

On Join Queries over XML and Probabilistic XML

E. Kharlamov

*Free University of Bozen-Bolzano
INRIA Saclay – Île-de-France*



FREIE UNIVERSITÄT BOZEN
LIBERA UNIVERSITÀ DI BOLZANO
FREE UNIVERSITY OF BOZEN · BOLZANO



Joint work with

S. Abiteboul, *INRIA Saclay – Île-de-France*

W. Nutt, *Free University of Bozen-Bolzano*

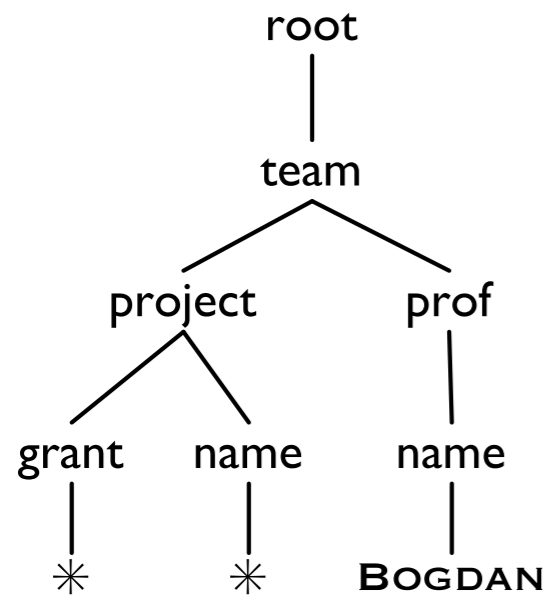
P. Senellart, *Télécom ParisTech*

DataRing Meeting, June 2010

Joins are important

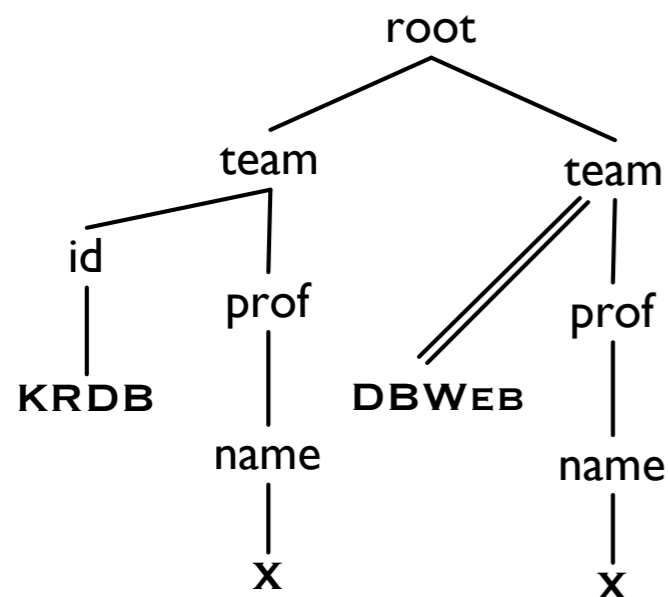
- For relational DBs
 - SPJ fragment of SQL is the most widely used
 - The only way to link different tables
- For XML DBs
 - to query documents with IDREF
 - to relate different fragments of a document

TPJ: Tree-Pattern queries with Joins



- *Is it the case that the team of prof. Bogdan is involved in a project with a grant?*

XPath{/, [, *}



- *Is there a prof. who works for both KRDB and DBWeb teams?*

XPath{/, //, [, Vars} = TPJ

MSOJ: MSO with Joins on Trees

1. variables: **node IDs**, unary **predicates**
 2. unary predicates for labels: $Label_a(n)$, $Label_b(n), \dots$
 3. vertical nav. relation: $Child(n, m)$
 4. ordering relations: $n < m$
 5. join predicate: **SameLabel(n,m)**
- } Navigation
in XML docs

Combined via Boolean operators and
first and second-order quantifiers $\exists n$ and $\exists S$

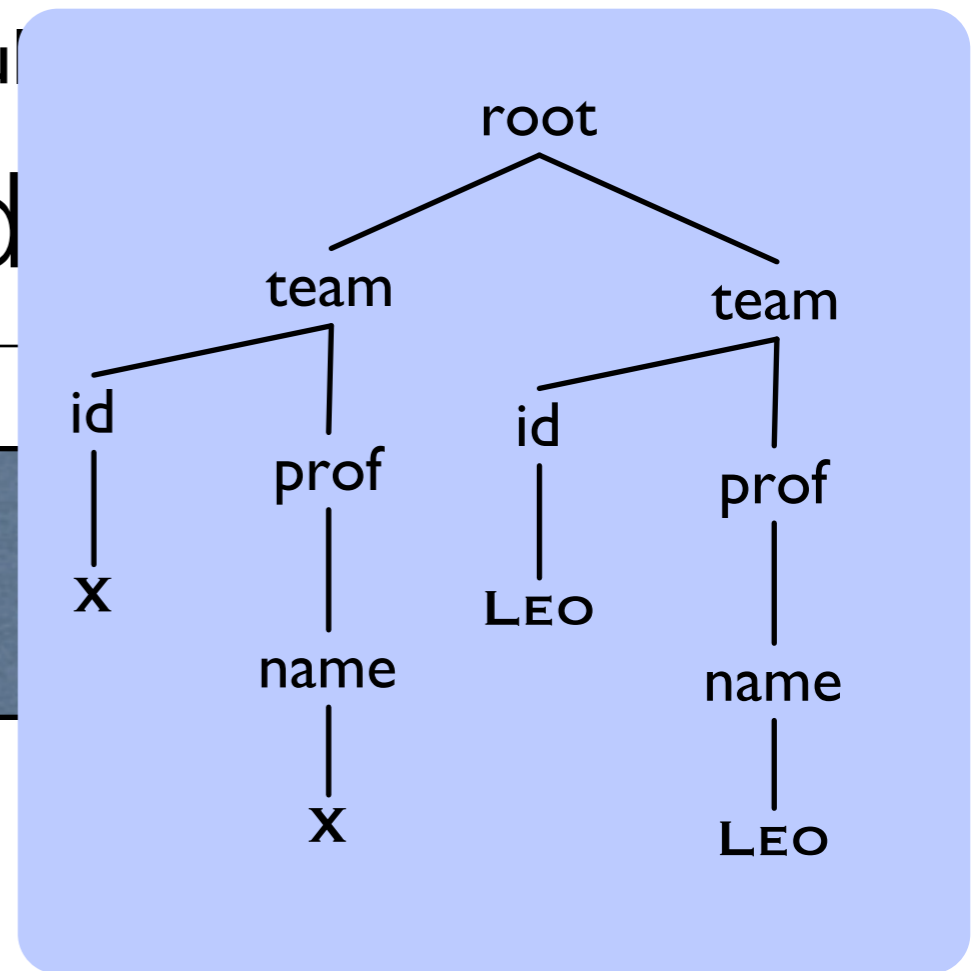
$$\text{SameLabel}(n, m) = \bigvee_{a \in \text{labels}} \text{Label}_a(n) \wedge \text{Label}_a(m)$$

Querying XML with TPJ and MSOJ

	TPJ	FOJ	MSOJ
Evaluation, no joins	Linear	Linear	Linear
Evaluation	LogSpace	LogSpace	Complete for all levels of PH
Deciding Joins	$\text{XPath}\{/, //, [], \text{Vars}\}$ - Π_2^P -complete $\text{XPath}\{/, [], \text{Vars}\}$ - NP-complete	undec.	undecidable

Deciding joins: check whether a query is equivalent to the one without a join

Querying XML with TPJ and



	TPJ			
Evaluation, no joins	Linear			
Evaluation	LogSpace		LogSpace	Complete for all levels of PH
Deciding Joins	XPath $\{/, //, [], Vars\}$ - Π_2^P -complete XPath $\{/, [], Vars\}$ - NP-complete		undec.	undecidable

Deciding joins:

check whether a query is equivalent to the one without a join

Querying XML with TPJ and MSOJ

	TPJ	FOJ	MSOJ
Evaluation, no joins	Linear	Linear	Linear
Evaluation	LogSpace	LogSpace	Complete for all levels of PH
Deciding Joins	XPath ^{/, //, [,], Vars} - Π_2^P -complete XPath ^{/, [,], Vars} - NP-complete	undec.	undecidable

Deciding joins: check whether a query is equivalent to the one without a join

Querying local PXML with TPJ and MSOJ

	TPJ	FOJ	MSOJ
Evaluation, no joins	Linear	Linear	Linear
Evaluation	#P - complete	#P-hard	#P-hard and hard for every level of PH

Our Goals 1: PXML vs. Probabilistic RDBs

- There is an extensive work on probabilistic RDBs
 - Trio project [Widom:07]
 - Block-independent databases (BIDs): MystiQ [Re&Succiu:06]
 - Probabilistic c-tables: MayBMS [Antova,Koch,Olteanu:06]
- How results on PRDBs are related to PXML?
 - BIDs are related to local PXML.
 - prob. c-tables are related to global PXML.

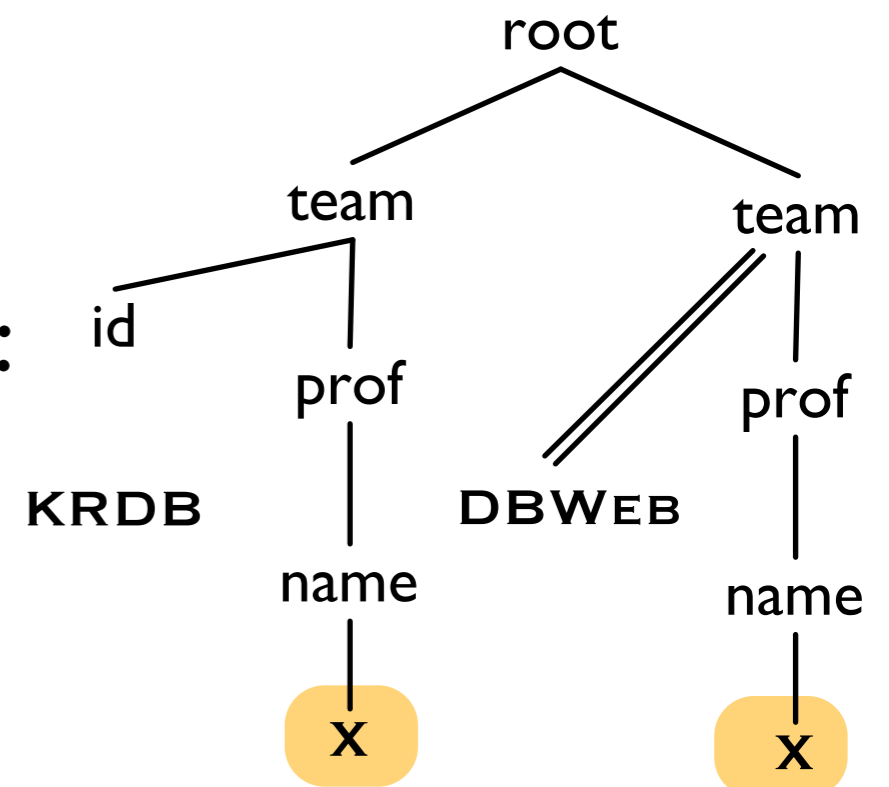
How to **translate** one model to another?

Our Goals 1: PXML vs. Probabilistic RDBs

- Dichotomy of Conjunctive Queries for BID:
 - **#P-hard**: 1) not hierarchical or
2) hierarchical with an inversion without erasers
 - **PTIME**: hierarchical and all inversion have erasers

Theorem: dichotomy of TPJs with **data value joins** for local PXML:

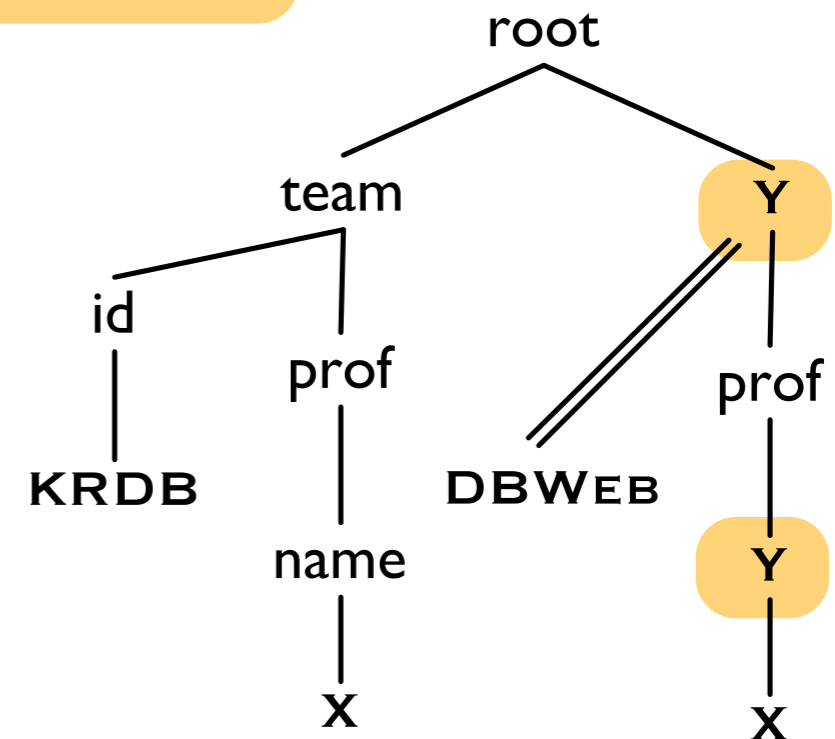
- **#P-hard**: with an **essential join**
- **PTIME**: essentially join-free



Our Goals 1: PXML vs. Probabilistic RDBs

Conjecture: dichotomy of TPJs for local PXML:

- **PTIME:** all essential joins are **hierarchical**
- **#P-hard:** with an essential and non-hierarchical join
- Holds for queries :
 - no descendant edges or
 - no branching ~ Single-Path queries



Our Goals 2: Tractable MSO for PRDBs

- Courcelle's Theorem:
Every MSO-definable property of graphs can be tested in Linear time if restricted to graphs with bounded tree-width

Extension to probabilistic data:

Conjecture:

The probability of every MSO query Q over BID D can be computed in polynomial time iff D has bounded tree-width and Q is tree-like

Our Goals 3: TPJ vs. XPath 1.0 and 2.0

- Which TPJ queries are in
 - XPath 1.0?
 - XPath 2.0 but not in XPath 1.0?
- Marx and de Rijke:
FO² is **equivalent** in expressiveness to navigational fragment of XPath 1.0

Our Goals 4: Tractable global PXML

- **TP** queries over PXML with **global** dependences are #P-complete
- **TPJ** queries over PXML with **local** dependences are #P-complete

Observation:

- Joins in **queries** behave similarly to global dependancies in probabilistic **data**
- Restricting joins in queries \Rightarrow tractability of QAnswering

To be understood:

- Restricting event-variables in data \Rightarrow TP tractability **How?**

Webdam

Webdam Project:

Foundations of Data Management

<http://webdam.inria.fr>



DataRing Project: P2P Data Sharing
for Online Communities

[http://www.lina.univ-nantes.fr/projets/
DataRing/](http://www.lina.univ-nantes.fr/projets/DataRing/)

Thank you

References

- [\[Abiteboul&al'10\]](#) - S. Abiteboul, T-H. H. Chan, E. Kharlamov, W. Nutt, and P. Senellart, Aggregate Queries for Discrete and Continuous Probabilistic XML. ICDT 2010
- [\[Abiteboul&al'95\]](#) - S. Abiteboul, R. Hull, V. Vianu: Foundations of Databases. Addison-Wesley 1995
- [\[Cohen&al'10\]](#) - S. Cohen, B. Kimelfeld, Y. Sagiv: Running tree automata on probabilistic XML. PODS 2009
- [\[Frick,Grohe'02\]](#) - M. Frick, M. Grohe: The Complexity of First-Order and Monadic Second-Order Logic Revisited. LICS 2002: 215-224
- [\[Kimelfeld&al'07\]](#) - B. Kimelfeld, Y. Sagiv: Matching Twigs in Probabilistic XML. VLDB 2007

References

- [\[Marx, deRijke'04\]](#) -Maarten Marx and Maarten de Rijke. “Semantic Characterizations of XPath”. In TDM'04. Workshop on XML Databases and Information Retrieval, Twente, The Netherlands, 2004.
- [\[Senellart&al'07\]](#) - P. Senellart, S. Abiteboul: On the complexity of managing probabilistic XML data. PODS 2007
- [\[Deutsch, Tannen'05\]](#) -A. Deutsch, V. Tannen: XML queries and constraints, containment and reformulation. Theor. Comput. Sci. 336(1): 57-87 (2005)
- [\[Senellart&al'07\]](#) - P. Senellart, S. Abiteboul: On the complexity of managing probabilistic XML data. PODS 2007