show complexity of reasoning, ranging from PTIME for weak variants of $\mathcal{EL}$ to undecidable for expressive variants of $\mathcal{ALC}$. Prob-$\mathcal{EL}$ is more closely studied in [10], where the authors show that reasoning is PTIME as long as (i) probability values are restricted to 0 and 1, and (ii) probabilistic annotations are only allowed on concepts; if (i) is dropped, then it becomes EXPTIME-complete, while if (ii) is dropped it becomes PSPACE-hard. The complexity results in our work are further testament to how intractable simple tasks become when probabilistic computations are involved, even when the starting point is a tractable logic.

Other recent efforts focused on extending $\mathcal{EL}$ DLs with probabilistic uncertainty include [5], [9], and [17]. In [5], a formalism is presented in which probability assessments are only allowed on ABoxes. The authors study the problem of satisfiability of KBs, which involves determining whether there exists a probability distribution that satisfies all the assignments over the ABoxes. The work of [9] is similar in spirit to [5], but there an MLN is employed in order to infer the probabilities of atoms in the ABox as a means to generate explanations as part of abductive inference. Though they use MLNs, this work is quite different from ours since only the ABox is assumed to be probabilistic, and the assertions are themselves part of the MLN instead of being annotated by external events as in our formalism. In [17], the authors present ELOG, which is $\mathcal{EL}^{++}$ without nominals or concrete domains combined with probabilistic log-linear models (a class which contains MLNs). The resulting probabilistic formalism basically assigns weights to axioms reflecting how likely the axiom is to hold. The main problem then corresponds to finding the most probable coherent ontology, a problem that is essentially different from the one tackled here.

Finally, a related formalism from the recent databases literature is that of probabilistic Datalog+/− [8, 7]. Datalog+/− is a language that arose from the generalization of rule-based constraints with the goal of expressing ontological axioms. The probabilistic extension in [8, 7] also makes use of MLNs, though the integration is loose in the sense that probabilistic annotations cannot refer to objects in the ontology, which leads to data-tractable algorithms but also limits the expressive power of the formalism.

6 Discussion, Conclusions, and Future Work

In this work, we have extended the DL $\mathcal{EL}^{++}$ with probabilistic uncertainty, based on the annotation of axioms. Such annotations refer to events whose probabilities are described by an associated MLN; one of the advantages of this formalism is that it is tightly coupled, which means that probabilistic annotations can refer to objects in the ontology. The proposed application of our formalism is in managing uncertainty in the Semantic Web, showing examples of how it can be applied in the analysis of Web forms, an important task in information extraction efforts.

Our focus here is on ranking queries, which request the set of atomic inferences sorted in descending order of probability. The algorithm we developed works in an anytime fashion, and therefore allows partial computations depending on the available resources; most importantly, we provide bounds on the error incurred by runs of this algorithm and conditions that allow to conclude when certain pairs in the output are correctly ordered. This algorithm works over cMLNs, in which only conjunctions of atoms are allowed. Regarding the expressivity of cMLNs, we can say that: (i) They are rich enough to simulate disjunction for a specific propositional subset. For instance, if the formula $p(X) \lor q(X)$ with a given weight needs to be enforced for the subset of individuals $\{a, b\}$, MLN learning algorithms can be directed to give corresponding weights to the specific worlds in which $(p(a) \lor q(a))$ and $(p(b) \lor q(b))$ hold. So, it is possible to represent certain special cases that may need to be handled. (ii) For the case of (atomic) negation, it is known to be representable via negative weights. (iii) Though material implications cannot be represented, this sort of constraint is more adequately placed on the ontology side, and the TBox is capable of representing them.

Regarding the practical applicability of the formalism, we can say that, despite probabilistic instance-checking being already intractable for cMLNs, it is often possible to bound the number of scenarios in practice. Consider, for instance, the problem of reasoning over data structures on the Web (related to the running example). Web pages are usually processed in isolation, since structured data do not usually span across pages and are often delimited by a certain DOM sub-tree. This implies that the number of possible worlds is bounded by a function of the constants appearing in a subset of the page. This bound can work together with the good computational behavior of cMLNs to allow tractability of reasoning in practice. Other applications where cMLNs can be leveraged are semantic and natural language-based Web search.

Future work involves investigating other DLs that can be extended in this manner, and pushing the known line between tractability and expressivity. We also need to empirically evaluate our approach both on synthetic and real-world data, as well as studying the application of other techniques such as random sampling, which may provide increased scalability at the cost of lost guarantees.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/J008346/1 ("PrOQAW"), the Oxford Martin School grant LC0910-019, the European Research Council (ERC) grant FP7/246858 ("DIADEM"), a Google Research Award, and a Yahoo! Research Fellowship.