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
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 $doc/A/C/D$

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.
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

<table>
<thead>
<tr>
<th></th>
<th>Local dependencies</th>
<th>Global dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TP</strong></td>
<td>PTIME</td>
<td>FP(^{#P}) -complete</td>
</tr>
<tr>
<td></td>
<td>[Kimelfeld et al., 2009]</td>
<td></td>
</tr>
<tr>
<td><strong>TPJ</strong></td>
<td>FP(^{#P}) -complete</td>
<td>FP(^{#P}) -complete</td>
</tr>
<tr>
<td><strong>MSO</strong></td>
<td>PTIME</td>
<td>FP(^{#P}) -complete</td>
</tr>
<tr>
<td></td>
<td>[Cohen et al., 2009]</td>
<td></td>
</tr>
</tbody>
</table>
Probabilistic XML Recap

What we have done in DataRing

What we are doing

What should we do next?
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]

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

\[ D: \text{directory} \]

\[ P: \text{person} \]

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].
Outline

Probabilistic XML Recap

What we have done in DataRing

What we are doing

What should we do next?
Current works

- A better understanding of the relation between complexity of a query and presence of value joins (cf. Evgeny’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.


