The Possibility Problem for Probabilistic XML

Antoine Amarilli

Télécom ParisTech, Paris, France

June 5, 2014
Probabilistic XML

We are unsure about the exact contents of an XML document.

Semantics: probability distribution over deterministic documents.
Local formalisms: possible worlds semantics

\[ (1 - \alpha)(1 - \beta) \quad \alpha(1 - \beta) \quad (1 - \alpha)\beta \quad \alpha\beta \]

\[ \begin{array}{c}
\alpha \\
\beta
\end{array} \quad \begin{array}{c}
a \\
b
\end{array} \quad \begin{array}{c}
\alpha \\
\beta
\end{array} \quad \begin{array}{c}
a \\
b
\end{array} \quad \begin{array}{c}
\alpha \\
\beta
\end{array} \quad \begin{array}{c}
a \\
b
\end{array} \]
Local formalisms: possible worlds semantics

\[
\begin{align*}
\text{r} \quad \Rightarrow \\
\text{ind} \quad \Rightarrow \\
\alpha \& \beta \\
\text{a} & \text{b}
\end{align*}
\]

\[
\begin{align*}
(1 - \alpha)(1 - \beta) & \quad \alpha(1 - \beta) & \quad (1 - \alpha)\beta & \quad \alpha\beta \\
\text{r} & \quad \text{r} & \quad \text{r} & \quad \text{r}
\end{align*}
\]

\[
\begin{align*}
\text{r} \quad \Rightarrow \\
\text{mux} \quad \Rightarrow \\
\alpha \& \beta \\
\text{a} & \text{b}
\end{align*}
\]

\[
\begin{align*}
1 - \alpha - \beta & \quad \alpha & \quad \beta \\
\text{r} & \quad \text{r} & \quad \text{r}
\end{align*}
\]

\[
\begin{align*}
\text{a} & \quad \text{b}
\end{align*}
\]
Local formalisms: possible worlds semantics

\[ \text{Caution: we impose } \alpha < 1, \beta < 1 \text{ in } \text{ind}. \]
Event formalisms

- Probability distribution on events
- Draw events independently
- Edges annotated with formulae on the events
- Edges with false formulae are removed

\[
\begin{array}{c}
\text{x} & 0.7 \\
\text{y} & 0.4 \\
\end{array}
\]

\[
\begin{array}{c}
\text{x} \\
\neg \text{x} \land \text{y} \\
\text{a} & \text{b} \\
\end{array}
\]

\[
\Rightarrow \text{mie}: \text{ multivalued events (see later)}
\]

\[
\Rightarrow \text{cie}: \text{ conjunctions of Boolean events}
\]

\[
\Rightarrow \text{fie}: \text{ formulae of Boolean events}
\]
Possibility problem (**Poss**)

- **Given:**
  - a probabilistic document $D$
  - a deterministic document $W$

- Is $W$ a possible world of $D$?

- If yes, with which probability?

- Diverse probabilistic formalisms, ordered and unordered

- Like query evaluation but:
  - Need inequality: “don’t collapse nodes”
  - Need negation: “no additional things”
  - Query depends on input $W$

- **Specific bounds** for this **Poss** problem?
<table>
<thead>
<tr>
<th></th>
<th>Introduction</th>
<th>Known results</th>
<th>Unordered documents</th>
<th>Unambiguous labels</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Known results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Unordered documents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Unambiguous labels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Conclusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In NP, in FP $\#P$

- Guess a *valuation* of the events
- Guess a *match* of $W$ in $D$
- Check that the match is *realized* by the valuation

$\Rightarrow$ Likewise, probability computation is in $FP^{\#P}$

$\Rightarrow$ Of course Poss is NP-hard for *fie*
Tractable for ordered local documents

- Local choices and ordered documents
- Possibility decision and computation are in PTIME
- Intuitively:
  - match each possible subsequences of siblings
  - dynamic algorithm for match at each level

⇒ Implied by determininistic tree automata on probabilistic XML: Cohen, Kimelfeld, and Sagiv 2009
⇒ Assumption of order is crucial
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Known results</th>
<th>Unordered documents</th>
<th>Unambiguous labels</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

1. Introduction

2. Known results

3. Unordered documents

4. Unambiguous labels

5. Conclusion
Computation is \#P-hard for \textit{ind} or \textit{mux}

\[
D \quad W
\]

\[
1/2 \quad 1/2 \quad 1/2 \quad 1/2 \quad 1/2
\]

\[
a1 \quad a2 \quad a3 \quad a2 \quad a3
\]

\Rightarrow \text{Probability of match times } 2^n: \text{ number of perfect matchings}

\Rightarrow \text{Computation is } \#P\text{-hard for unordered and } \textit{ind} \text{ or } \textit{mux}
Decision is in PTIME for \textit{ind} or \textit{mux}

- Compute bottom-up if a node has the \textit{empty} possible world
- Check \textit{dynamically} between all nodes of $D$ and $W$
  - $\Rightarrow$ Build bipartite graph based on child compatibility
  - $\Rightarrow$ Add \textit{dummy nodes} for deletions of nodes that can be deleted
  - $\Rightarrow$ Check in PTIME if graph has a \textit{perfect matching}
Decision is NP-hard for any two of \textit{ind}, \textit{mux}, \textit{det}

- With \textit{det}, reduction from \textit{exact cover}
  - $S = \{S_i\}, S_i = \{s_j^i\}$
  - Is there $T \subseteq S$ such that $\bigcup T = \bigcup S$ with no dupes?

\[
\begin{align*}
S &= \{\{a, b\}, \\
    &\quad \{a, c\}, \\
    &\quad \{b\}\}
\end{align*}
\]

\[
\begin{align*}
D &= r \\
    &/ 0.5 \\
    &/ 0.5 \\
    &/ 0.5 \\
    &/ \text{det} \\
    &/ \text{det} \\
    &/ \text{det} \\
    \quad \quad a \\
    \quad \quad b \\
    \quad \quad a \\
    \quad \quad c \\
    \quad \quad b
\end{align*}
\]
Decision is NP-hard for any two of *ind, mux, det* (cont’d)

- With *ind* and *mux*, reduction from SAT
- \[ F = (a \lor b \lor \neg c) \land (a \lor c) \land (\neg a) \]

Diagram:
```
D
  /\r
 / \mux
 /   \1/2
/     \ind
/       1/2
/         c1
       \2/1
     \2/c2
    \2/c3
```
```
W
  /\r
 / \r
 /   \c1
 /     \c2
/       \c3
```
Table of contents

1. Introduction
2. Known results
3. Unordered documents
4. Unambiguous labels
5. Conclusion
Unambiguity

- $D$ is **unambiguous** if node labels are unique
- Possible **refinements** (unique among siblings, etc.)
  - There is **at most one way** to match $\mathcal{M}$!
- All **local models** tractable (can impose order)
  - Can we have **correlations**?
Still NP-hard for \textit{cie}

- \( F = \bigwedge_i \bigvee_j \pm x^i_j \) in CNF
- Equivalently: \( \bigwedge_i \neg \bigwedge_j \pm x^i_j \)

\( W \) is a possible world of \( D \) iff \( F \) is satisfiable
\( \Rightarrow \) Decision for \text{Poss} is NP-hard
### The \textit{mie} class

<table>
<thead>
<tr>
<th>Var</th>
<th>Val</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>$x$</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>$x$</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>$x$</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>$y$</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>$y$</td>
<td>2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

- \textit{mie}: Multivalued independent events
- No \textit{conjunctions} allowed
- Captures \textit{mux}
- Doesn’t capture \textit{det} or \textit{ind} hierarchies
- Intractable if ambiguous

⇒ If non-ambiguous, do we have tractability?
tractable on non-ambiguous documents

<table>
<thead>
<tr>
<th>Var</th>
<th>Val</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>x</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>x</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>x</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>y</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>y</td>
<td>2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

- \( x \neq 2, x \neq 1, y = 2, y \neq 1 \)
- \( x \in \{3, 4\}, y \in \{2\} \)

\( \Rightarrow \) Probability 0.1.
Table of contents

1. Introduction
2. Known results
3. Unordered documents
4. Unambiguous labels
5. Conclusion
Conclusion

- **Ordered local models** are tractable
- **Unordered local models** are tractable
  - \( \Rightarrow \) For decision only, and
  - \( \Rightarrow \) With only \( mux \) or only \( ind \)

- \( mie \) is tractable on **unambiguous** documents
- Other cases are **hard**
Conclusion

- **Ordered local models** are tractable
- **Unordered local models** are tractable
  - For decision only, and
  - With only *mux* or only *ind*

- *mie* is tractable on unambiguous documents
- Other cases are hard

⇒ Height does not matter
⇒ Probabilities do not matter
⇒ Can we refine *mie*, unambiguity, *mux*–*ind* interaction?
⇒ What if *D* is partially ordered?
Conclusion

- Ordered local models are tractable
- Unordered local models are tractable
  - For decision only, and
  - With only mux or only ind

- mie is tractable on unambiguous documents
- Other cases are hard

- Height does not matter
- Probabilities do not matter
- Can we refine mie, unambiguity, mux–ind interaction?
- What if D is partially ordered?

Thanks for your attention!