Reasoning about XML Constraints

Sebastian Link

School of Information Management
Centre for Logic, Language and Computation
Victoria University of Wellington
New Zealand

(joint work with Sven Hartmann)

School of Informatics
Clausthal University of Technology
Germany

This research is supported by the Marsden Fund Council from Government funding, administered by the Royal Society of New Zealand.

Osaka, 16 June 2009
Motivation

- XML: de-facto standard for Web data exchange and integration
- high degree of syntactic flexibility, low degree of semantic capabilities
- challenge for computer scientists: provide full-fledged tools that can store, manage and process XML data in its native format


- integrity constraints enhance semantic capabilities:
  - find natural classes of constraints that can be maintained efficiently
  - balance trade-off between expressiveness and efficiency
An XML fragment

```xml
<db>
  <year calendar="2007">
    <semester no=1>
      <course name="maths">
        <teacher><title>Principal</title><name>Skinner</name></teacher>
        <student sid="007">
          <first>Bart</first><last>Simpson</last><grade>A+</grade></student>
      </course>
    </semester>
  </year>
  <year calendar="2007">
    <semester no=2>
      <course name="PE">
        <teacher><title>Principal</title><name>Skinner</name></teacher>
        <student sid="007">
          <first>Bart</first><last>Simpson</last><grade>A+</grade></student>
        <student sid="247">
          <first>Lisa</first><last>Simpson</last><grade>B-</grade></student>
      </course>
    </semester>
  </year>
</db>
```
Its corresponding XML Tree
Some Constraints

• relatively to each year: a semester is determined by its number

• each year node has either no semester nodes or precisely two

• relatively to each course: a student is determined by its sid

• each student deciding to study in a semester enrolls into at least two and at most 4 courses in that semester

• no student enrolls more than three times into the same course

• no course is offered more than once a year
Scenario 1: Data Cleaning

  - dirty date costs (US) businesses billions of dollars annually
  - data cleaning contributes to 30-80 percent of development time in datawarehouse projects
- data cleaning tools for removing inconsistencies and errors from data
- provide means that prevent dirty date from entering the database
- *constraints* restrict databases to those considered meaningful
  - above business rules capture properties of XML data
  - only allow those XML documents conforming to all business rules
Scenario 2: Predicting Query Costs

- *Bart Simpson* wishes to query XML source for results in 2007 using his cell-phone but only when costs reasonable
  - service provider: only retrieve data if service paid for

- consider the following XQuery query:

  ```
  for $s in doc("enrol.xml")/year[@calendar="2007"]//course/student
  where $s/sid="007"
  return ⟨grade⟩{$s/grade}⟨/grade⟩
  ```

- XML DBMSs capable of reasoning about cardinality constraints foresees maximal number of eight answers (without processing any data!)
- approximate costs returned to *Bart Simpson* who decides accordingly
- service provider minimizes costs for unpaid services

Osaka, 16 June 2009
Scenario 3: Query Optimisation

- teachers teach at most 3 courses per year
- students take up to 4 courses per semester, with two semester per year
- XML query optimiser should transform the XQuery query

```xquery
for $c in doc("enrol.xml")/year[@calendar="2007"]/semester/course
where $c/student/sid="247" and $c/teacher="Principal Skinner"
return ⟨course⟩{$c/name}⟨/course⟩
```

into

```xquery
for $c in doc("enrol.xml")/year[@calendar="2007"]/semester/course
where $c/teacher="Principal Skinner" and $c/student/sid="247"
return ⟨course⟩{$c/name}⟨/course⟩
```

- based on smaller selectivity

Osaka, 16 June 2009
Scenario 4: Predicting the No of Updates

- databases often subject to updates

- following XQuery updates last name of Lisa Simpson to Milhouse in all semester 2 courses of 2007:

  ```xml
  for $s in doc("enrol.xml")/year[@calendar="2007"]/semester[@no="2"]//student[@sid="247"]
  return do replace value of $s/last with "Milhouse"
  ```

- maximal number of four updates (decide implication efficiently)
Scenario 5: An encrypted XML fragment

```
<db>
  <year calendar="2007">
    <semester no="1">
      <course name="maths">
        <teacher>
          <title>Principal</title>
          <name>Skinner</name>
        </teacher>
        <student sid="007">
          <first>Bart</first>
          <last>Simpson</last>
        </student>
        <grade>
          <xenc:EncryptedData>
            <xenc:CipherData>
              <xenc:CipherValue>AbC234ndZ...</xenc:CipherValue>
            </xenc:CipherData>
          </xenc:EncryptedData>
        </grade>
      </course>
    </semester>
  </year>
</db>
```
Scenario 5: Predicting the No of Encryptions

- Web data often encrypted (grades of students)

- following XQuery query retrieves grades of Bart Simpson in 2007:

```xml
for $s in doc("enrol.xml")/year[@calendar="2007"]//course/student
where $s/sid="007"
return ⟨grade⟩{$s/grade}⟨/grade⟩
```

- efficient reasoning about constraints enables DBMS to infer that at most eight decryptions necessary
Scenario 6: XML Transformations

- **year**-nodes subsume between two and eight **course**-nodes containing **student/sid**-subnodes with same value

- Querying original XML tree for course info based on specific year and specific sid unnatural

- **sid**-updates difficult

- Create XML view
Scenario 6 continued: Query Rewriting

▶ rewrite

```
for $s$ in doc(“enrol.xml”)//course/student
where $s$/sid=”007”
return ⟨grade⟩{$s$/grade}⟨/grade⟩
```

▶ as

```
for $c$ in doc(“view.xml”)//student[@sid=”007”]/course
return ⟨grade⟩{$c$/grade}⟨/grade⟩
```

▶ early selection of student-elements based on their sid

▶ even more efficient in case @sid-values are encrypted

▶ easy updates of sid-values
Scenario 7: Consistent Query Answering

- approach to querying inconsistent databases without repairing them
- retrieves only certain answers, those present in all repairs

Bart, Lisa, Maggie: //course[@name = 'physics']/student/first

however:
- there mustn’t be different student-nodes with same @sid in any course
CQA: One Solution

repair may refer to removal of any offending student-node:

CQA returns only Maggie as a certain answer:

//course[@name = 'physics']/student/first
repair may refer to replacement of any offending \(@\text{sid}\)-value by a new \(@\text{sid}\)-value not present in the tree; infinitely many repairs, e.g.:

\[
\begin{align*}
\text{db} & \rightarrow \text{course} \\
& \quad \rightarrow \text{student} \quad \text{@sid} = \text{"007"} \\
& \quad \quad \rightarrow \text{student} \quad \text{@sid} = \text{"247"} \\
& \quad \quad \quad \rightarrow \text{student} \quad \text{@sid} = \text{"107"} \\
& \quad \rightarrow \text{teacher} \quad \text{@name} \rightarrow \text{"physics"} \\
& \quad \quad \rightarrow \text{teacher} \quad \text{@name} \rightarrow \text{"physics"} \\
& \quad \quad \quad \rightarrow \text{teacher} \quad \text{@name} \rightarrow \text{"physics"} \\
\end{align*}
\]

\(\text{Bart, Lisa and Maggie}\) present in all of them:

\[
\text{course}[@\text{name} = 'physics']/\text{student}/\text{first}
\]
Practical Benefits of Cardinality Constraints

- represent characteristics that XML data naturally exhibits
- restrict permitted XML database to those considered meaningful
- advance consistency of XML databases
- identify dirty data
- can be specified independently from schema specifications
- enable logical query optimization
- enable physical query optimization
- enable us to infer bounds on number of query answers (updates, en/decryptions) without any data processing
- unlock “good” XML database design, i.e., absence of data redundancies, processing difficulties and inefficient query processing
- new approach to dealing with inconsistent data (CQA)
**Fundamental Challenges and Outline**

- Introduce cardinality constraints based on XML tree model
  - Core CS problem: expressiveness vs. userfriendliness

- Investigate decision problems to unlock XML applications
  - Satisfiability problem and finite implication problem

- Identify sources of intractability: coNP-hardness results
  - Core CS problem: expressiveness vs. tractability

- Identify large tractable subclass
  - Finite axiomatisation
    - Core CS problem: syntax vs. semantics
  - Characterise implication in terms of shortest path
  - Devise algorithm to efficiently decide implication
    - Core CS problem: problems vs. algorithmic solutions
XML Trees

• XML documents can be illustrated as rooted trees
  • an XML tree is a 6-tuple \( T = (V, \text{lab}, \text{ele}, \text{att}, \text{val}, r) \)
  • connected digraphs, no cycles, unique path from \( r \) to each node
• we distinguish attribute nodes, element nodes, and text nodes
  • attribute nodes \( v \) have attribute name \( \text{lab}(v) \in A \)
  • element nodes \( v \) have element name \( \text{lab}(v) \in E \)
  • text nodes \( v \) have fixed name \( \text{lab}(v) = S \), root \( r \) is element node
• element nodes \( v \) have children
  • a set \( \text{att}(v) \subseteq V \) of attribute nodes
  • a list \( \text{ele}(v) \subseteq V \) of element and text nodes
  • each child \( w \) induces an edge \((v, w)\) in the XML tree
• attribute and text nodes have no children (they are leaves), but each attribute or text node \( v \) carries a value \( \text{val}(v) \in \text{String} \)
An Example of an XML Tree
Value Equality

• each node \( v \) is the root of a unique (maximal) sub-tree \( T(v) \) of \( T \)
• two nodes \( v_1 \) and \( v_2 \) are value-equal iff their induced sub-trees \( T(v_1) \) and \( T(v_2) \) are isomorphic
  • we denote this by \( v_1 =_v v_2 \)
• roughly speaking, we have:
  • \( \text{lab}(v_1) = \text{lab}(v_2) \)
  • if \( v_1, v_2 \) are attribute or text nodes, then \( \text{val}(v_1) = \text{val}(v_2) \)
  • if \( v_1, v_2 \) are element nodes, then they possess
    • sets of value-equal attribute children
    • lists of value-equal element and text children
• we define the value-intersection of two node sets \( V_1 \) and \( V_2 \)
  • \( V_1 \cap_v V_2 = \{(v_1, v_2) \mid v_1 \in V_1, v_2 \in V_2, v_1 =_v v_2\} \)
Example: Value-Equality

- student-nodes: $v_1 = v_3$, $v_2 \neq v_1$ and $v_2 \neq v_3$
- no two distinct course-nodes are value-equal
- for $V_1 = \{v_1, v_2\}$ and $V_2 = \{v_3, v_4\}$ we have $V_1 \cap_v V_2 = \{(v_1, v_3)\}$
Path languages for selecting nodes

- we need a suitable path language to select nodes in the XML tree
  - expressive enough to allow reasonable navigation
  - simple enough to allow efficient reasoning
- we use the path languages $PL$ and $PL_s$ here

<table>
<thead>
<tr>
<th>Path Language</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>$PL_s$</td>
<td>$P ::= \epsilon \mid P.\ell$</td>
</tr>
<tr>
<td>$PL$</td>
<td>$Q ::= \epsilon \mid Q.\ell \mid Q._*$</td>
</tr>
</tbody>
</table>

- $\ell \in E \cup A \cup \{S\}$ denotes any node name
- $*_*$ denotes a wildcard, $\epsilon$ denotes the empty path
- path expressions in $PL_s$ are simple path expressions
- each expression $Q$ in $PL$ represents set of simple path expressions
  - wildcards $*_*$ in $Q$ replacable by arbitrary simple path expressions
  - write $P \in Q$ for each simple $P$ that can be generated from $Q$
Example: Path Expressions

- **simple path expressions:**
  - year.semester
  - semester.student
  - year.semester.course.student.grade

- **path expressions:**
  - _*.year.semester
  - _*.course._*.grade

- **simple path expressions in the regular language of path expressions:**
  - year.semester ∈ _*.year.semester
  - year.semester.course.student.grade ∈ _*.course._*.grade
Selecting nodes in an XML tree

- path expression $Q$ said to be valid iff no attribute name and no $S$ occur in position other than terminal
  - attribute and text nodes are always leaves of the XML tree

- $u[Q]$: those nodes in tree $T$ reachable from node $u$ via simple $P \in Q$

- $Q_1 \subseteq Q_2$ iff $u[Q_1] \subseteq u[Q_2]$ for all trees $T$ and all nodes $u$ of $T$

- Path containment for $PL$:
  decide for arbitrary $Q_1, Q_2 \in PL$ whether $Q_1 \subseteq Q_2$ holds

- Path containment for $PL$ can be decided in quadratic time
Example: Navigation between Nodes

\[ u[\_\_\_\_\_\_\_\_.course\_\_\_\_\_\_\_.first] = \{v_1, v_2, v_3, v_4\} \]

\[ \text{semester.course}\_\_\_\_\_\_\_.first \subseteq \_\_\_\_\_\_.course\_\_\_\_\_\_.first \]

\[ \text{not } \_\_\_\_\_\_\_.course\_\_\_\_\_\_.first \subseteq \_\_\_\_\_\_.course\_\_\_\_\_\_.first \]
The Path Containment Problem

• some results on the complexity of the path containment problem:

PTIME $\{XP(/, //), XP(/, //, []), XP(/, [], *), XP(/, //, [])\}$

$\{XP(/, //, *, DTD)\}$

coNP-complete $\{XP(/, //, [], *), XP(/, //, [], *, |), XP(/, [], DTD)\}$

coNP-hard $\{XP(/, |), XP(/, //, |), XP(/, [], DTD)\}$

EXPTIME-complete $\{XP(/, //, [], *, |, DTD)\}$

EXPTIME-hard $\{XP(/, //, [], *, |, DTD), XP(/, //, |, DTD)\}$


• F. Neven, T. Schwentick. XPath containment in the presence of disjunction, DTDs, and variables. ICDT, PP. 315–329, 2003.
Cardinality constraints for XML

- A cardinality constraint $\varphi$ has the form

$$\text{card}(Q, (Q', \{Q_1, \ldots, Q_k\})) = (\text{min}, \text{max})$$

- $Q \in PL$ is the context path
- $Q' \in PL$ is the target path
- $Q_i \in PL$ is a key path (for $i = 1, \ldots, k$)
- $Q.Q'$ valid ($k = 0$) or $Q.Q'.Q_i$ valid path (for $i = 1, \ldots, k$)

- Two nodes $v_1, v_2$ are $\{Q_1, \ldots, Q_k\}$-confusable iff

$$v_1[Q_i] \cap v_2[Q_i] \neq \emptyset \text{ for all } i = 1, \ldots, k$$

- The XML tree $T$ satisfies $\varphi$ iff for all nodes $q \in r[Q]$ and all nodes $q' \in q[Q']$ there are no less than $\text{min}$ and no more than $\text{max}$ nodes in $q[Q']$ that are $\{Q_1, \ldots, Q_k\}$-confusable with $q'$

Osaka, 16 June 2009
Example 1

- In every semester every student who studies must enrol in 2 to 4 courses.

\[ T \text{ does not satisfy } card(\_*.semester,(course,\{\_*.sid\}))=(2,4) \]
Example 2

- $T$ satisfies $\text{card}(\text{year},(\_\ast . \text{course}, \{ \text{teacher} \})) = (3, 6)$
- $T$ violates $\text{card}(\_\ast . \text{semester}, (\text{course}, \{ \text{teacher} \})) = (2, 2)$
Example 3

- **XML keys covered by cardinality constraints:** $min = max = 1$

  
  - S. Hartmann, S. Link. Efficient Reasoning about a Robust XML Key Fragment. TODS 34(2:10), 2009

- $T$ satisfies $card(_*\text{.course},(\text{student},\{\text{sid}\}))=(1,1)$
Towards a Visual Specification Language

- allow users to specify their requirements graphically
- allow designers to present her specifications to users visually
- teachers that teach in a semester, teach between 2 and 4 different courses in that semester
- formally: \( \text{card}((\_*.semester, (\_*.course, \{\_*.title, teacher.name\}))) = (2,4) \)
- visually as a pattern:

```plaintext
semester C
/     \ 2 \= course \<= 4
/      \         \           
/       \          \           
title   teacher
        /       \  
        /         \   
        name
```
Decision Problems

◮ (finite) satisfiability problem:
   ➡ Given any finite $\Sigma$, is there (finite) $T$ satisfying all $\sigma \in \Sigma$?

◮ $\Sigma$ (finitely) implies $\varphi \in \Sigma^*$ (semantic closure):
   ➡ each (finite) XML tree that satisfies $\Sigma$ also satisfies $\varphi$

◮ (finite) implication problem:
   ➡ Given any finite $\Sigma \cup \{ \varphi \}$: does $\Sigma$ (finitely) imply $\varphi$?

◮ efficient solution to decision problems unlocks XML applications

◮ let $\Sigma^+_R$ be the syntactic closure of $\Sigma$
   • with respect to a fixed system $R$ of inference rules
   • $\varphi \in \Sigma^+_R$ can be inferred from $\Sigma$ using the rules in $R$

◮ $R$ is sound iff $\Sigma^+_R \subseteq \Sigma^*$, and complete iff $\Sigma^* \subseteq \Sigma^+_R$
A Computer Scientist’s Pilgrimage

Mark Musa, in his translation of Dante’s *Divine Comedy*:

Dante the Pilgrim, in order to arrive at the Divine Light, will come to an understanding of sin on his journey through Hell and will see the penance imposed on repentant sinners on the Mount of Purgatory.

- our journey through Hell (it’s not that bad!):
  - our sin: too much expressiveness
  - understanding: identifying intractabilities

- our Mount Purgatory:
  - restriction to tractable cases
  - do penance by studying properties

- our “Divine Light”:
  - finite axiomatization
  - efficient algorithms
Sources of Intractability: Lower and Upper Bounds

\[ C_1 = \{ \text{card}(\varepsilon, (P', \{P_1, \ldots, P_k\})) = (\min, \max) \mid k \geq 1 \} \]

Theorem:
The finite implication problem for the class \( C_1 \) of all simple absolute cardinality constraints with a non-empty set of key paths is coNP-hard.

suggests that computational intractability may result from the specification of both lower and upper bounds
Sources of Intractability: Empty sets of key paths

$\mathcal{C}_2 = \{ \text{card}(\epsilon, (P', \{P_1, \ldots, P_k\})) = (1, \max) \mid k \geq 0\}$

Theorem:

The finite implication problem for the class $\mathcal{C}_2$ of all simple absolute cardinality constraints where the lower bound is fixed to 1 is $\text{coNP}$-hard.

suggests that computational intractability may result from the permission of empty sets of key paths
Sources of Intractability: General key path expressions

\[ C_3 = \{ \text{card} (\epsilon, (Q', \{Q_1, \ldots, Q_k\})) = (1, \text{max}) \mid k \geq 1 \} \]

Theorem:
The finite implication problem for the class \( C_3 \) of all absolute cardinality constraints that have a non-empty set of key paths and where the lower bound is fixed to 1 is coNP-hard.

suggests that computational intractability may result from permission to have arbitrary path expressions in both target- and key paths
Numerical keys for XML

- A numerical key $\varphi$ has the form
  
  $$\text{card}(Q, (Q', \{P_1, \ldots, P_k\})) \leq \max$$

  - $Q \in PL$ is the context path
  - $Q' \in PL$ is the target path
  - $P_i \in PL$ is a key path (for $i = 1, \ldots, k$ and $k \geq 1$)
  - $Q.Q'.P_i$ valid path (for $i = 1, \ldots, k$)

- XML tree $T$ satisfies $\text{card}(Q, (Q', \{P_1, \ldots, P_k\})) \leq \max$ iff $T$ satisfies $\text{card}(Q, (Q', \{P_1, \ldots, P_k\})) = (1, \max)$

- Theorem:
  Every finite set of numerical keys is finitely satisfiable.

- Theorem:
  Implication and finite implication coincide for numerical keys.
### A finite axiomatisation for numerical keys

<table>
<thead>
<tr>
<th>Expression</th>
<th>Inequality</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{card}(Q, (Q', S)) \leq \infty$</td>
<td>(infinity)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', S)) \leq \max$</td>
<td></td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', S \cup {P})) \leq \max$</td>
<td>(superkey)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', S)) \leq \max$</td>
<td></td>
</tr>
<tr>
<td>$\text{card}(Q''', (Q', S)) \leq \max$</td>
<td>$Q'' \subseteq Q$</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', S \cup {\epsilon, P})) \leq \max$</td>
<td>(context-path-containment)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', S \cup {\epsilon, P.P'})) \leq \max$</td>
<td></td>
</tr>
<tr>
<td>$\text{card}(Q, (Q'.P, {P'})) \leq \max$</td>
<td>(prefix-epsilon)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', {P.P'})) \leq \max$</td>
<td>(subnodes-epsilon)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', {P.P'})) \leq \max$</td>
<td>(subnodes)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', {\epsilon, P.P'})) \leq \max$</td>
<td>(context target)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', S)) \leq \max$</td>
<td>$Q''. \subseteq Q'$</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q', S)) \leq \max$</td>
<td>(target-path-containment)</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q'.P, {P'})) \leq \max$</td>
<td></td>
</tr>
<tr>
<td>$\text{card}(Q, (Q'.P, {P_1, \ldots, P_k})) \leq \max$</td>
<td>$Q'. \subseteq Q'$</td>
</tr>
<tr>
<td>$\text{card}(Q, (Q'.P, {P_1, \ldots, P_k})) \leq \max$</td>
<td>(multiplication)</td>
</tr>
</tbody>
</table>

Osaka, 16 June 2009
Example

- sports league competition:
  - per season same home and away team both play in at most 3 months:
    \[ \sigma_1 = \text{card}(\_\_*.\text{season}, (\text{month}, \{\text{match.home,match.away}\})) \leq 3 \]
  - per month same home team plays same away team at most twice:
    \[ \sigma_2 = \text{card}(\_\_*.\text{season.month}, (\text{match}, \{\text{home,away}\})) \leq 2 \]
  - same home team plays same away team at most 5 times per season:
    \[ \varphi = \text{card}(\_\_*.\text{season}, (\text{month.match}, \{\text{home,away}\})) \leq 5 \]

- \( \varphi \) cannot be inferred from \( \Sigma = \{\sigma_1, \sigma_2\} \)

  How can we show that \( \varphi \) is not implied by \( \Sigma \) ?

  We construct a counter-example
Shortest Paths vs Derivability

- we construct a cardinality graph $G_{\Sigma, \varphi}$ from $\Sigma$ and $\varphi$
- if $d(q_\varphi, q'_\varphi) \leq \max_\varphi$ in $G_{\Sigma, \varphi}$, then $\varphi \in \Sigma^+$
- in other words:
  - if $\varphi$ cannot be inferred from $\Sigma$, then $d(q_\varphi, q'_\varphi) \geq \max_\varphi + 1$ in $G_{\Sigma, \varphi}$

- Example:
  - $\varphi = \text{card}(\_*.season, (\text{month.match}, \{\text{home,away}\})) \leq 5$
    - not derivable from
  - $\sigma_1 = \text{card}(\_*.season, (\text{month}, \{\text{match.home,match.away}\})) \leq 3$
  - $\sigma_2 = \text{card}(\_*.season.month, (\text{match}, \{\text{home,away}\})) \leq 2$
Completeness

\( T \) satisfies

- \( \sigma_1 = \text{card}(\_*.\text{season}, \text{month}, \{\text{match.home, match.away}\}) \leq 3 \)
- \( \sigma_2 = \text{card}(\_*.\text{season.month}, \text{match}, \{\text{home, away}\}) \leq 2 \)

\( T \) violates

- \( \varphi = \text{card}(\_*.\text{season}, \text{month.match}, \{\text{home, away}\}) \leq 5 \)
An Algorithm for the Implication Problem

- strong method: $\varphi \in \Sigma^*$ iff $d(q_\varphi, q'_\varphi) \leq \max_\varphi$ in $G_{\Sigma, \varphi}$
- this is how the algorithm works:
  - **Input:** $\Sigma$, $\varphi$
  - **Output:** YES if $\varphi \in \Sigma^*$; No otherwise
  - **Method:**
    1. construct the cardinality graph $G_{\Sigma, \varphi}$
    2. if $d(q_\varphi, q'_\varphi) \leq \max_\varphi$ then return(YES) else return(NO)
- shortest path: apply Dijkstra’s algorithm using start node $q_\varphi$ ($O(|\varphi|^2)$)
- cardinality graph constructable in time $O(||\Sigma|| \cdot |\varphi|^3)$
  - we use the evaluation algorithm for Core XPath queries
- in fact: cardinality graph constructable in time $O(||\Sigma|| \cdot |\varphi|)$
A finite axiomatisation for XML keys

• key $|=$ $\mathcal{T} (Q, (Q', \{P_1, \ldots, P_k\}))$ iff $|=$ $\mathcal{T} \text{card}(Q, (Q', \{P_1, \ldots, P_k\})) \leq 1$

<table>
<thead>
<tr>
<th>(Q, (ε, S'))</th>
<th>(Q, (Q', S \cup {ε, P}))</th>
<th>(Q, (Q', S))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q, (Q', {P_1, \ldots, P_k}))</td>
<td>(Q, (Q', S \cup {ε, P.P'}))</td>
<td>(Q, (Q', S \cup {P}))</td>
</tr>
<tr>
<td>(Q, (Q', {P.P'}))</td>
<td>(Q, (Q', S)) \subseteq Q</td>
<td>(Q, (Q'', S))</td>
</tr>
<tr>
<td>(Q, (Q.P, {P'}))</td>
<td>(Q'', (Q', S)) \subseteq Q</td>
<td>(Q, (Q'', S)) \subseteq Q'</td>
</tr>
<tr>
<td>(Q, (Q'.Q'', S'))</td>
<td>(Q, (Q'.P, {ε, P'}))</td>
<td>(Q, (Q', {P.P_1, \ldots, P.P_k}))</td>
</tr>
<tr>
<td>(Q.Q', (Q'', S'))</td>
<td>(Q, (Q', {ε, P.P'}))</td>
<td>(Q.Q', (P, {P_1, \ldots, P_k}))</td>
</tr>
<tr>
<td>(Q, (Q'.P, {P_1, \ldots, P_k}))</td>
<td>(Q, (Q'.P, {P_1, \ldots, P_k}))</td>
<td>(Q, (Q'.P, {P_1, \ldots, P_k}))</td>
</tr>
<tr>
<td>(Q, (Q'.Q'', S'))</td>
<td>(Q, (Q', {ε, P.P'}))</td>
<td>(Q, (Q'.P, {P_1, \ldots, P_k}))</td>
</tr>
</tbody>
</table>
Reachability instead of Shortest Paths

• all witness edges have weight 1,
• discard with labels to obtain reachability graph
• this is how the algorithm works:
  • **Input:** $\Sigma, \varphi$
  • **Output:** YES if $\varphi \in \Sigma^*$; No otherwise
  • **Method:**
    (1) construct the reachability graph $G_{\Sigma,\varphi}$
    (2) if $q'_{\varphi}$ reachable from $q_{\varphi}$ then return(YES) else return(NO)
• reachability test: depth-first search of $G_{\Sigma,\varphi}$ with root $q_{\varphi}$ in $O(|\varphi|^2)$
• reachability graph constructable in time $O(||\Sigma|| \cdot |\varphi|)$
  • we use the evaluation algorithm for Core XPath queries
Structural XML Keys

- keys: target nodes uniquely identified by values reachable via key paths
- structural keys identify target nodes independently of any data values
- \((\ast . course, (name, \emptyset))\), but not \((\ast . semester, (\ast . name, \emptyset))\)
- allow structural keys: XML key is expression \((Q, (Q', S))\) where
  - \(Q, Q'\) are \(PL\) expressions,
  - \(S\) is a finite set of \(PL_s\) expressions such that
  - \(S \neq \emptyset\): \(Q.Q'.P\) are valid \(PL\) expressions for all \(P \in S\) or
  - \(S = \emptyset\): \(Q.Q'\) is a valid \(PL\) expression.
- satisifiability of XML keys defined as before
### Axiomatising XML keys and Structural keys

<table>
<thead>
<tr>
<th>Expression</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(Q, (Q', S \cup {P, P.P'}))$</td>
<td>$(Q.Q', P, \emptyset)$</td>
</tr>
<tr>
<td>$(Q, (\varepsilon, S))$</td>
<td>(epsilon)</td>
</tr>
<tr>
<td>$(Q, (Q'.P, {P}))$</td>
<td>(subnodes)</td>
</tr>
<tr>
<td>$(Q, (Q', {P.P}))$</td>
<td>(context-path-containment)</td>
</tr>
<tr>
<td>$(Q, (Q', S), (Q''))$</td>
<td>$(Q, (Q', {\varepsilon, P})$</td>
</tr>
<tr>
<td>$(Q, Q', (Q''.S))$</td>
<td>(subnodes-epsilon)</td>
</tr>
<tr>
<td>$(Q, (Q'.Q'', S))$</td>
<td>$(Q, (Q'.Q', \emptyset))$</td>
</tr>
<tr>
<td>$(Q, (Q'.Q', S))$</td>
<td>(context-to-target)</td>
</tr>
<tr>
<td>$(Q, (Q', (Q'''.S)))$</td>
<td>$(Q, (\varepsilon, (Q'.Q', \emptyset)))$</td>
</tr>
<tr>
<td>$(Q, (\varepsilon, (Q'.Q', \emptyset)))$</td>
<td>(border)</td>
</tr>
<tr>
<td>$(Q, (\varepsilon, {P.P'.P''}))$</td>
<td>(neighbour)</td>
</tr>
</tbody>
</table>

Osaka, 16 June 2009
Deciding Implication

- method may fail: reachability graph can violate structural key
- pre-processing: merge distinct key paths to satisfy structural keys

**Input:** \( \Sigma, \varphi \)

**Output:** YES if \( \varphi \in \Sigma^* \); No otherwise

**Method:**
1. Construct \( \overline{T}_{\Sigma,\varphi} \) from \( \Sigma \) and \( \varphi \);
2. IF \( \exists (Q_\sigma, (Q'_\sigma, \emptyset)) \in \Sigma. \exists w_\sigma \in [Q_\sigma] \) in \( \overline{T}_{\Sigma,\varphi} \) and \( \exists w'_\sigma, w''_\sigma \in w_\sigma[Q'_\sigma] \) in \( \overline{T}_{\Sigma,\varphi} \) with \( w'_\sigma \neq w''_\sigma \)
3. THEN RETURN(yes)
4. ELSE Construct \( G_{\Sigma,\varphi} \) from \( \Sigma \) and \( \varphi \);
5. IF \( q_{\varphi} \) is reachable from \( q'_\varphi \) in \( G_{\Sigma,\varphi} \) THEN RETURN(yes)
6. ELSE RETURN(no).

- implication still decidable in time \( \mathcal{O}(|\varphi| \times (||\Sigma|| + |\varphi|)) \)
Findings for Cardinality Constraints

- form natural class of XML constraints that generalise XML keys
- useful for many XML applications:
  - XQuery, XPath, XML Encryption, XQuery update facility
  - cost estimation, query optimisation, query rewriting
- reasoning about cardinality constraints intractable in general
  - specification of both lower and upper bounds
  - empty set of key paths
  - general key path expressions
- identified numerical keys as large tractable subclass
  - always satisfiable
  - implication and finite implication coincide
  - established finite axiomatisation
  - characterised implication in terms of shortest paths
  - efficient solutions to decision problem unlocks applications
Future Work?
Future Work

▶ push expressiveness while maintaining tractability
▶ develop visual specification languages for cardinality constraints
▶ validity: is an XML document valid wrt a CC (tree automata)
▶ discover all CCs satisfied by a given XML document (data mining)
▶ interaction with schema specifications such as DTDs and XSDs
▶ interaction with other types of constraints
▶ automatic support for query optimization
▶ physical database tuning: indexes for efficient query performance
▶ investigate consistent query answering
▶ database design: transform poorly-designed databases into well-designed ones that avoid data redundandancies, processing difficulties and inefficient query processing
Questions?