ProTDB: Probabilistic Data in XML

Andrew Nierman and H. V. Jagadish

{andrewdn, jag}@eecs.umich.edu.

Electrical Engineering and Computer Science, University of Michigan
Motivation

- Databases are needed for efficient large-scale data management.
- Applications that use databases often involve uncertain data.
  - Some domains are inherently contradictory, uncertain, and incomplete.
  - Completely accurate information may be either too hard or too expensive to obtain.
  - Automated extraction of information is an error-prone process.
- XML is a flexible and widely used data model.
An Introductory Example

A sample domain without uncertainty

countries
country
countryName
government
countryName

United States

independence
country

United Kingdom

chiefOfState
country

George W. Bush

President

day

4

month

7

year

1776

title

name

age

spouse

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

President

Presiden
An Introductory Example

- Our source XML data contains probabilities
- A user queries the data normally
- A probability will be attached to each query result
- Node probabilities in the output are updated
An Introductory Example

Our source XML data contains probabilities

```xml
<countries>
  <country Prob='0.9'>
    <countryName>United States</countryName>
    <government>
      <independenceDay Prob='0.85'>07/04/1776</independenceDay>
      <chiefOfState>
        <Dist type="mutually-exclusive">
          <Val Prob='0.5'>
            <name>
              <Dist>
                <Val Prob='0.4'>George W. Bush</Val>
                <Val Prob='0.7'>George Bush</Val>
              </Dist>
            </name>
            <age>
              <Dist type="mutually-exclusive">
                <Val Prob='0.2'>54</Val>
                <Val Prob='0.3'>55</Val>
                <Val Prob='0.1'>56</Val>
              </Dist>
            </age>
          </Val>
        </Dist>
      </chiefOfState>
    </government>
  </country>
  ...
  <countryProb='0.2'>
    <title Prob='0.65'>President</title>
    <name>Bill Clinton</name>
    <age Prob='0.3'>55</age>
  </Val>
  ...
  <countryProb='0.2'>
    <countryName>Uruguay</countryName>
    ...
  </country>
</countries>
```
An Introductory Example

- **country**: United States
- **government**: independenceDay 07-04-1776
- **chiefOfState**: George W. Bush
- **name**: George W. Bush
- **age**: 55

- **country**: Uruguay
- **chiefOfState**: Bill Clinton
- **name**: Bill Clinton
- **age**: 55

**Countries**

- **countries**
  - **country**: United States
    - **countryName**: United States
    - **independenceDay**: 07-04-1776
    - **chiefOfState**: George W. Bush
    - **name**: George W. Bush
    - **age**: 55
  - **country**: Uruguay
    - **chiefOfState**: Bill Clinton
    - **name**: Bill Clinton
    - **age**: 55
An Introductory Example

“Return all chiefs of state having the name George Bush and an age of 55”:

F : $1.tag = chiefOfState
^ $2.tag = name
^ $2.content = George Bush
^ $3.tag = age
^ $3.content = 55

The query tree is transformed to account for the probabilistic constructs in the source data.
An Introductory Example

- **countries**
  - **United States**
  - **Uruguay**
    - `countryName`
      - `United States`
      - `Uruguay`
    - `independenceDay` [Prob=0.85]
      - `07-04-1776`
    - `chiefOfState`
      - `George W. Bush`
      - `Bill Clinton`
    - `name`
      - `Val [Prob=0.5]` George W. Bush
    - `age`
      - `Dist [type=mutually-exclusive]`
        - `Val [Prob=0.4]` George W. Bush
        - `Val [Prob=0.2]` George Bush
        - `Val [Prob=0.7]` George Bush
        - `Val [Prob=0.35]` Bill Clinton
        - `Val [Prob=0.5]` Bill Clinton
        - `Val [Prob=0.2]` Bill Clinton
        - `Val [Prob=0.1]` Bill Clinton
        - `Val [Prob=0.3]` Bill Clinton

28th International VLDB Conference – August 2002 – p.3/28
An Introductory Example

A probability will be attached to each query result

Result Probability = 0.11025
An Introductory Example

Node probabilities in the output are updated

Result Probability = 0.11025
Outline

- XML Constructs
- Probabilistic Model
- Probabilistic Query Evaluation
- Efficiency
- Related Work
- Contributions and Future Work
Outline

XML Constructs
- Prob – Element Probability
- Dist – Element Probability Distributions

Probabilistic Model

Probabilistic Query Evaluation

Efficiency

Related Work

Contributions and Future Work
XML Constructs – Prob

The `Prob` attribute associates a probability with an element:

```xml
<country Prob='0.9'>...</country>
<age Prob='0.3'>55</age>
```

If no probability is assigned to an element, then a default probability of 1.0 is used.
XML Constructs – Dist

Dist is used to represent a probability distribution for an element.

Distributions can assign probabilities to complex values (i.e. whole sub-trees)
XML Constructs – Dist

*Dist* is used to represent a probability distribution for an element

```xml
<age>
  <Dist type="mutually-exclusive">
    <Val Prob='..2'>54</Val>
    <Val Prob='..35'>55</Val>
    <Val Prob='..1'>56</Val>
  </Dist>
</age>

<name>
  <Dist>
    <Val Prob='..4'>George W. Bush</Val>
    <Val Prob='..7'>George Bush</Val>
  </Dist>
</name>
```
XML Constructs – Dist

Distributions can assign probabilities to *complex* values (i.e. whole sub-trees)

```
<chiefOfState>
    <Dist type="mutually-exclusive">
        <Val Prob='0.5'>
        ... 
        </Val>
        <Val Prob='0.2'>
            <title Prob='0.65'>President</title>
            <name>Bill Clinton</name>
            <age Prob='0.3'>55</age>
        </Val>
    </Dist>
</chiefOfState>
```
Outline

- XML Constructs
- Probabilistic Model
  - Hierarchy and Probabilities
  - Sibling Dependencies
  - Incompleteness
- Probabilistic Query Evaluation
- Efficiency
- Related Work
- Contributions and Future Work
Hierarchy and Probabilities

Probabilities in the source data are conditional probabilities

A probability $p$ for an element is assigned conditioned upon the parent, specifically:

<table>
<thead>
<tr>
<th>Parent</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>$p$</td>
</tr>
<tr>
<td>T</td>
<td>$1 - p$</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
</tr>
</tbody>
</table>

If a parent does not exist, a child cannot exist

If a child exists for certain, the parent does as well
Sibling Dependencies

- Sibling probabilities may be independent
  - A person can have multiple phone numbers, each with their own probability
- Sibling probabilities may be mutually exclusive
  - A person’s age is a unique non-negative integer
  - The probability of having two different ages is equal to 0
Sibling Dependencies – Example

```xml
<age>
  <Dist type="mutually-exclusive">
    <Val Prob='0.2'>54</Val>
    <Val Prob='0.35'>55</Val>
    <Val Prob='0.1'>56</Val>
  </Dist>
</age>

<name>
  <Dist>
    <Val Prob='0.4'>George W. Bush</Val>
    <Val Prob='0.7'>George Bush</Val>
  </Dist>
</name>
```
Incompleteness

- XML handles incomplete information gracefully.
- Information may be missing, so we should not force probabilistic distributions to be complete.
  - Probabilities must lie between 0 and 1.
  - Disjoint probabilities must not add up to more than 1.
- Assume that $Prob(\neg A) = 1 - Prob(A)$. 
Incompleteness – Example

There is total probability of 0.35 that is “unaccounted for” in the age distribution.

We assume that this 0.35 probability is associated with values for age other than 54, 55, and 56 – since $\text{Prob}(\neg A) = 1 - \text{Prob}(A)$.

```
<age>
    <Dist type="mutually-exclusive">
        <Val Prob='0.2'>54</Val>
        <Val Prob='0.35'>55</Val>
        <Val Prob='0.1'>56</Val>
    </Dist>
</age>
```
Outline

- XML Constructs
- Probabilistic Model
- Probabilistic Query Evaluation
  - TIMBER
  - Simple Probabilistic Queries
  - Conjunctive Probabilistic Queries
  - Disjunction
  - Negation
  - Computing Node Probabilities
- Efficiency
- Related Work
- Contributions and Future Work
TIMBER

- A tree (XML) database developed at the University of Michigan
- Queries involve the matching of query trees against data trees
- TAX (tree algebra for XML) provides *set-at-a-time processing* rather than *navigational processing*
TIMBER Architecture

XML Query → Query Parser → Query Optimizer → Metadata Manager → Data Storage Manager → Data Parser → Data Manager → Index Manager → Query Evaluator → Query Output API → Query Result → Data Loading Flow → Program Flow → ProTDB Functionality Added → Data
A Simple Query

“Return the independence day of each country”

$F : \$1.tag = \text{independenceDay}$
A Simple Query

“Return the independence day of each country”
A Simple Query

“Return the independence day of each country”

Result Tree Probability = 0.85 x 1.0 x 0.9 x 1.0
= 0.765
A Simple Query – continued

- We multiplied up the ancestor chain to find the probability

Given $A \rightarrow B \rightarrow C$, find $Prob(B)$:

$$Prob(B|A) = \frac{Prob(A|B) \times Prob(B)}{Prob(A)}$$

$$Prob(B|A) = \frac{1.0 \times Prob(B)}{Prob(A)} \text{ (since } Prob(A|B) = 1.0)$$

so, $$Prob(B) = Prob(B|A) \times Prob(A)$$

- Similarly, to find $Prob(C)$:

$$Prob(C|B) = \frac{Prob(C|B) \times Prob(C)}{Prob(B)}$$

$$Prob(C|B) = \frac{1.0 \times Prob(C)}{Prob(B)}$$

$$Prob(C|B) = \frac{Prob(C)}{Prob(B|A) \times Prob(A)}$$

so, $$Prob(C) = Prob(C|B) \times Prob(B|A) \times Prob(A)$$
A Conjunctive Query

“Return all chiefs of state having the name George Bush and an age of 55”:

\[ F : \begin{align*} 
&\text{key} = \text{chiefOfState} \\
&\land \text{value} = \text{name} \\
&\land \text{value} = \text{George Bush} \\
&\land \text{value} = \text{age} \\
&\land \text{value} = 55 
\end{align*} \]
A Conjunctive Query

We factor out the common node to root paths for conjunctive queries

Result Tree Probability = 0.7 \times 0.35 \times 0.5 \times 0.9 = 0.11025
A Conjunctive Query

Result Tree Probability = 0.7 x 0.35 x 0.5 x 0.9 = 0.11025
Disjunction

Convert query pattern tree function, $F$ into disjunctive normal form

$F = C_1 \lor \ldots \lor C_n$, and each $C_i$ is a sequence of conjunctions

Apply standard set-based notions of event intersection:

$$P \text{rob}(C_1 \lor C_2 \lor C_3) = P \text{rob}(C_1) + P \text{rob}(C_2) + P \text{rob}(C_3) - P \text{rob}(C_1 \land C_2) - P \text{rob}(C_1 \land C_3) - P \text{rob}(C_2 \land C_3) + P \text{rob}(C_1 \land C_2 \land C_3)$$

Only Probabilities of conjunctions are left
Disjunction – Example

\[ P(\)C_1 \lor C_2 \lor C_3) = P(C_1) + P(C_2) + P(C_3) -\]
\[ P(C_1 \land C_2) - P(C_1 \land C_3) - P(C_2 \land C_3) + \]
\[ P(C_1 \land C_2 \land C_3) \]
Disjunction – Example

\[ P_{rob}(C_1 \lor C_2 \lor C_3) = P_{rob}(C_1) + P_{rob}(C_2) + P_{rob}(C_3) - P_{rob}(C_1 \land C_2) - P_{rob}(C_1 \land C_3) - P_{rob}(C_2 \land C_3) + P_{rob}(C_1 \land C_2 \land C_3) \]
Disjunction – Example

\[
\text{Prob}(C_1 \lor C_2 \lor C_3) = \text{Prob}(C_1) + \text{Prob}(C_2) + \text{Prob}(C_3) - \\
\text{Prob}(C_1 \land C_2) - \text{Prob}(C_1 \land C_3) - \text{Prob}(C_2 \land C_3) + \\
\text{Prob}(C_1 \land C_2 \land C_3)
\]
Disjunction – Example

\[ \text{Prob}(C_1 \lor C_2 \lor C_3) = \text{Prob}(C_1) + \text{Prob}(C_2) + \text{Prob}(C_3) - \]
\[ \text{Prob}(C_1 \land C_2) - \text{Prob}(C_1 \land C_3) - \text{Prob}(C_2 \land C_3) + \]
\[ \text{Prob}(C_1 \land C_2 \land C_3) \]
Negation

Replace formulas containing negated conjuncts with equivalent formulas that do not contain negation

For example, \( Prob(A \land B \land \neg C) =
Prob(A \land B) - Prob(A \land B \land C) \)
Computing Node Probabilities

- Each node probability is updated in the output.
- We compute each node probability \( node_i \) as \( Prob(node_i|F) \), where \( F \) is the query pattern function.
- \( Prob(node_i|F) \) is equal to \( \frac{Prob(node_i \land F)}{Prob(F)} \), which we know how to compute.
Outline

- XML Constructs
- Probabilistic Model
- Probabilistic Query Evaluation
- Efficiency
- Related Work
- Contributions and Future Work
Efficiency

How much overhead do we incur for the probabilistic manipulations?

Compare against the non-probabilistic version of the queries

Machine specs and data characteristics:

- Pentium 4 @ 1.5 GHz with 256 MB RAM
- XML file was approximately 200 MB and had over 3,000,000 nodes (storage size in the database exceeded available RAM)
Efficiency

Query Timing - Probabilistic vs. Non-Probabilistic (selectivity = 0.05%, 50 subtrees)

- Number of Conjuncts in the Query:
  - 2
  - 4
  - 6

- Number of Disjuncts in the Query:
  - 2
  - 4
  - 6

- Time (seconds):
  - 0
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8

- Probabilistic Querying
- Non-Probabilistic Querying
Outline

- XML Constructs
- Probabilistic Model
- Probabilistic Query Evaluation
- Efficiency
- Related Work
  - Probabilistic Relational DBs
  - Information Retrieval for XML
  - Approximate XML Query Evaluation
- Contributions and Future Work
Related Work

- Probabilistic relational databases
  - 1NF Models – Large tables with redundant data
  - NFNF Models – Awkward querying

- Information retrieval (IR) with XML
  - Combines IR with tree structure
  - “Fuzzy” match with certain data vs. exact match against probabilistic data

- Approximate XML query evaluation
  - Modify the query tree and match these multiple (modified) query trees against the data
Contributions and Future Work

- Definition of a probabilistic model in XML
- Pattern tree selection queries in a probabilistic XML database

Future Work

- A complete probabilistic algebra
- Efficient probabilistic query evaluation in light of user, or data, constraints
- Enabling probabilistic data management in new application areas