Probabilistic XML via Markov Chains

Evgeny Kharlamov

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

Joint work with

Michael Benedikt *Oxford University*

Dan Olteanu *Oxford University*

Pierre Senellart *Télécom ParisTech*

VLDB. Singapore. Sept. 2010

Uncertain Data is Commonplace

- (Web) information extraction
- Processing manually entered data (such as census forms)
- Data integration, data cleaning
- Managing scientific data; sensor data
- Risk management / predictions
- ...

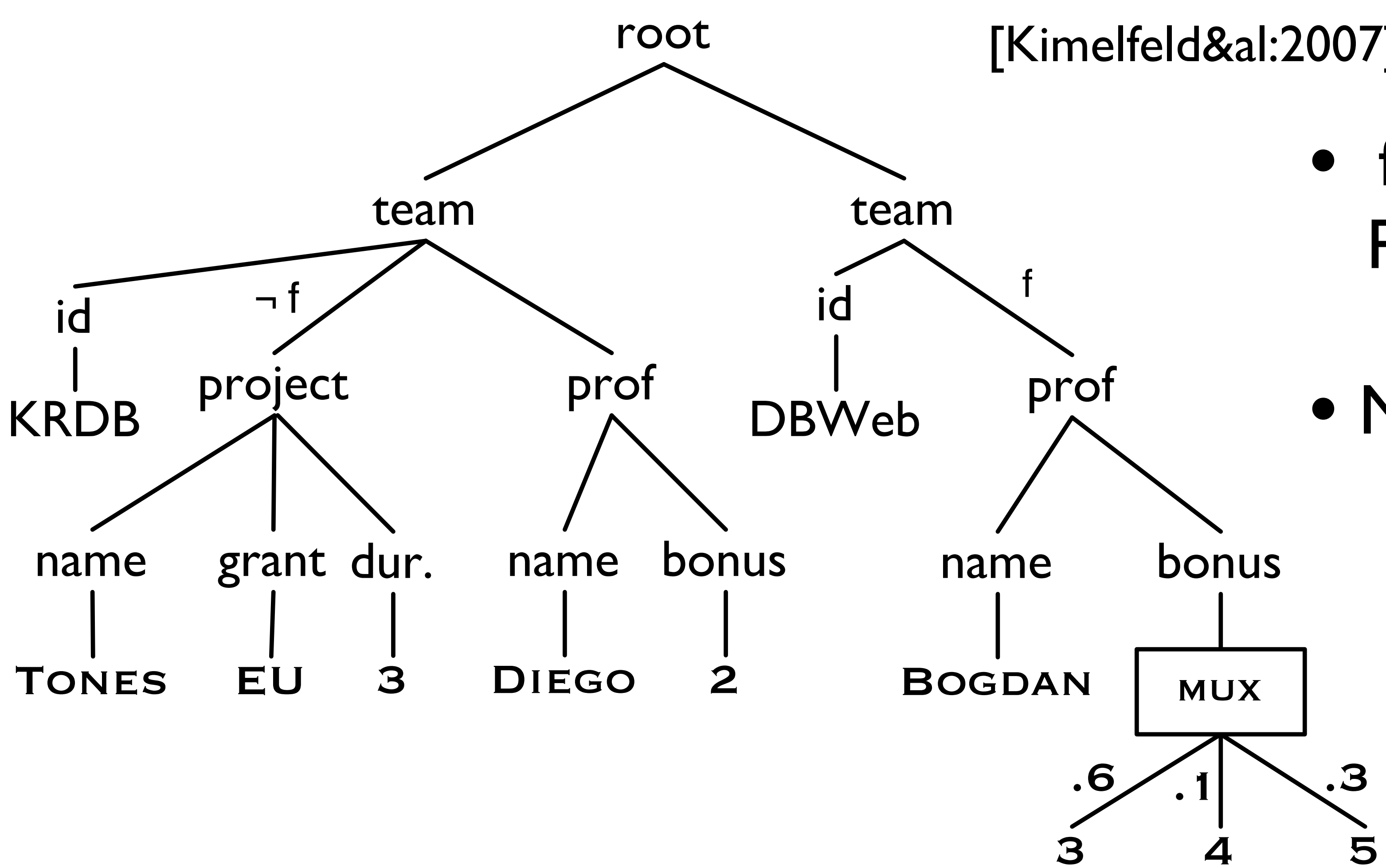
Probabilities are a way to deal with uncertain data

Dealing with Probabilistic Data

- Traditional DBMSs: not meant to deal with probabilistic data
- Ad hoc approaches: not very satisfactory
- Recent years: advances in developing
  - representation systems for incomplete/probabilistic data
  - uncertainty-aware query languages
  - ...
- Probabilistic relational DBMSs:
  MayBMS, MystiQ, PrDB, Trio, ...

Probabilistic XML Today: PrXML Model

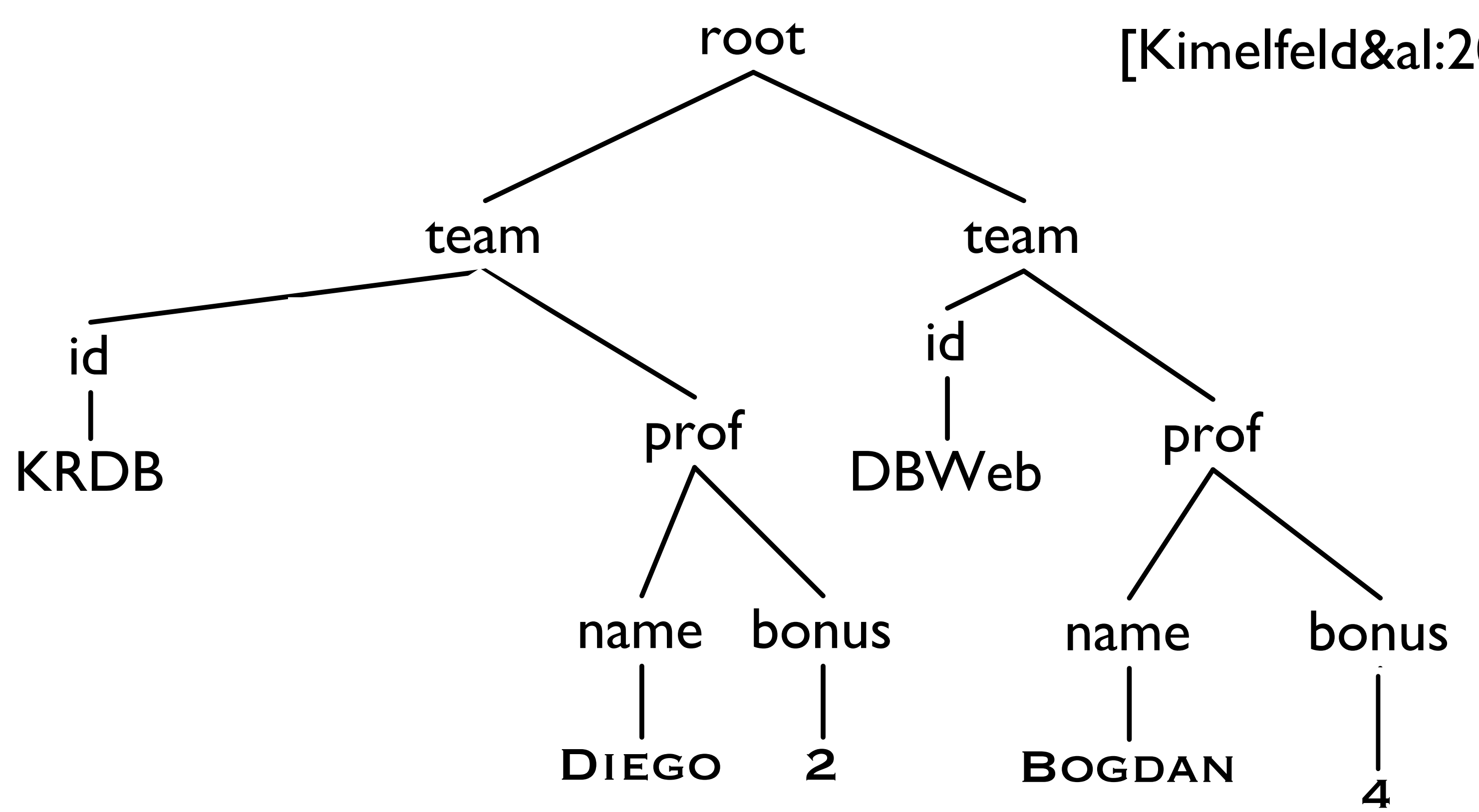
[Kimelfeld&al:2007] [Abiteboul&al:2009]



- f - event: “fresh”
  Pr(f) = 0.4
- MUX - distributional node, mutually exclusive options

Probabilistic XML Today: PrXML Model

[Kimelfeld&al:2007] [Abiteboul&al:2009]



- f - event: "fresh"
  Pr(f) = 0.4
- MUX - distributional node, mutually exclusive options
- Semantics: set of possible worlds.

- Example world:
  - f = true (the data is outdated), probability of this choice: 0.4
  - MUX: 4, probability of this choice: 0.1
- probability of this world is 0.4 x 0.1

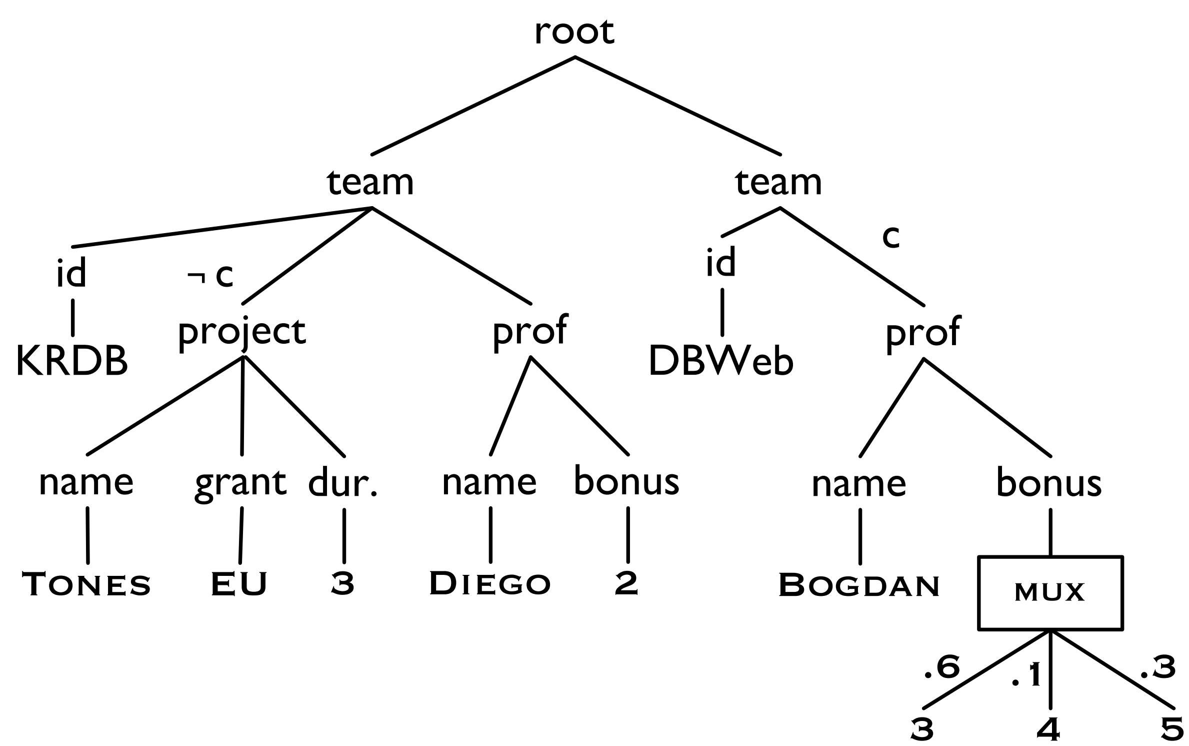
Probabilistic XML documents (compactly) represent probability spaces of ordinary XML documents

Probabilistic XML Today

- Trees enhanced with distributional nodes and event formulas that define the probabilistic process that generates random trees
- Proposed PrXML representation systems mirror the relational case
- Widely studied in recent years:
  - Query answering [Kimelfeld&al'09]
  - Aggregating [Abiteboul&al'10]
  - Constraints [Cohen&al'09]
  - Continuous models [Abiteboul&al'10]
  - Typing [Cohen&al'09]
  - Updates [Kharlamov&al'10]

Properties of PrXML Model

- Trees represented by PrXML document T have bounded height & width:
  - height: at most the height of T
  - width: at most the width of T
- Number of represented XML documents is bounded:
  - at most exp. many in |T|



Properties of PrXML Model

- Trees represented by PrXML document T have bounded height & width:
  - height: at most the height of T
  - width: at most the width of T
- Number of represented XML documents is bounded:
  - at most exp. many in |T|

Mailbox DTD

mailbox:	$(\text{thread})^*$
thread:	(message, id, subject)
message:	$(\text{from}, \text{to}, \text{content}, \text{message}^*)$
from:	#PCDATA
to:	#PCDATA
content:	#PCDATA
subject:	#PCDATA

- Try to make a probabilistic model of a mailbox with PrXML:
  - Unbounded # of threads /messages ~ unbounded width / height of docs
  - The deeper the thread, the lower its probability

No chance with PrXML ⇒ we need models akin to probabilistic DTDs

Goal of This Work

- Identify
  limitations of existing probabilistic representation systems
  - key limitations: expressiveness and succinctness
- Develop
  systems that naturally capture other formalisms
  for representing classes of XML documents
  - E.g. DTDs or XML schemas
- Understand
  what properties of new systems allow query tractability

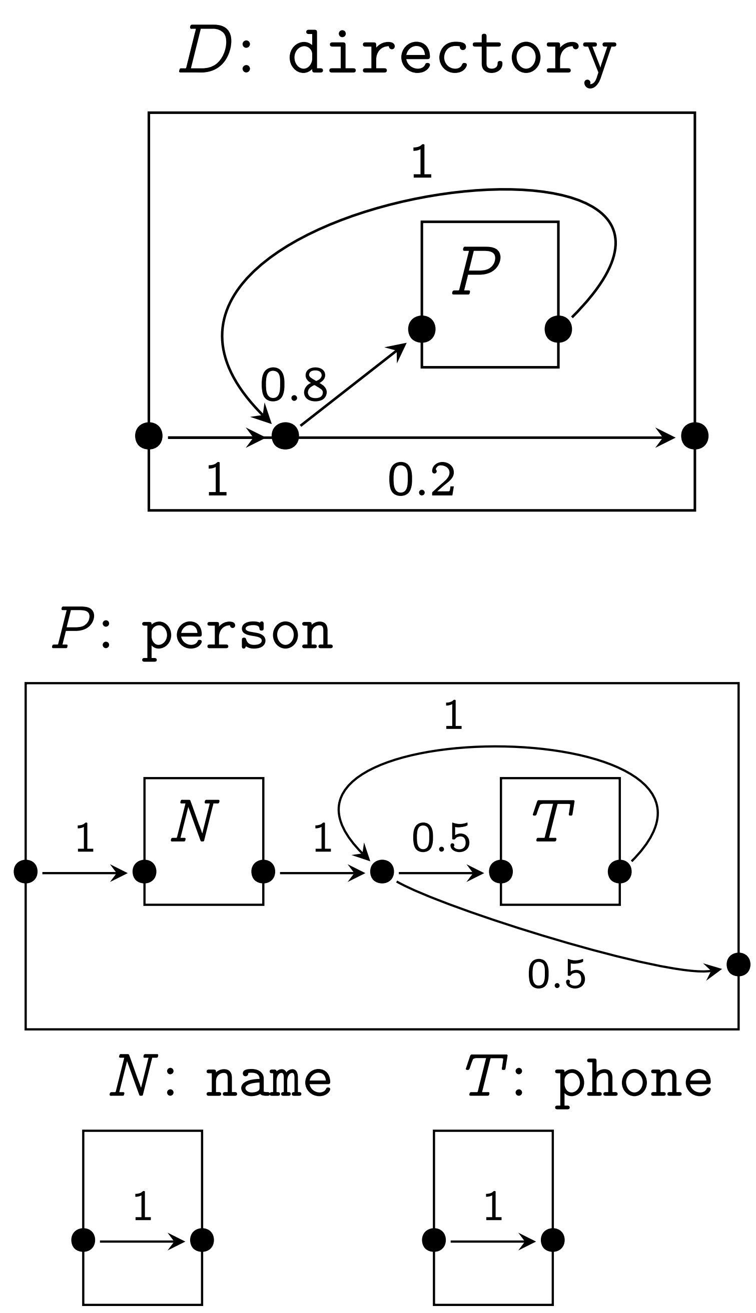
Outline

- Probabilistic Data and What We Want to Study
- Recursive Markov Chains (RMCs)
- Probabilistic XML via RMCs
- Querying RMCs

Recursive Markov Chains

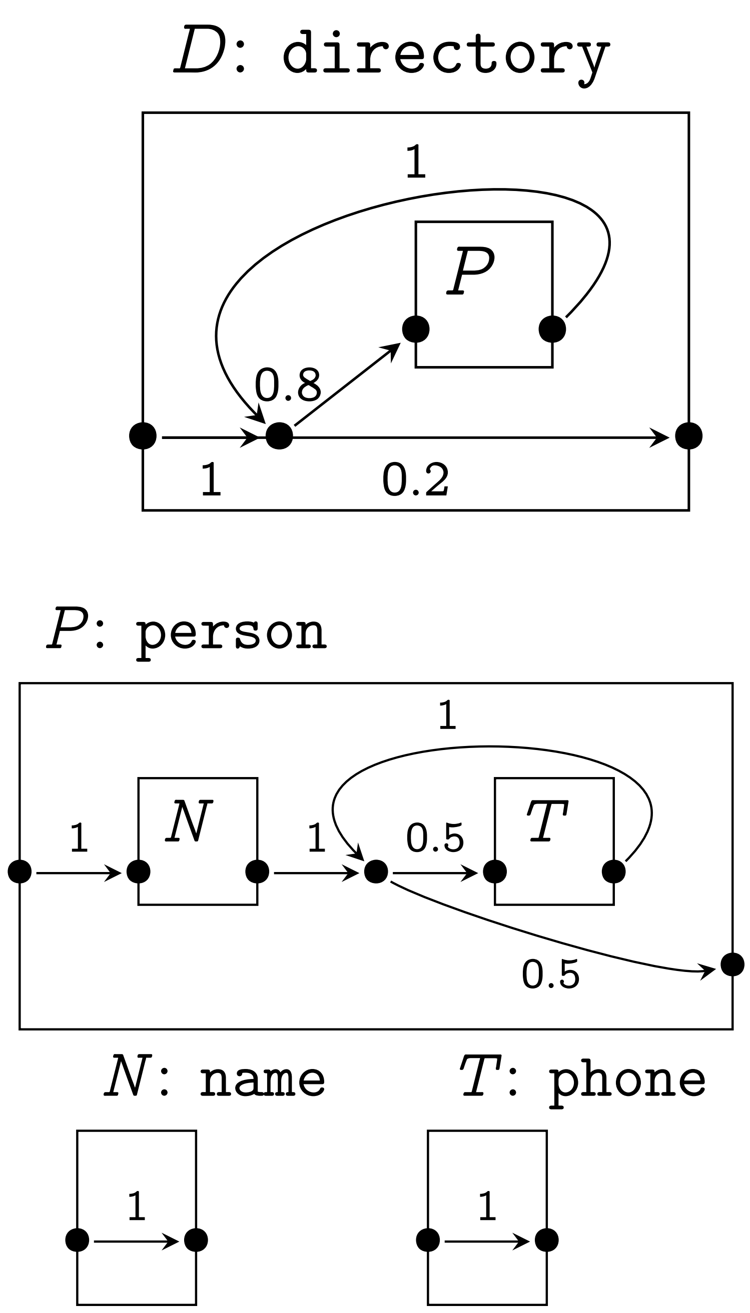
[Etessami, Yannakakis'09]

- Markov Chains
  - Graphs whose edges are labeled with probabilities
  - Define processes evolving via independent choices at nodes
- Recursive Markov Chains
  - Markov Chains with recursive calls
  - RMC runs have a natural hierarchical structure - nested words or trees



Recursive Markov Chains - Example

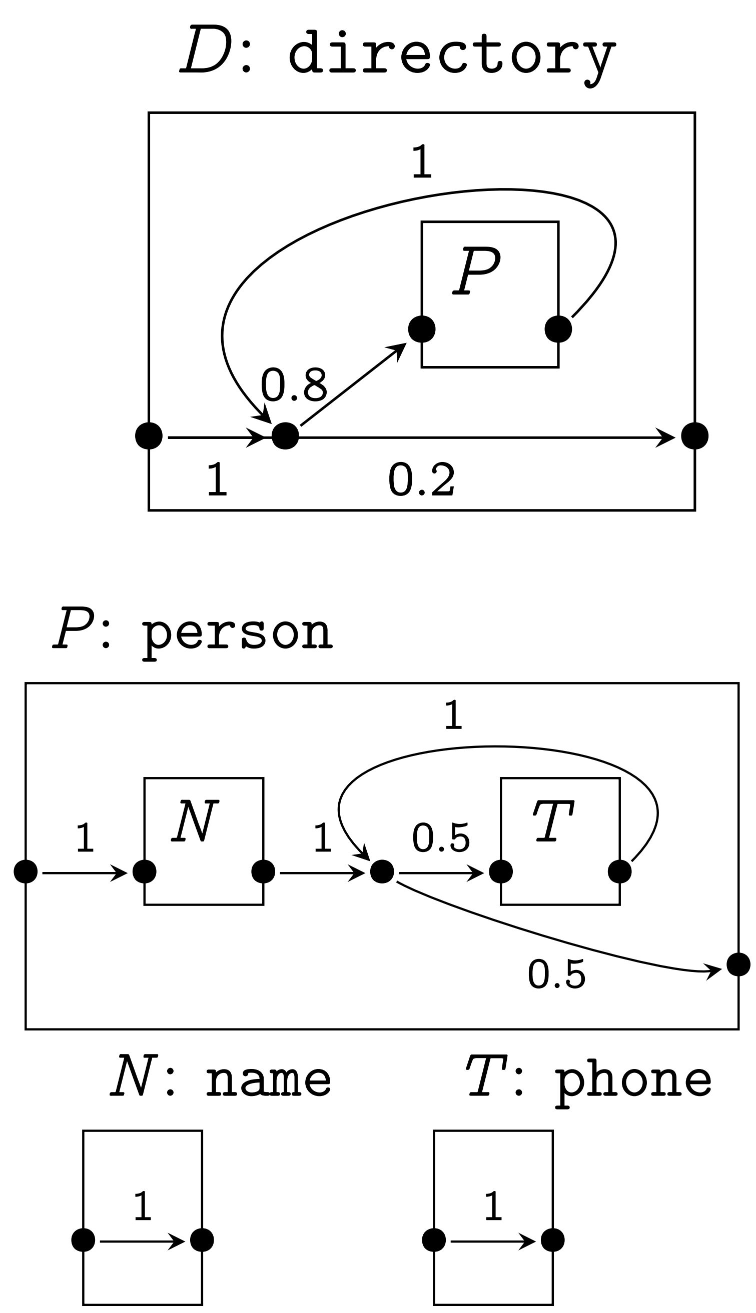
- RMC with four components D, P, N, and T
- Each component has
  - a label, e.g., "directory" is the label of D
  - nodes: entry, exit, call, return, others
  - boxes to simulate calls to other components, e.g., box P inside D
  - transitions $(u, p_{u,v}, v)$ from source u to destination v with probability $p_{u,v}$; For each source u: $$\sum_{\{v|(u,p_{u,v},v)\}} p_{u,v} = 1$$
- D is the start component, no calls to D are allowed.



Recursive Markov Chains - Applications

Variants of (R)MCs are well-understood and researched in

- Machine learning
  (e.g., hidden Markov models)
  [Bishop'06]
- Computational linguistics
  (e.g., stochastic CFGs)
  [Manning,Schuetze'99]
- Verification
  (e.g. probabilistic automata)
  [Kwiatkowska'03]



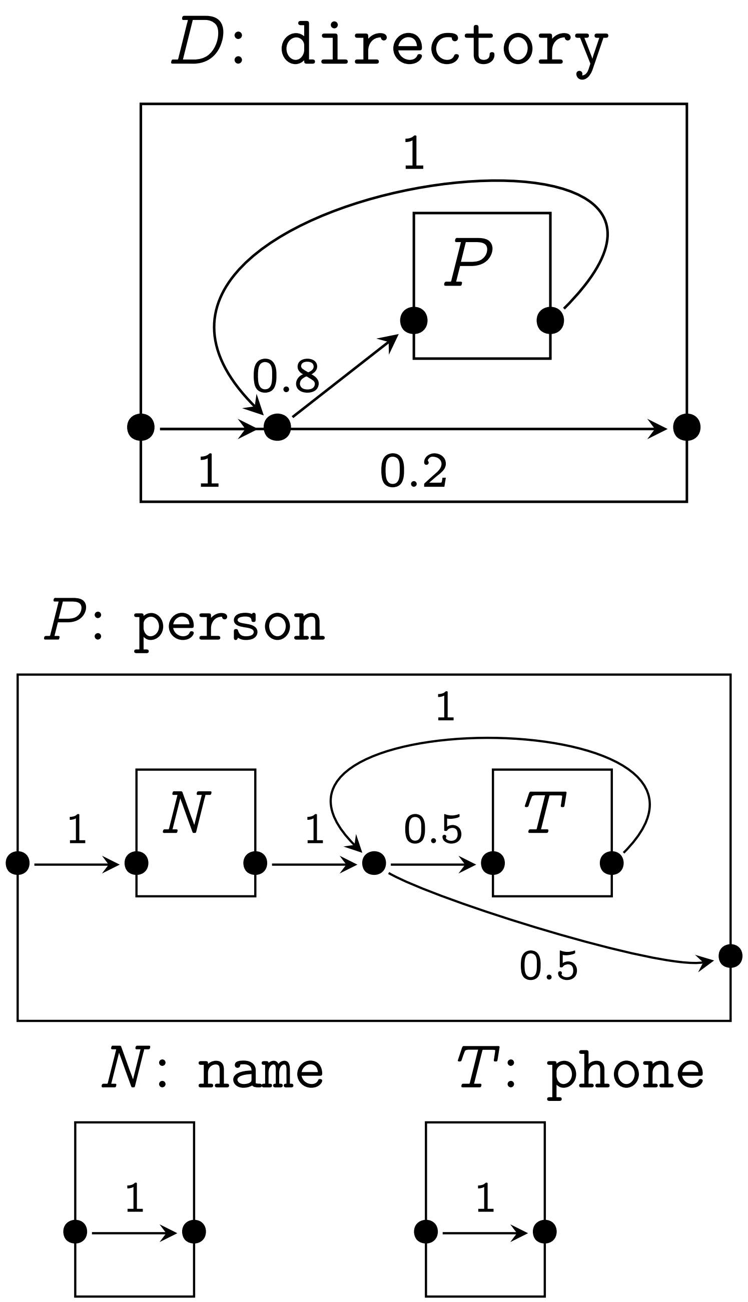
Outline

- Probabilistic Data and What We Want to Study
- Recursive Markov Chains (RMCs)
- Probabilistic XML via RMCs
- Querying RMCs

Recursive Markov Chains - Tree Generators

Document d	
<directory>	
<person>	$Pr = 1 \cdot 0.8$
<name>	$Pr = 1$
</name>	$Pr = 1$
<phone>	$Pr = 1 \cdot 0.5$
</phone>	$Pr = 1$
</person>	$Pr = 1 \cdot 0.5$
</directory>	$Pr = 1 \cdot 0.2$
Document d	$Pr(d) = 0.8 \cdot 0.5 \cdot 0.5 \cdot 0.2$

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

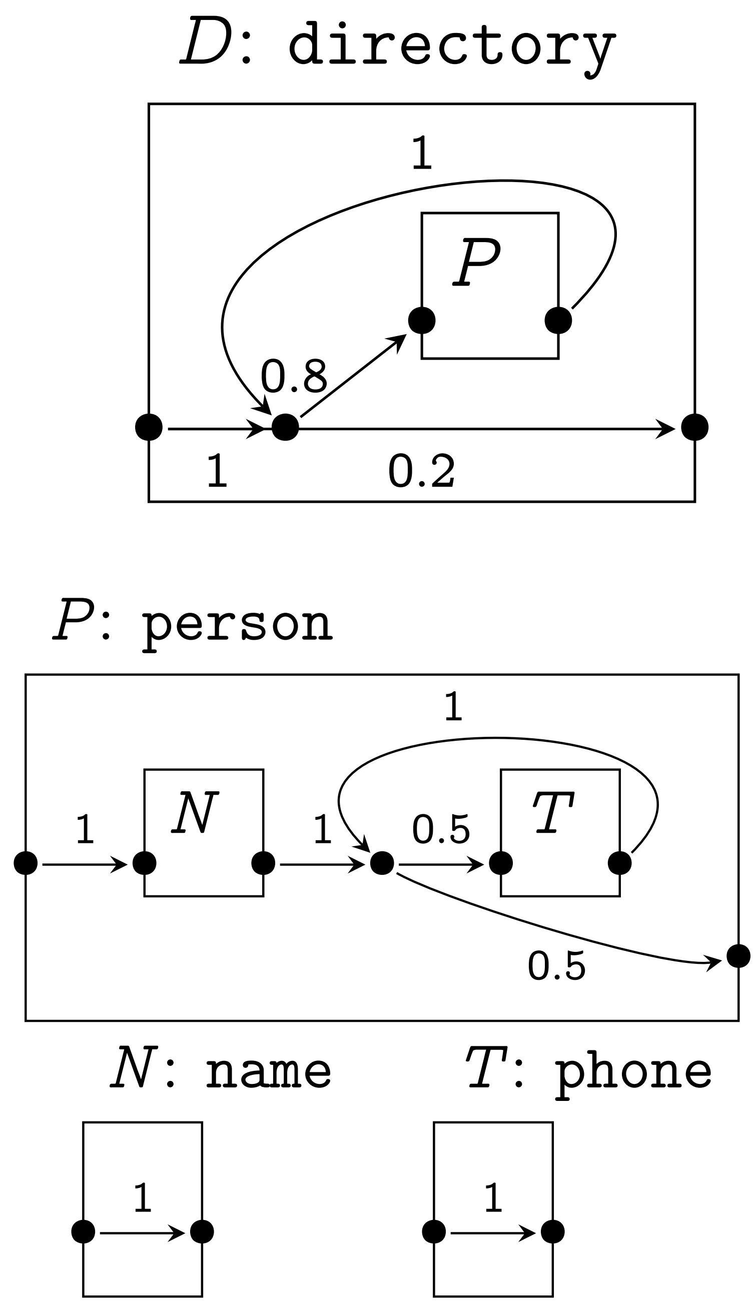


- Entering a component labeled L
  = generation of an opening tag <L>
- Exiting a component labeled L
  = generation of a closing tag </L>

Recursive Markov Chains - Tree Generators

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

Document d	$Pr(d) = 0.8 \cdot 0.5 \cdot 0.5 \cdot 0.2$
<directory>	
<person>	$Pr = 1 \cdot 0.8$
<name>	$Pr = 1$
</name>	$Pr = 1$
<phone>	$Pr = 1 \cdot 0.5$
</phone>	$Pr = 1$
</person>	$Pr = 1 \cdot 0.5$
</directory>	$Pr = 1 \cdot 0.2$



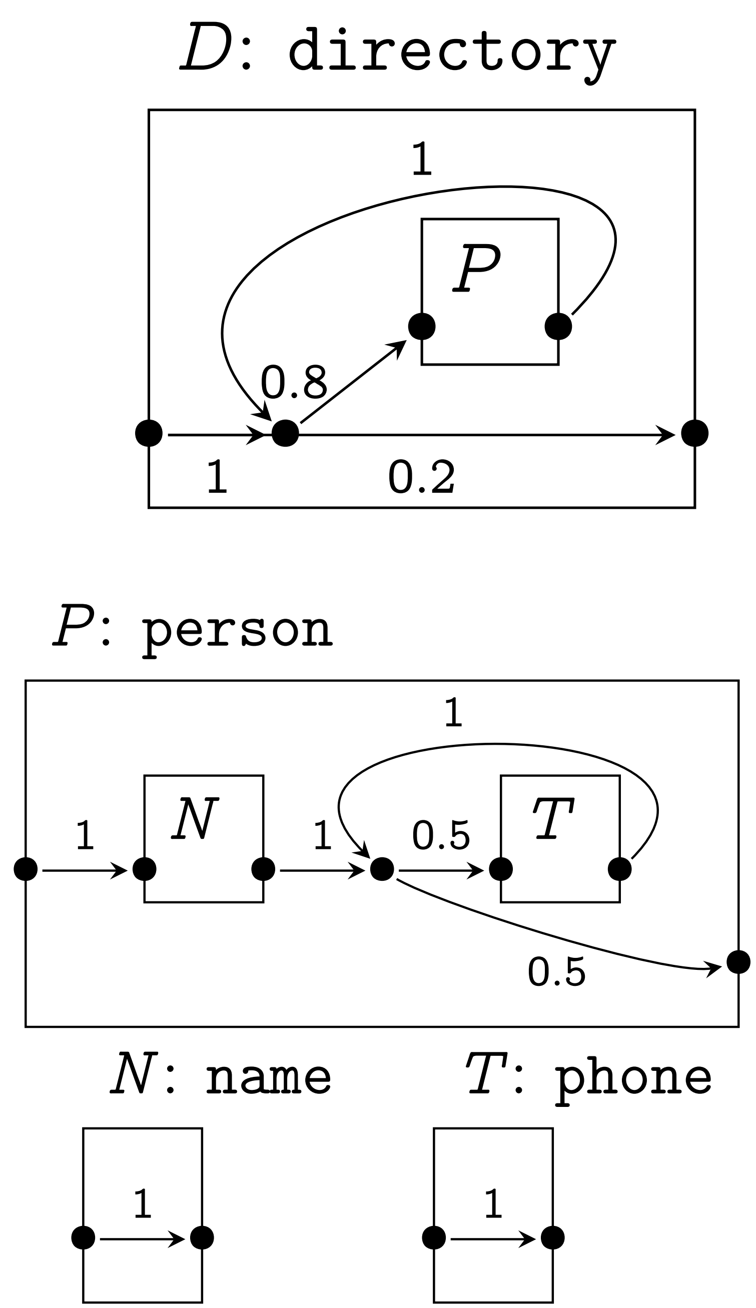
- A run generates a skeleton of a document
- Empty components N and D can model the actual data, i.e., names and telephone numbers of people

Recursive Markov Chains - Tree Generators

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

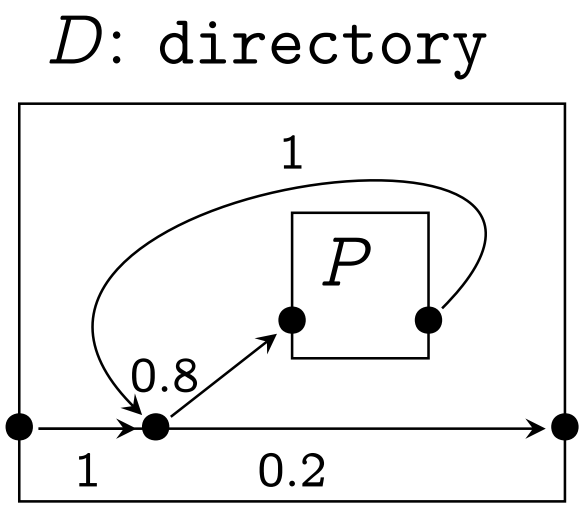
Document d	
<directory>	
<person>	$Pr = 1 \cdot 0.8$
<name>	$Pr = 1$
</name>	$Pr = 1$
<phone>	$Pr = 1 \cdot 0.5$
</phone>	$Pr = 1$
</person>	$Pr = 1 \cdot 0.5$
</directory>	$Pr = 1 \cdot 0.2$

Document d $\qquad Pr(d) = 0.8 \cdot 0.5 \cdot 0.5 \cdot 0.2$



- Advantages of RMCs over PrXML
  - More natural, e.g., akin to probabilistic DTDs
  - We connect questions on prob. XML to tools and techniques of Markov models

Probability Spaces of RMCs vs PrXML



- Size of generated documents:
  - RMC: could be
    - Unbounded width ~ cycles inside a component
    - Unbounded depth ~ cycles across components
  - PrXML: always
    linearly bounded by size of probabilistic document
- Probabilities of generated documents:

Comes from properties of RMCs

  - RMC: could be irrational, doubly exponentially small in the size of RMC
  - PrXML: always rational and at most exponentially small

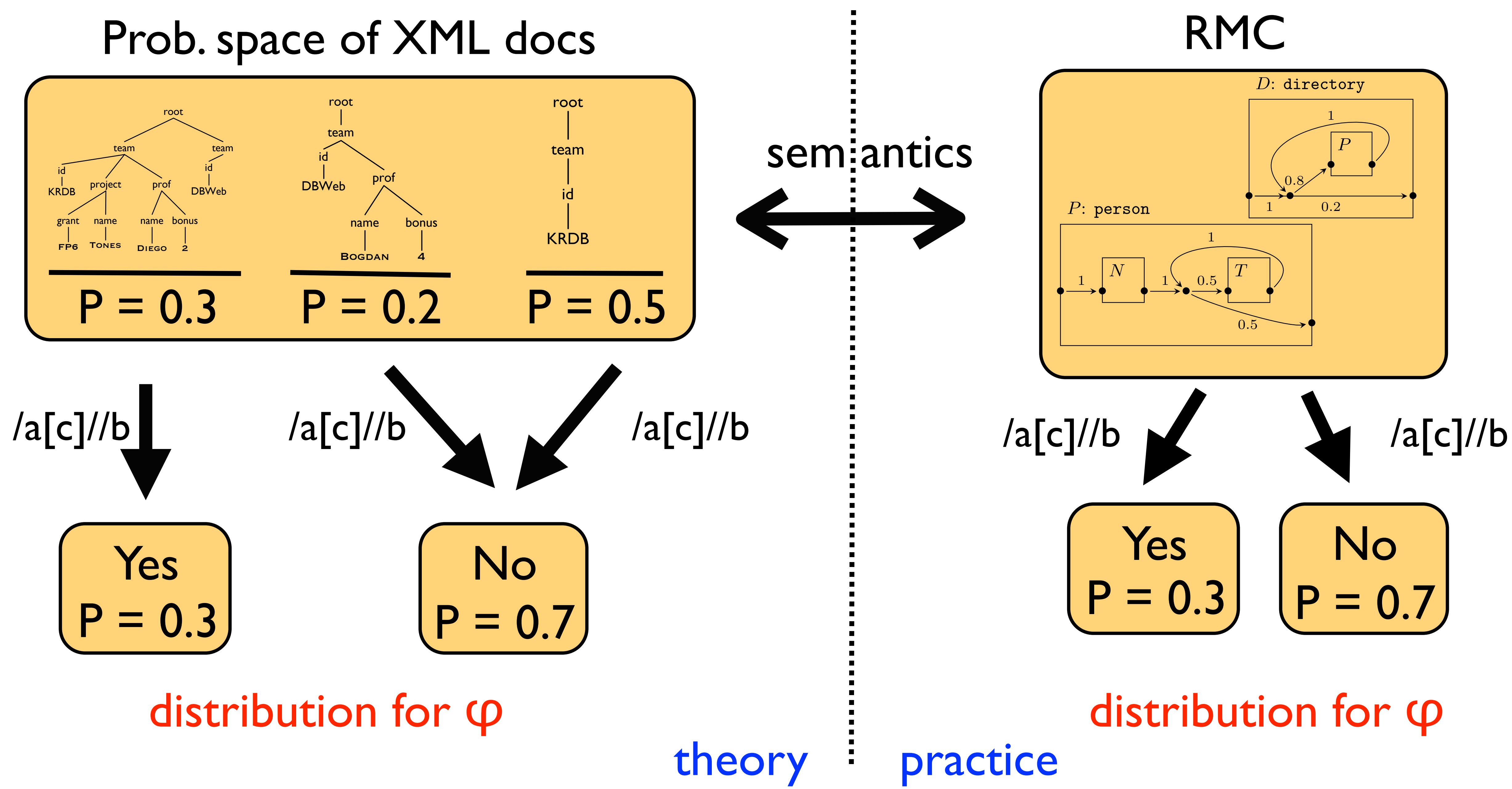
PrXML models with distributional nodes are subsumed by RMC

Outline

- Probabilistic Data and What We Want to Study
- Recursive Markov Chains (RMCs)
- Probabilistic XML via RMCs
- Querying RMCs

Querying RMC

Given: an RMC and a property, e.g., MSO formula, Boolean XPath query
Task: verify whether the RMC satisfies the property

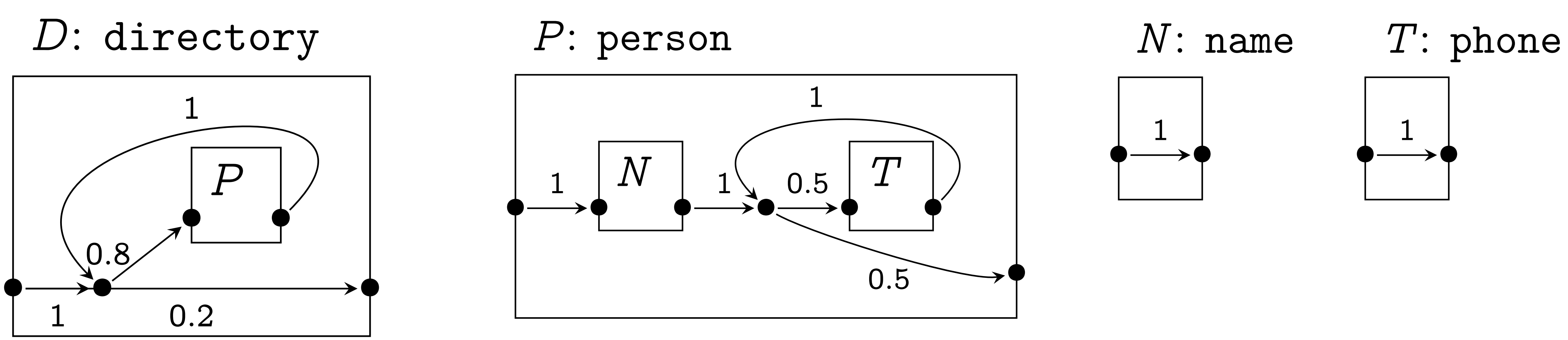


MSO Queries for RMCs

- Monadic Second Order (MSO) query language is very general
  - Subsumes: Tree-pattern queries, navigational XPath, ...
- Verifying MSO properties for unrestricted RMCs is
  - in PSPACE
  - as hard as SQRT-SUM: in PSPACE
    - lower bounds - long standing open problem
- We focus on RMC fragments to see the tension between
  - tractability of query evaluation ✓
  - expressiveness ✓
  - succinctness

RMC Fragments

- Hierarchical RMCs (HMC):
  - A component can not (eventually) call itself
- Tree-like RMCs (TLMC):
  - Every component can be called in one place only but possibly many times
  - special case of HMC



The directory RMC is in HMC and in TLMC

Tractability of RMC Fragments: TLMC

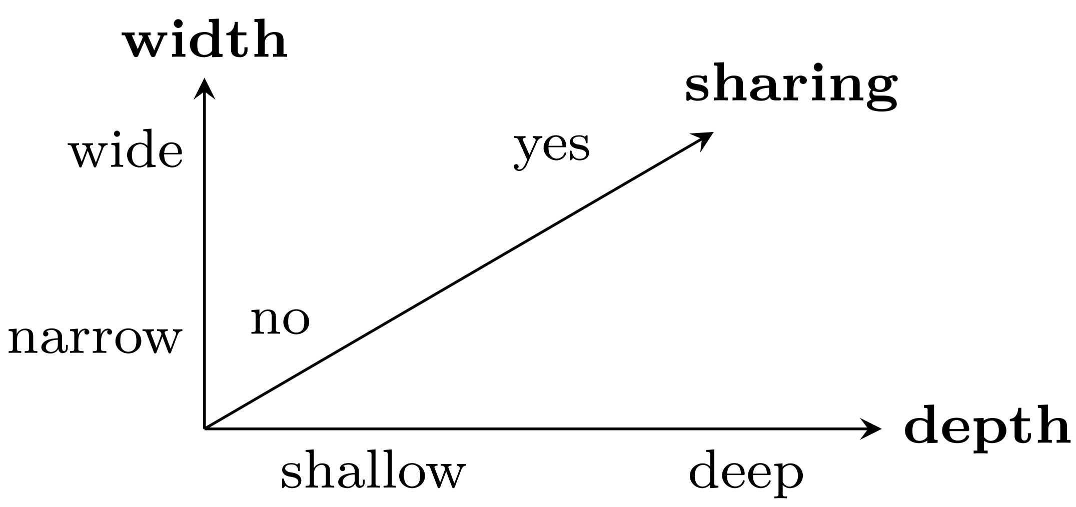
- Theorem:
  TLMC is tractable for MSO (in data complexity)
- Query evaluation algorithm : Given TLMC A and MSO φ
  - Pre-process TLMC:
    A ⇒ probabilistic push-down automaton (PPDA) B
  - Pre-process MSO:
    φ ⇒ tree automaton C (det. streaming tree automaton)
  - Compute a product PPDA automaton B×C — Computable in PTIME
  - Compute the termination probability for B×C — Computable in PTIME

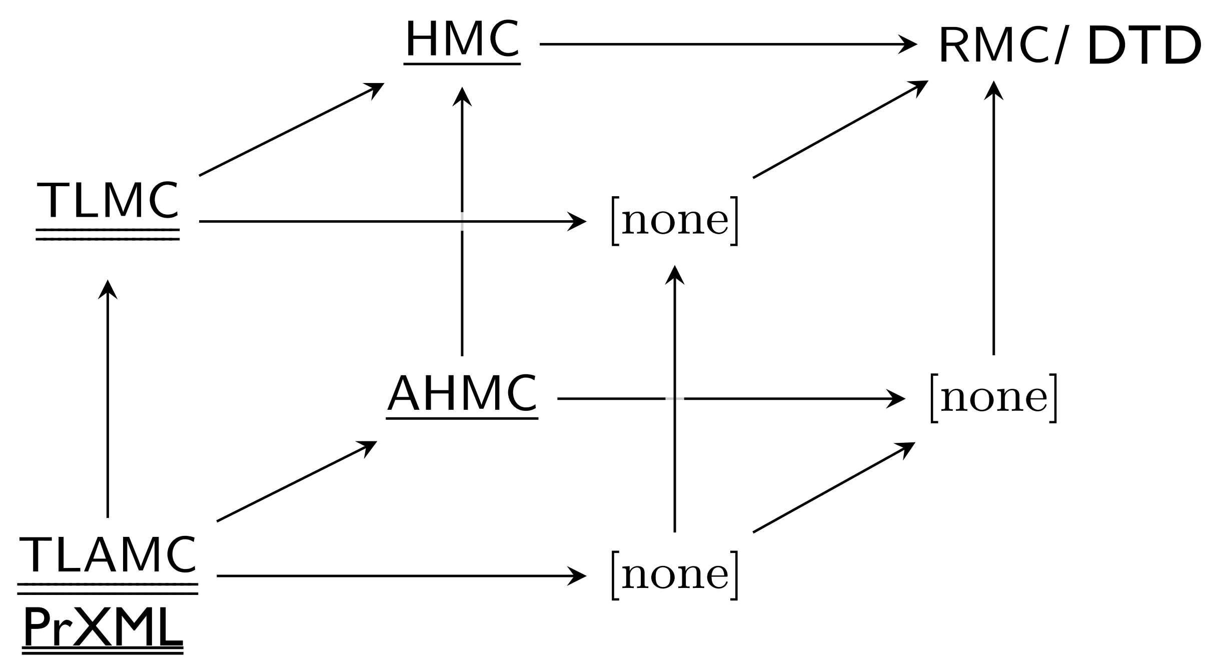
Probability that B×C terminates = Probability that φ holds in A

Tractability of RMC Fragments: HMC

- Theorem:
  HMC is ra-tractable for MSO (in data complexity)
- ra-tractability:
  - tractability in case of fixed-cost rational arithmetic
  - all arithmetic operations over rationals take unit time, no matter how large the numbers

Expressiveness of RMC Fragments

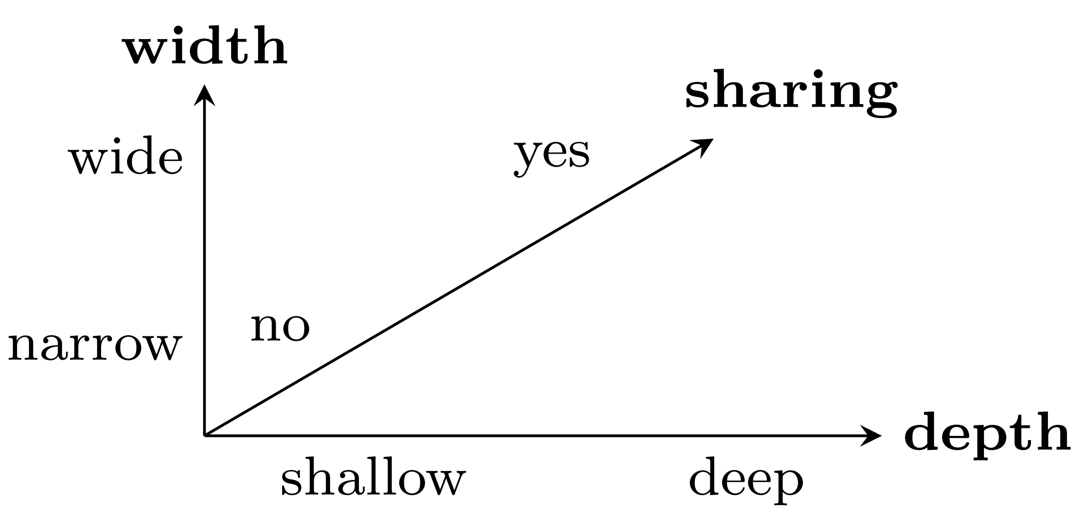


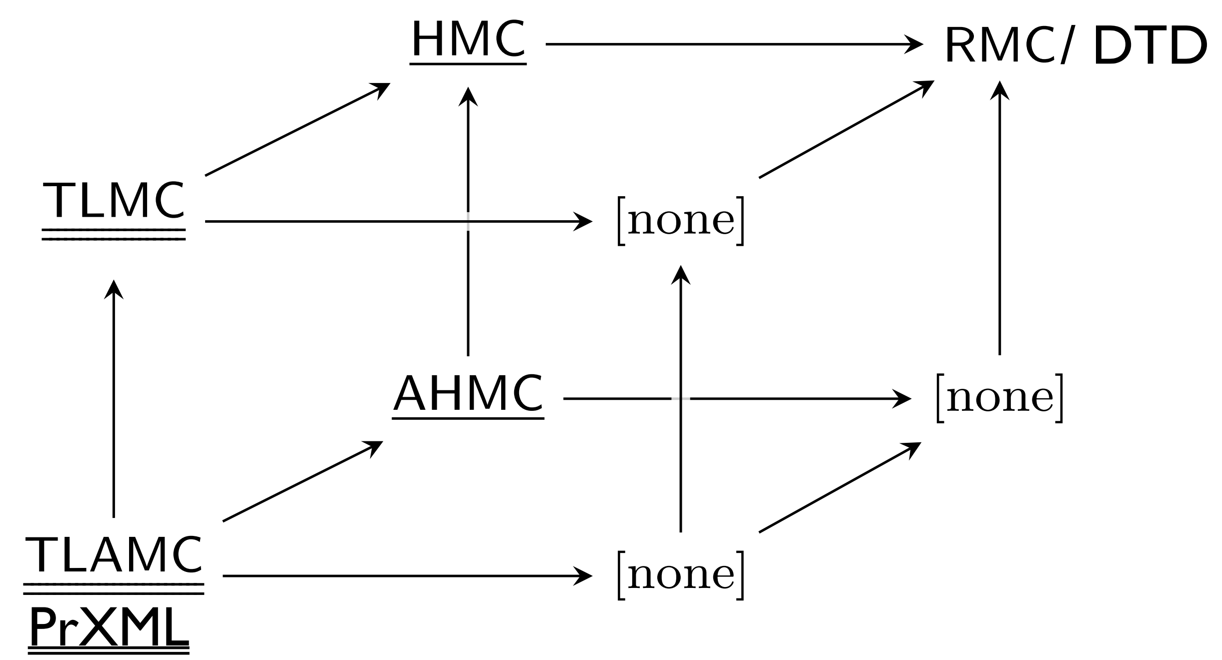


Three dimensions of expressiveness

- Width: wide vs. narrow
  Wide models: random trees of any width ~ recursion inside components
- Depth: deep vs. shallow
  Deep models: random trees of any depth ~ recursion across components
- Call sharing: yes vs. no
  Model with sharing: random trees with doubly exponentially many leaves
  ~ components can be called from multiple places

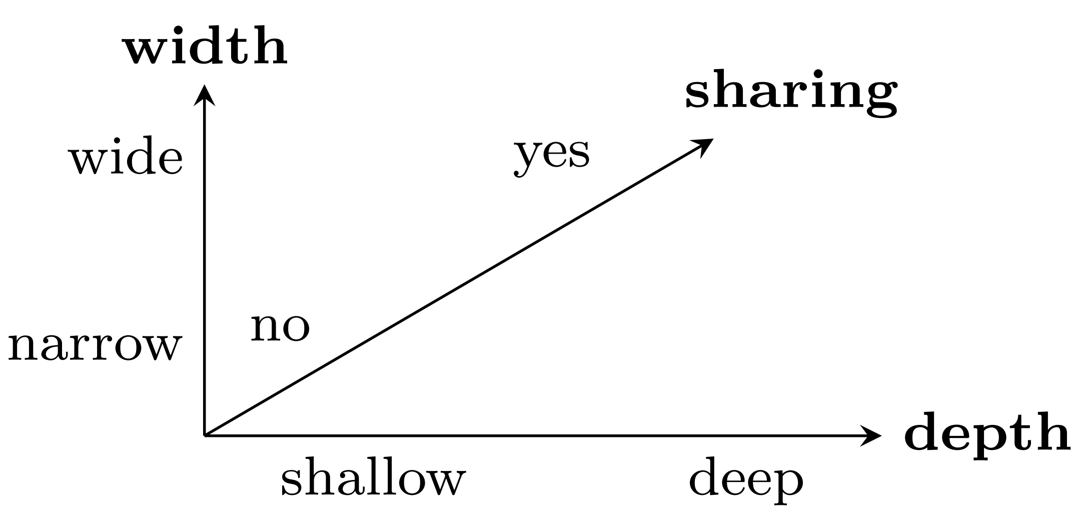
Expressiveness of RMC Fragments

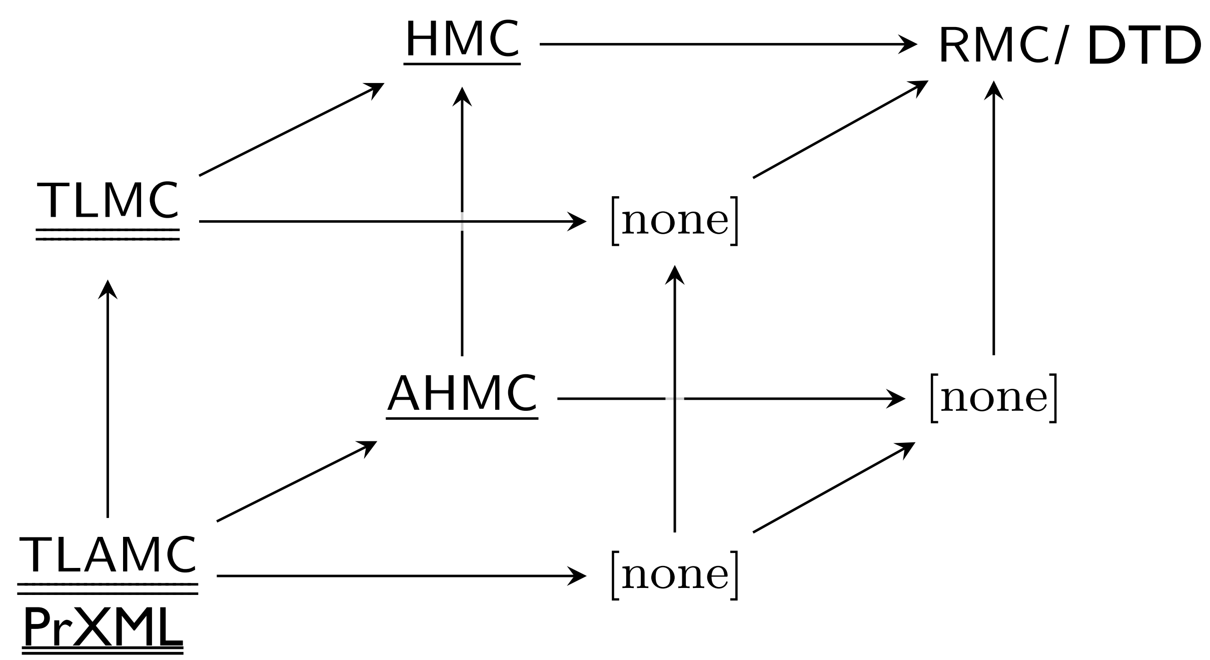




- [none] - no reasonable syntactic restriction for this class
- "A" = Acyclic.
  Each component is an acyclic graph

Expressiveness of RMC Fragments

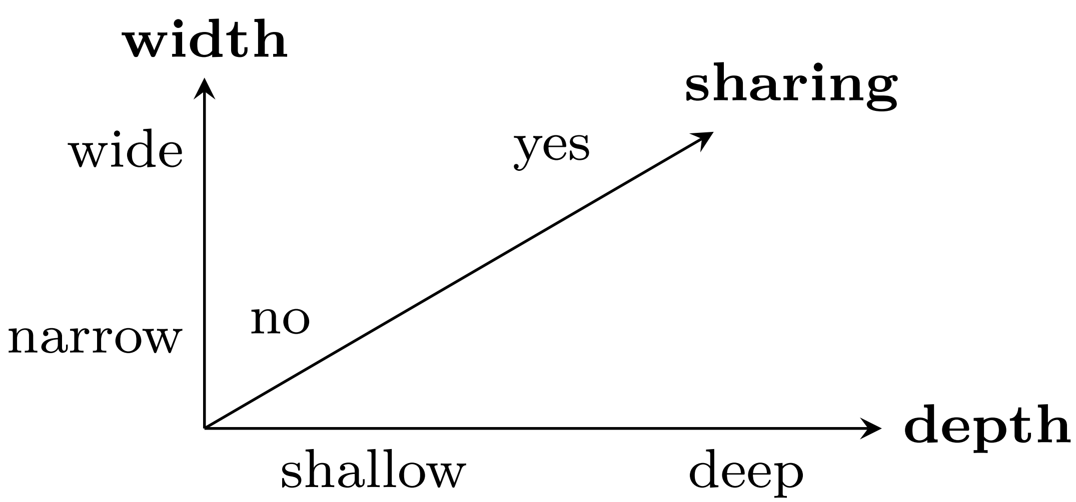


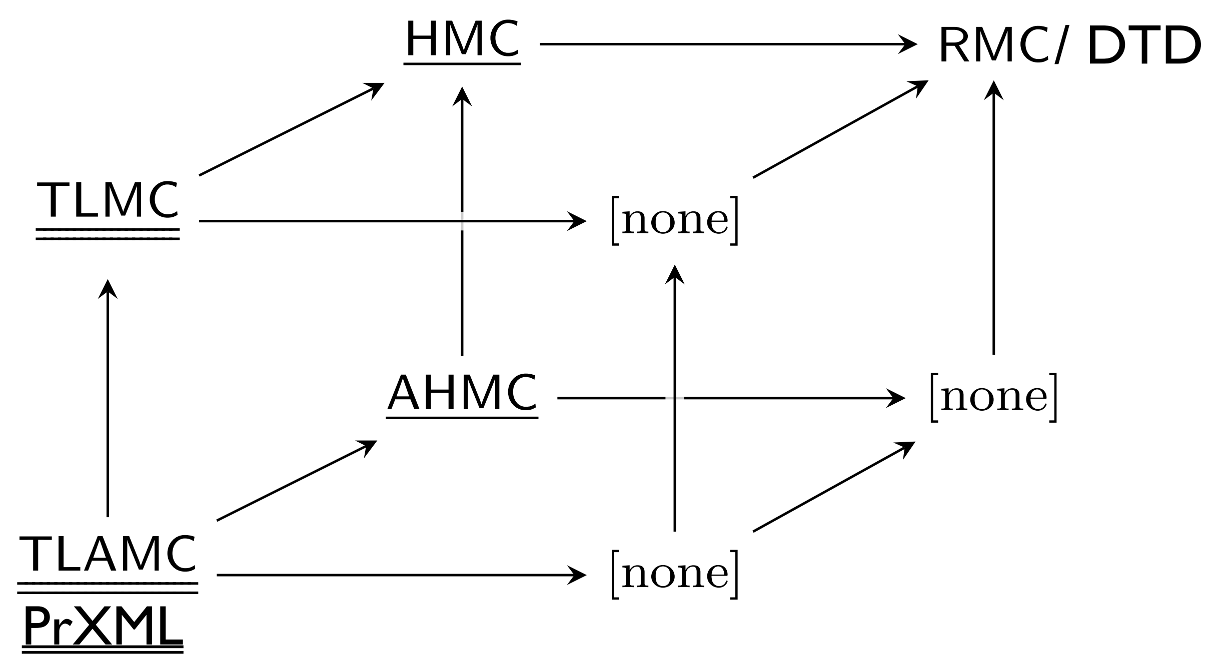


Existing PrXML models with distributional nodes:

- shallow, narrow, no sharing
- represent finite probability spaces only
- subsumed by TLAMC

Expressiveness vs Tractability

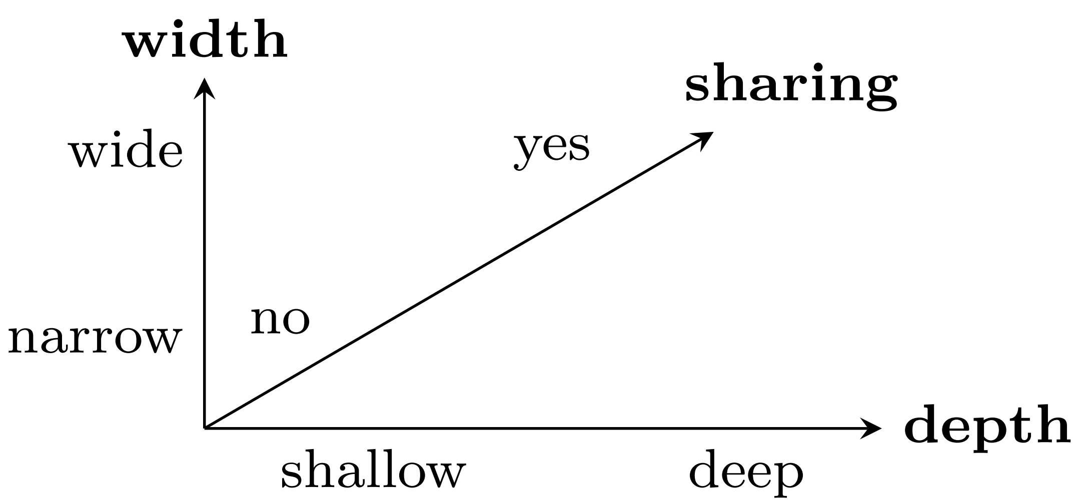


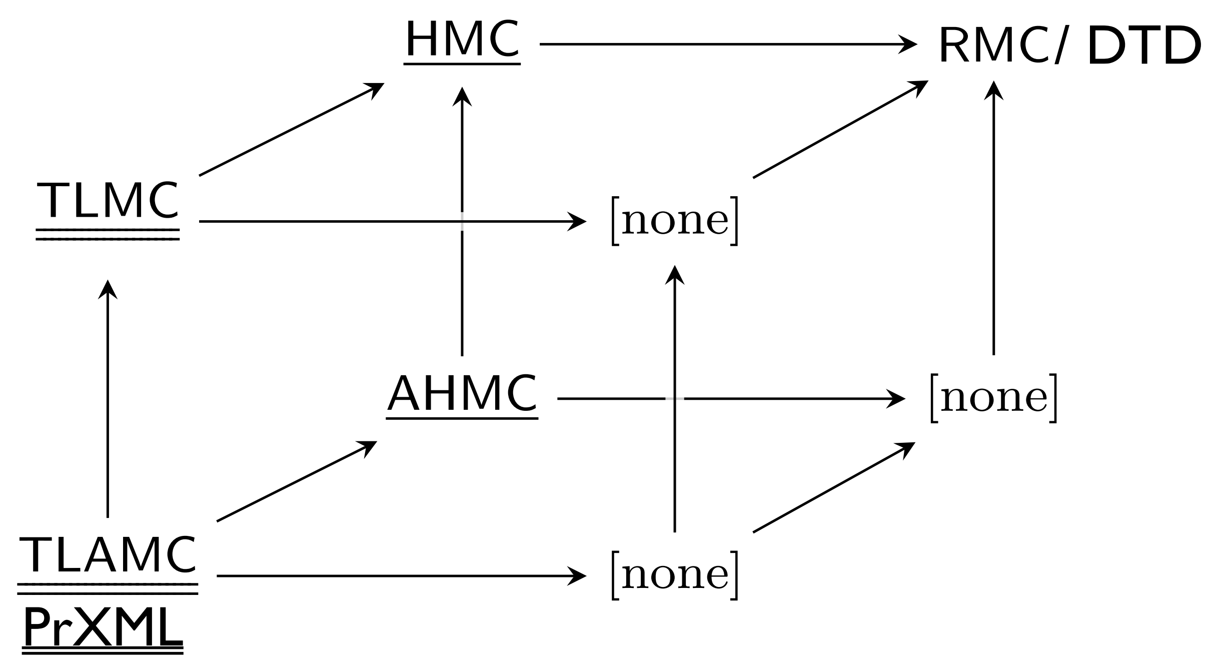


Tractability for MSO:

- double underlining ~ MSO evaluation is tractable
- single underlining ~ MSO evaluation is tractable under unit cost arithmetic
- no underlining ~ MSO evaluation is SQRT-SUM hard

Expressiveness vs Tractability

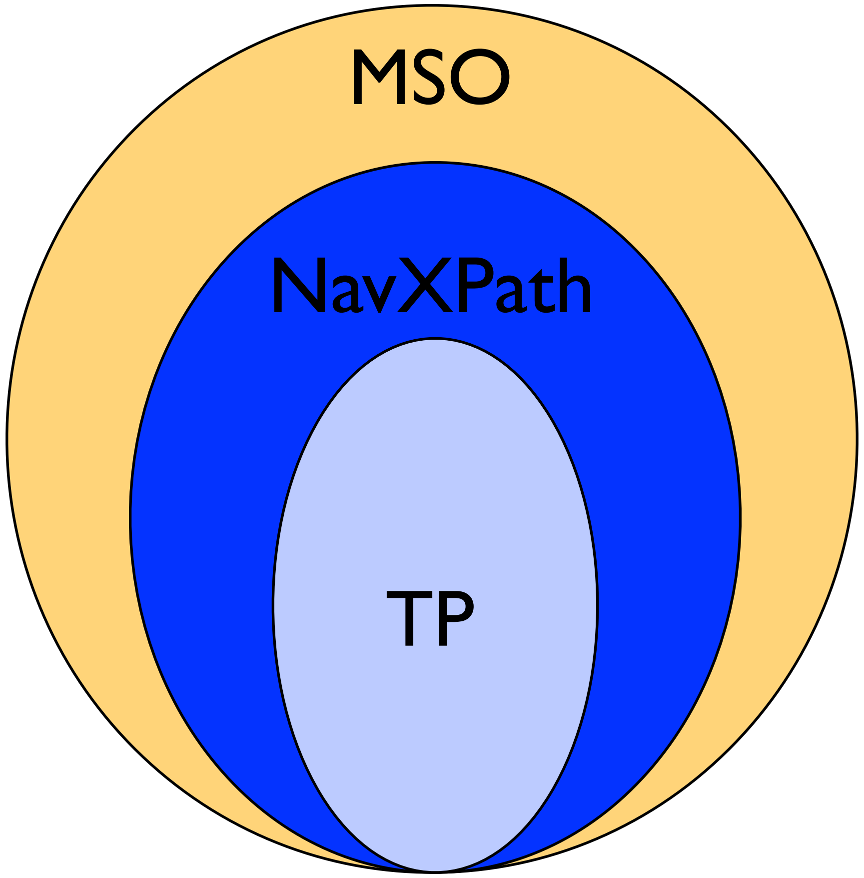




- Gain in width - no influence on tractability
- Gain in depth - loss in tractability
- Allowing sharing
  - tractability degrades to unit-cost arithmetics tractability

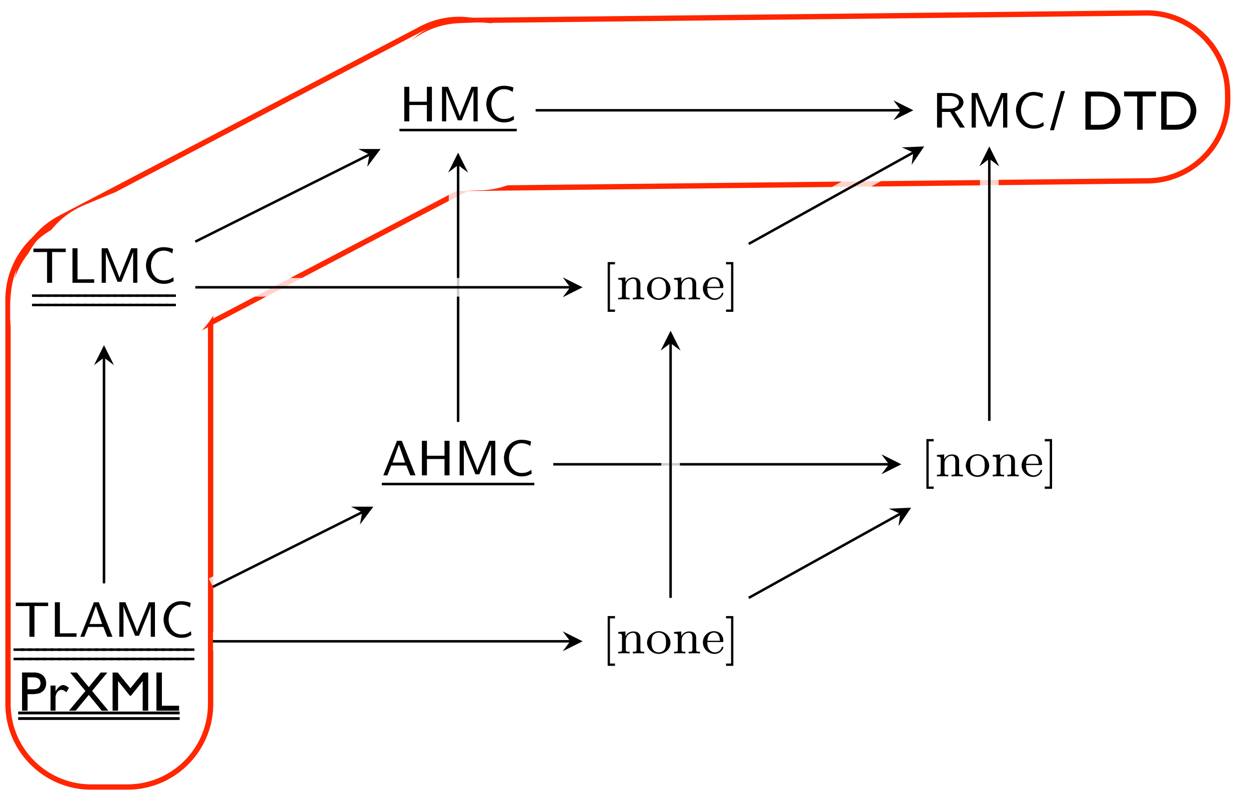
Combined Complexity of MSO Evaluation

- Tree-patterns over PrXML with distributional nodes: $FP^{\#P}$-complete
- Navigational XPath over AHMC: in PSACE
- MSO over
  - PrXML with distributional nodes and TLAMC: PSPACE-complete
  - AHMC: #EXP-hard and in EXPSPACE
  - Wide models, e.g., TLMC: even deciding whether an MSO query has a probability > 0 is not elementary



Conclusion

- We adopted a very general RMC model for probabilistic XML. RMC
  - Mimics DTDs with probabilities
  - Extends classical PrXML model with distributional nodes
- We studied
  - space of models between PrXML and RMC
  - complexity of MSO query answering



DataRing Project:
P2P Data Sharing for Online Communities
http://www.lina.univ-nantes.fr/projets/DataRing/

FOX Project: Foundations of XML
FP7-ICT-233599
http://fox7.eu/

ONTORULE: ONTOlogies meet business RULes
FP7-ICT-231875
http://ontorule-project.eu/

Webdam Project: Foundations of Data Management
ERC-FP7-226513
http://webdam.inria.fr

British EPSRC grant EP/G004021/1

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'09] - Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, Pierre Senellart: On the expressiveness of probabilistic XML models. VLDB J. 18(5): 1041-1064 (2009)
- [Antova&al'07] - Lyublena Antova, Christoph Koch, Dan Olteanu: 10106 Worlds and Beyond: Efficient Representation and Processing of Incomplete Information. ICDE 2007: 606-615
- [Bishop'06] - C. M. Bishop (2006), Pattern Recognition and Machine Learning.
- [Cohen&al'09] -Sara Cohen, Benny Kimelfeld, Yehoshua Sagiv: Running tree automata on probabilistic XML. PODS 2009: 227-236
- [Cohen&al'09] - S. Cohen, B, Kimelfeld, Y. Sagiv: Incorporating constraints in probabilistic XML. ACM Trans. Database Syst. 34(3): (2009)

References

- [Etessami&Yannakakis'05] - K. Etessami, M. Yannakakis: Recursive Markov Chains, Stochastic Grammars, and Monotone Systems of Nonlinear Equations. STACS 2005
- [Etessami'06] - Slides of talks at Dagstuhl. Available at http://homepages.inf.ed.ac.uk/kousha/etessami_wamt_tutorial.pdf
- [Kimelfeld&al'09] - Benny Kimelfeld, Yuri Kosharovsky, Yehoshua Sagiv: Query evaluation over probabilistic XML. VLDB J. 18(5): 1117-1140 (2009)
- [Kharlamov&al'10] - Evgeny Kharlamov, Werner Nutt, Pierre Senellart: Updating probabilistic XML. EDBT/ICDT Workshops 2010
- [Kwiatkowska'03]- M. Z. Kwiatkowska: Model checking for probability and time: from theory to practice. LICS 2003
- [Manning,Schuetze'99] - Christopher D. Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
- [Senellart&al'07] - P. Senellart, S. Abiteboul: On the complexity of managing

References

- [Senellart&al'07] - P. Senellart, S. Abiteboul: On the complexity of managing probabilistic XML data. PODS 2007