On Constrained Open-World Probabilistic Databases

Tal Friedman and Guy Van den Broeck
Department of Computer Science
University of California, Los Angeles
{tal, guyvdb}@cs.ucla.edu

Abstract
Increasing amounts of available data have led to a heightened need for representing large-scale probabilistic knowledge bases. One approach is to use a probabilistic database, a model with strong assumptions that allow for efficiently answering many interesting queries. Recent work on open-world probabilistic databases strengthens the semantics of these probabilistic databases by discarding the assumption that any information not present in the data must be false. While intuitive, these semantics are not sufficiently precise to give reasonable answers to queries. We propose overcoming these issues by using constraints to restrict this open world. We provide an algorithm for one class of queries, and establish a basic hardness result for another. Finally, we propose an efficient and tight approximation for a large class of queries.

1 Introduction
An ubiquitous pursuit in the study of knowledge base representation is the search for a model that can represent uncertainty while simultaneously answering interesting queries efficiently. The key underlying challenge is that these goals are at odds with each other. Modelling uncertainty requires additional model complexity. At the same time, the ability to answer meaningful queries usually demands restrictive model assumptions. Both of these properties are at odds with the key limiting factor of tractability: success in the first two goals is not nearly as impactful if it is not achieved efficiently. Unfortunately, probabilistic reasoning is often computationally hard, even on databases [Roth, 1996; Dalvi and Suciu, 2012].

One approach towards achieving this goal is to begin with a simple model such a probabilistic database (PDB) [Suciu et al., 2011; Van den Broeck and Suciu, 2017]. A PDB models uncertainty, but is inherently simple and makes very strong independence assumptions and closed-world assumptions allowing for tractability on a very large class of queries [Dalvi and Suciu, 2007, 2012]. However, PDBs can fall short under non-ideal circumstances, as their semantics are brittle to incomplete knowledge bases [Ceylan et al., 2016].

To bring PDBs closer to the desired goal, Ceylan et al. [2016] propose open-world probabilistic databases (Open-PDB), wherein the semantics of a PDB are strengthened to relax the closed-world assumption. While OpenPDBs maintain a large class of tractable queries, their semantics are so relaxed these queries lose their precision: they model further uncertainty, but in exchange give less useful query answers.

In this work, we aim to overcome these querying challenges, while simultaneously maintaining the degree of uncertainty modeled by OpenPDBs. To achieve this, we propose further strengthening the semantics of OpenPDBs by constraining the mean probability allowed for a relation. These constraints work at the schematic level, meaning no additional per-item information is required. They are practically motivated by knowledge of summary statistics, of how many tuples we expect to be true. A theoretical analysis shows that, despite their simplicity, such constraints fundamentally change the difficulty landscape of queries, leading us to propose a general-purpose approximation scheme.

The rest of the paper is organized as follows: Section 2 provides necessary background on relational logic and PDBs, as well as an introduction to OpenPDBs. Section 3 motivates and introduces our construction for constraining OpenPDBs. Section 4 analyses exact solutions subject to these constraints, providing a class of tractable queries along with an algorithm. It also shows that the problem is in general hard, even in some cases where standard PDB queries are tractable. Section 5 investigates an efficient and provably bounded approximation scheme. Section 6 discusses our findings, and summarizes interesting directions that we leave as open problems.

2 Background
This section provides background and motivation for probabilistic databases and their open-world counterparts. Notation and definitions are adapted from Ceylan et al. [2016].

2.1 Relational Logic and Databases
We now describe necessary background from function-free finite-domain first-order logic. An atom \( R(x_1, x_2, ..., x_n) \) consists of a predicate \( R \) of arity \( n \), together with \( n \) arguments. These arguments can either be constants or variables. A ground atom is an atom that contains no variables. A formula is a series of atoms combined with conjunctions (\( \land \)) or disjunctions (\( \lor \)), and with quantifiers \( \forall, \exists \). A substitution \( Q[x/t] \) replaces all occurrences of \( x \) by \( t \) in a formula \( Q \).
A relational vocabulary $\sigma$ is comprised of a set of predicates $\mathcal{R}$ and a domain $\mathcal{D}$. Using the Herbrand semantics [Hinrichs and Genesereth, 2006], the Herbrand base of $\sigma$ is the set of all ground atoms possible given $\mathcal{R}$ and $\mathcal{D}$. A $\sigma$-interpretation $\omega$ is then an assignment of truth values to every element of the Herbrand base of $\sigma$. We say that $\omega$ models a formula $Q$ whenever $\omega$ satisfies $Q$. This is denoted by $\omega \models Q$.

A reasonable starting point for the target knowledge base to construct would be to use a traditional relational database. Using the standard model-theoretic view [Abiteboul et al., 1995], a relational database for a vocabulary $\sigma$ is a $\sigma$-interpretation $\omega$. Less formally, a relational database consists of a series of relations, each of which corresponds to a predicate. Each relation consists of a series of rows, also called tuples, each of which corresponds to an atom of the predicate being true. Any atom not appearing as a row in the relation is considered to be $\text{false}$, following the closed-world assumption [Reiter, 1981]. Figure 1 shows an example database.

### 2.2 Probabilistic Databases

Despite the success of relational databases, their deterministic nature leads to a few shortcomings. A common way to gather a large knowledge base is to apply some sort of statistical model [Carlson et al., 2010; Suchanek et al., 2007; Peters et al., 2014; Dong et al., 2014] which returns a probability value for potential tuples. Adapting the output of such a model to a relational database involves thresholding on the probability value, discarding valuable information along the way. A probabilistic database (PDB) circumvents this problem by assigning each tuple a probability.

**Definition 1.** A (tuple-independent) probabilistic database $\mathcal{P}$ for a vocabulary $\sigma$ is a finite set of tuples of the form $(t : p)$ where $t$ is a $\sigma$-atom and $p \in [0, 1]$. Furthermore, each $t$ can appear at most once.

Given such a collection of tuples and their probabilities, we are now going to define a distribution over relational databases. The semantics of this distribution are given by treating each tuple as an independent random variable.

**Definition 2.** A probabilistic database $\mathcal{P}$ for vocabulary $\sigma$ induces a probability distribution over $\sigma$-interpretations $\omega$:

$$P_\mathcal{P}(\omega) = \prod_{t \in \omega} P_\mathcal{P}(t) \prod_{t \notin \omega} (1 - P_\mathcal{P}(t))$$

where $P_\mathcal{P}(t) = \begin{cases} p & \text{if } (t : p) \in \mathcal{P} \\ 0 & \text{otherwise} \end{cases}$

Notice this last statement is again making the closed-world assumption: any tuple that we have no information about is assigned probability zero. Figure 2 shows an example PDB.

**Probabilistic Queries** In relational databases, the fundamental task we are interested in solving is how to answer queries. The same is true for probabilistic databases, with the only difference being that we are now interested in probabilities over queries. In particular, we are interested in queries that are fully quantified - also known as Boolean queries. On a relational database, this corresponds to a query that has an answer of True or False.

For example, on the database given in Figure 1, we might ask if there is a scientist who is a coauthor:

$$Q_1 = \exists x. \exists y. S(x) \land CoA(x, y)$$

If we instead asked this query of the probabilistic database in Figure 2, we would be computing the probability by summing over the worlds in which the query is true:

$$P(Q_1) = \sum_{\omega \models Q_1} P_\mathcal{P}(\omega)$$

Queries of this form that are a conjunction of atoms are called conjunctive queries. They are commonly shortened as:

$$Q_1 = S(x), CoA(x, y).$$

A disjunction of conjunctive queries is known as a union of conjunctive queries (UCQ). UCQs have been shown to live in a dichotomy of efficient evaluation [Dalvi and Suciu, 2012]: computing the probability of a UCQ is either polynomial in the size of the database, or it is $\#P$-hard. This property can be checked through the syntax of a query, and we say that a UCQ is safe if it admits efficient evaluation. In the literature of probabilistic databases [Suciu et al., 2011; Dalvi and Suciu, 2012], as well as throughout the rest of this paper, UCQs are the primary query object studied.

**Efficient Query Evaluation** For probabilistic databases to be useful, we need to be able to efficiently compute the probabilities of queries: we now describe how to do this. Algorithm 1 does this in polynomial time for all queries that can be computed efficiently (known as safe queries). We now explain the steps in further detail.

We begin with the assumption that $Q$ has been processed to not contain any constant symbols, and that all variables appear in the same order in repeated predicate occurrences in $Q$. These preprocessing steps are known as shattering and ranking respectively, and can be done efficiently [Dalvi and Suciu, 2012].

**Step 0** covers the base case where $Q$ is simple a tuple, so it looks up in $\mathcal{P}$. **Step 1** attempts to rewrite the UCQ into a conjunction of UCQs to find decomposable parts. For example, the UCQ $$(R(x) \land S(y, z)) \lor (S(x, y) \land T(x))$$ can be
Algorithm 1 Lift$^\text{R}(Q, P)$, abbreviated by L(Q)

Require: UCQ Q, prob. database P with constants T.
Ensure: The probability $P_T(Q)$

1: Step 0 Base of Recursion
2: if $Q$ is a single ground atom $t$
3: return $p$ else return 0
4: Step 1 Rewriting of Query
5: Convert Q to conjunction of UCQ: $Q_\land = Q_1 \land \cdots \land Q_m$
6: Step 2 Decomposable Conjunction
7: if $m > 1$ and $Q_\land = Q_1 \land Q_2$ where $Q_1 \perp Q_2$
8: return $L(Q_1) \cdot L(Q_2)$
9: Step 3 Inclusion-Exclusion
10: if $m > 1$ but $Q_\land$ has no independent $Q_i$
11: (Do Cancellations First)
12: return $\sum_{s \subseteq [m]} (-1)^{|s|+1} \cdot L(\bigvee_{s \in s} Q_i)$
13: Step 4 Decomposable Disjunction
14: if $Q = Q_1 \lor Q_2$ where $Q_1 \perp Q_2$
15: return $1 - (1 - L(Q_1)) \cdot (1 - L(Q_2))$
16: Step 5 Decomposable Existential Quantifier
17: if $Q$ has a separator variable $x$
18: return $1 - \prod_{c \in T} (1 - L(Q[x/c]))$
19: Step 6 Fail (the query is #P-hard)

written as the conjunction of $(R(x)) \lor (S(x, y) \land T(x))$ and $(S(y, z)) \lor (S(x, y) \land T(x))$. When multiple conjuncts are found this way, there are two options. If they are symbolically independent (share no symbols, denoted $\perp$), then Step 2 applies independence and recurse. Otherwise, Step 3 recurses using the inclusion-exclusion principle, performing cancellations first to maintain efficiency [Dalvi and Suciu, 2012]. If there is only a single UCQ after rewriting, Step 4 tries to split it into independent parts, applying independence and recursing if anything is found.

Next, Step 5 searches for a separator variable, one which appears in every atom in $Q$. If $x$ is a separator variable for $Q$, and $a, b$ are different constants in the domain of $x$, this means that $Q[x/a]$ and $Q[x/b]$ are independent. This independence is again recursively exploited. Finally, if Step 6 is reached, then the algorithm has failed and the query cannot be computed efficiently [Dalvi and Suciu, 2012].

2.3 Open-World Probabilistic Databases

In the context of automatically constructing a knowledge base, as is done in for example NELL [Carlson et al., 2010] or Google’s Knowledge Vault [Dong et al., 2014], making the closed-world assumption is conceptually unreasonable. Conversely, it is also not feasible to include all possible tuples and their probabilities in the knowledge base. The resulting difficulty is that there are an enormous number of probabilistic facts that can be scraped from the internet, and by definition these tools will keep only those with the highest probability. As a result, knowledge bases like NELL [Carlson et al., 2010], PaleoDeepDive [Peters et al., 2014], and YAGO [Suchanek et al., 2007] consist almost entirely of probabilities above 0.95.

This tells us that the knowledge base we are looking at is fundamentally incomplete. In response to this problem, Ceylan et al. [2016] propose the notion of a completion for a probabilistic database.

Definition 3. A λ-completion of a probabilistic database $P$ is another probabilistic database obtained as follows. For each atom $t$ that does not appear in $P$, we add tuple $(t : p)$ to $P$ for some $p \in [0, \lambda]$.

Then, we can define the open world of possible databases in terms of the set of distributions induced by all completions.

Definition 4. An open-world probabilistic database (Open-PDB) is a pair $\mathcal{G} = (\mathcal{P}, \lambda)$, where $\mathcal{P}$ is a probabilistic database and $\lambda \in [0, 1]$. $\mathcal{G}$ induces a set of probability distributions $\mathcal{K}_\mathcal{G}$ such that a distribution $P$ belongs to $\mathcal{K}_\mathcal{G}$ iff $P$ is induced by some $\lambda$-completion of probabilistic database $\mathcal{P}$.

Open-World Queries Open-PDBs specify a set of probability distributions rather than a single one, meaning that a given query produces a set of possible probabilities rather than a single one. We focus on computing the minimum and maximum possible probability values that can be achieved by completing the database.

Definition 5. The probability interval of a Boolean query $Q$ in OpenPDB $\mathcal{G}$ is $K_\mathcal{G}(Q) = \{L_\mathcal{G}(Q), P_\mathcal{G}(Q)\}$, where

\[
P_\mathcal{G}(Q) = \min_{P \in \mathcal{K}_\mathcal{G}} P(Q) \quad \quad \quad \quad P_\mathcal{G}(Q) = \max_{P \in \mathcal{K}_\mathcal{G}} P(Q)
\]

In general, computing the probability interval for some first-order $Q$ is not tractable. As observed in Ceylan et al. [2016], however, the situation is different for UCQ queries, because they are monotone (they contain no negations). For UCQs, the upper and lower bounds are given respectively by the full completion (where all unknown probabilities are $\lambda$), and the closed world database. This is a direct result of the fact that OpenPDBs form a credal set: a closed convex set of probability measures, meaning that probability bounds always come from extreme points [Cozman, 2000].

Furthermore, Ceylan et al. [2016] also provide an algorithm for efficiently computing this upper bound corresponding to a full completion, and show that it works whenever the UCQ is safe.

3 Mean-Constrained Completions

This section motivates the need to strengthen the OpenPDB semantics, and introduces our novel probabilistic data model.

3.1 Motivation

The ability to perform efficient query evaluation provides an appealing case for OpenPDBs. They give a more reasonable semantics, better matching their use, and for a large class of queries they come at no extra cost in comparison to traditional PDBs. However, in practice computing an upper bound in this way tends to give results very close to 1. Intuitively, this makes sense: our upper bound comes from simultaneously assuming that every possible missing atom has some reasonable probability. While such a bound is easy to compute, it is too strong of a relaxation of the closed-world assumption.
Table 1: Comparison of upper bounds for the same query and database with different model assumptions: Closed-World (CW), Open-World (OW), and Constrained Open-World (CoOW).

<table>
<thead>
<tr>
<th>Query</th>
<th>CW</th>
<th>OW</th>
<th>CoOW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(LiLA(x), S(x))</td>
<td>(0 \quad 1 - 10^{-290})</td>
<td>(1 - 10^{-15})</td>
<td>(1 - 10^{-191})</td>
</tr>
<tr>
<td>(LiSpr(x), S(x))</td>
<td>(0 \quad 1 - 10^{-191})</td>
<td>(0.96)</td>
<td>(1 - 10^{-191})</td>
</tr>
</tbody>
</table>

### 3.2 Formalization

We begin here by defining mean-based constraints, before examining some immediate observations about the structure of the resulting constrained database.

**Definition 6.** Suppose we have a PDB \(\mathcal{P}\), and let \(\text{Tup}(R) \subseteq \mathcal{P}\) be the set of probabilistic tuples in relation \(R\). Let \(\bar{p}\) be a probability threshold. Then a *mean tuple probability constraint* (MTP constraint) \(\varphi\) is a linear constraint of the form

\[
\bar{p} \geq \frac{1}{|\text{Tup}(R)|} \sum_{(t:p) \in \text{Tup}(R)} p
\]

### 3.3 Completion Properties

A necessary property of OpenPDBs for efficient query evaluation is that they are credal – this is what allows us to consider only a finite subset of possible completions. MTP-constrained OpenPDBs maintain this property.

**Proposition 1.** Suppose we have an OpenPDB \(\mathcal{G} = (\mathcal{P}, \lambda)\) together with MTP constraints \(\Phi\). Then the induced set of probability distributions \(K_\Phi^G\) is credal.

This property allows us to examine only a finite subset of configurations when looking at potential completions, since query probability bounds of a credal set are always achieved at points of extrema [Cozman, 2000]. Next, we would like to characterize these points of extremum, by showing that the number of tuples not on their own individual boundaries (that is, 0 or \(\lambda\)) is given by the number of MTP constraints.

**Theorem 2.** Suppose we have an OpenPDB \(\mathcal{G} = (\mathcal{P}, \lambda)\) with MTP constraints \(\Phi\), and a UCQ \(Q\). Then there exists a non-boundary \(\lambda\)-correlation \(\mathcal{G}'\) for which \(\mathcal{P}_{\mathcal{G}'}(Q) = \overline{\mathcal{P}_{\mathcal{G}}}(Q)\) and that contains some \(T \subseteq \mathcal{P}' \setminus \mathcal{P}\) such that \(|T| \leq |\Phi|\), and

\[
\forall \langle t : p \rangle \in T : p \in [0, \lambda], \text{ and } \forall \langle t : p \rangle \in (\mathcal{P}' \setminus \mathcal{P}) \setminus T : p \in \{0, \lambda\}.
\]

That is, our upper bound is given by a completion that has at most \(|\Phi|\) added tuples with probability not exactly 0 or \(\lambda\). Intuitively, each MTP constraint contributes a single non-boundary tuple, which can be thought of as the “leftover” probability mass once the rest has been assigned in full.

This insight allows us to treat MTP query evaluation as a combinatorial optimization problem for the rest of this paper. Thus, we only consider the case where achieving the mean tuple probability exactly leaves us with every individual tuple at its boundary. To see that we can do this, we observe that

\[1\]Proofs of all theorems and lemmas are available in appendix of the full version of the paper at http://starai.cs.ucla.edu/papers/FriedmanIJCAI19.pdf
Theorem 2 leaves a single tuple per MTP constraint not necessarily on the boundary. But this tuple can always be forced to be on the boundary by very slightly increasing the mean \( \bar{p} \) of the constraint, as follows.

**Corollary 3.** Suppose we have an OpenPDB \( G = (\mathcal{P}, \lambda) \) with MTP constraints \( \Phi \), and a UCQ \( Q \). Suppose further that each relation in \( G \) has at most 1 constraint in \( \Phi \), and that each constraint allows adding open-world probability mass exactly divisible by \( \lambda \). Then there exists a \( \Phi \)-constrained \( \lambda \)-completion \( \mathcal{P}' \), where \( K_{\Phi}^{\lambda}(Q) = \{P_p(Q), P_{p'}(Q)\} \), and

\[
\forall (t : p) \in \mathcal{P}' \setminus \mathcal{P}: \ p \in \{0, \lambda\}.
\]

Our investigation into the algorithmic properties of MTP query evaluation will be focused on constraining a single relation, subject to a single combinatorial budget constraint.

### 4 Exact MTP Query Evaluation

With Section 3 formalizing MTP constraints and showing that computing upper bounds subject to MTP constraints is a combinatorial problem of choosing which \( \lambda \)-probability tuples to add in the completion, we now investigate exact solutions. With this now being a combinatorial problem, we slightly change our terminology: “adding” an open-world tuple \( t \) to a relation means we consider only completions where \( P(t) = \lambda \), and a “budget” \( b \) for a relation means we can add up to \( b \) tuples while still satisfying the MTP constraint.

#### 4.1 An Algorithm for Inversion-Free Queries

We begin by describing a class of queries that admits polynotable evaluation subject to an MTP constraint. We first need to define some syntactic properties of queries.

**Definition 10.** Let \( Q \) be a conjunctive query, and let \( at(x) \) denote the set of relations containing variable \( x \). We say that \( Q \) is hierarchical if for any \( x, y \), we have either \( at(x) \subseteq at(y) \), \( at(y) \subseteq at(x) \), or \( at(x) \cap at(y) = \emptyset \).

Intuitively, a conjunctive query being hierarchical indicates that it can either be separated into independent parts (the \( at(x) \cap at(y) = \emptyset \) case), or there is some variable that appears in every atom. This simple syntactic property is the basis for determining whether query evaluation on a conjunctive query can be done in polynomial time [Dalvi and Suciu, 2007]. We can further expand on this definition in the context of UCQs.

**Definition 11.** A UCQ \( Q \) is inversion-free if each of its disjunctions is hierarchical, and they all share the same hierarchy.\(^2\)

Inversion-free queries represent an especially tractable class of queries for general inference. Since they are hierarchical, they are also safe, meaning query evaluation is efficient. Moreover, they precisely characterize the class of queries that support compilation to a tractable form for performing more complex queries, such as computing any joint distribution [Jha and Suciu, 2011]. This query class remains tractable under MTP constraints.

**Theorem 4.** For any inversion-free query \( Q \), evaluating the probability \( P_{\mathcal{P}}(\lambda)(Q) \) subject to an MTP constraint is in PTIME.

\(^2\)See Jha and Suciu [2011] for a more detailed definition.

In order to prove Theorem 4, we provide a polytime algorithm for MTP query evaluation on inversion-free queries. As with OpenPDBs, our algorithm depends on Algorithm 1, the standard lifted inference algorithm for PDBs that was discussed in Section 2.

We now present an algorithm for doing exact MTP query evaluation on inversion-free queries. For brevity, we present the case of a binary relation; the general case follows similarly and can be found in appendix. Suppose that we have a probabilistic database \( \mathcal{P} \), a domain \( T \) of constants denoted \( c \), a query \( Q \), and an MTP constraint on relation \( R(x, y) \) allowing us to add exactly \( b \) tuples with probability \( \lambda \). Suppose that \( Q \) immediately reaches Step 5 of Algorithm 1 (other steps will be discussed later), implying that \( x \) and \( y \) are unique variables in the query. We let \( A(c_x, c_y, b) \) denote the upper query probability of \( Q(x/c_x, y/c_y) \) subject to an MTP constraint allowing budget \( b \) on \( R \) restricted to \( x = c_x, y = c_y \). That is, \( A \) tells us the highest probability we can achieve for a partial assignment given a fixed budget. Observe that we can compute all entries of \( A \) using a slight modification of Algorithm 1 where we compute probabilities with and without each added tuple. This will take time polynomial in \( |T| \).

Next, we impose an ordering \( c_1, \ldots, c_{|T|} \) on the domain. Then we let \( D(j, c_y, b) \) denote the upper query probability of

\[
\bigvee_{c \in \{c_1, \ldots, c_j\}} Q(x/c, y/c_y)
\]

with a budget of \( b \) on the relevant portions of \( R \). Then \( D(|T|, c_y, b) \) considers all possible substitutions in our first index, meaning we have effectively removed a variable. Doing this repeatedly would allow us to perform exact MTP query evaluation. However, \( D \) is non-trivial to compute, and cannot be done by simply modifying Algorithm 1. Instead, we observe the following recurrence:

\[
D(j + 1, y/c_y, b) = \max_{k \in \{1, \ldots, b\}} 1 - (1 - D(j, y/c_y, b - k) \cdot (1 - A(x/c_{j+1}, y/c_y, k))
\]

Intuitively, this recurrence says that since the tuples from each fixed constant are all independent, we do not need to store which budget configuration on the first \( j \) constants got us our optimal solution. Thus, when we add the \( j + 1 \)th constant, we just need to check each possible value we could assign to our new constant, and see which gives the overall highest probability. This recurrence can be implemented efficiently, yielding a dynamic programming algorithm that runs in time polynomial in the domain size and budget.

Finally, we would like to generalize this algorithm beyond the assumption that \( Q \) immediately reaches Step 5 of Algorithm 1. Looking at other cases, we see that Steps 0 and 1 have no effect on this recurrence, and that Steps 2 and 4 correspond to multiplicative factors. For a query that reaches Step 5 (inclusion-exclusion), we need to construct such \( A \) and \( D \) for each term in the inclusion-exclusion sum, and follow the analogous recurrence.

Notice that the modified algorithm would only work in the case where we can always pick a common variable for all
sub-queries to do dynamic programming on – that is, when the query is inversion-free, as was our assumption. If the sub-calls generated by inclusion-exclusion do not share a common variable hierarchy, and thus an order for using our dynamic programming algorithm, we suffer an exponential blowup.

4.2 Queries with Inversion

We now show that allowing for inversions in safe queries can cause MTP query evaluation to become NP-hard. Interestingly, this means that MTP constraints fundamentally change the difficulty landscape of query evaluation.

To show this, we investigate the following UCQ query.

\[ M_0 = \exists x \exists y \exists z \ (R(x, y, z) \land U(x)) \lor (R(x, y, z) \land V(y)) \lor (R(x, y, z) \land W(z)) \]

A key observation here is that the query \( M_0 \) is a safe UCQ. That is, if we ignore constraints and evaluate it subject to the closed- or open-world semantics, computing the probability of the query would be polynomial in the size of the database. We now show that this is not the case for open-world query evaluation subject to a single MTP constraint on \( R \).

**Theorem 5.** Evaluating the upper query probability bound \( P^*_Q(M_0) \) subject to an MTP constraint \( \Phi \) on \( R \) is NP-hard.

The full proof of Theorem 5 can be found in appendix, showing a reduction from the NP-complete 3-dimensional matching problem to computing \( P^*_Q(M_0) \) with an MTP constraint on \( R \). It uses the following intuitive correspondence.

**Definition 12.** Let \( X, Y, Z \) be finite disjoint sets representing nodes, and let \( T \subseteq X \times Y \times Z \) be the set of available hyperedges. Then \( M \subseteq T \) is a matching if for any distinct triples \((x_1, y_1, z_1) \in M, (x_2, y_2, z_2) \in M\), we have that \( x_1 \neq x_2, y_1 \neq y_2, z_1 \neq z_2 \). The 3-dimensional matching decision problem is to determine for a given \( X, Y, Z, T \) and positive integer \( k \) if there exists a matching \( M \) with \( |M| \geq k \).

The set of available tuples for \( R \) will correspond to all edges in \( T \). The MTP constraint on \( R \) forces a decision on which subset of \( T \) to add to the database.

However, if we simply queried to maximize \( P(R(x, y, z)) \), this completion need not correspond to a matching. Instead, we have the disjunct \( R(x, y, z) \land U(x) \) which is maximized when each tuple chosen from \( R \) has a different \( x \) value. Similar disjuncts for \( y \) and \( z \) ensure that the query is maximized when using distinct \( y \) and \( z \) values. Putting all of these together ensures that the query probability is maximized when the subset of tuples chosen to complete \( R \) form a matching.

Finally, the last part of the query \( (U(x) \land V(y)) \lor (U(x) \land W(z)) \) ensures that inference on \( M_0 \) is tractable, but it is unaffected by the choice of tuples in \( R \).

5 Approximate MTP Query Evaluation

With Section 4.2 answering definitively that a general-purpose algorithm for evaluating MTP query bounds is unlikely to exist, even when restricted to safe queries, an approximation is the logical next step. We now restrict our discussion to situations where we constrain a single relation, and dig deeper into the properties of MTP constraints to show their submodular structure. We then exploit this property to achieve efficient bounds with guarantees.

5.1 On the Submodularity of Adding Tuples

To formally define and prove the submodular structure of the problem, we analyze query evaluation as a set function on adding tuples. We begin with a few relevant definitions.

**Definition 13.** Suppose that we have an OpenPDB \( G \), with an MTP constraint \( \varphi \) on a single relation \( R \), and we let \( O \) be the set of possible tuples we can add to \( R \). Then the set query probability function \( P_{R \cup \{t\}}(Q) \) is defined as

\[ P_{R \cup \{t\}}(Q) = P_{R \cup \{t\}}(Q) \]
until adding another tuple would violate the MTP constraint. Finally, we say that \( \hat{P}_{\text{Greedy}}(Q) \) is the approximation given by this greedy algorithm and recall that the true upper bound is \( \hat{P}_G(Q) \). We observe that \( \hat{P}_{\text{Greedy}}(Q) \leq \hat{P}_G(Q) \). Furthermore, Nemhauser et al. [1978] tells us the following:

\[
P_{\text{Greedy}}(Q) - P(Q) \geq (1 - \frac{1}{e})(\hat{P}_G(Q) - P(Q))
\]

Combining these and multiplying through gives us the following upper and lower bound on the desired probability.

\[
P_{\text{Greedy}}(Q) \leq \hat{P}_G(Q) \leq \frac{e \cdot P_{\text{Greedy}}(Q) - P(Q)}{e - 1}
\]

It should be noted that depending on the query and database, it is possible for this upper bound to exceed 1.

6 Discussion, Future & Related Work

We propose the novel problem of constraining open-world probabilistic databases at the schema level, without having any additional ground information over individuals. We introduced a formal mechanism for doing this, by limiting the mean tuple probability allowed in any given completion, and then sought to compute bounds subject to these constraints. We now discuss remaining open problems and related work.

Section 4 showed that there exists a query that is NP-hard to compute exactly, and also presented a tractable algorithm for a class of inversion-free queries. The question remains how hard the other queries are — in particular, is the algorithm presented complete. Is there a complexity dichotomy, that is, a set of syntactic properties that determine the hardness of a query subject to MTP constraints? Questions of this form are a central object of study in probabilistic databases. It has been explored for conjunctive queries [Dalvi and Suciu, 2007], UCQs [Dalvi and Suciu, 2012], and a more general class of queries with negation [Fink and Olteanu, 2016].

The central goal of our work is to find stronger semantics based on OpenPDBs, while still maintaining their desirable tractability. This notion of achieving a powerful semantics while maintaining tractability is a common topic of study. De Raedt and Kimmig [2015] study this problem by using a probabilistic interpretation of logic programs to define a model, leading to powerful semantics but a more limited scope of tractability [Fierens et al., 2015]. Description logics [Nardi et al., 2003] are a knowledge representation formalism that can be used as the basis for a semantics. This is implemented in a probabilistic setting in, for example, probabilistic ontologies [Riguzzi et al., 2012, 2015], probabilistic description logics [Heinsohn, 1994], probabilistic description logic programs [Łukasiewicz, 2005], or the bayesian description logics [Ceylan and Peñaloza, 2014].

Probabilistic databases in particular are of interest due to their simplicity and practicality. Foundational work defines a few types of probabilistic semantics, and provides efficient algorithms as well as when they can be applied [Dalvi and Suciu, 2004, 2007, 2012]. These algorithms along with practical improvements are implemented as industrial level systems such as MystiQ [Ré and Suciu, 2008], SPROUT [Olteanu et al., 2009], MayBMS [Huang et al., 2009], and Trio which implements the closely related Uncertainty-Lineage Databases [Benjelloun et al., 2007].

Problems outside of simple query evaluation are also points of interest for PDBs, for example the most probable database problem [Gribkoff et al., 2014], or the problem of ranking the top-k results [Ré et al., 2007]. In the context of OpenPDBs in particular, Grohe and Lindner [2018] study the notion of an infinite open world, using techniques from analysis to explore when this is feasible. Borgwardt et al. [2017] study an orthogonal way to introduce constraints on OpenPDBs to make the probability bounds realistic, by adding logical constraints based on an ontology.

Finally, probabilistic databases are closely related to other statistical relational models such as Markov logic networks [Richardson and Domingos, 2006] and probabilistic soft logic [Kimmig et al., 2012]. These models implicitly support the open-world assumption, although inference will not be efficient in general given the large number of random variables induced by the open world.

Acknowledgements

The authors thank Yitao Liang and YooJung Choi for helpful feedback. This work is partially supported by NSF grants #IIS1657613, #IIS-1633857, #CCF-1837129, DARPA XAI grant #N66001-17-2-4032, NEC Research and gifts from Intel and Facebook Research.

References


Ismail Ilkan Ceylan and Raúl Peñaloza. The bayesian description logic \( \mathcal{BE}_L \). In Stéphane Demri, Deepak Kapur, and Christoph Weidenbach, editors, Automated Reasoning, pages 480–494, Cham, 2014. Springer International Publishing.


A Proofs of Theorems, Lemmas, and Propositions

A.1 Proof of Proposition 1

Proof. To prove this, we need to show that $K^{\Phi}_{G}$ is both closed and convex.

Due to the way our constraints are defined, we know that $K^{\Phi}_{G} = K_{G} \cap K^{\Phi}$, where $K^{\Phi}$ is the set of all completions satisfying $\Phi$ (but not necessarily having all tuple probabilities $\leq \lambda$). We already know that $K_{G}$ is credal, and thus closed and convex. $K^{\Phi}$ is a half-space, which we also know is closed and convex. The intersection of closed spaces is closed, and the intersection of convex spaces is convex, so $K^{\Phi}_{G}$ is credal. $\square$

A.2 Proof of Theorem 2

Lemma 8. If $S \subseteq \mathbb{R}^n$ is a set formed by the intersection of $k < n$ half-spaces, $S$ has no points of extrema.

Proof. Written as a set of linear equalities, the solution clearly must have at least 1 degree of freedom. This indicates that for any potential extrema point $x$, one can move in either direction along this degree of freedom to construct an open line intersecting $x$, but entirely contained in $S$. $\square$

Proof. Since $K^{\Phi}_{G}$ is credal, we are interested here in determining the point of extrema of $K^{\Phi}_{G}$, as this will tell us precisely which completions can represent boundaries.

Consider the construction of the set $K^{\Phi}_{G}$, and suppose that there are $d$ possible open-world tuples, meaning that $K^{\Phi}_{G} \subseteq \mathbb{R}^d$. As we observed in the proof of Theorem 1, $K^{\Phi}_{G} = K_{G} \cap K^{\Phi}$, where $K^{\Phi}$ is the set of all completions satisfying $\Phi$. We now make three key observations about these sets:

1. Each individual possible open-world tuple is described by the intersection of 2 half-spaces: that is, the tuple on dimension $i$ is described by $x_i \geq 0$ and $x_i \leq \lambda$. $K_{G}$ is the intersection of all 2$d$ of these half-spaces.

2. For any individual open-world tuple, the boundaries of the two half-spaces that describe it cannot intersect each other.

3. An MTP constraint is a linear constraint, and thus can be described by a single half-space. So $K^{\Phi}$ is described by the intersection of these $|\Phi|$ half-spaces.

Observations 1 and 3, together with Lemma 8 tell us that any point of extrema of $K^{\Phi}_{G}$ must be given by the intersection of the boundaries of at least $d$ of the half-spaces that form $K^{\Phi}_{G}$. Observation 3 tells us that at most $|\Phi|$ of these half-spaces come from MTP constraints, leaving the boundaries of at least $d - |\Phi|$ half-spaces which come from $K_{G}$. Finally, observation 2 tells us that each of these $d - |\Phi|$ half-spaces is describing a different open world tuple. But this means we must have at least $d - |\Phi|$ tuples which lie on the boundary of one of their defining half-spaces: they must be either 0 or $\lambda$. $\square$

A.3 Proof of Theorem 5

Before we present the formal proof, we state and prove 2 Lemmas we will need.

Lemma 9. Suppose we have two completions $P_1$ and $P_2$ of $R$, which only differ on a single triple, that is $P_1 = P_0 \cup \{x_1,y_1,z_1\}$ and $P_2 = P_0 \cup \{x_2,y_2,z_2\}$. Further suppose that $y_1 = y_2$, $z_1 = z_2$, and that $P_0$ contains no triples with $x$-value $x_1$, but does contain at least 1 triple with $x$-value $x_2$. Then $P_1(M_0) > P_2(M_0)$.

Proof. Let us directly examine $\Delta$, the logical formula found by grounding $M_0$. Since $M_0$ is a union of conjunctive queries, $\Delta$ must be a DNF. Each disjunct either does not contain $R$, in which case it does not vary with the choice of completion, or it contains it exactly once. Any disjunct containing an atom of $R$ not assigned probability by a completion is logically false.

In order to prove that $P_1(M_0) > P_2(M_0)$, let us compare the ground atoms that result from each. It is clear that the only spot on which they differ is on disjuncts involving $R(x_1,y_1,z_1)$ or $R(x_2,y_1,z_1)$. Any disjuncts involving one of these and $V$ or $W$ will also have an identical effect on the probability of the query, since the completions are identical over $y$ and $z$.

Finally this means we need to compare the term $R(x_1,y_1,z_1), U(x_1)$ with the term $R(x_2,y_1,z_1), U(x_2)$. Observe that we know $P_0$ contains triples with $x$-value $x_2$, which means the term only contributes new probability mass when $U(x_2)$ is true and none of the other triples involving $x_2$ are true. However, $P_0$ does not contain any triples with $x$-value $x_1$, so the term $R(x_1,y_1,z_1), U(x_1)$ contributes the maximum probability possible. Thus, for any choice of probabilities on $U$ such that $U(x_2) < 1$, we have that $P_1(M_0) > P_2(M_0)$. $\square$

Lemma 10. The upper bound $P(M_0)$ subject to an MTP constraint on $R$ allowing for $k \cdot \lambda$ total probability mass is maximized if and only if $P$ is a completion formed by a matching of size $k$, where $k$ is the maximum number of tuples with probability $\lambda$ that can be added to $R$ in $M_0$.

Proof. Observe that if we begin with a completion given by a matching, we can repeatedly apply Lemma 9 to arrive at any completion. Thus a completion given by a matching must have higher probability than any completion not given by a matching. $\square$

Finally, we are ready to present the proof of Theorem 5.

Proof. Suppose we are given an instance of a 3-dimensional matching problem $X,Y,Z,T$ and an integer $k$. Let $U(x), V(y), W(z)$ be 0.8 wherever $x \in X, y \in Y$, or $z \in Z$ respectively, and 0 everywhere else. Additionally, let $R(x,y,z)$ be unknown for any $(x,y,z) \in T$, and 0 otherwise. Finally, we place an MTP constraint on $R$ ensuring that at most $k$ tuples can be added, and let $\lambda = 0.8$. Then Lemma 10 tells us that $M_0$ evaluated on this database will be maximized if and only if the completion used corresponds to a matching of size $k$. We determine this probability $P_{\max}$ using a standard probabilistic database query algorithm, and
fixing $R$ to have entries 0.8 for some disjoint set of triples. Observe that this computation is easy if we allow any tuples from $R$ to have non-zero probability: this corresponds to a matching in the setting where all hyperedges are available, but will still give the same query probability.

Finally, we use our oracle for MTP constrained query evaluation to check $P(M_B)$ with the database we constructed from the matching problem. We compare the upper bound given by the oracle, and if it is equal to $P_{max}$, Lemma 10 tells us that a matching of size $k$ does exist. Similarly, if the upper bound given by the oracle is lower than $P_{max}$, Lemma 10 tells us a matching of size $k$ does not exist.

\begin{proof}{Proof of Theorem 6}
Without directly computing probabilities, let us inspect $\Delta$, the logical formula we get by grounding $Q$. $Q$ is a union of conjunctive queries, and thus $\Delta$ is a DNF. Each disjunct can contain our constrained relation $R$ at most once due to the query not having self-joins, and any one of these disjuncts containing an atom of $R$ not assigned any probability is logically false.

Next, to show that $S_{P,Q}$ is submodular, let $X \subseteq Y \subseteq O$, and let $x \in O \setminus Y$ be given. We assign names to the following subformulas of $\Delta$:

- $\alpha$ ($\beta$) is the disjunction of all disjuncts of $\Delta$ which are not logically false due to missing $R_B$ tuples in $X$ ($Y$)
- $\gamma$ is the disjunction of all disjuncts of $\Delta$ containing the tuple $x$

Additionally, since $X \subseteq Y$, we also know that $\alpha \Rightarrow \beta$.

Now, we make a few observations relating these quantities with our desired values for submodularity:

- $S_{P,Q}(X) = P(\alpha)$
- $S_{P,Q}(Y) = P(\beta)$
- $S_{P,Q}(X \cup \{x\}) = P(\alpha \lor \gamma)$
- $S_{P,Q}(Y \cup \{x\}) = P(\beta \lor \gamma)$

Finally, we have the following:

\[
S_{P,Q}(X \cup \{x\}) - S_{P,Q}(X) = P(\alpha \lor \gamma) - P(\alpha) \\
\geq P(\alpha \lor \gamma) \\
= P(\beta \lor \gamma) - P(\beta) \\
= S_{P,Q}(Y \cup \{x\}) - S_{P,Q}(Y)
\]

\end{proof}

\section{General Algorithm for Inversion-Free Queries}

We now present an algorithm for doing exact MTP query evaluation on inversion-free queries. Suppose that we have a probabilistic database $P$, a domain $T$ of constants denoted $c$, a query $Q$, and an MTP constraint on relation $R(x_1, x_2, \ldots, x_r)$ allowing us to add $B$ tuples. For any $I \subseteq \{1, \ldots, r\}$, we let $A(x_{i_1}/c_{i_1}, x_{i_2}/c_{i_2}, \ldots, x_{i_{|I|}}/c_{i_{|I|}}, b)$ denote the upper query probability of $Q(x_{i_1}/c_{i_1}, x_{i_2}/c_{i_2}, \ldots, x_{i_{|I|}}/c_{i_{|I|}})$ subject to an MTP constraint allowing budget $b$ on the relevant portion of $R$. That is, $A$ tells us the highest probability we can achieve for a partial assignment given a fixed budget. Observe that we can compute all entries of $A$ using a slight modification of Algorithm 1. This will take time polynomial in $|T|$.

Next, we impose an ordering $c_1, c_2, \ldots, c_{|T|}$ on the domain. For any $I \subseteq \{1, \ldots, r\}$, we let $D(j, x_{i_1}/c_{i_1}, x_{i_2}/c_{i_2}, \ldots, x_{i_{|I|}}/c_{i_{|I|}}, b)$ denote the upper query probability of $Q(x_{i_1}/c_{i_1}, x_{i_2}/c_{i_2}, \ldots, x_{i_{|I|}}/c_{i_{|I|}})$ with a budget of $b$ on the relevant portions of $R$. Then $D(|T|, x_{i_1}/c_{i_1}, x_{i_2}/c_{i_2}, \ldots, x_{i_{|I|}}/c_{i_{|I|}}, b)$ considers all possible substitutions in our first index, meaning we no longer need to worry about it. Doing this repeatedly would allow us to perform exact MTP query evaluation. However, $D$ is non-trivial to compute, and cannot be done by simply modifying Algorithm 1. Instead, we observe the following recurrence:

\[
D(j + 1, x_{i_1}/c_{i_1}, \ldots, x_{i_{|I|}}/c_{i_{|I|}}, b) = \max_{k \in \{1, \ldots, b\}} (1 - D(j, x_{i_1}/c_{i_1}, \ldots, x_{i_{|I|}}/c_{i_{|I|}}, b - k) \cdot A(x_{i_1}/c_{j+1}, x_{i_2}/c_{i_2}, \ldots, x_{i_{|I|}}/c_{i_{|I|}}, k))
\]

Intuitively, this recurrence is saying that since the tuples from each fixed constant are independent of each other, we can add a new constant to our vocabulary by simply considering all combinations of budget assignments. This recurrence can be implemented efficiently, yielding a dynamic programming algorithm that runs in time polynomial in the domain size and budget.

The keen reader will now observe that the above definition and recurrence only make sense if $Q$ immediately reaches Step 5 of Algorithm 1. While this is true, we see that Steps 0 and 1 have no effect on this recurrence, and Steps 2 and 4 correspond to multiplicative factors. For a query that reaches Step 5: inclusion-exclusion, we indeed need to construct such matrices for each sub-query. Notice that the modified algorithm would only work in the case where we can always pick a common $x_i$ for all sub-queries to do dynamic programming on - that is, when the query is inversion-free.