Open World Probabilistic Databases[⋆]

İsmail İlkan Ceylan[1], Adnan Darwiche[2], and Guy Van den Broeck[2]

[1] Theoretical Computer Science, TU Dresden, Germany
ceylan@tcs.inf.tu-dresden.de
[2] Automated Reasoning Lab, UCLA, United States
darwiche@cs.ucla.edu, guyvdb@cs.ucla.edu

Introduction and Motivation Driven by the need to learn from vast amounts of text data, efforts throughout natural language processing, information extraction, databases. and AI are coming together to build *large-scale knowledge bases*. Academic systems such as NELL [14], Reverb [7], Yago [11], and DeepDive [16] continuously crawl the web to extract relational information. Industry projects such as Microsoft's Probase [18] or Google's Knowledge Vault [6] similarly learn structured data from text to improve search products. Notably, such knowledge bases are inherently probabilistic and many of them [6, 16] are based on the foundations of *tuple-independent probabilistic databases* (PDBs) [17]. According to the PDB semantics, each database tuple is an independent Bernoulli random variable, and all other tuples have probability zero, enforcing a *closed-world assumption* (CWA) [15].

This paper revisits the choice for the CWA in probabilistic knowledge bases. We observe that the *CWA is violated* in their deployment, which makes it *problematic to reason, learn, or mine* on top of these databases. First, knowledge bases are part of a larger machine learning loop that continuously updates beliefs about facts based on new textual evidence. From a Bayesian learning perspective [2], this loop can only be principled when learned facts have an a priori non-zero probability. Hence, the CWA does not accurately represent this mode of operation and puts it on weak footing. Second, these issues are not temporary: it will never be possible to complete probabilistic knowledge bases of even the most trivial relations, as the memory requirements quickly become excessive. This already manifests today: statistical classifiers output facts at a high rate, but only the most probable ones make it into the knowledge base, and the rest is truncated, losing much of the statistical information. Third, query answering under the CWA does not take into account the effect the open world can have on the query probability. This makes it impossible to distinguish queries whose probability should intuitively differ. These issues stand in the way of some principled approaches to knowledge base completion and mining.

We propose an alternative semantics for probabilistic knowledge bases to address these problems, which results in open-world PDBs (OpenPDBs). We show that OpenPDBs provide more meaningful answers. Finally, we pinpoint limitations of OpenPDBs and discuss ontology based data access (OBDA) as promising approach to further strengthen this framework.

[⋆] This is an extended abstract of the paper to be presented at KR'16 [3]

Inmovie		P
w_smith	ali	0.9
w_smith	sharktale	0.8
j_smith	ali	0.6
arquette	scream	0.7
pitt	mr_ms_smith	0.5
jolie	mr_ms_smith	0.7
jolie	sharktale	0.9

Couple		P
arquette	cox	0.6
pitt	jolie	0.8
thornton	jolie	0.6
pitt	aniston	0.9
kunis	kutcher	0.7

Fig. 1: Probabilistic database tables

OpenPDBs Inspired by the *open-world assumption* (OWA) OpenPDBs assume that the knowledge of a domain may be incomplete. Hence, anything that is not in the DB remains possible: Tuples that are not present in the DB, called *open tuples*, are associated with a default probability interval of $[0, \lambda]$, where λ is a fixed threshold. Intuitively, we obtain a lower probability (0) and an upper probability (λ) for the open tuples as opposed to setting their probabilities to 0. Our proposal for OpenPDBsbuilds on the theory of imprecise probabilities, and credal sets in particular [13], to allow interval-based probabilities for open tuples. We briefly show that this framework provides more meaningful answers, in terms of upper and lower bounds on the query probability.

Query evaluation Consider the PDB given in Figure 1 and the queries

$$\begin{aligned} \mathsf{Q}_1(x, y) &= \texttt{Inmovie}(x, z) \wedge \texttt{Inmovie}(y, z) \wedge \texttt{Couple}(x, y) \\ \mathsf{Q}_2 &= \texttt{Inmovie}(x, z) \wedge \texttt{Inmovie}(y, z) \wedge \texttt{Couple}(x, y). \end{aligned}$$

First, note that $\mathrm{P}(\mathsf{Q}_2) = \mathrm{P}(\mathsf{Q}_1(\mathsf{pitt}, \mathsf{jolie})) = 0.28$ in the PDB of Figure 1. However, observe that $\mathsf{Q}_1(\mathsf{pitt}, \mathsf{jolie})$ entails Q_2 and thus intuitively, we would expect $\mathrm{P}(\mathsf{Q}_2) > \mathrm{P}(\mathsf{Q}_1(\mathsf{pitt}, \mathsf{jolie}))$ in an open-world setting. In fact, in the open-world setting, our estimate of $\mathrm{P}(\mathsf{Q}_2)$ should be higher, because there exist a large number of couples that could satisfy this query. This is indeed the case for OpenPDBs (for upper probabilities) since Q_2 is entailed in many worlds that now have non-zero probability. Of these, a large portion do not entail $\mathsf{Q}_1(\mathsf{pitt}, \mathsf{jolie})$.

Second, by the CWA we lose the ability to distinguish queries that are clearly not identical. Observe, for instance, that both $\mathrm{P}(\mathsf{Q}_1(\mathsf{thornton}, \mathsf{aniston}))$ and $\mathrm{P}(\mathsf{Q}_1(\mathsf{w\_smith}, \mathsf{j\_smith}))$ evaluate to 0. Taking into account the open world, however $\mathsf{Q}_1(\mathsf{w\_smith}, \mathsf{j\_smith})$ is supported by two facts in the PDB, while $\mathsf{Q}_1(\mathsf{thornton}, \mathsf{aniston})$ is supported by none, which should make it less likely. Taking these observations to the extreme, the query $\texttt{Inmovie}(x, y) \wedge \neg\texttt{Inmovie}(x, y)$ is unsatisfiable, yet it evaluates to the same probability as the satisfiable query $\mathsf{Q}_1(\mathsf{thornton}, \mathsf{aniston})$. Clearly, there is no attribution for being closer to satisfied. It is clear that in the open world setting the upper probability of a satisfiable query will be greater than the upper probability of an unsatisfiable query. In fact, any unsatisfiable query will still have the upper probability 0 in Open-

PDBs. These answers are clearly more in line with the incomplete projection of the world.

Overview of the Results Our open-world semantics is supported by a *query evaluation algorithm* for *unions of conjunctive queries* (UCQs). This class of queries, corresponding to monotone DNF, is particularly well-behaved and the focal point of database research. Perhaps the largest appeal of PDBs comes from a breakthrough dichotomy result by [5], perfectly delineating which UCQs can be answered efficiently in the size of the PDB. Their algorithm runs in polynomial time for all efficient queries, called *safe queries*, and recognizes all others to be #P-hard. Our OpenPDB algorithm extends the PDB algorithm of [5] and inherits its elegant properties: all safe queries run in polynomial time. When our algorithm fails, the query is #P-hard. Moreover, a careful analysis shows that both algorithms run in *linear time* in the number of (closed-world) tuples. Even though OpenPDBs model a polynomially larger set of random variables, these can be reasoned about as a whole, and there is no computational blow-up for open-world reasoning. Hence, both OpenPDBs and PDBs admit the same *data complexity dichotomy between linear time and #P*.[3]

For queries with negation, only a partial classification of PDB data complexity is known [8, 10]. We show that the complexity of open-world reasoning can go up significantly with negation. We identify a linear-time PDB query that becomes NP-complete on OpenPDBs. Moreover, there exists a PP-complete([4]) query on PDBs that becomes $\mathrm{NP}^{\mathrm{PP}}$-complete on OpenPDBs. Clearly, negation leads to a much richer data complexity landscape.

Ontologies and OpenPDBs OpenPDBs improve on PDBs and provide a more suitable setting for large probabilistic knowledge bases. However, default intervals are sometimes too loose and it is not always possible to distinguish queries that should intuitively differ. One of the ways of restricting the open world is to use an explicit formalism for restricting the models. More concretely, the worlds induced by an OpenPDB can be refined and restricted by using a background knowledge specified as a logical theory. Ontology based data access (OBDA) [1] has been introduced as a means of querying incomplete data sources with the help of a background knowledge provided by an ontology.

Probabilistic OBDA has been investigated; both in the context of Description Logics and Datalog$^{\pm}$ [4, 9, 12]. Closely related is the work by [12], where authors enrich PDBs with light-weight ontologies. This provides a delta on PDBs: A fact that is not in the ABox remains possible due to the axioms present in the TBox. However, this approach is limited to the encoding in the ontology: Any fact that is not entailed by the ontology still gets the probability 0. This is a major difference from our approach. Investigating the computational properties of OBDA in combination with OpenPDBs is left as future work.

[3] To the best of our knowledge, and to our surprise, the fact that safe PDB queries have linear-time data complexity, and that the dichotomy of [5] is between linear time (not PTime) and #P, has not previously been observed in the literature.

[4] Intuitively, PP is the decision version of the counting class #P.

References

1. Bienvenu, M., Cate, B.T., Lutz, C., Wolter, F.: Ontology-based data access: A study through disjunctive datalog, csp, and mmsnp. ACM Trans. Database Syst. 39(4), 33:1–33:44 (2014)
2. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
3. Ceylan, İ.İ., Darwiche, A., Van Den Broeck, G.: Open-world probabilistic databases. In: Proc. of KR'16. AAAI Press (2016)
4. Ceylan, İ.İ., Peñaloza, R.: Probabilistic Query Answering in the Bayesian Description Logic BEL. In: Proc. of SUM'15. LNAI, vol. 9310, pp. 21–35. Springer (2015)
5. Dalvi, N., Suciu, D.: The dichotomy of probabilistic inference for unions of conjunctive queries. J. of the ACM 59(6), 1–87 (2012), `http://dl.acm.org/citation.cfm?doid=2395116.2395119`
6. Dong, X.L., 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: Proc. of ACM SIGKDD'14. pp. 601–610. KDD'14, ACM (2014)
7. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545. Ass. for Computational Linguistics (2011)
8. Fink, R., Olteanu, D.: Dichotomies for Queries with Negation in Probabilistic Databases. ACM Transactions on Database Systems (TODS) (2015), (to appear)
9. Gottlob, G., Lukasiewicz, T., Martinez, M.V., Simari, G.I.: Query answering under Probabilistic Uncertainty in Datalog +/- Ontologies. Ann. Math. AI 69(1), 37–72 (2013)
10. Gribkoff, E., Van den Broeck, G., Suciu, D.: Understanding the Complexity of Lifted Inference and Asymmetric Weighted Model Counting. In: Proc. of UAI'14. pp. 280–289. AUAI Press (2014)
11. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: Yago2: A spatially and temporally enhanced knowledge base from wikipedia. In: In Proc. of IJCAI'2013. pp. 3161–3165. AAAI Press (2013)
12. Jung, J.C., Lutz, C.: Ontology-Based Access to Probabilistic Data with OWL QL. In: Proc. of ISWC'12. LNCS, vol. 7649, pp. 182–197. Springer Verlag (2012)
13. Levi, I.: The Enterprise of Knowledge. MIT Press (1980)
14. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M.: Never-Ending Learning. In: Proc. of AAAI'15. AAAI Press (2015)
15. Reiter, R.: On closed world data bases. Logic and Data Bases pp. 55–76 (1978)
16. Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., Ré, C.: Incremental knowledge base construction using deepdive. In Proc. of VLDB Endowment 8(11), 1310–1321 (2015)
17. Suciu, D., Olteanu, D., Ré, C., Koch, C.: Probabilistic Databases (2011)
18. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: A probabilistic taxonomy for text understanding. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. pp. 481–492. ACM (2012)