Ontology-Mediated Query Answering for Probabilistic Temporal Data with $\mathcal{EL}$ Ontologies*

Patrick Koopmann

Institute for Theoretical Computer Science, Technische Universität Dresden, Germany
patrick.koopmann@tu-dresden.de

Abstract. Especially in the field of stream reasoning, there is an increased interest in reasoning about temporal data in order to detect situations of interest or complex events. Ontologies have been proved a useful way to infer missing information from incomplete data, or simply to allow for a higher order vocabulary to be used in the event descriptions. Motivated by this, ontology-based temporal query answering has been proposed as a means for the recognition of situations and complex events. But often, the data to be processed do not only contain temporal information, but also probabilistic information, for example because of uncertain sensor measurements. While there has been a plethora of research on ontology-based temporal query answering, only little is known so far about querying temporal probabilistic data using ontologies. This work addresses this problem by introducing a temporal query language that extends a well-investigated temporal query language with probability operators, and investigating the complexity of answering queries using this query language together with ontologies formulated in the description logic $\mathcal{EL}$.

Keywords: Ontology-Based Query Answering · Temporal Query Answering · Probabilistic Reasoning

1 Introduction

Ontology-mediated query answering (OMQA) recently attracted considerable attention as a technique to query incomplete data. In OMQA, queries are evaluated with respect to an ontology, which specifies background knowledge about the current domain using a formal language such as a description logic (DL), so that, using reasoning procedures, also implicit information can be queried from the data. In the standard OMQA setting, the data to be queried is assumed to be both static and precise. However, a lot of applications encounter situations where this assumption fails, yet using ontologies could prove useful. The internet has become highly dynamic, with information being frequently added and changed, and new data being generated from a variety of sources. In addition, new technologies such as smartphones and the internet of things (IoT) frequently encounter a data environment that is constantly changing. To make use of these

* Supported by the DFG within the collaborative research center SFB 912 (HAEC).

data, there has been an increasing interest in investigating semantic and reasoning techniques that process not only static data, but streams of data, such as in the semantic stream reasoning paradigm [24]. As [24] illustrate, frequently, the data encountered in stream reasoning applications is not only temporal, but also probabilistic in nature.

As an example, consider a health or fitness monitoring application, for which one may want to use concepts from a medical ontology such as SNOMED CT [14] or Galen [25] to describe information about the health status of a patient. Specifically, such an application could be used on a smartphone in combination with a sensor that measures the diastolic blood pressure of the patient while he is exercising [21]. As the sensor might be imprecise in its measurements, it might report information about whether the blood pressure of the patient is high with an associated probability, and provide this information to the application in regular time intervals. If a too high blood pressure was observed for several times during a short period, the app should give a warning to the patient, and advise him to take a break from his exercise.

In order to properly take both the temporal and the probabilistic aspects into account when querying streams of data, we propose a query language for OMQA that comes with both temporal and probabilistic operators. For this, we assume a representation of the data in form of a sequence of probabilistic data sets, which may have been obtained using further preprocessing and windowing operations. An ontology expressed in a description logic (DL) gives additional background information about the domain to be queried, so that implicit information can be queried from incomplete data through reasoning. In the above scenario, the following query could for example be used to detect whether the patients blood pressure was at least twice recorded as high during the last 10 minutes.

$$\mathsf{P}_{>.8}(\bigcirc^{-10}\Diamond(\mathsf{HighBloodPressure}(x) \wedge \Diamond\mathsf{HighBloodPressure}(x)))$$

While there has been a lot of research on querying temporal data [1] and probabilistic data [7,19] using ontologies, we are not aware of any research were both aspects are combined in the specific setting we described. In this work, we focus on the setting where the ontology is formulated in $\mathcal{EL}$, a DL that is known for its good computational properties, such as polynomial decidability for most common reasoning problems. This DL, which underlies the OWL EL profile of the web ontology language standard OWL, is used for many large scale ontologies, especially in the bio-medical domain and for the semantic web, such as for the ontologies SNOMED CT and Galen mentioned above. However, our hardness results already apply for simpler description logics such as *DL-Lite*, as well as for the case where no ontology is used.

Related Work Our language is an extension of the temporal query language investigated in [3,8], which extends conjunctive queries with LTL operators. Other authors considered using these operators also as part of the DL, either to describe temporal concepts [16], or to make the axioms of the ontology itself temporal [4]. Recently, this work has been extended also to metric temporal logics,

in which temporal operators are annotated with numerical time intervals [2,9,18]. Temporal reasoning for streams of data has recently also been considered in the context of datalog [26]. Surveys on temporal reasoning and query answering with ontologies can be found in [1,23].

Our probabilistic query-answering framework is based on the OMQA framework for probabilistic data presented in [19]. Since this publication, several authors investigated OMQA in similar settings [7,6,12]. To our knowledge, the only work that combines both temporal and probabilistic query answering in the presence of description logic ontologies is [11]. Albeit, the authors consider a different setting, in which the flow of time is modelled by a Markov-process. In contrast, we consider temporal data that are provided as a sequence of probabilistic ABoxes. In addition to settings based on probabilistic databases, there is also research on extending DLs with probability operators, such as in **P**-$\mathcal{SHIF}(\mathcal{D})$/**P**-$\mathcal{SHOIN}(\mathcal{D})$ [22] or Prob-$\mathcal{ALC}$/Prob-$\mathcal{EL}$ [17]. While our DL does not support probability operators, the probability operator used in our query language syntactically and semantically corresponds to the probability operator in Prob-$\mathcal{ALC}$ and Prob-$\mathcal{EL}$.

Formal details of the proofs can be found in the extended version of the paper [20].

2 Preliminaries

We recall the DL $\mathcal{EL}$ [5] studied here. Let $\mathsf{N_C}$,$\mathsf{N_R}$ and $\mathsf{N_I}$ be countably infinite and pair-wise disjoint sets of respectively *concept names*, *role names* and *individual names*. An $\mathcal{EL}$ *concept* is of one of the forms

$$\top \mid A \mid C_1 \sqcap C_2 \mid \exists r.C$$

where $A \in \mathsf{N_C}$, $r \in \mathsf{N_R}$, and C_1, C_2 and C are $\mathcal{EL}$ concepts. An $\mathcal{EL}$ *axiom* is of the form $C_1 \sqsubseteq C_2$, where C_1 and C_2 are $\mathcal{EL}$ concepts. An $\mathcal{EL}$ *TBox* is a set of $\mathcal{EL}$ axioms. An *ABox* $\mathcal{A}$ is a set of *assertions* of the form $A(a)$ and $r(a,b)$, where $A \in \mathsf{N_C}$, $r \in \mathsf{N_R}$ and $a, b \in \mathsf{N_I}$. An $\mathcal{EL}$ knowledge base (KB) is a tuple $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$, where $\mathcal{A}$ is an ABox and $\mathcal{T}$ an $\mathcal{EL}$ TBox.

The semantics of KBs is defined in terms of *interpretations*, which are tuples $\mathcal{I} = \langle \Delta^\mathcal{I}, \cdot^\mathcal{I} \rangle$, $\Delta^\mathcal{I}$ being a set of *domain elements*, and $\cdot^\mathcal{I}$ an *interpretation function* that maps each $a \in \mathsf{N_I}$ to an element $a^\mathcal{I} \in \Delta^\mathcal{I}$ s.t. for $a \neq b \in \mathsf{N_I}$, $a^\mathcal{I} \neq b^\mathcal{I}$ (unique name assumption, UNA), each $A \in \mathsf{N_C}$ to a subset $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$, each $r \in \mathsf{N_R}$ to a subset $r^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}$. The interpretation function $\cdot^\mathcal{I}$ is extended to concepts and roles as follows:

$$\top^\mathcal{I} = \Delta^\mathcal{I} \qquad (C_1 \sqcap C_2)^\mathcal{I} = C_1^\mathcal{I} \cap C_2^\mathcal{I}$$
$$(\exists r.C_1)^\mathcal{I} = \{d \in \Delta^\mathcal{I} \mid \exists (d,e) \in r^\mathcal{I}, e \in C_1^\mathcal{I}\},$$

where C_1, C_2 are concepts and $r \in \mathsf{N_R}$. An interpretation is a model of a KB $\langle \mathcal{T}, \mathcal{A} \rangle$ (of an TBox) if for every $C \sqsubseteq D \in \mathcal{T}$, $C^\mathcal{I} \subseteq D^\mathcal{I}$, for every $A(a) \in \mathcal{A}$, $a^\mathcal{I} \in A^\mathcal{I}$, and for every $r(a,b) \in \mathcal{A}$, $(a^\mathcal{I}, b^\mathcal{I}) \in r^\mathcal{I}$.

A *conjunctive query* (CQ) takes the form $q = \exists \boldsymbol{y}.\phi(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$, $\boldsymbol{y}$ are vectors of variables and $\phi(\boldsymbol{x}, \boldsymbol{y})$ is a conjunction over atoms of the forms $A(t_1)$ and $r(t_1, t_2)$, where $A \in \mathsf{N_C}$, $r \in \mathsf{N_R}$, and t_1 and t_2 are *terms* taken from $\mathsf{N_I}$, $\boldsymbol{x}$ or $\boldsymbol{y}$. $\boldsymbol{x}$ are the *answer variables* of q. Given an interpretation $\mathcal{I}$ and a CQ q with answer variables $x_1, \ldots, x_n$, the vector $a_1 \ldots a_n \subseteq {\mathsf{N_I}}^n$ is an *answer of q in $\mathcal{I}$* if there exists a mapping $\pi : \mathsf{term}(q) \to \Delta^{\mathcal{I}}$ s.t. $\pi(x_i) = a_i$ for $i \in [\![1, n]\!]$ $\pi(b) = b^{\mathcal{I}}$ for $b \in \mathsf{N_I}$, $\pi(t) \in A^{\mathcal{I}}$ for every $A(t)$ in q, and $\langle \pi(t_1), \pi(t_2) \rangle \in r^{\mathcal{I}}$ for every $r(t_1, t_2)$ in q. $a_1 \ldots a_n$ is a *certain answer* of q in a KB $\mathcal{K}$ if it is an answer in every model of $\mathcal{K}$. If a query does not contain any answer variables, it is a *Boolean CQ*, and we say it is *entailed* by a KB $\mathcal{K}$ (an interpretation $\mathcal{I}$) if it has the empty vector as answer.

3 Temporal Probabilistic Knowledge Bases and Queries

We introduce temporal probabilistic knowledge bases (TPKBs) and temporal probabilistic queries (TPQs).

Temporal Probabilistic Knowledge Bases. Probabilistic information about a single time point is represented using a probabilistic ABox as introduced in [19]. For simplicity, we focus on assertion-independent probabilistic ABoxes (ipABoxes), though all results should easily extend to the more general case. ipABoxes are the ABox equivalent of tuple-independent probabilistic databases [13]. An *ipABox* is a set of *probabilistic ABox assertions* of the form $\alpha : p$, where α is an ABox assertion and $p \in [0, 1]$. Intuitively, $\alpha : p$ describes that the assertion α holds with a probability of at least p. Instead of $\alpha : 1$, we may just write α if the meaning is clear from the context. ipABoxes only specify a *lower bound* on the probability, to conform with the open-world semantics common in ontology-based representations.[1]

An $\mathcal{EL}$ TPKB is now a tuple $\langle \mathcal{T}, (\mathcal{A}_i)_{i \in [\![1,n]\!]} \rangle$, where $\mathcal{T}$ is an $\mathcal{EL}$ TBox and $(\mathcal{A}_i)_{i \in [\![1,n]\!]}$ is a sequence of n ipABoxes. Given a TPKB $\mathcal{K} = \langle \mathcal{T}, (\mathcal{A}_i)_{i \in [\![1,n]\!]} \rangle$, the set $\Omega_{\mathcal{K}}$ of *possible worlds of* $\mathcal{K}$ contains all sequences $w = (\mathcal{A}'_i)_{i \in [\![1,n]\!]}$ of classical ABoxes such that for every $i \in [\![1, n]\!]$ and $\alpha \in \mathcal{A}'_i$, $\mathcal{A}_i$ contains an axiom of the form $\alpha : p$. Each TPKB uniquely defines a probability space $\langle \Omega_{\mathcal{K}}, \mu_{\mathcal{K}} \rangle$, where $\mu_{\mathcal{K}} : 2^{\Omega_{\mathcal{K}}} \to [0, 1]$ satisfies

$$\mu_{\mathcal{K}}(\{(\mathcal{A}'_i)_{i \in [\![1,n]\!]}\}) = \prod_{\substack{i \in [\![1,n]\!] \\ \alpha : p \in \mathcal{A}_i \\ \alpha \in \mathcal{A}'_i}} p \cdot \prod_{\substack{i \in [\![1,n]\!] \\ \alpha : p \in \mathcal{A}_i \\ \alpha \notin \mathcal{A}'_i}} (1 - p)$$

and for $W \subseteq \Omega_{\mathcal{K}}$, $\mu_{\mathcal{K}}(W) = \sum_{w \in W} \mu(\{w\})$. Intuitively, $\mu_{\mathcal{K}}(W)$ gives the probability of being in one of the possible worlds in W, by summing up the probabilities of each possible world. The definition of $\mu_{\mathcal{K}}(W)$ reflects the assumption that all probabilities in the TPKB are statistically independent.

[1] Note that this is different to the open-world semantics for probabilistic databases suggested in [10], which assumes a fixed upper probability for facts absent in the data.

$\Omega_\mathcal{K}$	$\mathcal{A}'_1$	$\mathcal{A}'_2$	$\mathcal{A}'_3$	$\mathcal{A}'_4$	$\mathcal{A}'_5$	$\mu_\mathcal{K}$
w_1	$\{\mathsf{hasBP}(p,b), \mathsf{HighBP}(b)\}$	$\emptyset$	$\{\mathsf{HighBP}(b)\}$	$\{\mathsf{HighBP}(b)\}$	$\emptyset$	0.378
w_2	$\{\mathsf{hasBP}(p,b)\}$	$\emptyset$	$\{\mathsf{HighBP}(b)\}$	$\{\mathsf{HighBP}(b)\}$	$\emptyset$	0.162
w_3	$\{\mathsf{hasBP}(p,b), \mathsf{HighBP}(b)\}$	$\emptyset$	$\emptyset$	$\{\mathsf{HighBP}(b)\}$	$\emptyset$	0.042
w_4	$\{\mathsf{hasBP}(p,b)\}$	$\emptyset$	$\emptyset$	$\{\mathsf{HighBP}(b)\}$	$\emptyset$	0.018
w_5	$\{\mathsf{hasBP}(p,b), \mathsf{HighBP}(b)\}$	$\emptyset$	$\{\mathsf{HighBP}(b)\}$	$\emptyset$	$\emptyset$	0.252
w_6	$\{\mathsf{hasBP}(p,b)\}$	$\emptyset$	$\{\mathsf{HighBP}(b)\}$	$\emptyset$	$\emptyset$	0.108
w_7	$\{\mathsf{hasBP}(p,b), \mathsf{HighBP}(b)\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	0.028
w_8	$\{\mathsf{hasBP}(p,b)\}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	0.012

Table 1. Probability space of example TPKB.

Example 1. We define the TPKB $\mathcal{K} = \langle \mathcal{T}, (\mathcal{A}_i)_{i\in[\![1,5]\!]} \rangle$ where $\mathcal{T}$ contains the GCI

$$\mathsf{HighBloodPressurePatient} \equiv \exists \mathsf{hasBloodPressure}.\mathsf{HighBloodPressure}$$

and the ABoxes $\mathcal{A}_1 = \{\mathsf{hasBP}(p,b), \mathsf{HighBP}(b) : 0.7\}$, $\mathcal{A}_2 = \emptyset$, $\mathcal{A}_3 = \{\mathsf{HighBP}(b) : 0.9\}$, $\mathcal{A}_4 = \{\mathsf{HighBP}(b) : 0.6\}$ and $\mathcal{A}_5 = \emptyset$, where BP is short for $\mathsf{BloodPressure}$. Every possible world $w = (\mathcal{A}'_i)_{i\in[\![1,5]\!]}$ with $\mathsf{hasBP}(p,b) \notin \mathcal{A}'_1$ has probability $\mu_\mathcal{K}(w) = 0$. The remaining possible worlds are shown in Table 1, with the probability measure $\mu_\mathcal{K}$ shown in the last column.

A *model* of a TPKB $\mathcal{K} = \langle \mathcal{T}, (\mathcal{A}_i)_{i\in[\![1,n]\!]} \rangle$ is a mapping ι from possible worlds $w = (\mathcal{A}'_i)_{i\in[\![1,n]\!]} \in \Omega_\mathcal{K}$ to sequences $(\iota(w)_i)_{i>0}$ of (classical) models of $\mathcal{T}$ s.t. for all $i \in [\![1,n]\!]$, $\iota(w)_i$ is a model of the classical knowledge base $\langle \mathcal{T}, \mathcal{A}'_i \rangle$, and all $\iota(w)_i$ have the same set Δ^ι of domain elements (constant domain assumption).

Rigid Names. As typical for temporal knowledge bases, we may assume in addition a set $\mathsf{N_{rig}}$ of *rigid names*, containing the set $\mathsf{N_{Crig}} \subseteq \mathsf{N_C}$ of *rigid concept names* and the set $\mathsf{N_{Rrig}} \subseteq \mathsf{N_R}$ of *rigid role names*. Rigid names denote names whose interpretation is independent of the flow of time. We say that a model ι of a TPKB $\mathcal{K} = \langle \mathcal{T}, (\mathcal{A}_i)_{i\in[\![1,n]\!]} \rangle$ *respects rigid names* iff for all $w \in \Omega_\mathcal{K}$, $i, j \in [\![1,n]\!]$ and $X \in \mathsf{N_{rig}}$, $X^{\iota(w)_i} = X^{\iota(w)_j}$. Allowing for rigid names often has a direct impact on complexity and decidability of common reasoning problems, which is why typically different cases based on whether $\mathsf{N_{Crig}} = \emptyset$ or $\mathsf{N_{Rrig}} = \emptyset$ are studied for complexity.

Example 2. In the above example, the relation hasBP is rigid, as its interpretation should be independent of time, while the concept HighBP is not rigid, as the blood pressure of a patient can change from high to not high. As a consequence, the individual p will be related to the blood pressure b at all time points, even though the assertion $\mathsf{hasBP}(p,b)$ is only placed in the ipABox $\mathcal{A}_1$.

Temporal Probabilistic Queries. A *temporal probabilistic query* (TPQ) is of one of the following forms, where q is a CQ, ϕ_1 and ϕ_2 are a TPQs and $p \in [0,1]$.

$$q \mid \neg\phi_1 \mid \phi_1 \wedge \phi_2 \mid \phi_1 \vee \phi_2 \mid \bigcirc\phi_1 \mid \Diamond\phi_1 \mid \Box\phi_1 \mid \phi_1 \mathcal{U} \phi_2$$
$$\bigcirc^-\phi_1 \mid \Diamond^-\phi_1 \mid \Box^-\phi_1 \mid \phi_1 \mathcal{S} \phi_2 \mid \mathsf{P}_{>p}\phi_1 \mid \mathsf{P}_{=p}\phi_1 \mid \mathsf{P}_{<p}\phi_1$$

ϕ	$\iota, w, i \models \phi$ iff	ϕ	$\iota, w, i \models \phi$ iff
$\exists \boldsymbol{y}.\psi(\boldsymbol{y})$	$\iota(w), i \models \exists \boldsymbol{y}.\psi(\boldsymbol{y})$	$\neg\phi$	$\iota, w, i \not\models \phi$
$\phi_1 \wedge \phi_2$	$\iota, w, i \models \phi_1$ and $\iota, w, i \models \phi_2$	$\phi_1 \vee \phi_2$	$\iota, w, i \models \phi_1$ or $\iota, w, i \models \phi_2$
$\bigcirc\phi_1$	$\iota, w, i+1 \models \phi_1$	$\Diamond\phi_1$	$\iota, w, j \models \phi_1$ for some $j \geq i$
$\Box\phi_1$	$\iota, w, j \models \phi_1$ for all $j \geq i$	$\phi_1 \mathcal{U} \phi_2$	$\iota, w, j \models \phi_2$ for some $j \geq i$ and $\iota, w, k \models \phi_1$ for all $k \in [\![i, j-1]\!]$
$\bigcirc^- \phi_1$	$\iota, w, i-1 \models \phi_1$ and $i > 0$	$\Diamond^- \phi_1$	$\iota, w, j \models \phi_1$ for some $j \leq i$
$\Box^- \phi_1$	$\iota, w, j \models \phi_1$ for all $j \leq i$	$\phi_1 \mathcal{S} \phi_2$	$\iota, w, j \models \phi_2$ for some $j \leq i$ and $\iota, w, k \models \phi_1$ for all $k \in [\![j+1, i]\!]$
$\mathsf{P}_{\sim p}\phi$	$\mu_{\mathcal{K}}(\{w \in \Omega_{\mathcal{K}} \mid \iota, w, i \models \phi\}) \sim p$, where $\sim \in \{<, =, >\}$		

Table 2. Entailment of Boolean TPQs in the possible world w at time point i under interpretation ι.

The operators $\bigcirc$ (next), $\Diamond$ (eventually), $\mathcal{U}$ (until) and their inverses are temporal operators of LTL, while $\mathsf{P}_>$, $\mathsf{P}_=$ and $\mathsf{P}_<$ are the probability operators that we add to this language. TPQs without probability operators corresponds to *temporal queries (TQs)* investigated in [8]. Note that due the disjunction operator, we can also express *unions of conjunctive queries* (UCQs), which are simply disjunctions of CQs. The answer variables of a TPQ ϕ are the answer variables of the CQs occurring in ϕ. A TPQ ϕ is *Boolean* if every variable in ϕ is bound by an existential quantifier.

We define the semantics of TPQs. Note that each possible world $w \in \Omega_{\mathcal{K}}$ has its own time line, while a model of $\mathcal{K}$ contains a sequence of models for every possible world. For a given model, we define the semantics of temporal operators with respect to a single time line, that is, with respect to a current possible world. Probabilistic expressions $\mathsf{P}_{\sim p}\phi$ are the only expressions that are interpreted with respect to other possible worlds.

Let ι be a model of $\mathcal{K}$, and ϕ a Boolean TPQ. For a single possible world $w \in \Omega_{\mathcal{K}}$ and a time point i, we say that ϕ is satisfied at w, i under ι, in symbols $\iota, w, i \models \phi$ iff the conditions in Table 2 are satisfied. Note that the temporal operators refer to the time line of a single possible world, for which they are defined as in [8]. In contrast, the probabilistic operator refers to the current time point in multiple possible worlds, and is defined similar to the probabilistic concept constructor in the DL Prob-$\mathcal{ALC}$ [17]. A Boolean TPQ ϕ is *satisfied* in an interpretation ι at i, in symbols $\iota, i \models \phi$, iff $\iota, w, i \models \phi$ for all $w \in \Omega_{\mathcal{K}}$. It is *entailed* by the TPKB $\mathcal{K}$ at i iff $\iota, i \models \phi$ for all models ι of $\mathcal{K}$. ϕ is *satisfiable in* $\mathcal{K}$ *at* i iff there exists a model ι of $\mathcal{K}$ s.t. $\iota, i \models \phi$.

Now given a TPKB $\mathcal{K}$, a TPQ ϕ with answer variables $\boldsymbol{x}$, a time point $i > 0$, and a mapping $\sigma : \boldsymbol{x} \to \mathsf{N_I}$, σ is a *certain answer for* ϕ *in* $\mathcal{K}$ *at* i iff $\mathcal{K}, i \models \phi'$, where ϕ' is the result of applying σ on ϕ. As common, since computing answers for TPQs can be seen as a search problem that uses Boolean TPQ entailment,

we focus on the decision problem of query entailment, and may refer to Boolean TPQs simply as TPQs.

Example 3. If we consider a slight variation of the query from the introduction.

$$\mathsf{P}_{>.8}(\bigcirc^{-5}\Diamond(\mathsf{HighBPPatient}(x) \wedge \Diamond\mathsf{HighBPPatient}(x)))$$

For $x = p$ and time point 5, the query below the probability operator is entailed in every model of the possible worlds w_1, w_2, w_3 and w_5, which together have a probability of 0.834. Consequently, b is an answer to the query at time point 5. Now consider the variation where the probability operators are moved inside:

$$\bigcirc^{-5}\Diamond(\mathsf{P}_{>.8}(\mathsf{HighBPPatient}(x)) \wedge \Diamond\mathsf{P}_{>.8}(\mathsf{HighBPPatient}(x)))$$

This corresponds to the situation where at least twice in the last 5 minutes, the probability of having a high blood pressure was above 0.8. As this probability is only once above this bound, this query is not entailed.

4 Lower Complexity Bound

Temporal query answering without probabilities is PSPACE-complete in combined complexity if $\mathsf{N}_{\mathsf{Rrig}} = \emptyset$, and otherwise coNEXPTIME-complete [8]. On the other hand, computing the probability of a CQ from an ipABox is $\mathrm{PP}^{\mathrm{NP}}$-complete (see extended version of the paper), and thus also in PSPACE [27]. It turns out that, if both the temporal and the probabilistic dimension are combined, we obtain an increase to EXPSPACE in complexity. This complexity increase already happens without any rigid symbols, and for TPKBs without TBox and with only one ABox, so that the DL is in fact irrelevant for this result.

A query ϕ is entailed by a TPKB $\mathcal{K}$ iff $\neg\phi$ is not satisfiable in $\mathcal{K}$. As the complexity class EXPSPACE is closed under complement, we can therefore focus on the problem of query satisfiability. We obtain EXPSPACE-hardness by reduction of the exponential variant of the corridor tiling problem [15]. In this problem, we are given a set T of *tile types*, two special tile types t_s,$t_e \in T$, a natural number n, and two functions v and h of *compatibility constraints* $v : T \to 2^T$ (vertical) and $h : T \to 2^T$ (horizontal). The input is an *instance* of the exponential corridor tiling problem if there exists a number $m \in \mathbb{N}$ and a *tiling* $f : [\![0, m]\!] \times [\![0, 2^n - 1]\!] \to T$ such that $f(0, 0) = t_s$, $f(m, 0) = t_e$, and for all $x \in [\![0, m]\!]$ and $y \in [\![0, 2^n - 1]\!]$, if $x < m$, $f(x + 1, y) \in h(f(x, y))$ and if $y < 2^n - 1$, $f(x, y + 1) \in v(f(x, y))$.

We only sketch the idea of the construction here, and leave the details to the long version of the paper. We use n concept names A_i to mark the different possible worlds $w \in \Omega_{\mathcal{K}}$ with a counter, such that in interpretations ι that satisfy both the TPQ and the TPKB, $\iota, w, j \models A_i(a)$ iff the ith bit of the counter is 1 at time point j, and $\iota, w, j \not\models A_i(a)$ iff the ith bit is 0 at time point j. Furthermore, we make sure that each possible world is at each time point uniquely determined by its counter value. For this, we use the ipABox $\mathcal{A}_1 = \{A_i(a) : 0.5 \mid i \in [\![1, n]\!]\}$. The query then makes sure that the counter values are increased for each time

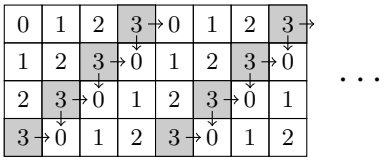

Fig. 1. Illustration of the tilings represented by the possible worlds.

point. Figure 1 illustrates this idea. Intuitively, each possible world corresponds to a row in the tiling, with its counter value at time point 1 denoting the row number.

At each time point, there are two possible worlds that can be most easily recognised by a query: the one whose counter value is 0 (which satisfies the query $\bigwedge_{1\leq i\leq n} \neg A_i(a)$), and the one whose counter value is $2^n - 1$ (which satisfies the query $\bigwedge_{1\leq i\leq n} A_i(a)$). Unless the latter one represents the last row, both these possible worlds correspond to neighbouring rows, which means at each time point we can recognise the vertical neighbour relation for two rows easily, and thus enforce tiling conditions in that direction with the following query, where $L(a)$ is an assertion that marks the last row, and for a tile type $t \in T$, $B_t(a)$ expresses that the current cell has a tile of type t.

$$\Box \bigwedge_{t_1 \in T} \left(\left(B_{t_1}(a) \wedge \bigwedge_{i \in [\![1,n]\!]} A_i(a) \wedge \neg L(a) \right) \rightarrow \bigvee_{t_2 \in v(t_1)} \mathsf{P}_{=1} \left(\left(\bigwedge_{i \in [\![1,n]\!]} \neg A_i(a) \right) \rightarrow B_{t_2}(a) \right) \right)$$

As we can only check the vertical tiling conditions for one pair of rows at a time, we represent each cell by up to 2^n succeeding time points in each possible world, performing a switch only when the counter reaches $2^n - 1$. The remaining reduction is described in the extended version of the paper.

Lemma 1. *Entailment of TPQs is* ExpSpace*-hard in combined complexity, even for TKBs* $\langle \mathcal{T}, (\mathcal{A}_i)_{i\in[\![1,n]\!]} \rangle$ *where* $\mathcal{T} = \emptyset$*,* $n = 1$ *and* $\mathsf{N}_{\mathsf{Crig}} = \mathsf{N}_{\mathsf{Rrig}} = \emptyset$*.*

5 Upper Complexity Bound

We show that the complexity result presented in the last section is indeed tight, even if $\mathsf{N}_{\mathsf{Rrig}} \neq \emptyset$. We sketch here only the case without rigid symbols. How rigid symbols are integrated is then discussed in the extended version of the paper. Our construction is based on an abstraction of a temporal probabilistic model, which we call quasi-model, which collects for each time point and possible world the CQs occurring in the input query that are entailed, as well as the CQs that are not entailed. We focus on satisfiability of a TPQ ϕ in a TPKB $\mathcal{K} = \langle \mathcal{T}, (\mathcal{A}_i)_{i\in[\![1,n]\!]} \rangle$,

where we say ϕ is satisfiable in $\mathcal{K}$ iff ϕ is satisfiable in $\mathcal{K}$ at 1. In other words, we ignore the time point to make things simpler. Since ϕ is satisfiable in $\mathcal{K}$ at i iff $\bigcirc^{i-1}\phi$ is satisfiable, this is sufficient for our complexity analysis.

We can assume without loss of generality that ϕ contains only the operators $\wedge$, $\neg$, $\mathcal{U}$, $\mathcal{S}$ and $\mathsf{P}_{\sim p}$, since the remaining operators can be linearly encoded using known equivalences. Denote by $\mathsf{sub}(\phi)$ the sub-queries of ϕ and set $T(\phi) = \{\psi, \neg\psi \mid \psi \in \mathsf{sub}(\phi)\}$. A *quasi-state* is a mapping $Q : \Omega_{\mathcal{K}} \to T(\phi)$ that satisfies the following conditions:

S1 $\neg\psi \in Q_i(w)$ iff $\psi \notin Q_i(w)$,
S2 for all $\psi_1 \wedge \psi_2 \in T(\phi)$: $\psi_1 \wedge \psi_2 \in Q_i(w)$ iff $\psi_1 \in Q_i(w)$ and $\psi_2 \in Q_i(w)$, and
S3 for all $\mathsf{P}_{\sim p}(\psi) \in T(\phi)$: $\mathsf{P}_{\sim p}(\psi) \in Q_i(w)$ iff $\mu_{\mathcal{K}}(\{w \mid \psi \in Q_i(w)\}) \sim p$.

The quasi-state abstracts probabilistic interpretations at a single time point by assigning queries to each possible world according to the semantics of the atemporal operators in our query language. To incorporate the temporal dimension, we consider unbounded sequences of quasi-states $(Q_i)_{i\geq 1}$, which we call *quasi-models for ϕ in $\mathcal{K}$*, and which have to satisfy the following conditions for $i \geq 1$ and $w = (\mathcal{A}'_i)_{i\in[\![1,n]\!]} \in \Omega_{\mathcal{K}}$.

Q1 $\phi \in Q_1(w)$,
Q2 if $i \in [\![1, n]\!]$, $\mathcal{A}'_i \models \bigwedge_{\psi\in X} \psi$, where $X = \{\psi \in Q_i(w) \mid \psi$ is a CQ or a negated CQ$\}$.
Q3 for all $\bigcirc\psi \in T(\phi)$, $\bigcirc\psi \in Q_i(w)$ iff $\psi \in Q_{i+1}(w)$,
Q4 for all $\bigcirc^-\psi \in T(\phi)$, $\bigcirc^-\psi \in Q_{i+1}(w)$ iff $\psi \in Q_i(w)$,
Q5 for all $\psi_1 \mathcal{U} \psi_2 \in T(\phi)$, $\psi_1 \mathcal{U} \psi_2 \in Q_i$ iff there exists $j \geq i$ s.t. $\psi_2 \in Q_j(w)$ and for all $k \in [\![i, j-1]\!]$, $\psi_1 \in Q_k(w)$, and
Q6 for all $\psi_1 \mathcal{S} \psi_2 \in T(\phi)$, $\psi_1 \mathcal{S} \psi_2 \in Q_i$ iff there exists $j \leq i$ s.t. $\psi_2 \in Q_j(w)$ and for all $k \in [\![j-1, i]\!]$, $\psi_1 \in Q_k(w)$.

Again, the intuition behind these conditions follows directly from the semantics of the temporal operators. As we show in the extended version of the paper, quasi-models are indeed sufficient to witness the satisfiability of a TPQ in a TPKB. Moreover, it is sufficient to consider quasi-models that are of a certain regular form, which is the crucial element for our complexity bound.

Lemma 2. *ϕ is satisfiable in $\mathcal{K}$ with $\mathsf{N}_{\mathsf{Crig}} = \mathsf{N}_{\mathsf{Rrig}} = \emptyset$ iff there exists a quasi-model for ϕ in $\mathcal{K}$ wrt. $\mathcal{S}$ and a which is of the form*

$$Q_1, \dots Q_m (Q_{m+1}, \dots Q_{m+o})^\omega,$$

where m and o are both double-exponentially bounded in the size of $\mathcal{K}$ and ϕ.

Exploiting the fact that ExpSpace = NExpSpace, we obtain our space bounds by a non-deterministic decision procedure that can be roughly sketched as follows. We first guess the numbers m and o. While m and o are double-exponentially bounded, they can be stored in exponential space using binary encoding. We now guess the quasi-states Q_1, …, Q_{m+o} one after the other,

where we carefully make sure that all conditions of quasi-models are satisfied. In particular, we keep track of $\mathcal{U}$- and $\mathcal{S}$-formulae that have to be satisfied, and we keep the state Q_{m+1} in memory to test that it is compatible to Q_{m+o}, and that all $\mathcal{U}$-formulae in Q_{m+1} are satisfied before we reach Q_{m+o}.

In the extended version of the paper, we present a refined version of quasi-models, which also have the above regularity property, but additionally take into consideration rigid predicates. The main idea is to use for each possible world an additional structure that determines which sets of CQs and their negations can be entailed at any time point under the rigidity constraints. This structure takes exponential space per possible world, and can be computed in non-deterministic exponential time.

Theorem 1. *Entailment of TPQs from $\mathcal{EL}$-TPKBs can be decided in* ExpSpace*, even if $\mathsf{N}_{\mathsf{Rrig}} \neq \emptyset$ and $\mathsf{N}_{\mathsf{Crig}} \neq \emptyset$.*

6 Removing Negation

The complexity increase discussed in the last sections can be avoided if we restrict ourselves to *positive TPQs*, which are TPQs that do not use the operators $\neg$, $\mathsf{P}_{<p}$ and $\mathsf{P}_{=p}$. Note that the probability operators $\mathsf{P}_{<p}\phi$ and $\mathsf{P}_{=p}\phi$ can be seen as implicit negation operators, as they express the non-entailment of ϕ in some possible worlds, whereas $\mathsf{P}_{>p}\phi$ only expresses the positive entailment of ϕ in possible worlds. The example queries shown in this paper are all positive queries.

In the absence of negation, it is possible to evaluate the probabilities of subqueries "inside-out", starting from queries of the form $\mathsf{P}_{>p}\phi$ where ϕ contains no probabilistic operators. For non-probabilistic temporal queries, it can be decided in P data and NP combined complexity whether they are entailed. This allows to decide the entailment of $\mathsf{P}_{>p}\phi$ at any time point in PP, by using a probabilistic Turing machine that guesses all possible worlds of the TPKB. Using closure properties of the complexity class PP, we can thus obtain tight complexity bounds for the case where the nesting depth of probability operators is bounded, and otherwise inclusion in P^{PP}, a complexity class that is still contained in PSpace.

Theorem 2. *Entailment of positive TPQs from $\mathcal{EL}$-TPKBs is* PP*-complete wrt. data complexity. Regarding combined complexity, it is $\mathrm{PP}^{\mathrm{NP}}$-complete if the nesting depth of probability-operators in the query is bounded, and otherwise in P^{PP}. The results already hold for $\mathsf{N}_{\mathsf{Rrig}} \neq \emptyset$.*

7 Conclusion

We investigated the complexity of querying temporal probabilistic data using a combination of LTL and conjunctive queries with probability operators. While pure temporal and pure probabilistic query answering are both in PSpace for most cases, combining both dimensions yields completeness for ExpSpace. This increase in complexity already happens without TBoxes and just with a single

ABox, so that the hardness result is in fact independent of DL reasoning. This increase of complexity can be avoided if we restrict ourselves to positive TPQs, in which case the temporal dimension comes at no cost or almost no cost compared to pure probabilistic query answering. While this paper presented a theoretical study of the setting of temporal probabilistic query answering, the methods presented give no clear idea how a practical implementation would look like. For description logics that enjoy first-order rewritability such as *DL-Lite*, a solution could be to rewrite temporal queries into SQL and use a probabilistic database system to compute their probabilities. However, this approach would only work for queries that do not use negation, and it is not clear whether it can be used with rigid symbols [8]. Another open question is how the data complexity looks like in the case where we allow for negation, and whether the complexities further change if we admit more expressive DLs, or even DLs that support temporal and probabilistic operators themselves.

References

1. Artale, A., Kontchakov, R., Kovtunova, A., Ryzhikov, V., Wolter, F., Zakharyaschev, M.: Ontology-mediated query answering over temporal data: A survey (invited talk). In: Proceedings of TIME 2017. pp. 1:1–1:37 (2017). https://doi.org/10.4230/LIPIcs.TIME.2017.1
2. Baader, F., Borgwardt, S., Koopmann, P., Ozaki, A., Thost, V.: Metric temporal description logics with interval-rigid names. In: Proceedings of FroCoS 2017. pp. 60–76 (2017). https://doi.org/10.1007/978-3-319-66167-4_4
3. Baader, F., Borgwardt, S., Lippmann, M.: Temporal query entailment in the description logic $\mathcal{SHQ}$. J. Web Sem. **33**, 71–93 (2015). https://doi.org/10.1016/j.websem.2014.11.008
4. Baader, F., Ghilardi, S., Lutz, C.: LTL over description logic axioms. ACM Trans. Comput. Log. **13**(3), 21:1–21:32 (2012). https://doi.org/10.1145/2287718.2287721
5. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017)
6. Baader, F., Koopmann, P., Turhan, A.: Using ontologies to query probabilistic numerical data. In: Proceedings of FroCoS 2017. pp. 77–94 (2017). https://doi.org/10.1007/978-3-319-66167-4_5
7. Borgwardt, S., Ceylan, İ.İ., Lukasiewicz, T.: Ontology-mediated queries for probabilistic databases. In: Proceedings of AAAI 2017. pp. 1063–1069 (2017), http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14365
8. Borgwardt, S., Lippmann, M., Thost, V.: Temporalizing rewritable query languages over knowledge bases. J. Web Sem. **33**, 50–70 (2015). https://doi.org/10.1016/j.websem.2014.11.007
9. Brandt, S., Kalayci, E.G., Kontchakov, R., Ryzhikov, V., Xiao, G., Zakharyaschev, M.: Ontology-based data access with a Horn fragment of metric temporal logic. In: Proceedings of AAAI 2017. pp. 1070–1076 (2017), http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14881
10. Ceylan, İ.İ., Darwiche, A., den Broeck, G.V.: Open-world probabilistic databases: An abridged report. In: Proceedings of IJCAI 2017. pp. 4796–4800 (2017). https://doi.org/10.24963/ijcai.2017/669

11. Ceylan, İ.İ., Peñaloza, R.: Dynamic Bayesian description logics. In: Proceedings of DL 2015 (2015), http://ceur-ws.org/Vol-1350/paper-48.pdf
12. Ceylan, İ.İ., Peñaloza, R.: The Bayesian ontology language $\mathcal{BEL}$. J. Autom. Reasoning **58**(1), 67–95 (2017). https://doi.org/10.1007/s10817-016-9386-0
13. Dalvi, N.N., Schnaitter, K., Suciu, D.: Computing query probability with incidence algebras. In: Proceedings of PODS 2010. pp. 203–214 (2010). https://doi.org/10.1145/1807085.1807113
14. Elkin, P.L., Brown, S.H., Husser, C.S., Bauer, B.A., Wahner-Roedler, D., Rosenbloom, S.T., Speroff, T.: Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin. Proc. **81**(6), 741–748 (2006)
15. van Emde Boas, P.: The convenience of tilings. Lecture Notes in Pure and Applied Mathematics pp. 331–363 (1997)
16. Gabbay, D.M., Kurucz, A., Wolter, F., Zakharyaschev, M.: Many-dimensional modal logics: Theory and applications. Elsevier (2003)
17. Gutiérrez-Basulto, V., Jung, J.C., Lutz, C., Schröder, L.: Probabilistic description logics for subjective uncertainty. J. Artif. Intell. Res. **58**, 1–66 (2017). https://doi.org/10.1613/jair.5222
18. Gutiérrez-Basulto, V., Jung, J.C., Ozaki, A.: On metric temporal description logics. In: Proceedings of ECAI 2016. pp. 837–845 (2016). https://doi.org/10.3233/978-1-61499-672-9-837
19. Jung, J.C., Lutz, C.: Ontology-based access to probabilistic data with OWL QL. In: The Semantic Web - ISWC 2012. Lecture Notes in Computer Science, vol. 7649, pp. 182–197. Springer (2012). https://doi.org/10.1007/978-3-642-35176-1_12
20. Koopmann, P.: Ontology-mediated query answering for probabilistic temporal data with EL ontologies (extended version). LTCS-Report 18-07, Chair of Automata Theory, Institute for Theoretical Computer Science, Technische Universität Dresden, Dresden, Germany (2018), see http://lat.inf.tu-dresden.de/research/reports.html
21. Kumar, N., Khunger, M., Gupta, A., Garg, N.: A content analysis of smartphone–based applications for hypertension management. Journal of the American Society of Hypertension **9**(2), 130–136 (2015)
22. Lukasiewicz, T.: Expressive probabilistic description logics. Artif. Intell. **172**(6-7), 852–883 (2008). https://doi.org/10.1016/j.artint.2007.10.017
23. Lutz, C., Wolter, F., Zakharyaschev, M.: Temporal description logics: A survey. In: Proceedings of TIME 2008. pp. 3–14. IEEE Press (2008). https://doi.org/10.1109/TIME.2008.14
24. Margara, A., Urbani, J., van Harmelen, F., Bal, H.: Streaming the web: Reasoning over dynamic data. Web Semantics: Science, Services and Agents on the World Wide Web **25**, 24 – 44 (2014). https://doi.org/https://doi.org/10.1016/j.websem.2014.02.001
25. Rector, A., Gangemi, A., Galeazzi, E., Glowinski, A., Rossi-Mori, A.: The GALEN CORE model schemata for anatomy: Towards a re-usable application-independent model of medical concepts. In: Proc. MIE'94. pp. 229–233 (1994)
26. Ronca, A., Kaminski, M., Grau, B.C., Motik, B., Horrocks, I.: Stream reasoning in temporal datalog. CoRR **abs/1711.04013** (2017), http://arxiv.org/abs/1711.04013
27. Toda, S.: PP is as hard as the polynomial-time hierarchy. SIAM J. Comput. **20**(5), 865–877 (1991). https://doi.org/10.1137/0220053