The Possibility Problem for Probabilistic XML

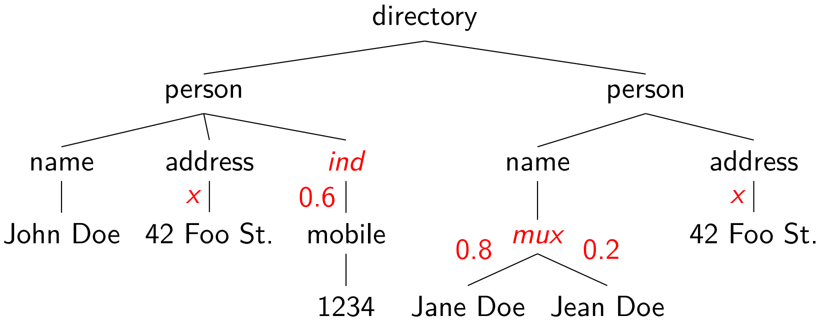
Antoine Amarilli

Télécom ParisTech, Paris, France

June 5, 2014

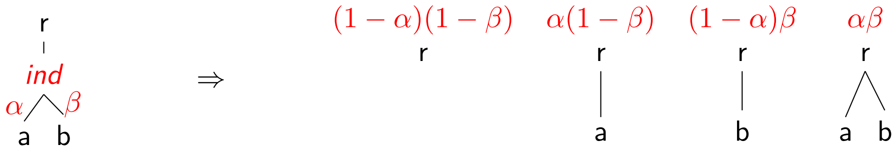
Probabilistic XML

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

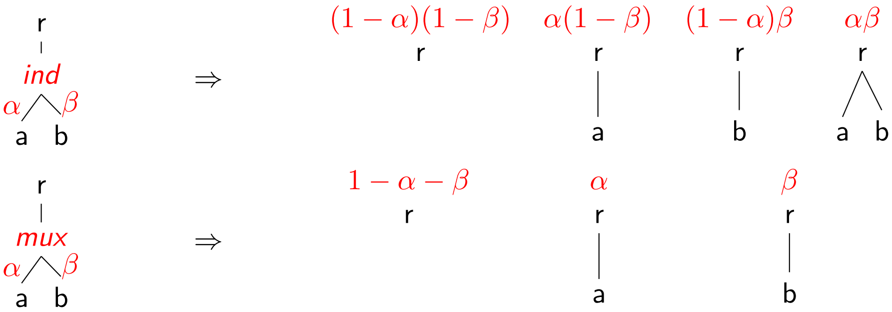


Semantics: probability distribution over deterministic documents.

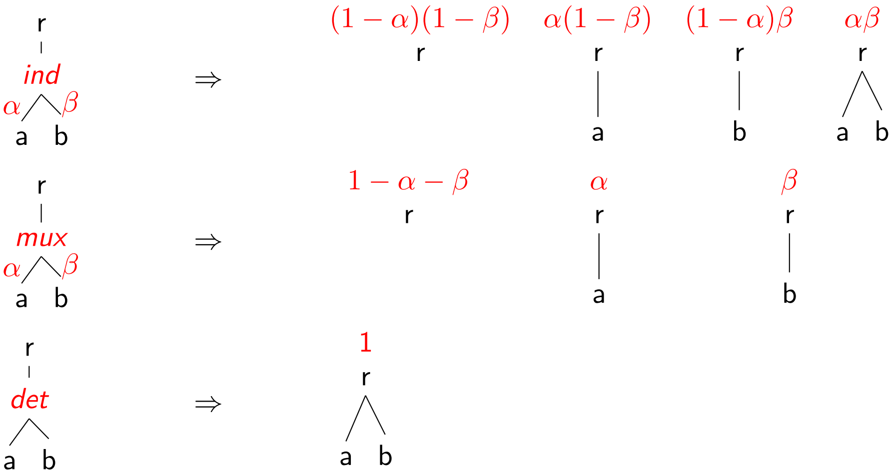
Local formalisms: possible worlds semantics



Local formalisms: possible worlds semantics



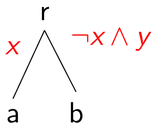
Local formalisms: possible worlds semantics



Caution: we impose $\alpha < 1$, $\beta < 1$ in *ind*.

Event formalisms

x	0.7
y	0.4



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

⇒ *mie*: multivalued events (see later)

⇒ *cie*: conjunctions of Boolean events

⇒ *fie*: 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:

  $\Rightarrow$ Need inequality: "don't collapse nodes"

  $\Rightarrow$ Need negation: "no additional things"

  $\Rightarrow$ Query depends on input W

$\Rightarrow$ Specific bounds for this Poss problem?

Table of contents

In NP, in $\mathrm{FP}^{\#\mathrm{P}}$

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

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

⇒ Of course $\textsc{Poss}$ is NP-hard for *fie*

Tractable for ordered local documents

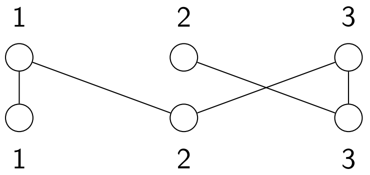
- Local choices and ordered documents
- Possibility decision and computation are in PTIME
- Intuitively:
  - match each possible subsequences of siblings
  - dynamic algorithm for match at each level

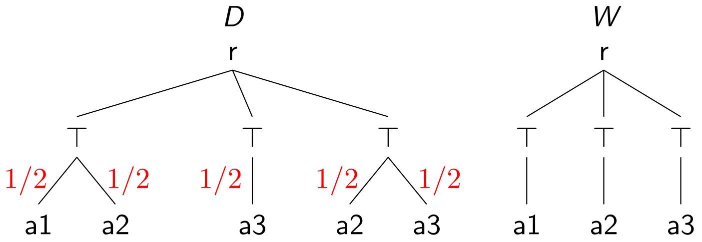
⇒ Implied by determininstic tree automata on probabilistic XML: Cohen, Kimelfeld, and Sagiv 2009

⇒ Assumption of order is crucial

Table of contents

Computation is #P-hard for *ind* or *mux*



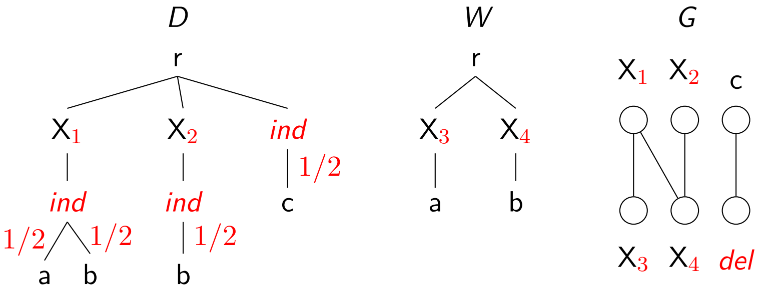


$\Rightarrow$ Probability of match times 2^n: number of perfect matchings

$\Rightarrow$ Computation is #P-hard for unordered and *ind* or *mux*

Decision is in PTIME for *ind* or *mux*

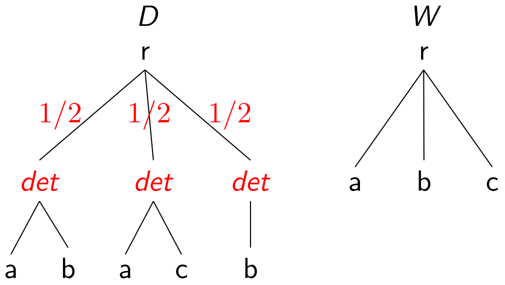
- Compute bottom-up if a node has the empty possible world
- Check dynamically between all nodes of D and W
  - ⇒ Build bipartite graph based on child compatibility
  - ⇒ Add dummy nodes for deletions of nodes that can be deleted
  - ⇒ Check in PTIME if graph has a perfect matching



Decision is NP-hard for any two of *ind*, *mux*, *det*

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

$$S = \{\{a, b\}, \{a, c\}, \{b\}\}$$



Decision is NP-hard for any two of *ind*, *mux*, *det* (cont'd)

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

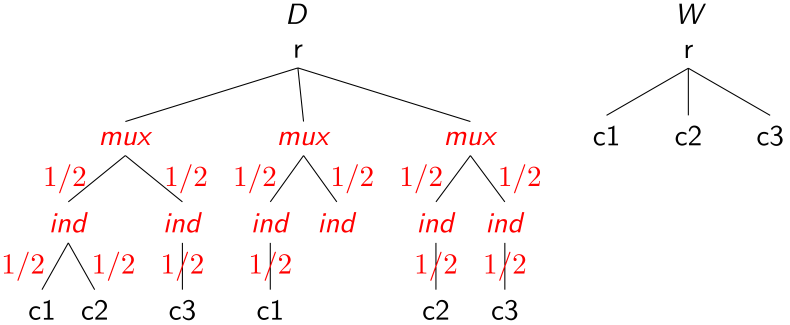


Table of contents

Unambiguity

- D is unambiguous if node labels are unique
- Possible refinements (unique among siblings, etc.)

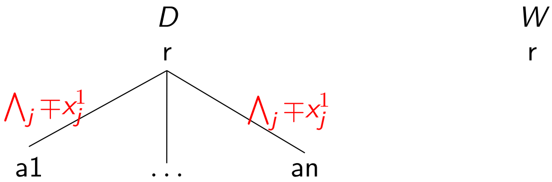
$\Rightarrow$ There is at most one way to match W!

- All local models tractable (can impose order)

$\Rightarrow$ Can we have correlations?

Still NP-hard for *cie*

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



$\Rightarrow$ W is a possible world of D iff F is satisfiable

$\Rightarrow$ Decision for Poss is NP-hard

The *mie* class

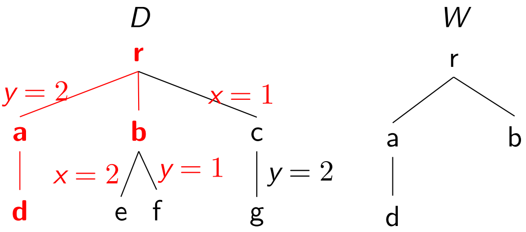
Var	**Val**	**Prob**
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5

- *mie*: Multivalued independent events
- No conjunctions allowed
- Captures *mux*
- Doesn't capture *det* or *ind* hierarchies
- Intractable if ambiguous

$\Rightarrow$ If non-ambiguous, do we have tractability?

mie tractable on non-ambiguous documents

Var	Val	Prob
x	1	0.6
x	2	0.2
x	3	0.1
x	4	0.1
y	1	0.5
y	2	0.5



- $x \neq 2$, $x \neq 1$, $y = 2$, $y \neq 1$
- $x \in \{3, 4\}$, $y \in \{2\}$.

$\Rightarrow$ Probability 0.1.

Table of contents

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

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!

References

Cohen, Sara, Benny Kimelfeld, and Yehoshua Sagiv (2009). "Running tree automata on probabilistic XML". In: *Proc. PODS*. ACM, pp. 227–236.