Query-Directed Width Measures for Probabilistic Databases

PhD topic proposal, Inria Lille, team LINKS

Advisors: Antoine Amarilli, Mikaël Monet, Sylvain Salvati

Abstract: We are interested in designing algorithms for *query evaluation over uncertain data*, where the typical tasks are, given as input a query and a probabilistic database, to compute the probability that the data satisfies the query, or to compute provenance information that can help explain the query results. The complexity of these problem has so far mostly been studied from two different angles: the first one is to fix the query and to not restrict the shape of the data, while the second one is to restrict to databases of a certain shape (e.g., bounded treewidth), which allows for more queries to be tractable. The goal of this PhD is to develop algorithms and techniques that combine both approaches.

Keywords: Probabilistic databases, Treewidth, Logic, Complexity theory

1 Topic presentation

Context. Data in real life is rarely certain: it can be subject to human error when manually collected, or generated from physical sensors having imperfect precision, or extracted from sources of various quality from the Web via natural language processing (NLP) techniques (that are often themselves of a probabilistic nature), and so on. The traditional way of dealing with this uncertainty is to simply ignore it, for instance, by removing all rows containing NULL values in a database before evaluating a query. This naive approach can however lead to incorrect or incomplete answers and is thus not always desirable. *Uncertain data management* is a field of research that aims at developing models and algorithms that take this uncertainty into account in a principled way. For instance, one of the simplest and most commonly studied formalism in that field is that of *probabilistic databases* [SORK11]: a probabilistic database is a relational database in which every tuple (row) is annotated with a probability value, that intuitively represents the degree of certainty that one has about that specific data item. Given a *Boolean query* q, whose answer on a "traditional", non-probabilistic database would be YES or NO, one can then compute *the probability that the probabilistic database satisfies the query*, assuming independence across tuples.

Exemple. *A toy example of a probabilistic database is shown in Table 1, describing in a single table with two attributes which person likes which person, and the estimated probability of these facts. Consider then the query* $q :=$ *"there exist two different people that like the same person", which can be written in first-order logic as* $\exists x\, \exists y\, \exists z\ : \mathrm{Likes}(x, z) \wedge \mathrm{Likes}(y, z) \wedge x \neq y$. *One could then compute that the probability that* q *is*

Likes		Prob.
Alice	Bob	0.5
Alice	John	0.9
Bob	Bob	0.2
John	Bob	0.7

Table 1: Example probabilistic database.

satisfied by the example probabilistic database is equal to $1 - \Big[(1-0.5)(1-0.2)(1-0.7) + 0.5(1-0.2)(1-0.7) + (1-0.5)0.2(1-0.7) + (1-0.5)(1-0.2)0.7\Big]$.

Related work. Existing work on the complexity of this problem mainly focused on two approaches:

- **Restricting the query.** In [DS12] the authors study the computational complexity of the probabilistic query evaluation problem for a fixed Boolean query q, written PQE(q) for short: given as input a probabilistic database, compute the probability that it satisfies q. It is shown that for the class of queries corresponding to the SELECT FROM WHERE fragment of SQL this problem admits a dichotomy: either the query q is what they call *safe*, and PQE(q) can be solved in polynomial time, or q is not safe and PQE(q) is provably intractable under usual complexity-theoretic assumptions. The main inconvenient of this approach is that very few queries are tractable in this sense.

- **Restricting the data.** In [ABS16] the authors study the problem PQE(q) but where the databases are restricted to have *bounded treewidth*. Intuitively, treewidth is a parameter that indicates how close the database is to being “treelike”. They show that for a very large class of queries, PQE(q) is always tractable when restricted to such databases. Hence, they obtain tractability of the problem for more queries than the previous approach, but at the cost of restricting the shape of the data.

In addition to performing query evaluation, both approaches allow to compute so-called *provenance representations* of the query result, which intuitevely are objects that capture the trace of the computation and that can be used, to explain the query result, or to rank the input tuples by importance [DFKM22].

The goal of this PhD is then to reconcile these two approaches, and to study the complexity of PQE and of computing provenance representations by taking into account the interaction between the query and the data. For instance one could want to define a width notion that would be parameterized by a query, which would ensure that when this quantity is bounded then PQE is tractable. We would also want to obtain lower bounds, in the spirit of [AM22] that shows that for some queries, bounding the treewidth is actually a *necessary* criterion to obtain tractability, or again as in [ACMS19] that shows the same kind of lower bounds on the size of provenance representation of the query results in the form of *knowledge compilation* formalisms. The PhD candidate could, for example, start to study these questions in the simpler setting where all relations have arity two, which corresponds to computing the probability of existence of homomorphisms between *graphs*. Practical implementation of the obtained algorithms could then be developed to test the applicabitliy of such approaches.

2 Context and advisors

The PhD will be carried out in LINKS[1], which is a joint research team between Inria Lille[2], the University of Lille[3], and the CRIStAL laboratory[4]. It will be supervised by Sylvain Salvati[5] and co-supervised by Mikaël Monet[6] and Antoine Amarilli[7]. Mikaël Monet is an Inria full-time researcher in LINKS working on theoretical aspects of uncertain data management, knowledge compilation, and formal explainability. Antoine Amarilli is an associate professor at Télécom Paris and works on database theory, knowledge compilation, and enumeration complexity. Sylvain Salvati is a professor in LINKS working on formal methods and programming languages.

3 Candidates

Candidates to this PhD proposal should have a good background in the following areas of computer science: discrete mathematics, complexity theory, formal languages, logic, and databases.

References

[ABS16] Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Tractable lineages on treelike instances: Limits and extensions. In *PODS*, pages 355–370, 2016.

[ACMS19] Antoine Amarilli, Florent Capelli, Mikaël Monet, and Pierre Senellart. Connecting knowledge compilation classes and width parameters. *Theory of Computing Systems*, pages 1–54, 2019.

[AM22] Antoine Amarilli and Mikaël Monet. Weighted Counting of Matchings in Unbounded-Treewidth Graph Families. In *MFCS*, volume 241 of *LIPIcs*, pages 9:1–9:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[DFKM22] Daniel Deutch, Nave Frost, Benny Kimelfeld, and Mikaël Monet. Computing the Shapley value of facts in query answering. In *Proceedings of the 2022 International Conference on Management of Data*, pages 1570–1583, 2022.

[DS12] Nilesh N. Dalvi and Dan Suciu. The dichotomy of probabilistic inference for unions of conjunctive queries. *J. ACM*, 59(6):30, 2012.

[SORK11] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Morgan & Claypool, 2011.

[1] https://team.inria.fr/links/
[2] https://www.inria.fr/en/centre-inria-lille-nord-europe
[3] https://www.univ-lille.fr/home/
[4] https://www.cristal.univ-lille.fr/en/
[5] https://pro.univ-lille.fr/sylvain-salvati
[6] https://mikael-monet.net/
[7] https://a3nm.net/