MSc Internship

Querying Probabilitistic Data via Tree Decompositions

Pierre Senellart

Télécom ParisTech & National University of Singapore

Topic description

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).

Supervision and Environment

This Master's internship will have a duration of between 4 and 6 months and will be supervised by Pierre Senellart[2], professor at Télécom ParisTech and senior research fellow at the National University of Singapore (within IPAL, a joint French–Singaporean lab), with the help of Antoine Amarilli[3], PhD candidate at Télécom ParisTech.

The student may be based in Paris, in Singapore, or partly in both locations (e.g., 2 months in Singapore and 3 months in Paris), to be discussed before the internship depending on the student's preference and the best opportunities to organize the research.

[1] `http://www.brics.dk/mona/index.html`
[2] `http://pierre.senellart.com/`
[3] `http://a3nm.net/`

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*. Available online: `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. Available online: `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. Available online: `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.