Probabilities and Provenance on Trees and Treelike Instances

Antoine Amarilli

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay

In many practical data management settings, it is necessary to manage data that may be *incorrect*, and where we only have some degree of confidence in each fact. The recent field of *probabilistic databases* [SORK11] accordingly studies how to generalize to probabilistic data the known results and techniques on relational data management. Sadly, query evaluation is often hard in the probabilistic context: evaluating simple fixed conjunctive queries may be already #P-hard in the data [DS07]. Thus, existing work has characterized the (quite limited) subclass of unions of conjunctive queries where the task can be performed efficiently, culminating in the dichotomy result of [DS12].

To mitigate this intractability, we have recently proposed an alternative approach: impose structural restrictions on the *data* to support more expressive queries. Indeed, in the restricted context of *probabilistic XML* [KS11] documents (i.e., probabilistic trees), query evaluation is tractable on some document classes [CKS09]. In a recent work at ICALP'15 [ABS15], we showed that tractability also holds on bounded-treewidth data, by leveraging Courcelle's well-known theorem [Cou90, FFG02]. Formally, we showed that the evaluation of monadic second-order (MSO) queries is tractable (in data complexity) over several probabilistic database formalisms, assuming that the treewidth of the input data is bounded by a constant.

Our proof technique relies on the notion of *lineage* of a query on a database: the lineage is Boolean formula describing which sets of facts of the data can suffice to make the query true, and we can compute the query probability from the lineage expression (though this is not always efficient). In our setting, we showed that we can efficiently compute the lineage of MSO queries on instances, and that the resulting lineages were in classes for which probability computation could also be performed efficiently: bounded-treewidth circuits, and d-DNNFs [Dar01]. To compute the lineages, we compile the MSO query to an automaton that tests it on bounded-treewidth instances using Courcelle's result, and we then show how to compute the query lineage from the automaton, defining a natural notion of automaton lineage on uncertain trees as a Boolean circuit. The use of this result goes beyond probability computation, as it extends to the abstract setting of *semiring provenance* [GKT07] which generalizes many other data management tasks.

In our recent work [ABS16], to be presented at PODS'16, we show a converse to this tractability result, to achieve an instance-based dichotomy for query evaluation on probabilistic data. Specifically, we show that the evaluation of FO queries on *any* restricted class C of probabilistic instances is #P-hard under RP reductions whenever the treewidth of C is unbounded, under some technical hypotheses (the instances are arity-two, and they are efficiently constructible). We show this hardness result using the technique of graph minor extraction, following Robertson and Seymour's grid minor theorem [RS86]: we leverage recent polynomial bounds on this problem [CC14] which also allow us to strengthen similar hardness results for MSO in the non-probabilistic context [GHL$^+$14]. It is especially surprising that our probabilistic hardness result applies to FO queries rather than MSO queries, so that it would not hold in the non-probabilistic context [Kre08].

The proposed talk will review our approach for probability evaluation and provenance computation using tree automata from [ABS15], and our recent corresponding lower bounds [ABS16]. It will hint at our ongoing work in these directions on practical experimentation with these techniques for query evaluation on road networks, and efficient compilation of restricted query classes to tree automata.

Coauthors. This is joint work with Pierre Bourhis (CNRS CRIStAL) and Pierre Senellart (LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay & IPAL, CNRS, NUS).

References

[ABS15] Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Provenance circuits for trees and treelike instances. In *Proc. ICALP*, 2015.

[ABS16] Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Tractable lineages on treelike instances: Limits and extensions. In *PODS*, 2016. To appear.

[CC14] Chandra Chekuri and Julia Chuzhoy. Polynomial bounds for the grid-minor theorem. In *Proc. STOC*, 2014.

[CKS09] Sara Cohen, Benny Kimelfeld, and Yehoshua Sagiv. Running tree automata on probabilistic XML. In *Proc. PODS*, 2009.

[Cou90] Bruno Courcelle. The monadic second-order logic of graphs. I. Recognizable sets of finite graphs. *Inf. Comput.*, 85(1), 1990.

[Dar01] Adnan Darwiche. On the tractable counting of theory models and its application to truth maintenance and belief revision. *J. Applied Non-Classical Logics*, 11(1-2), 2001.

[DS07] Nilesh Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. *VLDBJ*, 16(4), 2007.

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

[FFG02] Jörg Flum, Markus Frick, and Martin Grohe. Query evaluation via tree-decompositions. *JACM*, 49(6), 2002.

[GHL+14] Robert Ganian, Petr Hliněnỳ, Alexander Langer, Jan Obdržálek, Peter Rossmanith, and Somnath Sikdar. Lower bounds on the complexity of MSO1 model-checking. *JCSS*, 1(80), 2014.

[GKT07] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proc. PODS*, 2007.

[Kre08] Stephan Kreutzer. Algorithmic meta-theorems. In *Parameterized and Exact Computation*. 2008.

[KS11] Benny Kimelfeld and Pierre Senellart. Probabilistic XML: Models and complexity, 2011.

[RS86] Neil Robertson and Paul D. Seymour. Graph minors. V. Excluding a planar graph. *J. Comb. Theory, Ser. B*, 41(1), 1986.

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