March 2003

Propagating XML Constraints to Relations

Susan B. Davidson  
*University of Pennsylvania, susan@cis.upenn.edu*

Wenfei Fan  
*Temple University*

Carmem Hara  
*Universidade Federal do Parana*

Jing Qin  
*Temple University*

Follow this and additional works at: [http://repository.upenn.edu/cis_papers](http://repository.upenn.edu/cis_papers)

Recommended Citation

Propagating XML Constraints to Relations

Abstract
We present a technique for refining the design of relational storage for XML data based on XML key propagation. Three algorithms are presented: one checks whether a given functional dependency is propagated from XML keys via a predefined view; the others compute a minimum cover for all functional dependencies on a universal relation given XML keys. Experimental results show that these algorithms are efficient in practice. We also investigate the complexity of propagating other XML constraints to relations, and the effect of increasing the power of the transformation language. Computing XML key propagation is a first step toward establishing a connection between XML data and its relational representation at the semantic level.

Comments

This material is posted here with permission of the IEEE. Such permission of the IEEE does not in any way imply IEEE endorsement of any of the University of Pennsylvania's products or services. Internal or personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE by writing to pubs-permissions@ieee.org. By choosing to view this document, you agree to all provisions of the copyright laws protecting it.
Propagating XML Constraints to Relations

Susan Davidson*  Wenfei Fan†  Carmem Hara  Jing Qin
U. of Pennsylvania  Bell Labs  U. Federal do Parana, Brazil  Temple U.

Abstract

We present a technique for refining the design of relational storage for XML data based on XML key propagation. Three algorithms are presented: one checks whether a given functional dependency is propagated from XML keys via a predefined view; the others compute a minimum cover for all functional dependencies on a universal relation given XML keys. Experimental results show that these algorithms are efficient in practice. We also investigate the complexity of propagating other XML constraints to relations, and the effect of increasing the power of the transformation language. Computing XML key propagation is a first step toward establishing a connection between XML data and its relational representation at the semantic level.

1 Introduction

Over the past five years, XML has become enormously popular as a data exchange format. A common paradigm is for a data provider to export its data using XML; on the other end, the data consumer imports some or all of the XML data and stores it using database technology. Since the XML data being transmitted is often large in size and fairly regular in structure, the database technology used is frequently relational.

A problem with XML is that it is only syntax and does not carry the semantics of the data. To address this problem, a number of constraint specifications have recently been proposed for XML which include a notion of keys; such proposals have also found their way into XML-Data [18] and XML Schema [28]. A natural question to ask, therefore, is how information about constraints can be used to determine when an existing consumer database design is incompatible with the data being imported, or to generate de-novo a good consumer database. We illustrate the problem below.

Example 1.1: Suppose that the XML data (represented as a tree) in Fig. 1 is being exchanged and that the initial design of the consumer database has a single table Chapter with fields bookTitle, chapterNum

<table>
<thead>
<tr>
<th>bookTitle</th>
<th>chapterNum</th>
<th>chapterName</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML</td>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>XML</td>
<td>10</td>
<td>Conclusion</td>
</tr>
<tr>
<td>XML</td>
<td>1</td>
<td>Getting Acquainted</td>
</tr>
</tbody>
</table>

(a) Chapter: the initial design

<table>
<thead>
<tr>
<th>isbn</th>
<th>chapterNum</th>
<th>chapterName</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>123</td>
<td>10</td>
<td>Conclusion</td>
</tr>
<tr>
<td>234</td>
<td>1</td>
<td>Getting Acquainted</td>
</tr>
</tbody>
</table>

(b) Chapter: a refined design

Figure 2. Sample relational instances

and chapterName (written Chapter(bookTitle, chapterNum, chapterName)). The table is populated from the XML data as follows: For each book element, the value of the title subelement is extracted. A tuple is then created in the Chapter relation for each chapter subelement containing the title value for bookTitle, the number value for chapterNum, and the name value for chapterName (see Fig. 2(a) for the resulting relational instance.) The key of the Chapter table has been specified as bookTitle and chapterNum. While importing this XML data, violations of the key are detected because two different books have the same title (“XML”) and disagree on the name of chapter one (“Introduction” versus “Getting Acquainted”). After digging through the documentation accompanying the XML data, the database designers decide to change the schema to Chapter(isbn, chapterNum, chapterName) with a key of isbn and chapterNum (populated in the obvious way from the XML data). The resulting relational instance is shown in Fig. 2(b). While importing the XML data, no violations of the key constraint are detected. However, the designers are not sure whether they were lucky with this particular XML data set, or whether such violations will never occur.

It turns out that given the following keys on the XML data, the designers of the consumer database could prove that the key of Chapter in their modified design is correct:

1. isbn uniquely identifies a book element.
2. Within each book, number is a key for chapter, i.e., number is a key for chapter relative to book.
3. Each book has a unique title, and within each

---

*Research supported by NSF DBI-9975206.
†Research supported in part by NSF Career Award IIS-0093168. Currently on leave from Temple University.
book, each chapter has a unique name.

That is, if these XML keys hold on the data being imported, then isbn,chapterNum \to chapterName is a functional dependency (FD) that is guaranteed to hold on the Chapter relation generated (in other words, isbn, chapterNum) is a key of the relation. We refer to the FD as one that is propagated from these XML keys.

In general, given a transformation to a predefined relational schema and a set \( \Sigma \) of XML keys, one wants to know whether or not an FD is propagated from \( \Sigma \) via the transformation. Let us refer to this problem as XML key propagation. The ability to compute XML key propagation is important in checking the consistency of a predefined relational schema for storing XML data.

On the other hand, suppose that the relational database is designed from scratch or can be re-designed to fit the constraints (and thus preserve the semantics) of the data being imported. A common approach to designing a relational database is to start with a rough schema and refine it into a normal form (such as BCNF or 3NF [1]) using FDs. In our scenario, we assume that the designer specifies the rough schema by a mapping from the XML document. The FDs over that rough schema must then be inferred from the keys of the XML document using the mapping. However, it is impractical to compute the set \( F \) of all the FDs propagated since \( F \) is exponentially large in the number of attributes. We would therefore like to find a minimum cover [1] of \( F \), that is, a subset \( F_m \) of \( F \) that is equivalent to \( F \) (i.e., all the FDs in \( F \) can be derived from \( F_m \) using Armstrong’s Axioms) and is non-redundant (i.e., none of the FDs in \( F_m \) can be derived from other FDs in \( F_m \)).

**Example 1.2:** Returning to our example, suppose that the database designers decide to start from scratch and initially propose a schema of Chapter(isbn, bookTitle, author, chapterNum, chapterName), with the obvious mapping from the data in Fig. 1. From the three keys given earlier, the following minimum cover for Chapter can be derived: 1) isbn \to bookTitle, and 2) isbn,chapterNum \to chapterName. Taking advantage of these FDs, the following BCNF decomposition of the initial design would be produced: Book(isbn, bookTitle), Chapter(isbn, chapterNum, chapterName), and Author(isbn, author). Note that isbn \to author is not mapped from the keys since a book may have several authors.

**Contributions.** In this paper, we propose a framework for improving consumer relational database design. Our approach is based on inferring functional dependencies from XML keys through a given mapping (transformation) of XML data to relations. The class of XML keys considered includes those commonly found in practice, and is a subset of those in XML Schema [27]. More specifically, we make the following contributions:

- A polynomial time algorithm for checking whether an FD on a predefined relational database is propagated from a set of XML keys via a transformation.
- A polynomial-time algorithm that, given a universal relation specified by a transformation rule and a set of XML keys, finds a minimum cover for all the functional dependencies mapped from XML keys.
- Undecidability results that show the difficulty of XML constraint propagation.
- Experimental results which show that the algorithms are efficient in practice.

Note that the polynomial-time algorithm for finding a minimal cover from a set of XML keys is rather surprising, since it is known that a related problem in the relational context – finding a minimum cover for functional dependencies embedded in a subset of a relation schema – is inherently exponential [16].

The undecidability results give practical motivation for the restrictions adopted in this paper. In particular, one result shows that it is impossible to effectively propagate all forms of XML constraints supported by XML Schema, which include keys and foreign keys, even when the transformations are trivial. This motivates our restriction of constraints to a simple form of XML keys. Another undecidability result shows that when the transformation language is too rich, XML constraint propagation is also not feasible, even when only keys are considered. Since XML to relational transformations are subsumed by XML to XML transformations expressible in query languages such as XQuery [8], this negative result applies to most popular XML query languages.
2 XML Keys and Transformations

XML keys. To define a key we specify three things: 1) the context in which the key must hold; 2) a target set on which we are defining a key; and 3) the values which distinguish each element of the target set. For example, the second key specification of Example 1.1 has a context of a chapter of a book, and a single key value, @number. Specifying the context node and target set involve path expressions.

The path language we adopt is a common fragment of regular expressions [17] and XPath [10]:

\[ Q ::= \epsilon | l | Q/Q | // \]

where \( \epsilon \) is the empty path, \( l \) is a node label, “/” denotes concatenation of two path expressions (child in XPath), and “//” means descendant-or-self in XPath. To avoid confusion we write \( P//Q \) for the concatenation of \( P \), “//” and \( Q \). A path \( \rho \) is a sequence of labels \( l_1, \ldots, l_n \). A path expression \( Q \) defines a set of paths, while “//” can match any path. We use \( \rho \in Q \) to denote that \( \rho \) is in the set of paths defined by \( Q \). For example, book/author/name \( \in //name \).

Following the syntax of [6] we write an XML key as:

\[ K : \{ C, (T, \{ P_1, \ldots, P_p \}) \} \]

where \( K \) is the name of the key, \( C \) is the context of the key, path expressions \( C \) and \( T \) are the context and target path expressions respectively, and \( P_1, \ldots, P_p \) are the key paths. For the purposes of this paper, we restrict the key paths to be simple attributes @A_1, \ldots, @A_p, and denote this class of keys as \( K_a \). A key is said to be absolute if the context path \( C \) is the empty path \( \epsilon \), and relative otherwise.

Example 2.1: Using this syntax, the sample constraints from Section 1 and others can be written as follows:

- \( KS_1 : (\epsilon, //book( @isbn)) \): within the context of the entire document (\( \epsilon \) denotes the root) a book element is identified by its @isbn attribute. The book node can occur anywhere in the tree.
- \( KS_2 : //book( chapter, @number) \): within the context of any subtree rooted at a book node, a chapter is identified by its @number attribute. The chapter node must be immediately under the book node.
- \( KS_3 : //book( title, {}) \): each book has at most one title; similarly,
- \( KS_4 : //book/ chapter( name, {}) \) for the name of a chapter, and
- \( KS_5 : //book/ chapter/ section( name, {}) \) for section name.
- \( KS_6 : //book/ chapter/ section( section, @number) \): within the context of a chapter of a book, each section is identified by its @number attribute.
- \( KS_7 : //book( author/ contact, {}) \): a book can have multiple authors, but at most one has contact information (the contact author).

To define the meaning of an XML key, we use the following notation: in an XML document (tree), \( r[P] \) denotes the set of node identifiers that can be reached by following path expression \( P \) from the node with identifier \( r \). \( [r] \) is an abbreviation for \( r[P] \), where \( r \) is the root node of the tree.

Example 2.2: In Fig. 1, \{book\} = \{119\}, \{1\} = \{67\} and \{119, 67\} = \{119, 123\}.

Definition 2.1: An XML tree \( T \) satisfies an XML key \( \varphi : (C, (T, \{ A_1, \ldots, A_n \})) \), denoted \( T \models \varphi \), iff for any \( n \in [C] \) and any \( n_1, n_2 \in n[T] \), (1) \( n_1 \) and \( n_2 \) each has a unique attribute \( A_i \) for all \( i \in [1, p] \), and (2) if \( val(n_1, A_i) = val(n_2, A_i) \) for all \( i \in [1, p] \) then \( n_1 = n_2 \). For example, if \( val(1, @isbn) \neq val(19, @isbn) \). One can check \( KS_2 \) by verifying the absolute key

\footnote{We adopt this syntax for keys because it is more concise than that of XML Schema.}
This definition of keys has several salient features: First, keys can be scoped within the context of the entire document (an absolute key), or within the context of a sub-document (a relative key). Second, the specification of keys is orthogonal to the typing specification for the document (e.g. DTD or XML Schema). The type of documents will therefore be ignored throughout this paper. Combining keys with schema information, as is done in XML Schema, adds complexity to the inference problem. As demonstrated in [25], the type of documents is orthogonal to the typing specification for the document and thereby the correspondence between keys and schema information.

Transformation Language. The transformation language forms a core of many common transformations found throughout the literature, in particular those of [25].

Definition 2.2: A transformation $\sigma$ from XML data to relations of schema $R = (R_1, \ldots , R_n)$ is specified as $\langle \text{Rule}(R_1), \ldots , \text{Rule}(R_n) \rangle$, where each $\text{Rule}(R_i)$, referred to as the rule for $R_i$, is defined with:

- a set $X_i$ of variables, in which $x_r$ is a distinguished variable, referred to as the root variable;
- a set of field rules $\{ l : \text{value}(x) \mid l \in \text{attr}(R_i) \}$, where $x$ is a variable in $X_i$, and $\text{attr}(R_i)$ denotes the set of attributes in the schema of relation $R_i$;
- a set of variable mapping rules of the form $x \leftarrow y/P$, where $x, y \in X_i$ and $P$ is a path expression.

In addition, each variable $x \in X_i$ is connected to the root $r$; that is, $x$ is specified with either $x \leftarrow x_r/P$ in the rule, or $x \leftarrow y/P$ and $y$ is connected to the root $r$; moreover, for any $x \leftarrow y/P$, $P$ is a simple path (i.e. without //) unless $y$ is $x_r$, and 2) no field rule is defined as $x \leftarrow y/P$. 

Example 2.4: Expanding on Example 1.1, consider the following schema $R$ (with keys underlined):

- book(isbn, title, author, contact),
- chapter(inBook, number, name),
- section(inChapt, number, name).

A transformation $\sigma$ from the XML data of Fig. 1 to $R$ could be specified as:

\[
\sigma = (\text{Rule(book)}, \text{Rule(chapter)}, \text{Rule(section)})
\]

\text{Rule(book)} = \{ isbn: value(z_1), title: value(z_2), author: value(z_3), contact: value(z_4),
\]

\[ x_b \leftarrow x_r/\text{book}, x_1 \leftarrow x_b/@isbn, x_2 \leftarrow x_b/title,
\]

\[ x_3 \leftarrow x_b/\text{author}, x_4 \leftarrow x_b/name, x_4 \leftarrow x_b/\text{name}; \]

\text{Rule(chapter)} = \{ inBook: value(y_1), number: value(y_2), name: value(y_3),
\]

\[ y_b \leftarrow x_r/\text{book}, y_1 \leftarrow y_1/@isbn, y_2 \leftarrow y_1/\text{chapter},
\]

\[ y_3 \leftarrow y_1/@number, y_3 \leftarrow y_1/\text{name}; \]

\text{Rule(section)} = \{ inChapt: value(z_1), number: value(z_2),
\]

\[ z_c \leftarrow x_r/\text{book/chapter}, z_1 \leftarrow z_c/@number, z_2 \leftarrow z_c/@section, z_2 \leftarrow z_c/\text{name}; \]

\[ z_3 \leftarrow z_c/\text{name}, z_3 \leftarrow z_c/\text{name}. \]

Table trees. Throughout the remainder of the paper, we will use an abstract representation of a table rule called a table tree. The idea is that by treating "//" as a special node label, each table rule can be represented as a node-labeled tree. For example, Fig. 3 depicts the table trees for Rule(book) and Rule(section) in Example 2.4. In a table tree $T_R$ representing Rule(R), each variable in Rule(R) corresponds to a unique node, and each node corresponds to at most one variable.

Semantics. Given an XML tree $T$, each Rule($R_i$) maps $T$ to an instance $I_i$ of $R_i$. More specifically, given a variable specification $x \leftarrow y/P$, $x$ ranges over $y/P$, and $y$ is always interpreted as the root $r$. A field rule $l : \text{value}(x)$ populates the field $x$ with values in $\{ \text{value}(x) \mid x \in y/P \}$, where function value returns a string representing the pre-order traversal of the subtree rooted at $x$. Let attr($R_i$) = \{ $l_1, \ldots , l_k$ \} and each variable $x$ be specified with $x \leftarrow x'/P_x$. Then the instance $I_i$ is generated by $I_i = (l_1 : \text{value}(x_1), \ldots , l_k : \text{value}(x_k)) \mid x_r = r, x \in x'/P_x, x \in X_i$. 

Example 2.5: Rule(section) is interpreted as:

\[
\{ (\text{inChapt}=\text{value}(z_1), \text{name}=\text{value}(z_2)) \mid z_1 \in Z_1[@number], z_2 \in Z_2[@section], \}
\]

Referring to the XML tree $T$ in Fig. 1, value(6) returns $\{ (@number=1, \text{name}=\text{Fundamentals}) \}$.
Several subtleties are worth mentioning. First, since XML data is semistructured it is possible that for $x \leftarrow y/P$, $y[P]$ is empty. In this case $\text{value}(x)$ is defined to be $\text{null}$. Second, if $y[P]$ has multiple elements, then to generate the relation, an implicit Cartesian product is computed so that all nodes in $y[P]$ are covered in the relation.

3 Problem Statement and Limitations

Key propagation. The question of key propagation asks if given a transformation $\sigma$ from XML data to relations of a fixed schema $R$ and an XML tree $T$ satisfying a set $\Sigma$ of XML keys, whether $\sigma(T)$ satisfies an FD $\varphi$ (on a schema $R$ in $R$). We write $\Sigma \models_\sigma R : \varphi$ if the implication holds for all XML trees satisfying $\Sigma$, and refer to $\varphi$ as an FD propagated from $\Sigma$. With respect to a transformation specification language, the key propagation problem is to determine, given any $\sigma$ expressed in the language, any XML keys $\Sigma$ and an FD $\varphi$, whether or not $\Sigma \models_\sigma R : \varphi$. Note that we do not require the XML data to conform to any type specification.

A subtle issue arises from null values in $\sigma(T)$, the relations generated from an XML tree $T$ via $\sigma$. In particular, there may exist $R$ tuples in $\sigma(T)$ with FD $X \rightarrow Y$ such that their $X$ or $Y$ fields contain null. The presence of null complicates FD checking since comparisons of null with any value do not evaluate to a Boolean value [23]. A brutal solution is to restrict the semantics of the transformation $\sigma$ so that a tuple is not included if it has a null field. Since XML is semistructured, this could exclude a large number of “incomplete” tuples from $\sigma(T)$. We therefore adopt the following semantics of FDs: $\sigma(T)$ satisfies FD $X \rightarrow Y$, denoted by $\sigma(T) \models X \rightarrow Y$, iff (1) for any tuple $t$ in $R$, if $\pi_X(t)$ contains null then so does $\pi_Y(t)$; and (2) for tuples $t_1, t_2$ in $R$, if neither $t_1$ nor $t_2$ contains null and $\pi_X(t_1) = \pi_Y(t_2)$, then $\pi_Y(t_1) = \pi_Y(t_2)$. The motivation behind the first condition is that an FD is possibly treated as a key when normalizing the relational schema, and an “incomplete key” $X$ cannot determine complete $Y$ fields.

Another issue we should address is the simplicity of the transformation language, which can only express projection ($\pi$), Cartesian product ($\times$) and a limited form of set union (U). One might be tempted to develop a richer language which can express all relational algebra operators: projection, selection ($\sigma$), Cartesian product, set union and difference ($\setminus$). Although these operators can be generalized to XML trees, the following negative result holds:

**Theorem 3.1:** The key propagation problem from XML to relational data is undecidable when the transformation language can express all relational algebra operators.

The undecidability is established by reduction from the equivalence problem for relational algebra queries (see [11] for a proof); the latter is a well-known undecidable problem [1]. In contrast, for our transformation language there is a polynomial time algorithm in the size of $\Sigma$ and $\sigma$.

**Figure 4. Rule (U)**

**Minimum cover.** The problem of finding a minimum cover is to compute, given a universal relation $U$ and a set $\Sigma$ of XML keys, a minimum cover $F_m$ for the set $F^+$ of all FDs on $U$ propagated from $\Sigma$. Guided by $F_m$, one can then decompose $U$ into a normal form as illustrated by Example 1.2. This is analogous to techniques for designing relational databases [1]. In our context, a universal relation is simply the collection of all the fields of interest, along with a table rule that defines these fields.

**Example 3.1:** Recall the schema $R$ and the transformation given in Example 2.4. A universal relation $U$ here is the collection of all the fields of $R$, defined as follows:

$U = \{\text{bookIsbn}, \text{bookTitle}, \text{bookAuthor}, \text{authContact}, \text{chapNum}, \text{chapName}, \text{secNum}, \text{secName}\}$

**Rule (U)** = \{\text{bookIsbn: value(x1), bookTitle: value(x2), bookAuthor: value(x3), authContact: value(x4), chapNum: value(y1), chapName: value(y2), secNum: value(z1), secName: value(z2)}\},

$X_5 \leftarrow X_r/\text{book}, X_1 \leftarrow x_5/@\text{isbn}, x_2 \leftarrow x_5/\text{title}, x_3 \leftarrow x_5/@\text{author}, x_3 \leftarrow x_3/@\text{name}, x_4 \leftarrow x_3/@\text{contact}, y_1 \leftarrow y_1/@\text{number}, y_2 \leftarrow y_2/@\text{name}, z_1 \leftarrow z_1/@\text{number}, z_2 \leftarrow z_2/@\text{name}$

The table rule of Rule (U) is depicted in Fig. 4. From the set of XML keys of Example 2.1 the following minimum cover for the FDs on $U$ can be computed:

- \text{bookIsbn} \rightarrow \text{bookTitle}, \text{bookIsbn} \rightarrow \text{authContact}, \text{bookIsbn} \rightarrow \text{chapNum} \rightarrow \text{chapName}, \text{bookIsbn} \rightarrow \text{secNum} \rightarrow \text{secName}.

Guided by these FDs, we can decompose $U$ into BCNF:

book(bookIsbn, bookTitle, authContact),
author(bookIsbn, bookAuthor),
chap(bookIsbn, chapNum, chapName),
section(bookIsbn, chapNum, secNum, secName)

Although in the relational context algorithms have been developed for computing a minimum cover for a set of FDs [4, 16, 20], they cannot be used in our context since the FDs must be computed from the XML keys $\Sigma$ via the transformation $\sigma$, instead of being provided as input for those relational algorithms. Furthermore, relational FDs are not capable of expressing XML keys and vice versa.
Propagation of other XML constraints. XML Schema supports keys and foreign keys. Although it is tempting to develop algorithms to compute the propagation of both keys and foreign keys, we have the following negative result:

**Theorem 3.2:** The propagation problem for XML keys and foreign keys is undecidable for any transformation language that can express identity mapping.

The “identity” mapping is one in which the XML representation of relations is mapped to the same relations (in our language this corresponds to a small class of transformations defined with paths of length 3). The undecidability result is established by reduction from implication of relational keys and foreign keys, which is undecidable [15] (see [11] for a reduction). Because of this we restrict our attention to the propagation of XML keys.

4 Checking Key Propagation

Checking key propagation is nontrivial for a number of reasons: First, XML data is semistructured in nature, which complicates the analysis of key propagation by the presence of null values. Second, XML keys which are not in \(\Sigma\) but are consequences of \(\Sigma\) may yield FDs on a relational view. Thus key propagation involves XML key implication. Third, XML data is hierarchically structured and thus XML keys are relative in their general form – they hold on a sub-document. However, its relational view collapses the hierarchical structures into a flat table and thus FDs are “absolute” – they hold on the entire relational view. Thus one needs to derive a unique identification of a sub-document from a set of relative keys.

Before presenting our polynomial-time algorithm for checking XML key propagation (Algorithm propagation), we first discuss the notion of a “keyed” node and the implication of XML keys.

**Transitive set of XML keys.** To uniquely identify a node within the entire document we need a set of XML keys identifying unique contexts up to the root. To formalize this, we use the following notion [7]: \((Q_1, (Q'_1, S_1))\) immediately precedes \((Q_2, (Q'_2, S_2))\) if \(Q_2 = Q_1/Q'_1\). The precedes relation is the transitive closure of the immediately precedes relation. A set \(\Sigma\) of keys is transitive if for any relative key \((Q_1, (Q'_1, S_1))\) in \(\Sigma\) there is an absolute key \((\epsilon, (Q'_2, S_2))\) in \(\Sigma\) which precedes \((Q_1, (Q'_1, S_1))\). We say that a node is keyed if there exists a transitive set of keys to uniquely identify the node.

**Example 4.1:** The set \(\{K_S_1, K_S_2\}\) is transitive since any chapter in the document can be identified by providing \#isbn\_of a book and \#number\_of a chapter. Thus every chapter node is keyed. In contrast, \(\{K_S_2\}\) is not transitive since with it alone there is no way to uniquely identify a book in the document, which is necessary before identifying a chapter of that book.

**Implication of XML keys.** One aspect of key propagation is to determine whether an XML key \(\varphi\) must hold provided that a set \(\Sigma\) of XML keys holds, denoted by \(\Sigma \vdash \varphi\). In other words, \(\Sigma \vdash \varphi\) if for any XML tree \(T\), \(T\) satisfies \(\varphi\) as long as \(T\) satisfies all the keys in \(\Sigma\). An algorithm for implication analysis, implication, can be found in [11]. The algorithm takes as input a set \(\Sigma\) and \(\varphi\) of XML keys of \(\mathcal{K}_a\) and returns true iff \(\Sigma \vdash \varphi\). It is based on a set of inference rules that, along the same lines as the Armstrong’s Axioms for implication of FDs in relational databases, allows one to derive key implication systematically. One example of the rules is target-to-context: if \((Q, (Q_1/Q_2, S))\) is a key then so is \((Q/Q_1, (Q_2, S))\). Intuitively, the rule states that if \(S\) can uniquely identify a set \(N\) of nodes in the entire tree \(T\), then it can also identify nodes of \(N\) in any subtree of \(T\); observe that for any nodes \(n \in [Q]\) and \(n' \in n[Q_1]\), the subtree rooted at \(n'\) is a subtree of the one rooted at \(n\). Another example of a trivial rule is epsilon: for any path \(Q_i\), it is true that \((Q, (\epsilon, \{\})}\). Intuitively, it states that any subtree has a unique root node. Algorithm implication determines whether or not \(\Sigma \vdash \varphi\) in \(O(|\Sigma|^2 \cdot \varphi^2)\) time, where \(|\Sigma|\) and \(|\varphi|\) are the sizes of \(\Sigma\) and \(\varphi\).

**Table tree.** Algorithm propagation uses the tree representation of a transformation to bridge the gap between XML keys and the FD \(\phi\) to be checked. Without loss of generality, assume that \(\phi\) is of the form \(Y \rightarrow l\) with \(l \in \text{attr}(\mathcal{R})\) and \(Y \subseteq \text{attr}(\mathcal{R})\), and that for the relation \(R\), \(\text{Ruam}(R)\) is \([ \{i \mid \text{value}(x_i) \mid i \in [1, m]\} ]\) along with a set \(X\) of variables and mappings \(x \leftarrow y/P\) for each \(x \in X\). In the table tree \(T_R\) representing \(\text{Ruam}(R)\), any variable \(x\) in \(X\) has a unique node corresponding to it, referred to as the \(x\)-node. In particular, the \(x_r\)-node is the root of \(T_R\). Observe that for any \(x, y \in X\), if the \(x\)-node is a descendant of the \(y\)-node in \(T_R\), then there is a unique path in \(T_R\) from the \(y\)-node to \(x\)-node, which is a path expression. We denote the path by \(P(y, x)\), which exists only if there are variables \(x_1, \ldots, x_k \in X\) such that \(x_1 = y, x_k = x\) and for each \(i \in [1, k-1], x_{i+1} \leftarrow x_i/P_i\) is a mapping in \(\text{Ruam}(R)\). We use \(\text{descendants}(y)\) to denote the set of all the variables that are descendants of \(y\); we define \(\text{ancestors}(y)\) similarly. In particular, if \(x\) is specified with \(x \leftarrow y/P\) then the variable \(y\) is called the parent of \(x\), denoted by \(\text{parent}(x)\). Referring to Fig. 3 (b), for example, \(x_r\) is the parent of \(Z_o\), and \(P(x_rZ_o)\) is //book/chapter.

**Algorithm.** The intuition behind Algorithm propagation is as follows. Given an FD \(\phi \Rightarrow Y \rightarrow l\) on \(R\), assume that \(l\) is specified with \(\text{value}(x)\), and that the table tree representing \(\text{Ruam}(R)\) is \(T_R\). Then \(\Sigma \vdash \phi\) iff (1) either \(\phi\) is trivial, that is, \(l \in Y\), or there exists an ancestor target of \(x\) in \(T_R\) such that target is keyed with fields of \(Y\) and moreover, \(x\) is unique under target; that is, there is a set of transitive keys that uniquely identifies target with only those attributes which define
Algorithm propagation

Input: XML keys $\Sigma$, FD $\phi = Y \rightarrow l$ over $R$, and Rule$(R)$ in transformation $\sigma$, in which $l : value(x)$.
Output: true iff $\Sigma \models_{R} : \phi$.
1. ancestor[$x$]$ := \text{nil}$;
2. $w := x$;
3. while $w \neq x$, do
4. $w := \text{parent}(w)$;
5. ancestor[$y$] := $w$ :: ancestor[$x$];
6. $Ycheck := Y - \{l\}$;
7. if $l \notin Y$
8. then keyFound := true;
9. else keyFound := false;
10. context := $x$;
11. while ancestor[$y$] $\neq \text{nil}$ do
12. target := head(ancestor[$y$]);
13. $S := \{\text{a} | l' \in Y, l' : value(y) \in \text{Rule}(R),
       y \leftarrow target/\text{a} \text{ is a variable mapping}\}$
14. if not keyFound
15. then if implication($\Sigma$, (P(x, context),
16. (P(context, target), $S$)))
17. then context := target;
18. if implication($\Sigma$, (P(x, target),
19. (P(target, x), $\{\}$)))
20. then keyFound := true;
21. if exists(P(x, target), $S$)
22. then $L := \{l' \mid l' \in Y, l' : value(y) \in \text{Rule}(R),
       y \leftarrow target/\text{a} \text{ is a variable mapping}\}$
23. $Ycheck := Ycheck - L$;
24. ancestor[$x$] := fail(ancestor[$x$]);
25. return keyFound and (Ycheck = $\{\}$);

function exist(Q, $S$)
Input: Q: path expression; $S$: a set of attributes.
Output: true iff for all $l \in S$ and $n \in Q$, $n @ l$ exists.
1. $X := S$;
2. for each key $\phi = (Q_1, (Q_1', S_1))$ in $\Sigma$ do
3. if $Q \subseteq Q_1 / Q_1'$
4. then $X := X - S_1$;
5. return ($X = \{\}$);

Figure 5. XML key propagation algorithm

fields of $Y$, and $\Sigma \models_{R} := (P(x, target), (P(target, x), \{\}))$; 
(2) every field of $Y$ is defined with an attribute of some ancestor of $x$ that is required to exist. The first condition asserts that for any $R$ tuples $t_1$ and $t_2$, if they agree on their $Y$ fields and do not contain null, then they agree on their $l$ fields. The second condition excludes the possibility that in some $R$ tuple $t$, the $l$ field is defined while some of their $Y$ fields are null.

Putting everything together, Algorithm propagation is shown in Fig. 5. The algorithm first computes the list of all the ancestors of $x$ (Lines 1 to 5); it then traverses the table-tree $T_R$ top-down along the ancestor path from the root $x_r$ to $x$ (Lines 11 to 22), and for each ancestor $target$ in this path, checks if $target$ is keyed (Lines 15). The central part of the algorithm is to check whether there is a set of transitive keys for $target$. To do so, it uses variable $context$ to keep track of the closest ancestor for which a key has been found, and collects the attributes of $target$ that populate fields in $Y$ in a set $S$. Thus $target$ is keyed iff $\Sigma \models_{R} := (P(x, context), (P(context, target), S))$, i.e., $S$ is a key of $target$ relative to its closest ancestor with a key. XML key implication is checked by invoking Algorithm implication mentioned above. If it holds, the algorithm moves $context$ down to $target$ (Line 16; the correctness of this step is ensured by the $target$-to-context rule given above); then, it sets the Boolean flag $keyFound$ to true if $x$ is unique under $target$ (Line 17). To ensure that all the fields of $Y$ are defined with attributes of ancestors of $x$ that are required to exist, it uses a variable $Ycheck$ (with an initial value of $Y - \{l\}$) and removes from $Ycheck$ the field names that correspond to the set $S$ of attributes (Lines 19 to 21). The algorithm returns true iff $keyFound$ and $Ycheck$ becomes empty, i.e., the two conditions given above are satisfied.

Example 4.2: To illustrate the algorithm, recall the transformation $\sigma$ of Example 2.4 and the set $\Sigma$ of XML keys of Example 2.1. Consider FD: $\text{isbn} \rightarrow \text{contact}$ over relation book defined by Rule(book), which is depicted in Fig. 3 (a). Note that the field $\text{contact}$ in the FD is specified with variable $x_4$. Given $\Sigma, \sigma$ and the FD, the algorithm computes the ancestors of $x_4$, which consists of $x_r, x_6, x_5$ and $x_4$. Then, it first checks if $x_r$ is keyed by inspecting $\Sigma \models_{R} := (\epsilon, \{\})$. Since this holds by the $\text{epsilon}$ rule given above, the algorithm then checks whether $x_6$ is keyed by inspecting $\Sigma \models_{R} := (//\text{book}[@\text{isbn}])$. Since this is also true, the algorithm proceeds to check whether $x_5$ is unique under $x_4, i.e., whether $\Sigma \models_{R} := (//\text{book}[@\text{author}/@\text{contact}, \{\}])$. This is also the case. In addition, the field $\text{isbn}$ in the FD is defined in terms of an attribute of $x_4$ that is required to exist. That is, by the semantics of keys, $//\text{book}[@\text{isbn}]$ requires every book element to have an $\text{isbn}$ attribute. Thus the algorithm concludes that the FD is derived from $\Sigma$ via $\sigma$ and returns true.

Next, let us consider Rule(section) of Example 2.4, represented by the table tree of Fig. 3 (b), and let $\phi$ be an FD: $\text{inChapt, number} \rightarrow \text{name}$ over relation section. After successfully verifying that $x_r$ is keyed, the algorithm checks whether its next ancestor is keyed, i.e., whether $\Sigma \models_{R} := (//\text{book}/@\text{chapter}, \{\text{number}\})$. This fails. Thus it attempts to verify another key relative to the root: $\Sigma \models_{R} := (//\text{book}/@\text{chapter}/@\text{section}, \{\text{number}\})$, which fails again. At this point the algorithm concludes that the FD cannot be derived from $\Sigma$ and returns false. □

The complexity of the algorithm is $O(m^2n^3)$, where $m$ and $n$ are the sizes of XML keys $\Sigma$ and table tree $T_R$, re-
respectively (see [11] for details as well as for a proof of correctness of the algorithm).

5 Computing Minimum Cover

In this section we present two algorithms for finding a minimum cover for FDs propagated from XML keys. The first algorithm is a direct generalization of Algorithm propagation of Fig. 5, and always takes exponential time. We use this naive algorithm to illustrate the difficulties in connection with finding a minimum cover. The second algorithm takes polynomial time in the size of input, by reducing the number of FDs generated in the following way: a new FD is inserted in the resulting set only if it cannot be implied from the FDs already generated, using the inference rules for FDs. To the best of our knowledge, this is the first effective algorithm for finding a minimum cover for FDs propagated from XML keys.

A Naive Algorithm. Algorithm propagation given in the last section allows us to check XML key propagation. Thus a naive algorithm for finding a minimum cover is to generate each possible FD on U, check whether or not it is in \( F^+ \), the set of all the FDs mapped from the XML keys, using Algorithm propagation, and then eliminate both extraneous attributes and redundant FDs from \( F^+ \) using standard relational database techniques; this yields a minimum cover \( F_m \) for \( F^+ \). The algorithm, Algorithm naive, can be found in [11]. It takes exponential time in the size of U for any input since it computes all possible FDs on U. It should be mentioned that the function invoked by the algorithm for eliminating redundancy, Function minimize given below [4], takes quadratic time in the size of its input FDs, since FD implication can be checked in linear time using the Armstrong’s Axioms; but when invoked in naive, the set of input FDs is exponentially large.

```plaintext
function minimize(F)
    Input: F: a set of FDs.
    Output: A non-redundant cover of F.
    1. for each \( Y \rightarrow l \) in \( F \) do /* eliminate extra attributes */
       1.1. for each \( \ell \) in Y do
          1.1.1. if \( F \models (Y \setminus \{ \ell \}) \rightarrow l \)
              1.1.2. then \( Y := Y \setminus \{ \ell \} \);
       1.2. \( G := F \); /* eliminate redundant FDs */
    2. for each \( \phi \) in F do
       2.1. if \( (G \setminus \{ \phi \}) \models \phi \)
          2.2. then \( G := G \setminus \{ \phi \} \);
    3. return G;
```

Obviously, Algorithm naive is too expensive to be practical. The problem is that it needs to compute \( F^+ \), which is exponential in the size of U even with trivial FDs removed. This observation motivates us to develop an algorithm that directly finds \( F_m \) without computing \( F^+ \).

A Polynomial-Time Algorithm. We next present a more efficient algorithm for finding a minimum cover for all the propagated FDs. The algorithm takes \( O(m n^3) \) time, where \( m \) and \( n \) are the sizes of XML keys \( \Sigma \), and the transformation \( \sigma \), respectively. The algorithm works as follows. Recall that the transformation \( U \rightarrow \Sigma \) can be depicted as a table tree \( T \), in which each variable \( x \) is in the set \( X \) of \( U \) is represented by a unique node, referred to as the \( x \)-node. The algorithm traverses \( T \) top-down starting from the root of \( T \), \( x_r \), and generates a set \( F \) of FDs that is a cover of \( F^+ \), i.e., a superset of \( F_m \). More specifically, at each \( x \)-node encountered, it expands \( F \) by including certain FDs propagated from \( \Sigma \). It then removes redundant FDs from \( F \) to produce a minimum cover \( F_m \).

The obvious question is what new FDs are added at each \( x \)-node. As in Algorithm propagation, at each \( x \)-node a new FD \( Y \rightarrow l \) is included into \( F \) only if \( (x \in Y) \) \( Y \) is a key with a set of attributes that define the fields in \( Y \); (2) the field \( l \) is defined by the value of a node \( y \) and \( y \) is unique under \( x \).

Example 5.1: Recall the universal relation \( U \) defined by the transformation \( \sigma \) of Example 3.1, the table tree depicted in Fig. 4, and the set \( \Sigma \) of XML keys of Example 2.1. An FD derived from \( \Sigma \) at the \( z \)-node is \( \text{book} \rightarrow \text{name} \rightarrow \text{name} \rightarrow \text{name} \rightarrow \text{chapter} \rightarrow \text{chapter} \rightarrow \text{chapter} \rightarrow \text{chapter} \). The left-hand side of the FD corresponds to a transitive set of keys for the \( z \)-node consisted of a section \( @number \) which is an attribute of \( \text{z}_s \), as well as a chapter \( @number \) and a book \( @isbn \), which are a key of \( \text{z}_s \)’s ancestor \( y_e \). The right-hand side of the FD is defined by a node \( z \) unique under \( \text{z}_s \) by \( \text{K}S_0 \) in \( \Sigma \). Thus the key for the \( z \)-node actually consists of the key of its ancestor \( y_e \) as well as a key for section \( @number \) relative to \( y_e \).

Critical to the performance of the algorithm is to minimize the number of FDs added at each \( x \)-node while ensuring that no FDs in \( F_m \) are missed. This is done in two ways: First, we reduce our search for candidate FDs to those whose left-hand side corresponds to attributes of keys in \( \Sigma \). Second, we observe that an ancestor \( target \) of an \( x \)-node may have several keys, but that in creating a transitive key for \( x \) only one of them needs to be selected as long as the following property is enforced: for any two transitive keys \( K_1 \) and \( K_2 \) of the \( x \)-node, \( F \) includes \( Y_i \rightarrow l \) for each \( l \in Y_2 \) and \( Y_2 \rightarrow l' \) for each \( l' \in Y_1 \), where \( Y_1 \) and \( Y_2 \) are sets of \( U \) fields defined by \( K_1 \) and \( K_2 \), respectively. Given this, \( Y_1 \) and \( Y_2 \) are equivalent by Armstrong’s Axioms.

There is a subtlety caused by the troublesome null value. Let \( K_2 \) be a transitive key for an \( x \)-node, \( K_1 \) be a transitive key for an ancestor \( y \) of \( x \), \( Y_1 \) and \( Y_2 \) be the sets of \( U \) fields defined by \( K_1 \) and \( K_2 \), respectively, and \( Z \) be another set of \( U \) fields. Then the following is a rule for populating \( F \); if \( (Y_1 \cup Z \rightarrow l) \) is in \( F \) and \( l \) is a \( U \) field defined by a descendant \( z \) of \( x \), then \( (Y_2 \cup Z \rightarrow l) \) should also be included in \( F \). The intuition behind this rule is that a key for \( x \) is also a key for its ancestor \( y \), provided that the existence of \( x \) under \( y \) is assured. This is because
there is a unique ancestor $y$ of $x$ in a tree that connects to $x$ via the path $P(y, x)$. Thus, provided the existence of $x$ under $y$, we have $Y_2 \rightarrow l'$ for any $l' \in Y_1$. As a result, if $Y_1 \cup Z \rightarrow l$ then $Y_2 \cup Z \rightarrow l$, by the transitivity of FD implication. The existence of $x$ under $y$ is ensured by the existence of $z$: if $Y_2 \cup Z \rightarrow l$, then by the definition of FDs, $z$ must exist under $y$; hence, $x$ must be on the path from $y$ to $z$, i.e., $x$ exists. It is worth remarking that when $l$ is not defined by a descendant of $x$, the rule may not be sound since $Y_2 \rightarrow l'$ may not hold for an $l' \in Y_1$; more specifically, $Y_2$ consists of null if $x$ does not exist under $y$, while $l'$ may not be. As an example, consider two transitive keys at a node $x$: $K_1 = \{(Q_1, \{A_1\}), (Q_1, (Q_2, \{A_3\}))\}$ and $K_2 = \{(Q_1/Q_2, \{A_4\})\}$, with each attribute $A_i$ populating a field $l_i$ in $U$. Note that $(Q_1, \{A_1\})$ is a key of the ancestor $y$ of $x$ that connects to $x$ via the path $Q_2$. Consider the following FDs: $f_1 = (l_1, l_2) \rightarrow l_3$, $f_2 = l_3 \rightarrow l_4$, and $f_3 = l_2 \rightarrow l_3$. Although $f_1$ and $f_2$ are indeed propagated from $K_1$ and $K_2$, $f_3$ is not. This is because the existence of attribute $A_2$ of node $y$ in $(Q_1)$ does not guarantee the existence of attribute $A_3$ of $x$ in $y(Q_2)$; therefore, $l_4$ can have a non-null value even when $l_2$ is null, violating the FD. Thus $f_2$ should not be included in $F$. However, if $F$ contains $l_1 \rightarrow l_4$ for some $l_4$ defined with a descendant of $x$, then $F$ should also include $l_3 \rightarrow l_4$.

To keep track of the information needed to generate FDs at each $x$-node, we associate the following with each variable $x$ in Rule($U$):

- $keys[x]$: a list of sets of $U$ fields, each set mapped from a transitive key of the $x$-node;
- $keyAnce[x]$: a set of sets of $U$ fields, each mapped from a transitive key of an ancestor of $x$;
- $desc[x]$: the set of all descendants of the $x$-node;
- $unique[x]$: the set of all unique descendents of the $x$-node;
- $att[x]$: the set of attributes of the $x$-node;
- $ancestor[x]$: the list of all the ancestors of $x$ starting from the root:

Note that $unique[x]$ is a subset of $desc[x]$, and $att[x]$ is a subset of $unique[x]$ since any node in an XML tree has at most one attribute labeled with a particular name.

Using this notation, at each $x$-node, we expand the cover $F$ of FDs as follows: First, for each $(Q, (Q', S))$ in $\Sigma$ we compute $K$, the set of fields of $U$ defined by attributes in $S$. We check whether $S$ is contained in $att[x]$ and whether every attribute of $S$ defines a $U$ field (by comparing the cardinals of $S$ and $K$). If it is the case then we traverse the ancestor path of $x$ starting from the root. We find the first ancestor $target$ of $x$ that is keyed, and check whether $S$ is a key for the $x$-node relative to $target$ using Algorithm implication. If these conditions are met we construct a transitive key $K'$ for $x$ by combining $S$ with an arbitrary transitive key for the ancestor $target$ of the $x$-node, which is in $keys[target]$. We increment $keys[x]$ by adding

---

**Figure 6. Computing minimum cover**
this transitive key, and insert $keys[target]$ in $keyAne[x]$. Second, we expand $F$ by including $Y \rightarrow l$ for each $l$ in $unique[x]$ (excluding $Y$), where $Y$ is a set of $U$ fields defined by $K'$. That is, the transitive key of the $x$-node determines the unique descendants of $x$. After the set $F$ is computed by traversing all variables $x$ in $Rule(U)$, it is expanded by applying transitivity on keys in $keyAne[x]$, which includes a key of $x$'s closest keyed ancestor. That is, if $K_1 \in keyAne[x]$, we inspect each $Y \rightarrow l$ in $F$, checking if $K_1$ is a subset of $Y$, i.e., whether there exists $Z$ such that $Y = K_1 \cup Z$. If this is the case and $l \in dese[x]$, then for each $K_2$ in $keys[x]$ we add $(K_2 \cup Z) \rightarrow l$ to $F$. One can show that the sizes of $keys[x]$ and $keyAne[x]$ are quadratic in the size of $\Sigma$, $unique[x]$ is bounded by the size of $Rul\epsilon(U)$, and the set $X$ is no larger than the size of $Rul\epsilon(U)$; thus the set $F$ is bounded by $m^4n^5$, i.e., the size (thus the cardinality) of $F$ is at most $O(m^4n^5)$, a polynomial in the input size.

**Algorithm.** Based on these observations, we show Algorithm $minimumCover$ in Fig. 6. After computing $dese[x]$, $att[x]$, $unique[x]$ and initializing $keys[x]$, and $keyAne[x]$ for each variable $x$ in $Rul\epsilon(U)$ (Lines 1 to 10), the algorithm initializes these variables for the root node (Lines 11, 12), and inserts in $F$ FDS of the form $\emptyset \rightarrow l$ for each unique field under the root (Line 13). It then invokes a recursive procedure $genFDs$ to process the children of the root node (Line 14). Procedure $genFDs$ expands $F$ given an input $x$-node as described above, and recursively processes the children of the $x$-node. After $F$ is computed, Algorithm $minimumCover$ expands it by applying the transitivity rule (Lines 15 to 22) and invokes function $minimize$ given in the last section to eliminate redundant FDS from $F$, and thus yields a minimum cover $F_m$ (Lines 23). The correctness of the algorithm is established in [11].

**Example 5.2:** Given the transformation $\sigma$ of Example 3.1 and the set $\Sigma$ of XML keys of Example 2.1, Algorithm $minimumCover$ returns the FDS given in Example 3.1, which are a minimum cover for all the FDS propagated from $\Sigma$ via $\sigma$. Specifically, the algorithm traverses the table tree of Fig. 4 (a) top-down starting at the root. At node $x_{\text{book}}$, two FDS are generated: one is $\text{bookIbm} \rightarrow \text{bookTitle}$, and the other is $\text{bookIbm} \rightarrow \text{authContact}$. Here $keys[x_{\text{book}}] = \{[\text{isbn}]\}$. At node $y_{\text{chap}}$, FD $\text{bookIbm}, \text{chapNum} \rightarrow \text{chapName}$ is included in $F$ and $keys[y_{\text{chap}}]$ is changed to $\{[\text{isbn}, \text{number}]\}$, which is constructed by combining $\text{number}$, a key of $y_{\text{chap}}$ relative to $x_{\text{book}}$, and the key in $keys[x_{\text{book}}]$. Similarly, at node $x_{\text{sec}}$, FD $\text{bookIbm}, \text{chapNum}, \text{secNum} \rightarrow \text{secName}$ is inserted into $F$. No FDS are generated at any other nodes.

The complexity of the algorithm is $O(m^8n^8)$ time, where $m$ and $n$ are the sizes of XML keys $\Sigma$ and table tree $T_R$, respectively (see [11] for details). Since $\Sigma$ and $Rul\epsilon(U)$ are usually small, this algorithm is efficient in practice. The experimental results of the next section also show that it substantially outperforms Algorithm naive.

A final remark is that, although one can generalize Algorithm $minimumCover$ to check XML key propagation instead of using Algorithm propagation, there are good reasons for not doing so. The complexity of Algorithm $minimumCover$ is much higher than that of Algorithm propagation ($O(m^8n^8)$ vs. $O(m^2n^3)$). In short, Algorithm propagation is best used to inspect a predefined relational schema, whereas Algorithm $minimumCover$ helps normalize a universal relation at the early stage of relational design.

### 6 Experimental Study

The various algorithms presented in this paper have been implemented, and a number of experiments performed. The results of these experiments show that despite their $O(m^2n^3)$ and $O(m^8n^8)$ worst-case performance, both Algorithms propagation and $minimumCover$ work well in practice: they take merely a few seconds even given large transformation and XML keys. For computing minimum cover, Algorithm $minimumCover$ is several orders of magnitude faster than Algorithm naive, and for checking key propagation Algorithm propagation significantly outperforms the generalization of Algorithm $minimumCover$. Our results also reveal that Algorithm $minimumCover$ is more sensitive to the number of XML keys than to the size of the transformation. This is nice since in many applications the number of keys does not change frequently, whereas a relational schema may define tables with a variety of different arities (number of fields). Our results also show that Algorithm propagation has a surprisingly low sensitivity to the size of the transformation, and that its execution time grows linearly with the size of XML keys.

To perform these experiments, we synthetically generated transformations and XML keys based on the number of fields in a relation, the depth of a table-tree, and the number of XML keys. All experiments were conducted on the same 1.6GHz Pentium 4 machine with 512MB memory. The operating system is Linux RedHat v7.1 and the program was implemented in C++.

The first experiment evaluates the performance of the two algorithms for computing minimum cover (see Fig. 7(a)). These results tell us the following. First, the average complexity of Algorithm $minimumCover$ in practice is much better than its $O(m^8n^8)$ worst-case complexity. Consider, for example, the execution time of the algorithm for $depth = 10$ and $key = 10$. When the number of fields is increased (which corresponds roughly to increasing the size of the transformation), the execution time grows in the power of two in average instead.
of in the power of six. Second, the algorithm needs less than 35 seconds for 200 fields, and a little over 2 minutes even for 500 fields. Since in most applications the number of fields in a relation is much less than 500, we can say that Algorithm \textit{minimumCover} performs well in practice. Third, the performance of Algorithm \textit{minimumCover} is much better than Algorithm \textit{naive}. For example, when the number of fields is incremented by 5, the execution time of \textit{minimumCover} at most doubles, while for \textit{naive} it grows almost two-hundred-fold.

We next consider checking XML key propagation. An algorithm for doing so, Algorithm \textit{propagation}, was presented in Section 4. An alternative algorithm can also be developed by means of Algorithm \textit{minimumCover} as follows: Given a transformation $\sigma$, a set of keys $\Sigma$, and an FD $\phi = Y \rightarrow I$, the algorithm first invokes \textit{minimumCover}($\Sigma, \sigma$) to compute a minimum cover $F_m$ of all the FDs propagated; it then checks whether or not $F_m$ implies $\phi$ using relational FD implication, and whether all the fields in $Y$ are guaranteed to have a non-null value when $I$ is not null. It returns true iff these conditions are met. In what follows, we refer to this generalized algorithm as \textit{GminimumCover} since the performance is roughly comparable to the original algorithm.

Our second experiment serves two purposes: to compare the effectiveness of these two algorithms for checking key propagation, and to study the impact of the depth of table-tree ($\text{depth}$) on the performance of Algorithms \textit{propagation} and \textit{GminimumCover}. Fig. 7(b) depicts the execution time of these algorithms for field $= 15$ and keys $= 10$ with depth varying from 2 to 15. (These parameters were chosen based on the average tree depth found in real XML data [9].) The results in Fig. 7(b) reveal the following. First, Algorithm \textit{propagation} works well in practice: it takes merely 0.05 second even when the table tree is as deep as 15. Second, these algorithms are rather insensitive to the change to depth. Third, \textit{propagation} is much faster then \textit{GminimumCover} for checking key propagation, as expected. Although the actual execution times of the algorithms are quite different, the ratios of increase when the depth of the table-tree grows are similar. This is because in both algorithms the depth determines how many times Algorithm \textit{implication} is invoked, and because the complexity of Algorithm \textit{implication} is a function of the size of the XML keys, which grows when the depth of the table tree gets larger.

Our third experiment demonstrates how the number of XML keys ($\text{keys}$) influences the performance of Algorithms \textit{propagation} and \textit{GminimumCover} when checking key propagation. The results (Fig. 7(c)) show that increasing the number of keys has a bigger impact on Algorithm \textit{GminimumCover} than on \textit{propagation},

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7.png}
\caption{Experimental results}
\end{figure}
in which the growth of the execution time is almost linear. In fact, additional experiments tell us that for depth = 10 and keys = 50, Algorithm $\text{GminimumCover}$ runs in under 2 minutes for 200 fields, but when increasing the number of keys to 100, its execution time is over 4 minutes for relations with 150 fields. In contrast, Algorithm $\text{propagation}$ runs in both settings in less than 5 seconds. In addition, for 1000 fields, which is the maximum number of fields allowed by Oracle [22], the execution time of $\text{propagation}$ is 85 seconds on average for 50 keys, and 142 seconds for 100 keys.

A closer look at Algorithm $\text{propagation}$ reveals that the constant ratio of increase is based on the time needed for executing calls to Algorithm implication. That is, if the depth of the table-tree is fixed, the number of calls is roughly the same for the whole experiment; the increase in running time is based on the the performance of Algorithm implication, which depends on the size of the XML keys. The performance of implication also has an impact on the Algorithm $\text{GminimumCover}$. However, the number of keys has a bigger influence in this algorithm because for each node in the table-tree all the keys are analyzed. Also, by increasing the number of XML keys, the number of FDs in the resulting set is likely to grow, increasing the execution time for eliminating redundant FDs by calling $\text{minimize}$.

7 Conclusion

We have proposed a framework for refining the relational design of XML storage based on XML key propagation. For this purpose we have developed algorithms for checking whether a functional dependency is propagated from XML keys, and for finding a minimum cover for all functional dependencies propagated from XML keys, along with complexity results in connection with XML constraint propagation. Our experimental results show that these algorithms are efficient and effective in practice. These algorithms can be generalized and incorporated into relational storage techniques published in the literature (e.g. [25, 26, 22]). Our results are also useful in optimizing queries and in understanding XML to XML transformations.

Topics for future work include studying the propagation of other forms of integrity constraints, and re-investigating constraint propagation in the presence of types (e.g., XML Schema).

References