Ontology-Mediated Query Answering over Log-Linear Probabilistic Data (Abstract)[⋆]

Stefan Borgwardt[1], İsmail İlkan Ceylan[2], and Thomas Lukasiewicz[2]

[1] Faculty of Computer Science, Technische Universität Dresden, Germany
stefan.borgwardt@tu-dresden.de
[2] Department of Computer Science, University of Oxford, UK
firstname.lastname@cs.ox.ac.uk

Advances in automated knowledge base construction have led to successful systems, such as DeepDive [17], NELL [13], and Google's Knowledge Vault [7]. They extract *structured knowledge* from multiple sources, through a chain of statistical techniques, and produce *probabilistic knowledge bases* (PKBs). The basic data model underlying these systems is given by *probabilistic databases* (PDBs) [18]; see recent surveys focusing on PKBs [2, 4]. PKBs are inherently *incomplete*, which makes reasoning more challenging. A common way to deal with incompleteness is to add *commonsense knowledge*, in the form of logical theories, to allow for deductions that go beyond existing facts in the knowledge base.

Statistical relational models are *concise*, and *lifted* representations of probabilistic graphical models [9]. Well-known examples include Markov logic networks (MLNs) [15], relational Bayesian networks [10], and approaches to probabilistic logic programming (PLP). All these models can encode commonsense knowledge, but they are based on the *closed-domain assumption (CDA)* that requires the set of relevant objects to be *finite*, and *known* at design-time, which is not always an easy condition to be met. And contrary to the intuition, the CDA does not necessarily imply efficiency in comparison to open-domain models.

Example 1. In the closed domain $\mathbf{C} = \{c_1, \dots, c_n\}$, the following are equivalent:

$$\forall x\, \mathsf{Employee}(x) \to \exists y\, \mathsf{Address}(x, y), \tag{1}$$

$$\forall x\, \mathsf{Employee}(x) \to \mathsf{Address}(x, c_1) \lor \dots \lor \mathsf{Address}(x, c_n). \tag{2}$$

That is, all employee's addresses must be one of the *known objects* in the database. If the address of a new employee is still unknown, in an interpretation they will be randomly assigned the address of another employee. A common remedy is to introduce a number of auxiliary objects into $\mathbf{C}$ that can serve as “unknown addresses”. However, it is unclear *how many* additional objects are needed.

Another problem is the large disjunction in (2), which introduces a huge amount of *nondeterminism*. For MLNs, this is a known problem, and more sophisticated techniques to eliminate existential quantification exist [19]. However,

[⋆] This is an abstract of a paper presented at AAAI 2019 [3]. This work was supported by the German Research Foundation (DFG) within the project BA 1122/19-1 (GOASQ), by The Alan Turing Institute under the UK EPSRC grant EP/N510129/1, and by the EPSRC grants EP/R013667/1, EP/L012138/1, and EP/M025268/1.

in the worst case, these techniques also cannot avoid the nondeterminism over the fixed domain. For this reason, almost all MLN implementations only support universal quantification [6,14]. This inefficiency appears also in ontology languages. For example, (1) can be formulated in $\mathcal{EL}$, where reasoning is P-complete, but becomes NP-complete in a closed domain [8]. ■

Probabilistic models that can encode commonsense knowledge while allowing an *open domain* include PLP with function symbols [5, 16], the probabilistic programming language BLOG [12], and ontology-based approaches [1, 11]. The latter are further distinguished from the rest by the *open-world assumption*, i.e., they do not interpret the absence of facts as the negation of these facts; this assumption means that the incomplete nature of the PKB is respected.

Another problem that is inherent to knowledge-based probabilistic models is related to *inconsistent worlds*, which are usually removed, and the resulting probability distribution is *renormalized*.

Example 2. Consider the following tuple-independent PDB $\mathcal{P}$ and theory $\mathcal{T}$:

$$\mathcal{P} \coloneqq \{\langle \mathsf{A}(a) : 0.5\rangle, \langle \mathsf{B}(a) : 0.5\rangle\} \qquad \mathcal{T} \coloneqq \{\forall x\, \mathsf{A}(x) \rightarrow \mathsf{B}(x)\}.$$

The possible worlds are

$$\mathcal{W}_1 \coloneqq \{\mathsf{A}(a), \mathsf{B}(a)\},\ \mathcal{W}_2 \coloneqq \{\mathsf{A}(a), \neg\mathsf{B}(a)\},\ \mathcal{W}_3 \coloneqq \{\neg\mathsf{A}(a), \mathsf{B}(a)\},\ \mathcal{W}_4 \coloneqq \dots$$

Without $\mathcal{T}$, each of these worlds has the probability 0.25, by the independence assumptions of $\mathcal{P}$. However, since $\mathcal{W}_2$ is inconsistent with $\mathcal{T}$, its probability is reduced to 0, and the probability of the remaining worlds is renormalized to add up to 1, yielding a probability of 0.33 each. As an undesired side effect, the probabilities for $\mathsf{A}(a)$ and $\mathsf{B}(a)$ change to 0.33 and 0.66, respectively. ■

We argue that the observed probabilities of 0.5 should be preserved, and try to find a log-linear distribution that deviates from these input values as little as possible. By assigning both $\mathcal{W}_2$ and $\mathcal{W}_3$ a probability of 0 and the remaining worlds 0.5 each, we obtain a model that satisfies the constraints of both $\mathcal{T}$ and $\mathcal{P}$. This approach respects both the probabilistic and the logical input, and does not favor one over the other. In our new approach, we assume the database given as a set of facts with associated *weights*, which is then interpreted as a log-linear model. As in MLNs, we restrict the probability distribution to the *known* objects, but additionally use a first-order theory interpreted over arbitrary, *possibly infinite* domains, whereby we achieve open-world, open-domain reasoning.

We show that reasoning in our model can be reduced (via polynomial rewriting techniques) to inference in MLNs, or PDBs. These results are significant given the expressive nature of our formalism. As a consequence of the above reductions, many computational complexity results from previous models carry over. We also describe a new approach to learn the weights for our model, based on the principle of *maximum entropy*, to establish the connection with existing PKBs. This approach is independent of the other results, however—in principle, we could use any other weight learning method, e.g., using standard renormalization.

The full paper can be found at `https://tu-dresden.de/inf/lat/papers`.

References

1. Borgwardt, S., Ceylan, İ.İ., Lukasiewicz, T.: Ontology-mediated queries for probabilistic databases. In: AAAI (2017)
2. Borgwardt, S., Ceylan, İ.İ., Lukasiewicz, T.: Recent advances in querying probabilistic knowledge bases. In: IJCAI-ECAI (2018)
3. Borgwardt, S., Ceylan, İ.İ., Lukasiewicz, T.: Ontology-mediated query answering over log-linear probabilistic data. In: AAAI (2019), to appear
4. Van den Broeck, G., Suciu, D.: Query processing on probabilistic data: A survey. FTD **7**(3/4), 197–341 (2017)
5. De Raedt, L., Kimmig, A., Toivonen, H.: ProbLog: A probabilistic Prolog and its application in link discovery. In: IJCAI (2007)
6. Domingos, P., Lowd, D.: Markov Logic: An Interface Layer for Artificial Intelligence. Morgan & Claypool (2009)
7. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. In: SIGKDD (2014)
8. Gaggl, S.A., Rudolph, S., Schweizer, L.: Fixed-domain reasoning for description logics. In: ECAI (2016). https://doi.org/10.3233/978-1-61499-672-9-819
9. Getoor, L., Taskar, B.: Introduction to Statistical Relational Learning. The MIT Press (2007)
10. Jaeger, M.: Relational Bayesian Networks. In: UAI (1997)
11. Jung, J.C., Lutz, C.: Ontology-based access to probabilistic data with OWL QL. In: ISWC (2012)
12. Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D.L., Kolobov, A.: Blog: Probabilistic models with unknown objects. In: IJCAI (2005)
13. Mitchell *et al.*, T.: Never-ending learning. In: AAAI (2015)
14. Niu, F., Ré, C., Doan, A., Shavlik, J.: Tuffy: Scaling up statistical inference in Markov Logic Networks using an RDBMS. PVLDB **4**(6), 373–384 (2011)
15. Richardson, M., Domingos, P.: Markov logic networks. ML **62**(1), 107–136 (2006)
16. Sato, T., Kameya, Y.: PRISM: A language for symbolic-statistical modeling. In: IJCAI. pp. 1330–1335 (1997)
17. Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., Ré, C.: Incremental knowledge base construction using DeepDive. PVLDB **8**(11), 1310–1321 (2015)
18. Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic Databases. Morgan & Claypool (2011)
19. Van den Broeck, G., Meert, W., Darwiche, A.: Skolemization for weighted first-order model counting. In: KR (2014)