



# DataRing: Model and Language

Past, Current, and Future Work

Pierre Senellart





Probabilistic XML Recap

What we have done in DataRing

What we are doing

What should we do next?



# Uncertain data

Numerous sources of **uncertain data**:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemata
- Imprecise automatic process (information extraction, natural language processing, etc.)
- Imperfect human judgment



# Managing this imprecision

## Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Especially:

- Use **probabilities** to represent the confidence in the data
- Query data and retrieve **probabilistic** results
- Allow adding, deleting, modifying data in a **probabilistic** way



# Managing this imprecision

## Objective

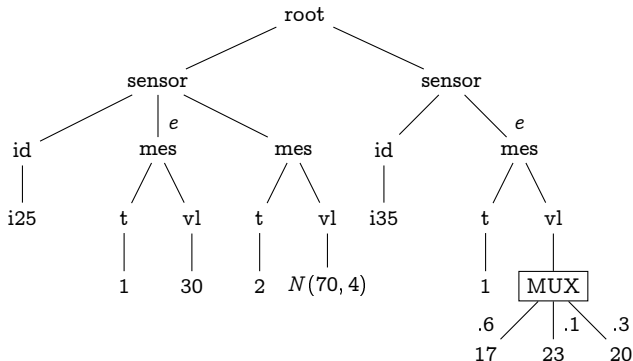
Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Especially:

- Use **probabilities** to represent the confidence in the data
- Query data and retrieve **probabilistic** results
- Allow adding, deleting, modifying data in a **probabilistic** way

# A General Probabilistic XML Model

[Abiteboul et al., 2009]



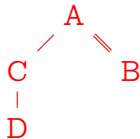
- $e$ : event “it did not rain” at time 1
- MUX: mutually exclusive options
- $N(70, 4)$ : normal distribution

- Compact representation of a **set of possible worlds**
- Two kinds of dependencies: global ( $e$ ) and local (MUX)
- Generalizes **all existing models** of the literature



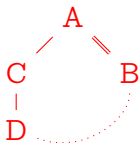
# Query languages on trees

Tree-pattern queries (TP) `/A[C/D]//B`



Tree-pattern queries with joins (TPJ) for `$x` in `$doc/A/C/D`

`return $doc/A//B[.= $x]`



Monadic second-order queries (MSO) generalization of TP, do not cover TPJ unless the size of the alphabet is bounded



# Querying probabilistic XML

Semantics of a (Boolean) query = **probability**:

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

**EXPTIME** algorithm! Can we do better, i.e., can we apply directly the algorithm on the probabilistic document?

We shall talk about **data complexity** of query answering.





# Querying probabilistic XML

Semantics of a (Boolean) query = **probability**:

1. Generate **all possible worlds** of a given probabilistic document (possibly exponentially many)
2. In each world, **evaluate the query**
3. **Add up** the probabilities of the worlds that make the query true

**EXPTIME** algorithm! Can we do better, i.e., can we apply directly the algorithm on the probabilistic document?

We shall talk about **data complexity** of query answering.



# Complexity of Boolean Query Evaluation

	Local dependencies	Global dependencies
TP	<b>P</b> TIME [Kimelfeld et al., 2009]	<b>FP</b> <sup>#P</sup> -complete
TPJ	<b>FP</b> <sup>#P</sup> -complete	<b>FP</b> <sup>#P</sup> -complete
MSO	<b>P</b> TIME [Cohen et al., 2009]	<b>FP</b> <sup>#P</sup> -complete



Probabilistic XML Recap

What we have done in DataRing

What we are doing

What should we do next?



## More General PXML Data Model

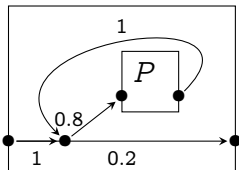
**Continuous distributions** As presented above. For sensor network data, unknown values, etc. [Abiteboul et al., 2010]

**Recursive Markov Chains** (between other things, probabilistic versions of DTDs) [Benedikt et al., 2010]

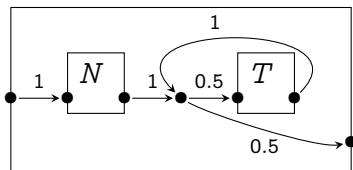
```
<!ELEMENT directory (person*)>
```

```
<!ELEMENT person (name,phone*)>
```

*D*: directory



*P*: person



On such simple RMCs, **MSO queries are tractable!**



## More General Operations on PXML

**Aggregate queries** (count, sum, max, avg, etc.): (somewhat) tractable on local dependencies when the aggregate function is a **monoid** function; **continuous distributions** do not add complexity [Abiteboul et al., 2010]; Evgeny's talk at the meeting in Nantes.

**Updates** (insertions, deletions) **Not the same kind of updates** are tractable for **local** and **global** dependencies [Kharlamov et al., 2010]; more precise picture of the complexity of updates in PXML models, extends the first characterization of updates from [Senellart and Abiteboul, 2007, Abiteboul et al., 2009].



Probabilistic XML Recap

What we have done in DataRing

What we are doing

What should we do next?



- A better understanding of the relation between complexity of a query and presence of **value joins** (cf. Evgeny's talk).
- An **actual system** for querying probabilistic XML data [Senellart and Souihli, 2010] (cf. Asma's talk).
- Application to **mining probabilistic XML data**: association rules, trend analysis. Writing up of a book chapter in progress.
- Using probabilistic XML to represent a corpus of XML documents: **probabilistic schema extraction** from tree-structured documents. Work in progress with Tova Milo.



# Outline

Probabilistic XML Recap

What we have done in DataRing

What we are doing

What should we do next?





## Other questions on PXML

- **Killer application**, killer example, killer dataset! Still looking. Data integration in a peer-to-peer setting?
- Support of updates for continuous distributions and RMCs: **partial global dependencies**?
- Better connections with the theoretical works and actual systems for **probabilistic relational data**. A little on that in Evgeny's and Asma's talks.
- System issues: indexing, distribution.



## Going beyond PXML

- Initial problem: Data model and query language for the DataRing.
- Partial solution: Probabilistic XML model and techniques.
- What else do we need?
- Connection with the other work packages: Querying Views, Representing Graph Data, Integration.

Merci.



DataRing Project: P2P Data Sharing for Online Communities

Webdam



## References I

- Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, and Pierre Senellart. On the expressiveness of probabilistic XML models. *VLDB Journal*, 18(5):1041–1064, October 2009.
- Serge Abiteboul, T-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Aggregate queries for discrete and continuous probabilistic xml. In *Proc. ICDT*, Lausanne, Switzerland, March 2010.
- Michael Benedikt, Evgeny Kharlamov, Dan Olteanu, and Pierre Senellart. Probabilistic XML via Markov chains, March 2010. Preprint.
- Sara Cohen, Benny Kimelfeld, and Yehoshua Sagiv. Running tree automata on probabilistic XML. In *Proc. PODS*, Providence, RI, USA, June 2009.



## References II

- Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Updating probabilistic XML. In *Proc. Updates in XML*, Lausanne, Switzerland, March 2010.
- Benny Kimelfeld, Yuri Kosharovskiy, and Yehoshua Sagiv. Query evaluation over probabilistic XML. *VLDB Journal*, 18(5): 1117–1140, October 2009.
- Pierre Senellart and Serge Abiteboul. On the complexity of managing probabilistic XML data. In *Proc. PODS*, pages 283–292, Beijing, China, June 2007.
- Pierre Senellart and Asma Souihli. Un système de gestion de données XML probabilistes, May 2010. Preprint.