Master Internship 2014

# Querying Probabilitistic Data via Tree Decompositions

Supervisor

Prof. Pierre Senellart, Télécom ParisTech & National University of Singapore – pierre.senellart@telecom-paristech.fr

## Presentation of the project

Probabilistic databases are compact representations of probability distributions over regular databases. A number of models have been proposed for probabilistic data, both relational [7] and XML [4]. Evaluating a Boolean query over such a probabilistic database means computing the probability that the query is true in the probability distribution represented by the database. While query evaluation is usually tractable on regular databases, evaluating queries in this sense on probabilistic databases is often intractable.

A number of research works have looked at characteristics of queries that can make them tractable. For instance, queries without self-joins are tractable over tuple-independent databases if and only if they are hierarchical [2], while tree-pattern queries on XML data with a single join are tractable if and only if they are equivalent to a join-free query [3].

By contrast, our recent work [1] has shown that, as long as the data and probabilistic correlations jointly have bounded treewidth [6] in a certain sense, query evaluation of monadic second-order queries remains tractable. This result is, however, mostly of theoretical interest. We have not investigated the extent to which real-world probabilistic data can be modeled with bounded treewidth databases, or whether the tree-automata constructions from [1] can be effectively used for real applications. Another of our recent work [5] has shown that, even when the data does not have bounded treewidth, partial tree decompositions may help query evaluation.

The objective of this internship is to explore concrete applications of the results of [1], perhaps inspired by partial decompositions as in [5], on real-world uncertain datasets. This may include the study of theoretical problems left open in [1] that are relevant for practical implementation: e.g., extending the constructions of this work to on-the-fly variants. These techniques should then be implemented on concrete query classes, perhaps with the help of MONA[1], and evaluated on applications (e.g., routing in transportation networks with uncertain delays).

Image & Pervasive Access Lab
1 Fusionopolis Way
#21-01 Connexis, South Tower
Singapore 138632

Tel. (65) 6408 2542
Director. (65) 6408 2536
Fax. (65) 6776 1378

secretariat@ipal.cnrs.fr
www.ipal.cnrs.fr

1 http://www.brics.dk/mona/index.html

## Expected deliverables

This internship will lead to the implementation of tree decompositions for probabilistic query evaluations. Experiment will be run to test the viability of the approach, and results will be published, with target a major conference or journal in the data management area.

## Keywords

Probabilistic Data, Tree Automata, Bounded Treewidth, Tree Decompositions.

## Applicant profile

- Currently enrolled in a Master's programme in Computer Science, or in an engineering school
- Strong background in theoretical computer science (automata theory, complexity, logics)
- Programming skills
- Strong motivation towards this challenging project.
- Availability for 4 to 6 months starting in the first semester of 2014.

Gratification: About 800€ net per month

## References

[1] A. Amarilli, P. Bourhis, and P. Senellart. Probabilities and provenance via tree decompositions, Oct. 2014. Preprint available at http://pierre.senellart.com/publications/amarilli2015probabilities.pdf.
[2] N. N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. VLDB Journal, 16(4), 2007.
[3] E. Kharlamov, W. Nutt, and P. Senellart. Value joins are expensive over (probabilistic) XML. In LID. http://pierre.senellart.com/publications/kharlamov2011value.pdf.
[4] B. Kimelfeld and P. Senellart. Probabilistic XML: Models and complexity. In Z. Ma and L. Yan, editors, Advances in Probabilistic Databases for Uncertain Information Management. Springer-Verlag, May 2013. http://pierre.senellart.com/publications/kimelfeld2013probabilistic.pdf.
[5] S. Maniu, R. Cheng, and P. Senellart. ProbTree: A query-efficient representation of probabilistic graphs. In Proc. BUDA, Snowbird, USA, June 2014. Workshop without formal proceedings. http://pierre.senellart.com/publications/maniu2014probtree.pdf .
[6] N. Robertson and P. D. Seymour. Graph minors. III. Planar tree-width. Journal of Combinatorial Theory, Series B, 36(1), 1984.
[7] D. Suciu, D. Olteanu, C. Ré, and C. Koch. Probabilistic Databases. Morgan & Claypool, 2011.

Image & Pervasive Access Lab
1 Fusionopolis Way
#21-01 Connexis, South Tower
Singapore 138632

Tel. (65) 6408 2542
Director. (65) 6408 2536
Fax. (65) 6776 1378

secretariat@ipal.cnrs.fr
www.ipal.cnrs.fr