Tractable Query Answering under Probabilistic Constraints

Antoine Amarilli

Télécom ParisTech, Institut Mines-Télécom, CNRS LTCI

Large knowledge bases such as YAGO [SKW07] or DBpedia [BLK$^+$09] can be used to answer queries in various domains. However, as they are automatically harvested from Web sources, they may be *incomplete*: important facts may be missing because they were not materialized in the original sources, or could not be extracted correctly. To mitigate this problem, approaches such as association rule mining [GTHS13] can extract statistical rules from the data which hold in most situations. For instance, people are usually nationals of the country where they are born; people who died in a place are often buried there. The application of such rules allows us to infer some of the missing facts, which may help mitigate the issue of incompleteness. Hence, we study the problem of query answering on large-scale knowledge bases under the constraints of such probabilistic deduction rules.

As such rules only represent statistical tendencies, one needs to keep track of uncertainty on rule consequences when reasoning about them. There is a large body of work on probabilistic data management [SORK11]; yet, in that setting, many important tasks are intractable. For example, fixed conjunctive queries may be #P-hard [DS07] to evaluate on a probabilistic instance, even in the very simple tuple-independent database (TID) model [LLRS97]. To work around such hardness results, existing work has already investigated which query classes are tractable over all data instances, with a complex dichotomy between safe and unsafe queries [DS12].

Yet, there has been no attempt to generalize the observation that query evaluation is tractable, for *all* queries and for much more expressive query languages, on some instances such as probabilistic XML trees [CKS09]. Our work follows this intuition and revisits the probabilistic inference problem by studying *instance classes* that ensure tractability. More precisely, we study complexity as a function of instance *treewidth*, which is motivated by well-known tractability results on evaluating monadic second-order (MSO) queries on non-probabilistic bounded-treewidth instances [FFG02] and counting queries on bounded-treewidth graphs [ALS91]. This approach is also practically relevant, as the treewidth of real-world data is usually much less than its size. We thus show that, for the TID model, MSO query evaluation has linear data complexity if the treewidth of the instance is fixed.

The TID model is not sufficient to represent the consequences of uncertain deduction rules, however: it assumes *independence* of all facts, whereas rule application imposes *correlations* between cause and consequence facts. Correlations are usually represented by probabilistic events shared between multiple facts, yet their presence makes it generally intractable to evaluate even the simplest queries, both in the relational [GT06] and XML [KS11] setting. However, we show that query evaluation is tractable if the instance has bounded width under a new notion of tree decomposition that accounts for probabilistic events; intuitively, we enforce their compatibility with the tree structure. This result implies, for example, that it is tractable to evaluate queries on the block-independent disjoint [BGMP92] probabilistic relational model, if the underlying instance has bounded treewidth in the usual sense and if the size of blocks is bounded by a constant. In the XML setting, it implies that query evaluation is tractable whenever there are only a bounded number of relevant events to propagate at any point along the tree.

We last turn to our original problem of query evaluation on probabilistic instances under *uncertain deduction rules*: the goal is to determine the answers of a query on a knowledge base, annotated by their probability, when some of the instance facts are uncertain, and when new uncertain facts can be generated using the provided rules. We formally define this problem as computing the probability of query answers in the universal model obtained by a probabilistic version of the standard chase procedure. We then show its tractability for the language of guarded tuple-generating dependencies [CGK13] under the assumption that the chase terminates, which may be enforced, e.g., by syntactic conditions such as weak acyclicity [One13]. Indeed, we show that the rule consequences and resulting correlation events have a bounded-width decomposition in this case.

A possible avenue for future work is to generalize the rule language to non-terminating guarded rules, or to, e.g., disjunctive rules. Another important extension for practical knowledge bases would be to account for equality-generating dependencies, or for possible element reuse among the nulls created during the chase; however, we suspect that such extensions would lead to undecidability. Last, it would be useful to study how to implement the proposed methods in practice, including techniques such as sampling or pruning irrelevant facts.

Coauthors. This is joint work with Pierre Bourhis (CNRS Lille, Université Lille 1) and Pierre Senellart (Télécom ParisTech, Institut Mines-Télécom, CNRS LTCI).

References

[ALS91] Stefan Arnborg, Jens Lagergren, and Detlef Seese. Easy problems for tree-decomposable graphs. *J. Algorithms*, 12(2):308–340, 1991.

[BGMP92] Daniel Barbará, Hector Garcia-Molina, and Daryl Porter. The management of probabilistic data. *TKDE*, 4(5), 1992.

[BLK$^+$09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the Web of Data. *J. Web Semantics*, 7(3):154–165, September 2009.

[CGK13] Andrea Calì, Georg Gottlob, and Michael Kifer. Taming the infinite chase: Query answering under expressive relational constraints. *JAIR*, 48, 2013.

[CKS09] Sara Cohen, Benny Kimelfeld, and Yehoshua Sagiv. Running tree automata on probabilistic XML. In *Proc. PODS*, pages 227–236. ACM, 2009.

[DS07] Nilesh Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. *VLDBJ*, 16(4):523–544, 2007.

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

[FFG02] Jörg Flum, Markus Frick, and Martin Grohe. Query evaluation via tree-decompositions. *JACM*, 49(6):716–752, 2002.

[GT06] Todd J. Green and Val Tannen. Models for incomplete and probabilistic information. In *Proc. EDBT Workshops, IIDB*, March 2006.

[GTHS13] Luis Galárraga, Christina Teflioudi, Katja Hose, and Fabian M. Suchanek. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proc. WWW*, pages 413–422, 2013.

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

[LLRS97] Laks V. S. Lakshmanan, Nicola Leone, Robert B. Ross, and V. S. Subrahmanian. ProbView: A flexible probabilistic database system. *TODS*, 22(3), 1997.

[One13] Adrian Onet. The chase procedure and its applications in data exchange. In *Data Exchange, Information, and Streams*, 2013.

[SKW07] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proc. WWW*, pages 697–706, 2007.

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