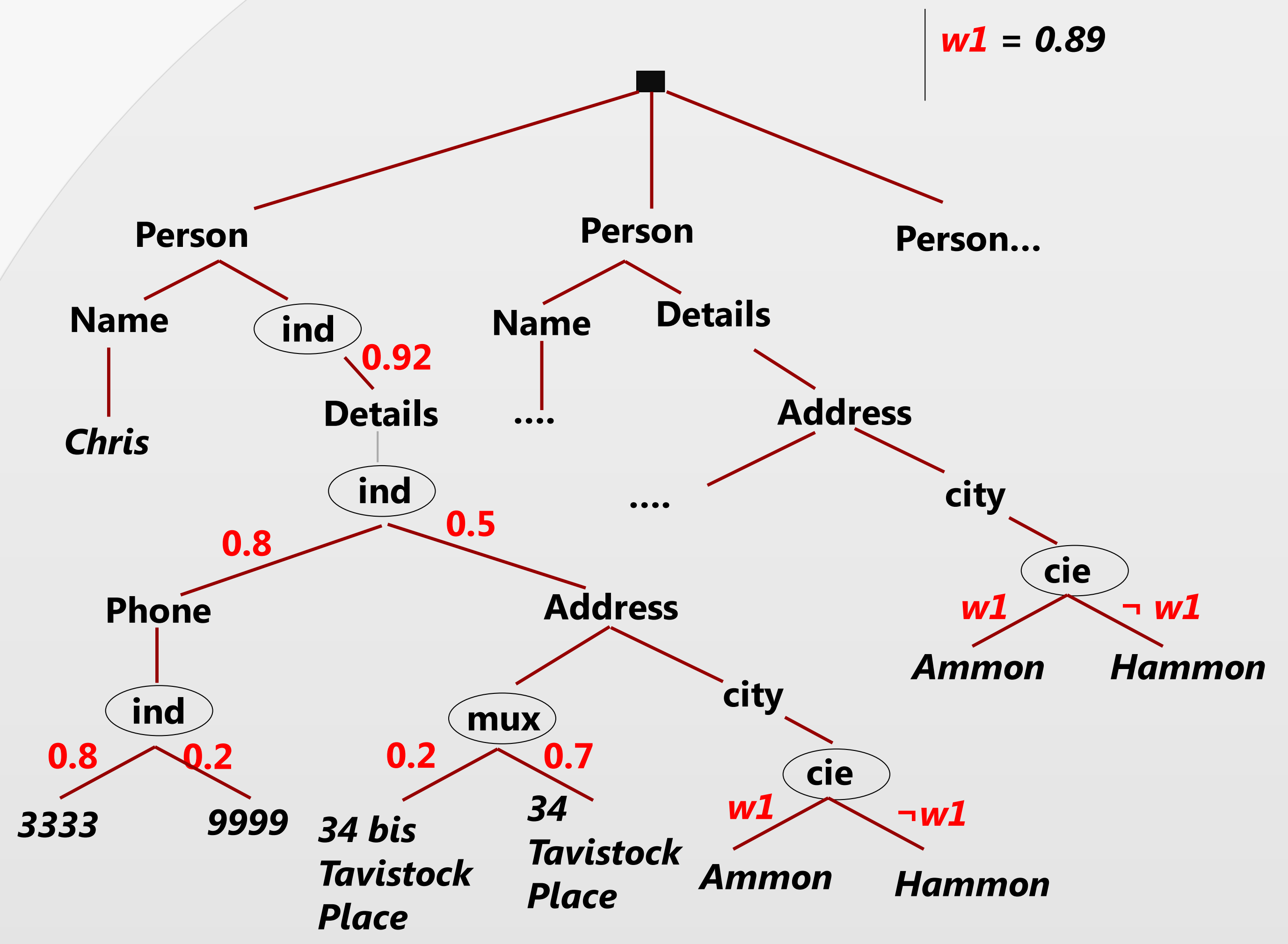
Querying and Managing Probabilistic XML

ProApproX: A Lightweight Approximation Query Processor over Probabilistic Trees

Asma Souihli Pierre Senellart

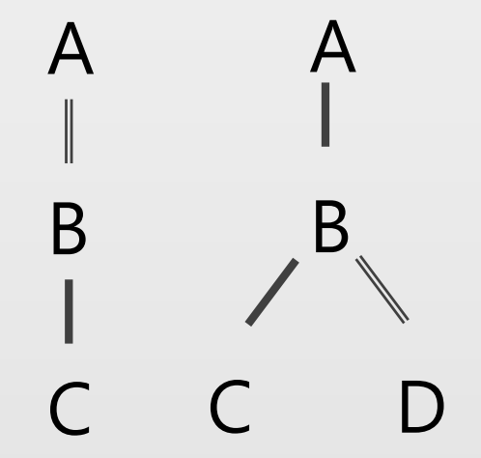
ProApprox is a query processor over probabilistic trees that represents a first step towards building a fully-fletched probabilistic semistructured data management system. It relies on a generalization of the different uncertain data models in XML proposed in the literature and allows for efficient data querying with a subset of the XPath query language, through techniques of exact calculations or efficient approximations of the result.

Probabilistic Data

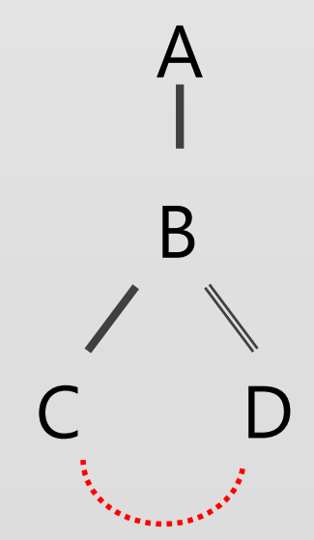


Boolean query languages on trees :

Tree Pattern Queries



Tree Pattern Queries with joins

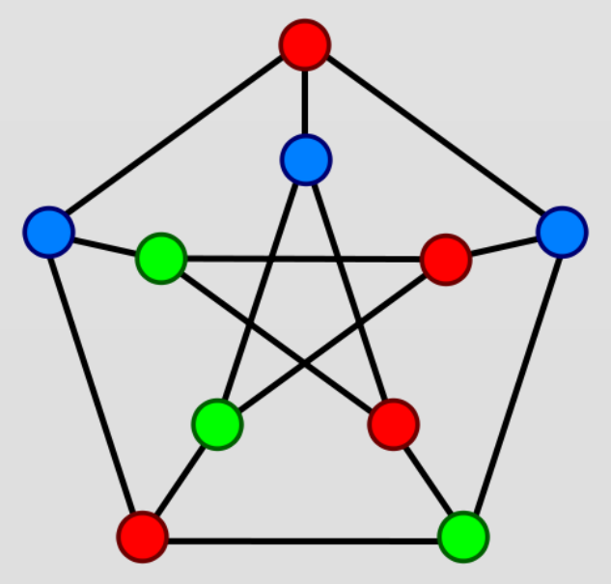


- Semantics of a (Boolean) **query = probability.**
- If the patterns are not independent up to intersection, the computation is a hard mathematical problem.

- Naïve but exact solution:
  1. Generate all possible worlds of a given probabilistic document
  2. In each world, evaluate the query
  3. Add up the probabilities of the worlds that make the query true

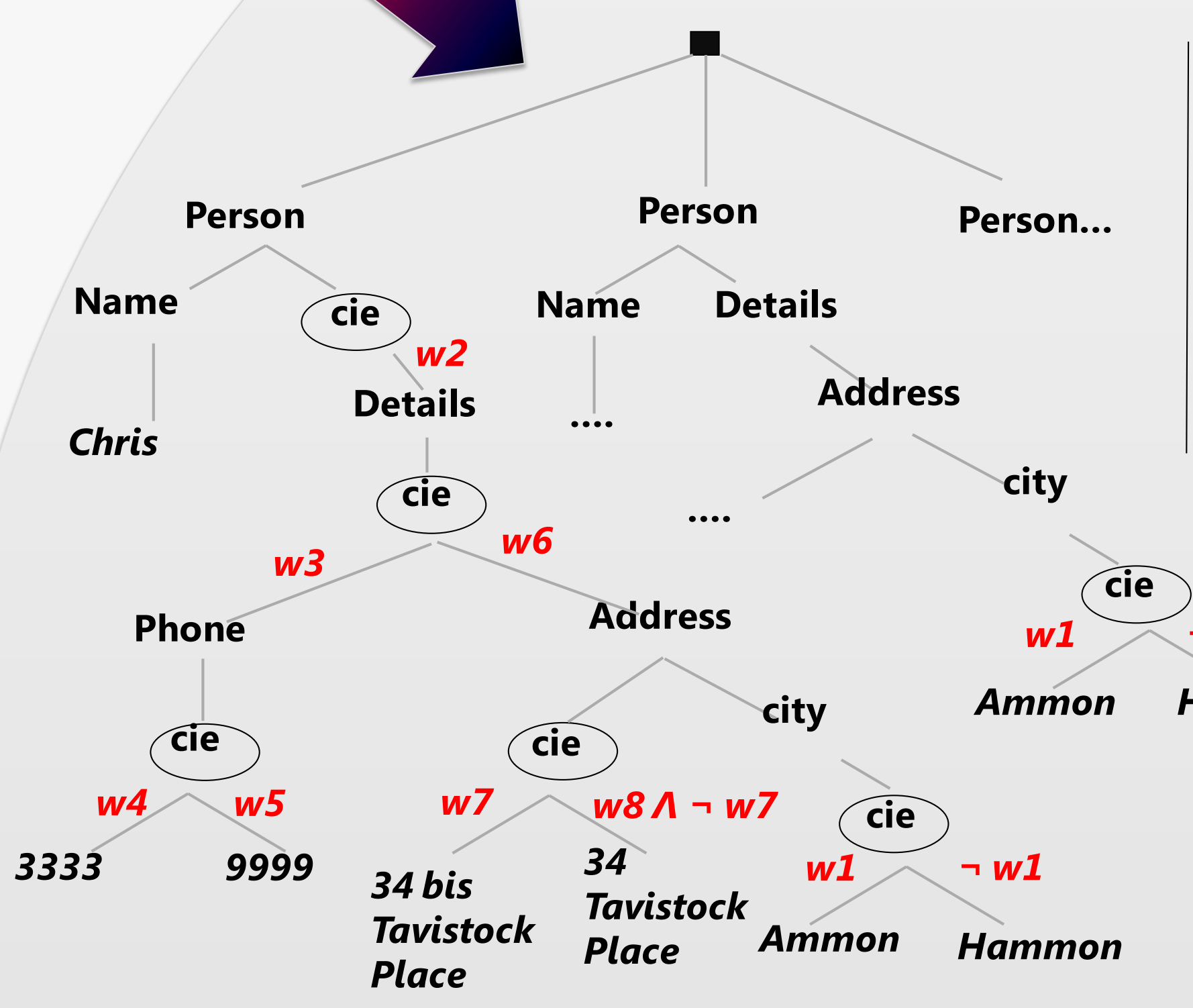
A #P-complete problem and an EXPTIME algorithm!

→#P problems ask "how many" rather than "are there any".
How many graph colorings using k colors are there for a particular graph G?



How to run Approximations efficiently??

Equivalent translation to a global dependencies model



w1 = 0.89
w2 = 0.92
w3 = 0.8
w4 = 0.8
w5 = 0.2
w6 = 0.5
w7 = 0.2
w8 = 0.875

Q1:
//Person[Name='Chris']//Address/text()

1. **Encoding the Matches:**
   Xpath Parser: *Rewrites the query (Q1' in XQuery) so as to return the sequence of wi's along a pattern to the query*

   →***Mappings for Q1:***
   ⟨w2,w6,w7⟩
   ⟨w2,w6,w8,¬w7⟩

2. Processing the Query:

- **Exact Computation:**
  Use the w_i's of the mappings to run exact computation (whenever possible).
- **Additive Approximation:**
  Draw values for the w_i's of the mappings.
  Evaluate the mappings.
  →Use this process n times running Additive App.
- **Multiplicative Approximation:**
  Evaluate Multiplicative App. formula using mapping sequences for draws and evaluation.

- **What we aim for: A probabilistic DBMS using XML technology capable of efficiently querying discrete probabilistic data models.**
- **And... deal with aggregate queries. The result of a query that make uses of aggregate functions is a set of possible values (for each possible document), each with its probability.**
- **Also... move to a distributed framework to manage probabilistic data, i.e., in a open file sharing environment or in the case of data integration.**
- **Update operations also belong to future implementation perspectives.**