MSc Internship

Characterizing Tractable Queries and Probabilistic Instance Families

Antoine Amarilli

Télécom ParisTech

Topic description

A probabilistic database is a compact representation of a probability distribution over a regular database. A number of models have been proposed for probabilistic data, both relational [8] and XML [7]; see e.g. [4] for a gentle introduction. The simplest model is that of *tuple-independent databases* (TID), where each database fact is annotated with a probability in $[0, 1]$ of being present or absent, and where we assume independence across all facts.

Evaluating a Boolean query over a TID database means computing exactly the probability that the query is true in the probability distribution represented by the database: the complexity of this task is measured *as a function of the data*, i.e., of the input probabilistic database, with the query being fixed. Surprisingly, although query evaluation is usually tractable on regular databases, evaluating queries in this sense on TID is often intractable.

Several lines of research have tried to understand this intractability. The first line of research has looked at characteristics of *queries* that can make them tractable over *all TID instances*. For instance, queries without self-joins are tractable over TID instances if and only if they are hierarchical [5]. This has led to a dichotomy result on the complexity of unions of conjunctive queries on arbitrary TID instances [6].

In a second line of research, we have investigated the characteristics of *instance families* that make them tractable when fixing an *arbitrary query* in a sufficiently expressive language. We showed that, on families of TID instances whose *treewidth* is bounded by a constant, Boolean monadic first-order queries are tractable to evaluate [2]. We also showed a converse result: there are first-order queries whose evaluation is intractable on *any* unbounded treewidth instance family, provided that the family has arity two and is constructible in some sense [3]. Hence, bounded treewidth is the right condition to make instances tractable for probabilistic query evaluation.

However, these two lines of work do not classify completely the landscape of tractable queries and instances. Indeed, when we fix a family of queries and a family of instances, it may be the case that probabilistic query evaluation is tractable even though the queries are unsafe and the instances have unbounded treewidth. This may happen for trivial reasons (e.g., the queries and instances are using a disjoint vocabulary), but it may also happen for more interesting reasons (e.g., the query does not "see" the unbounded-treewidth correlations in the instance family).

The goal of this internship is to achieve a deeper understanding of the complexity of probabilistic query evaluation, and to obtain tractability and intractability results that take into account the shape of both the queries and the instance families. One possible direction is to investigate the *lineage* of the queries on the instances, and show intractability when this lineage is sufficiently complex, e.g., when it has unbounded treewidth even after it has been simplified by applying the *unfolding* constructions of [3]. Another possible direction is to extend the OBDD lower bounds of [3] to broader query classes or to more expressive lineage representations.

More generally, the internship will give the candidate an opportunity to investigate many other problems related to our ongoing study of the complexity of query evaluation on probabilistic treelike data, e.g., extensions of our result on efficient compilation of queries to tree automata [1], of our ongoing work about efficient *combined complexity* for probabilistic query evaluation tasks, etc.

Supervision and Environment

This Master's internship is intended for students with a solid background in theoretical computer science. Familiarity with database theory is appreciated but not required, as the basic notions can be learned during the internship.

The internship will have a duration of between 4 and 6 months and will be supervised by Antoine Amarilli[1], maître de conférences at Télécom ParisTech. The internship is in the context of an ongoing collaboration between Antoine Amarilli and Pierre Senellart[2], Professor at École normale supérieure; Pierre Senellart will also be involved in the research.

The internship will be based at Télécom ParisTech, in Paris (13th district), in the DBWeb team[3].

References

[1] A. Amarilli, P. Bourhis, M. Monet, and P. Senellart. Combined tractability of query evaluation via tree automata and cycluits. In *Proc. ICDT*, 2017. `http://pierre.senellart.com/publications/amarilli2017combined.pdf`. To appear.

[2] A. Amarilli, P. Bourhis, and P. Senellart. Provenance circuits for trees and treelike instances. In *Proc. ICALP*, 2015. `https://arxiv.org/abs/1511.08723`.

[3] A. Amarilli, P. Bourhis, and P. Senellart. Tractable lineages on treelike instances: Limits and extensions. In *Proc. PODS*, 2016. `https://a3nm.net/publications/amarilli2016tractable.pdf`.

[4] A. Amarilli and P. Senellart. On the connections between relational and XML probabilistic data models. In *Proc. BNCOD*, 2013. `http://pierre.senellart.com/publications/amarilli2013connections.pdf`.

[5] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *VLDB Journal*, 2007. `http://www.vldb.org/conf/2004/RS22P1.PDF`.

[6] N. Dalvi and D. Suciu. The dichotomy of probabilistic inference for unions of conjunctive queries. *J. ACM*, 2012. `https://homes.cs.washington.edu/~suciu/jacm-dichotomy.pdf`.

[7] 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`.

[8] D. Suciu, D. Olteanu, C. Ré, and C. Koch. *Probabilistic Databases*. Morgan & Claypool, 2011.

[1] `https://a3nm.net/`

[2] `http://pierre.senellart.com/`

[3] `http://dbweb.enst.fr/`