24]. To do this, we consider a variation of the Diophantine problem, referred as the positive Diophantine quadratic system problem. An instance of the problem is

\[
P_i(x_1, \ldots, x_k) = Q_1(x_1, \ldots, x_k) + c_1 \\
P_2(x_1, \ldots, x_k) = Q_2(x_1, \ldots, x_k) + c_2 \\
\vdots \\
P_n(x_1, \ldots, x_k) = Q_n(x_1, \ldots, x_k) + c_n
\]

where for \(1 \leq i \leq n\), \(P_i\) and \(Q_i\) are polynomials in which all coefficients are positive integers; the degree of \(P_i\) is at most 2 and the degree of each of its monomial is at least 1; each polynomial \(Q_i\) satisfies the same condition, and each \(c_i\) is a nonnegative integer constant. The problem is to determine, given any positive Diophantine quadratic system, whether it has a nonnegative integer solution.

The positive Diophantine quadratic system problem is undecidable. To prove this, it is straightforward to reduce to it another variation of the Diophantine problem, the positive Diophantine equation problem, which is known to be undecidable. An instance of this problem is \(R(y) = S(y)\), where \(R\) and \(S\) are polynomials in which all coefficients are positive integers, and the problem is to determine whether it has a nonnegative integer solution.

In what follows, we show a reduction from the positive Diophantine quadratic system problem to SAT(\(R_C_{K,FK}\)). More precisely, given a quadratic equation we show how to represent it by using a DTD and a set of constraints. It is straightforward to extend this representation to consider an arbitrary number of quadratic equations.

Consider the following equation:

\[
\sum_{i=1}^{m} a_i x_{\alpha_i} + \sum_{i=m+1}^{n} a_i x_{\alpha_i} x_{\beta_i} = \sum_{i=1}^{p} b_i x_{\gamma_i} + \sum_{i=p+1}^{q} b_i x_{\gamma_i} x_{\delta_i} + a.
\]

(4.1)

In this equation, for every \(i \in [1, n]\) and \(j \in [m+1, n]\), \(a_i\) is a positive integer and \(x_{\alpha_i}, x_{\beta_i}\) represent variables, where \(\alpha_i, \beta_i \in [1, k]\). Furthermore, for every \(i \in [1, q]\) and \(j \in [p+1, q]\), \(b_i\) is a positive integer and \(x_{\gamma_i}, x_{\delta_i}\) are variables, where \(\gamma_i, \delta_i \in [1, k]\). Finally, \(a\) is a nonnegative integer.

To code the previous equation, we need to define a DTD \(D = (E, A, P, R, r)\) and a set of \(R_C_{K,FK}\)-constraints \(\Sigma\). Here \(D\) includes the following elements and attributes:

\[
E = \{r, X, Y\} \cup \bigcup_{i=1}^{k} \{\alpha_i\} \cup \bigcup_{i=1}^{n} \{\alpha_i\} \cup \bigcup_{i=m+1}^{n} \{\alpha_i, \beta_i, c_i, d_i, e_i\} \cup \bigcup_{i=1}^{q} \{\gamma_i\} \cup \bigcup_{i=p+1}^{q} \{\gamma_i, \delta_i, f_i, g_i, h_i\},
\]

\[
A = \{@v\}.
\]

In this DTD, \(r\) is the root. Intuitively, referring to an XML tree conforming to \(D\), we use \(\text{ext}(n_i)\) to code the value of the variable \(x_i\), and we use \(\text{ext}(X)\) and \(\text{ext}(Y)\) to code the values of the left and the right hand sides of (4.1), respectively.

We define \(P(r)\) as follows:

\[
P(r) = n_1^s, \ldots, n_k^s, \alpha_1^s, \ldots, \alpha_m^s, \epsilon(\alpha_{m+1}), \ldots, (\epsilon(\alpha_n), \gamma_1^s, \ldots, \gamma_p^s, (\epsilon|\gamma_{p+1}), \ldots, (\epsilon|\gamma_q), Y, \ldots, Y\text{ }a\text{ }times
\]

We define the function \(P\) on \(\alpha_i\) and \(\beta_i\) as follows:

\[
P(\alpha_i) = X, \ldots, X \text{ }a_i\text{ }times
\]

\[
P(\alpha_i) = (\beta_i, c_i, c_i, X, \ldots, X)^s, \alpha_i' \text{ }m + 1 \leq i \leq n
\]

\[
P(\gamma_i) = Y, \ldots, Y \text{ }b_i\text{ }times
\]

\[
P(\gamma_i) = (\delta_i, f_i, f_i, Y, \ldots, Y)^s, \gamma_i' \text{ }p + 1 \leq i \leq q
\]
To code (4.1) we need to capture the multiplication operator. To do this, we use $\alpha_i'$ and $\gamma_i'$:

$$P(\alpha_i') = (\beta_i, d_i, d_i)^*, (\alpha_i, e_i, e_i)^* m + 1 \leq i \leq n$$

$$P(\gamma_i') = (\beta_i, g_i, g_i)^*, (\gamma_i, f_i, h_i)^* p + 1 \leq i \leq q$$

For all the other element types $\tau$ in $D$, $P(\tau)$ is defined as $\varepsilon$:

$$P(\beta_i) = \varepsilon m + 1 \leq i \leq n$$
$$P(\alpha_i) = \varepsilon m + 1 \leq i \leq n$$
$$P(d_i) = \varepsilon m + 1 \leq i \leq n$$
$$P(g_i) = \varepsilon p + 1 \leq i \leq q$$
$$P(h_i) = \varepsilon p + 1 \leq i \leq q$$

Finally, we include the following attributes:

$$R(\gamma_i) = \emptyset$$
$$R(\alpha_i) = \{ @v \} 1 \leq i \leq k$$
$$R(X) = R(Y) = \{ @v \} m + 1 \leq i \leq n$$
$$R(\gamma_i) = \emptyset p + 1 \leq i \leq q$$

To ensure that XML documents that conform to $D$ indeed code equation (4.1) we need to define a set of $\mathcal{RC}_{K, FK}$-constraints $\Sigma$. This set contains the following absolute keys:

$$r(X, @v \rightarrow X)$$
$$r(\alpha_i, @v \rightarrow \alpha_i) \text{ for every } 1 \leq i \leq n$$
$$r(\beta_i, @v \rightarrow \beta_i) \text{ for every } m + 1 \leq i \leq n$$
$$r(\gamma_i, @v \rightarrow \gamma_i) \text{ for every } p + 1 \leq i \leq q$$

$$r(\alpha_i, @v \rightarrow @v) \text{ for every } 1 \leq i \leq n$$
$$r(\beta_i, @v \rightarrow @v) \text{ for every } m + 1 \leq i \leq n$$
$$r(\gamma_i, @v \rightarrow @v) \text{ for every } p + 1 \leq i \leq q$$

$\Sigma$ contains the following absolute foreign keys:

$$r(n_i, @v \rightarrow n_i) \text{ for every } 1 \leq i \leq k$$

We show next that there is an XML tree $T$ such that $T \models (D, \Sigma)$ if and only if there exists a nonnegative integer solution for (4.1). To do this, we prove that every XML tree $T$ satisfying $D$ and $\Sigma$ codifies equation (4.1). More precisely, if the value of every variable $x_i$ is $\alpha_i$ and $|\text{ext}(n_i)| = \alpha_i$, for $i \in [1, k]$, then

$$|\text{ext}(X)| = \sum_{i=1}^{m} a_i v_{\alpha_i} + \sum_{i=m+1}^{n} a_i v_{\alpha_i} v_{\beta_i}$$

$$|\text{ext}(Y)| = \sum_{i=1}^{p} b_i v_{\gamma_i} + \sum_{i=p+1}^{q} b_i v_{\gamma_i} v_{\delta_i} + o$$

Let $T$ be an XML tree conforming to $D$. Then every node of type $X$ in $T$ appears as a child of some node of type $\alpha_i$ ($i \in [1, n]$). Thus, to prove (4.2) it suffices to show that the number of $X$-nodes that are children of some node of type $\alpha_i$ ($i \in [1, n]$) is equal to the $i$-th term of (4.2), that is,

$$[\{x \mid x \text{ is an } X \text{-node in } T \text{ and } x \text{ is a child of a node of type } \alpha_i\}] = a_i v_{\alpha_i} 1 \leq i \leq m,$$

$$[\{x \mid x \text{ is an } X \text{-node in } T \text{ and } x \text{ is a child of a node of type } \alpha_i\}] = a_i v_{\alpha_i} v_{\beta_i} m + 1 \leq i \leq n.$$
Analogously, to show that (4.3) holds, we have to prove that the number of $Y$-nodes that are children of some node of type $\tau_i$ ($i \in [1, q]$) is equal to the $i$-th term of (4.3). We will only consider here the case of $X$-nodes, being the other case similar.

First, let $i \in [1, m]$ and $s$ be the value of $\alpha_i$ in (4.2). Given that $r(n_s, @v) \subseteq_{FK} \alpha_i, @v)$, $r(\alpha_i, @v) \subseteq_{FK} n_s, @v)$ are in $\Sigma$, by the definition of $P(\alpha_i)$ the total number of $X$-nodes that are children of a node of type $\alpha_i$ is equal to $a_i \nu_{\alpha_i}$. Second, let $i \in [m+1, n]$ and $s, t$ be the values of $\alpha_i$ and $\beta_i$ in (4.1), respectively. Next we prove that $|\{x \mid x \text{ is an } X \text{-node in } T \text{ and } x \text{ is a child of a node of type } \alpha_i\}| = a_i \nu_{\alpha_i} \nu_{\beta_i}$. Thus, in $T$ there are exactly $\nu_{x_i}$ nodes of type $\alpha_i$, each of them having exactly one child of type $\alpha'_i$. Hence, there are exactly $\nu_{v_i}$ nodes of type $\alpha'_i$, being the last one of the form shown in Figure 4.1 (see node $r_4$). By the definition of $P(\alpha'_i)$, $|\{x \mid x \text{ is a child of } r_4 \text{ of type } \epsilon_i\}| = |\{x \mid x \text{ is a child of } r_4 \text{ of type } \epsilon_i\}|$. Given that $r(n_t, @v) \subseteq_{FK} \alpha_i, @v)$, $r(\epsilon_i, @v) \subseteq_{FK} n_t, @v)$ are in $\Sigma$ and that every node of type $\epsilon_i$ in $T = \alpha_i \nu_{\epsilon_i} \nu_{\beta_i}$. Thus, since $r_4$ is a node of type $\alpha'_i$ and $\alpha'_i(\beta_i, @v) \subseteq_{FK} \epsilon_i, @v)$, $\alpha'_i(\epsilon_i, @v) \subseteq_{FK} \beta_i, @v)$ are in $\Sigma$. $|\{x \mid x \text{ is a child of } r_4 \text{ of type } \beta_i\}| = |\text{ext}(n_t)| = \nu_{\epsilon_i}$. In addition, by the definition of $P(\alpha'_i)$, the number of children of $r_4$ of type $d_i$ is $2\nu_{\epsilon_i}$.

Given that $r_3$ is a node of type $\alpha_i$ and $\alpha_i(\beta_i, @v) \subseteq_{FK} d_i, @v)$, $\alpha_i(d_i, @v) \subseteq_{FK} \beta_i, @v)$ are in $\Sigma$, $|\{x \mid x \text{ is a child of } r_3 \text{ of type } \beta_i\}| = \nu_{d_i}$, since there are $2\nu_{d_i}$ descendants of $r_3$ of type $d_i$ and $\nu_{d_i}$ children of $r_4$ of type $\beta_i$. Furthermore, by the definition of $P(\alpha_i)$, the number of children of $r_3$ of type $X$ is $a_i \nu_{d_i}$ and the number of children of $r_3$ of type $\epsilon_i$ is $2\nu_{d_i}$. We can use the same argument to prove that the number of children of $r_2$ of types $\beta_i$ and $d_i$ are $\nu_{d_i} \nu_{\epsilon_i}$ and $2\nu_{d_i}$, respectively. Thus, the number of children of $r_1$ of type $X$ is $a_i \nu_{d_i} \nu_{\epsilon_i}$ and the number of descendants of $r_1$ of type $X$ is $2\nu_{d_i} \nu_{\epsilon_i}$. If we continue with this process we can prove, by induction, that the number of $X$-nodes in $T$ that are children of some node of type $\alpha_i$ is $\nu_{\epsilon_i} \nu_{\alpha_i} \nu_{\beta_i}$, since there are $\nu_{d_i}$ nodes of type $\alpha_i$ in $T$. This conclude the proof, since $|\{x \mid x \text{ is an } X \text{-node in } T \text{ and } x \text{ is a child of a node of type } \alpha_i\}| = a_i \nu_{\alpha_i} \nu_{\beta_i}$. In the proof of Theorem 4.1, all relative keys are primary. Thus, we obtain:

**Corollary 4.2.** SAT(RC_{FK,K}), the restriction of SAT(RC_{FK,K}) to primary keys, is undecidable. $\Box$

5. **Extended DTDs.** In this section, we consider a slight extension of DTDs which captures unranked tree automata. An extended DTD [32, 29] $ED$ is a tuple $(D', f, E)$, where $D' = (E', A', P', R', r')$ is a DTD, $E$ is a finite set of element types such that $E \cap E' = \emptyset$ and $f$ is a surjective mapping $f : E' \to E$ such that for every $\tau_1, \tau_2 \in E'$, we have that $R'(\tau_1) = R'(\tau_2)$ if $f(\tau_1) = f(\tau_2)$. We say that a tree $T$ conforms to $ED$ if there exists a tree $T'$ that conforms to $D'$ such that $T = f(T')$, that is, $T$ can be obtained by replacing each label $\tau$ in $T'$ by $f(\tau)$. Extended DTDs support a subtyping mechanism (specialization), and have proven useful in data migration and integration, among other things (see, e.g., [29] for examples of extended DTDs and their applications in data integration). It is also known that extended DTDs capture precisely MSO definable trees and the regular tree languages of finite unranked trees [26, 29].

A constraint $\varphi$ is said to be defined over extended DTD $ED$ if every element type $\tau$ mentioned in $\varphi$ is in $E$.

The consistency problem for extended DTDs is defined exactly as in the case of DTDs: Given a specification $(ED, \Sigma)$, the problem is to verify whether there exists a tree $T$ conforming to $ED$ and satisfying $\Sigma$. 

![Figure 4.1. Part of the XML tree used in the proof of Theorem 4.1.](image)
Next we show that the consistency problem for extended DTDs can be efficiently reduced to the consistency problem for DTDs.

Let $ED = (D', f, E)$ be an extended DTD, where $D' = (E', A', P', R', r')$, and $\Sigma$ be either a set of $\mathcal{AC}_{FK}^{*,*}$-constraints over $ED$ or a set of $\mathcal{AC}_{FK}^{reg}$-constraints over $ED$. We consider here only absolute constraints since, by Theorem 4.1, the consistency problem for extended DTDs and relative constraints is already undecidable for unary keys and foreign keys without regular expressions. We define DTD $D(ED) = (ED', A', P_{ED}, R_{ED}, r')$ as follows. Let $E_{ED} = E' \cup E$ and

$$P_{ED}(\tau) = \begin{cases} f(\tau), & \tau \in E' \\ P(\tau), & \tau \in E \end{cases} \quad R_{ED}(\tau) = \begin{cases} \emptyset, & \tau \in E' \\ R'(\tau'), & \tau \in E \text{ and } f(\tau') = \tau \end{cases}$$

Notice that $R_{ED}$ is well defined since $f$ is a surjective mapping such that $R'(\tau_1) = R'(\tau_2)$ whenever $f(\tau_1) = f(\tau_2)$. Moreover, we define a set of keys and foreign keys $\Gamma(ED, \Sigma)$ over DTD $D(ED)$ as follows. If $\Sigma$ is a set of $\mathcal{AC}_{FK}^{*,*}$-constraints, then $\Gamma(ED, \Sigma) = \Sigma$. Otherwise, $\Sigma$ is a set of $\mathcal{AC}_{FK}^{reg}$-constraints and for every $\varphi \in \Sigma$ we have the following constraint $\psi$ in $\Gamma(ED, \Sigma)$. For every element type $\tau \in E$, define the image of substitution $h$ on $\tau$ as $h(\tau) = (\tau_1 \cup \cdots \cup \tau_n)$, where $\{\tau_1, \ldots, \tau_n\}$ is the set of element types $\tau' \in E'$ such that $f(\tau') = \tau$. If $\varphi$ is a key $\beta.\varphi.l_1 \rightarrow \beta.\tau$, then $\psi = h(\beta.\tau).\varphi.l_1 \rightarrow h(\beta.\tau).\tau$. If $\varphi$ is a foreign key $\beta_1.\varphi.l_1 \subseteq \varphi.l_2 \rightarrow \beta_2.\tau$, then $\psi = h(\beta_1.\varphi.l_1 \subseteq \varphi.l_2 \rightarrow h(\beta_2.\tau), \tau_2 \rightarrow \beta_2.\tau_2, \beta_2.\tau_2 \rightarrow \beta_2.\tau_2$. The following simple lemma shows that the consistency problems for $(ED, \Sigma)$ and $(D(ED), \Gamma(ED, \Sigma))$ are equivalent.

**Lemma 5.1.** Let $ED = (D', f, E)$ be an extended DTD and $\Sigma$ either a set of $\mathcal{AC}_{FK}^{*,*}$-constraints over $ED$ or a set of $\mathcal{AC}_{FK}^{reg}$-constraints over $ED$. Then $(ED, \Sigma)$ is consistent if and only if $(D(ED), \Gamma(ED, \Sigma))$ is consistent.

We note that if $C$ is one of $\mathcal{AC}_{FK}, \mathcal{AC}_{FK}^{-1}, \mathcal{AC}_{FK}^{*,1}$, and $\mathcal{AC}_{FK}^{reg}$, and $\Sigma$ is a set of $C$-constraints over an extended DTD $ED$, then $\Gamma(ED, \Sigma)$ is a set of $C$-constraints over $D(ED)$. Thus, given that $D(ED)$ and $\Gamma(ED, \Sigma)$ can be constructed in polynomial time from $(ED, \Sigma)$, we obtain the following corollary from the previous lemma, Theorem 4.7 and Corollary 4.8 in [16], Corollary 3.5 and Theorem 3.7.

**Corollary 5.2.**

a) The consistency problem for extended DTDs and $\mathcal{AC}_{FK}$-constraints ($\mathcal{AC}_{FK}^{-1}$-constraints) is NP-complete.

b) The consistency problem for extended DTDs and $\mathcal{AC}_{FK}^{+,1}$-constraints is NP-hard, and can be solved in NEXPTIME.

c) The consistency problem for extended DTDs and $\mathcal{AC}_{FK}^{reg}$-constraints is PSPACE-hard, and can be solved in 2-NEXPTIME.

6. Conclusion. We have studied the problem of statically checking XML specifications, which may include various schema definitions as well as integrity constraints. As observed earlier, such static validation
is quite desirable as an alternative to dynamic checking, which would attempt to validate each document; indeed, in the case of repeated failures, one does not know whether the problems lies in the documents or in the specification. Our main conclusion is that, however desirable, the static checking is hard: even with very simple document definitions given by DTDs, and (foreign) keys as constraints, the complexity ranges from NP-hard to undecidable.

The main results are summarized in Figures 6.1, 6.2 (we also included the main results from [16] in those figures for completeness). When one deals with absolute constraints, which hold in an entire document, the general consistency problem is undecidable. It is solvable in NEXPTIME if foreign keys are single-attribute, and is NP-complete if so are all the keys as well. However, if regular expressions are allowed in single-attribute constraints, the lower bounds becomes at least PSPACE. For relative constraints, which are only required to hold in a part of a document, the situation is quite bleak, as even the very simple case of single-attribute constraints is undecidable.

Although most of the results of the paper are negative, the techniques developed in the paper help study consistency of individual XML specification with type and constraints. These techniques include, e.g., the connection between regular expression constraints and integer linear programming and automata.

One open problem is to close the complexity gaps. However, these are by no means trivial: for example, SAT($\mathcal{A}_K, FK$) was proved to be equivalent to a problem related to Diophantine equations whose exact complexity remains unknown. In the case of SAT($\mathcal{A}_K, FK$), we think that it is more likely that our lower bounds correspond to the exact complexity of those problems. However, the algorithms are quite involved, and we do not yet see a way to simplify them to prove the matching upper bounds.

Another topic for future work is to study the interaction between more complex XML constraints, e.g., those defined in terms of XPath [37], and more complex schema specifications such as XML Schema [38] and the type system of XQuery [39]. Our lower bounds apply to those settings, but it is open whether upper bounds remain intact.

Acknowledgments. We are grateful to anonymous referees for their comments.

Most of this work was done while Arenas and Libkin were at the University of Toronto, supported by grants from NSERC, CITO, and PREA grants. M. Arenas was also supported by ONTDECYT grant 1050701, W. Fan by NSF Career Award IIS-0093168, and grants EPSRC GR/S63205/01, EPSRC GR/T27433/01 and NSFC 60228006, and L. Libkin by the European Commission Marie Curie Excellence grant MEXC-CT-2005-024502 and EPSRC grant E005039.

REFERENCES


