Provenance Analysis: A Perspective for Description Logics?

Katrin M. Dannert and Erich Grädel(✉)

RWTH Aachen University, Aachen, Germany
{dannert,graedel}@logic.rwth-aachen.de

For Franz Baader on the occasion of his 60th birthday

Abstract. Provenance analysis aims at understanding how the result of a computational process with a complex input, consisting of multiple items, depends on the various parts of this input. In database theory, provenance analysis based on interpretations in commutative semirings has been developed for positive database query languages, to understand which combinations of the atomic facts in a database can be used for deriving the result of a given query. In joint work with Val Tannen, we have recently proposed a new approach for the provenance analysis of logics *with negation*, such as first-order logic and fixed-point logic. It is based on new semirings of dual-indeterminate polynomials or dual-indeterminate formal power series, which are obtained by taking quotients of traditional provenance semirings by congruences that are generated by products of positive and negative provenance tokens. This provenance approach has also been applied to fragments of first-order logics such as modal and guarded logics. In this paper, we explore the question whether, and to what extent, the provenance approach might be useful in the field of description logics.

1 Introduction

This paper is intended as an account, written for the description logics community, of recent developments in semiring provenance, that make provenance analysis applicable to logical formalisms with negation. In particular, we discuss the question whether provenance analysis could be a fruitful perspective for description logics.

Provenance analysis is an algebraic approach to abstract from a computation with multiple input items, such as the evaluation of a database query, mathematical information on how the result of the computation depends on the various input data. In database theory, provenance analysis based on interpretations in commutative semirings has been successfully developed for query languages such as unions of conjunctive queries, positive relational algebra,

K. M. Dannert—Supported by the DFG RTG 2236 UnRAVeL.

C. Lutz et al. (Eds.): Baader Festschrift, LNCS 11560, pp. 266–285, 2019.
https://doi.org/10.1007/978-3-030-22102-7_12

nested relations, Datalog, XQuery, SQL-aggregates and several others, and it has been implemented in software systems such as Orchestra and Propolis, see e.g. [2,5,6,11,13,17]. In this approach, atomic facts are interpreted not just by true or false, but by values in an appropriate semiring, where 0 is the value of false statements, whereas any element $a \neq 0$ of the semiring stands for some shade of truth. These values are then propagated from the atomic facts to arbitrary queries in the language, which permits to answer questions such as the minimal cost of a query evaluation, the confidence one can have that the result is true, the number of different ways in which the result can be computed, or the clearance level that is required for obtaining the output, under the assumption that some facts are labelled as confidential, secret, top secret, etc. We refer to [14] for a recent account on the semiring framework for database provenance.

We argue that provenance analysis may have a strong potential for useful applications also in the context of description logics. We shall propose notions of provenance semantics for ABoxes and TBoxes where concept and role assertions take values in a commutative semiring, and concept inclusions $C \sqsubseteq D$ translate into comparisons of such provenance values. The common reasoning problems in description logics, such as subsumption, consistency, or query answering get a new twist, generalizing Boolean reasoning to algebraic reasoning in a commutative semiring. Potential applications of this approach include *cost computations* of concept assertions (by means of provenance evaluations in the tropical semiring), the study of required *clearance levels* for accessing confidential or secret data (using valuation in an access control semiring), or reasoning about *confidences* achievable in ontology-mediated query evaluations. We shall discuss these notions in more detail in Sects. 4 and 5 below.

For a long time, an essential limitation of the semiring provenance approach has been its confinement to *positive* query languages. There have been algebraically interesting attempts to cover difference of relations [1,7,8,12] but they have not resulted in systematic tracking of *negative information*, and until recently there has been no convincing provenance analysis for languages with full negation. For applications to description logics, the inability to deal with negation and absent information would certainly be a major obstacle. However, a new approach for the provenance analysis of logics with negation, such as first-order logic and fixed-point logic, has now been proposed in [9,10] based on the following ingredients:

- Negation is dealt with by transformation to negation normal form. This is a common approach in logic, but while this is often just a matter of convenience and done for simplification, its seems indispensable for provenance semantics. Indeed, beyond Boolean semantics, negation is not a compositional logical operation: the provenance value of $\neg\varphi$ is not necessarily determined by the provenance value of φ.
- On the algebraic side, new provenance semirings of polynomials and formal power series have been introduced, which take negation into account. They are obtained by taking quotients of traditional provenance semirings by congruences generated by products of positive and negative provenance tokens; they

are called semirings of dual-indeterminate polynomials or dual-indeterminate power series.
– Provenance analysis of logics is closely connected to provenance analysis of games. In [10], the provenance approach to logics with negation is described from the perspective of the associated model checking games. In fact provenance analysis of games is of independent interest, and provenance values of positions in a game provide detailed information about the number and properties of the strategies of the players, far beyond the question whether or not a player has a winning strategy from a given position. However, in the interest of a reasonably compact presentation, we do not use the game perspective in this paper, but describe the approach in purely algebraic and logical terms.

In this paper we propose to study the potential of the semiring provenance approach as a perspective for description logics. Although we ourselves are certainly not experts in description logics and their applications, we believe that there are good reasons why this might be interesting and useful. Given that most description logics use negation in an essential way, the new provenance approach for dealing with negation could help to combine provenance analysis and description logics in a fruitful way. A point in favour is that description logics are, as it is put in the textbook [3], 'cousins of modal logics', and that the new approach to provenance analysis has already been applied to modal and guarded logics in [4]. On the other side, the application of provenance to description logics certainly also poses nontrivial problems. Indeed the standard scenario of provenance analysis is formula evaluation in a fixed finite structure. In most applications of description logics, however, a knowledge base is considered that, logically speaking, axiomatizes a class of structures, and the main reasoning problems are variants of satisfiability, validity, and entailment problems. Nevertheless, notions developed in [9] of provenance tracking interpretations by means of dual-indeterminate polynomials permit to deal with multiple models, and with reverse provenance analysis, constructing appropriate models from a given specification, at least in the case of a fixed universe. Further it also seems a quite promising project to generalize the tableaux-based reasoning techniques that are so popular in description logic to provenance semantics based on semirings. Thus, while differences and difficulties exist, they do not seem unsolvable. We thus hope that the description logic community will take an interest in these new developments in provenance analysis, and that a fruitful collaboration between the two fields will emerge.

This paper does not assume that the reader is already familiar with semiring provenance. However, we do assume that the reader knows basic definitions and results about description logics. Our notation and terminology is largely based on [3].

2 Commutative Semirings

Definition 1. A *commutative semiring* is an algebraic structure $(K, +, \cdot, 0, 1)$, with $0 \neq 1$, such that $(K, +, 0)$ and $(K, \cdot, 1)$ are commutative monoids, $\cdot$ distributes over $+$, and $0 \cdot a = a \cdot 0 = 0$. A semiring is *+-positive* if $a + b = 0$ implies

$a = 0$ and $b = 0$. This excludes rings. A semiring is *root-integral* if $a \cdot a = 0$ implies $a = 0$. All semirings considered in this paper are commutative, $+$-positive and root-integral. Further, a commutative semiring is *positive* if it is $+$-positive and has no divisors of 0. The standard semirings considered traditionally in provenance analysis are positive, but for the treatment of negation we need semirings (of dual-indeterminate polynomials or power series) that have divisors of 0.

Notice that a semiring K is positive if, and only if, the unique function $h : K \to \{0, 1\}$ with $h^{-1}(0) = \{0\}$ is a homomorphism from K into the Boolean semiring $\mathbb{B} = (\{0, 1\}, \vee, \wedge, 0, 1)$. A semiring K is $(+)$-idempotent if $a + a = a$, for all $a \in K$, and $(+, \cdot)$-idempotent if, in addition, $a \cdot a = a$ for all a. Further, K is *absorptive* if $a + ab = a$, for all $a, b \in K$. Obviously, every absorptive semiring is $(+)$-idempotent.

In provenance analysis, elements of a commutative semiring are used as truth values for logical statements. The intuition is that $+$ describes the *alternative use* of information, as in disjunctions or existential quantifications whereas $\cdot$ stands for the *joint use* of information, as in conjunctions or universal quantifications. Further, 0 is the value of false statements, whereas any element $a \neq 0$ of a semiring K stands for a 'nuanced' interpretation of true.

2.1 Application Semirings

We briefly discuss some specific semirings that provide interesting information about about a logical statement.

- The *Boolean semiring* $\mathbb{B} = (\{0, 1\}, \vee, \wedge, 0, 1)$ is the domain of standard logical truth values.
- The semiring $\mathbb{N} = (\mathbb{N}, +, \cdot, 0, 1)$ can be used for counting successful strategies for query evaluation. It also plays an important role for *bag semantics* in databases.
- $\mathbb{T} = (\mathbb{R}_+^\infty, \min, +, \infty, 0)$ is called the *tropical* semiring. It has many applications for cost computations, for instance for query evaluation.
- The *Viterbi* semiring $\mathbb{V} = ([0, 1], \max, \cdot, 0, 1)$ is isomorphic to $\mathbb{T}$ via $x \mapsto e^{-x}$ and $y \mapsto -\ln y$. We will think of the elements of $\mathbb{V}$ as *confidence scores* and use it to describe the confidence assigned to a logical statement.
- The access control semiring is $\mathbb{A} = (\{\mathsf{P} < \mathsf{C} < \mathsf{S} < \mathsf{T} < 0\}, \min, \max, 0, \mathsf{P})$ where P is 'public', C is 'confidential', S is 'secret', T is 'top secret', and 0 is 'so secret that nobody can access it!'. The valuation of a statement in $\mathbb{A}$ describes the *minimal clearance level* that is needed to establish it.
- The *max-min* semiring on a totally ordered set $(A, \leq)$ with least element a and greatest element b is the semiring $(A, \max, \min, a, b)$. The class of max-min semirings includes, of course, the Boolean semiring and the access control semiring but also infinite ones, for instance the one on the real interval $[0, 1]$ which is sometimes called the fuzzy semiring.

2.2 Provenance Semirings

Beyond such application semirings, there are important universal provenance semirings of polynomials and formal power series that are used for a general provenance analysis. They admit to compute provenance values once in a general semiring and then to specialise these via homomorphisms to specific application semirings as needed.

Let X be a set of abstract *provenance tokens*, i.e. variables that we use to label atomic data (such as concept or role assertions in description logics). The commutative semiring that is freely generated by the set X is $\mathbb{N}[X] = (\mathbb{N}[X], +, \cdot, 0, 1)$, the semiring of multivariate polynomials in indeterminates from X and with coefficients from $\mathbb{N}$.

Computing provenance values of a statement φ (from some appropriate logical formalism) in $\mathbb{N}[X]$ gives us precise information about which combinations of the atomic facts can be used to derive φ. Indeed, each monomial $cx_1^{e_1} \dots x_k^{e_k}$ that occurs in the provenance polynomial $\pi[\![\varphi]\!] \in \mathbb{N}[X]$ indicates that we have c different evaluation strategies that make use of precisely those atomic facts that are labelled by $x_1, \dots x_k$ and use the fact labelled by x_i precisely e_i times. Evaluation strategies can be understood either as 'proof trees' (as in [9,13]) or as winning strategies in the model checking game associated with φ (as in [10]).

There are a number of other polynomial semirings that can be obtained from $\mathbb{N}[X]$ by dropping coefficients, dropping exponents, or absorption laws, in which provenance polynomials are less informative, but possibly easier to compute. This includes the $+$-idempotent semiring $\mathbb{B}[X]$, the so-called why semiring $\mathbb{W}[X]$, the absorptive semiring $\mathbb{S}[X]$ and the free distributive lattice $\mathsf{PosBool}(X)$, see e.g. [10,13,14] for more information.

However, in none of these semirings there is an adequate treatment of negation, or tracking of missing information, because either negative atoms are not represented at all, or an atom and its negation are labelled by two different tokens without any algebraic connection between them. To address this issue, a new approach has been proposed in [9], and further developed in [10].

2.3 Dual-Indeterminate Polynomials and Formal Power Series

Here is the algebraic construction to make provenance analysis available for logics with negation. Let $X, \bar{X}$ be two disjoint sets of provenance tokens, together with a bijection $X \to \bar{X}$, that maps each 'positive' token $p \in X$ to a corresponding 'negative' token $\bar{p} \in \bar{X}$. We call p and $\bar{p}$ complementary tokens. By convention, if we annotate an atomic fact by p then $\bar{p}$ can only be used to annotate its negation, and vice versa.

Definition 2. The semiring $\mathbb{N}[X, \bar{X}]$ of *dual-indeterminate polynomials* is the quotient of the semiring of polynomials $\mathbb{N}[X \cup \bar{X}]$ by the congruence generated by the equalities $p \cdot \bar{p} = 0$ for all $p \in X$. This is the same as quotienting by the ideal generated by the polynomials $p\bar{p}$ for all $p \in X$. Two polynomials $f, g \in \mathbb{N}[X \cup \bar{X}]$ are congruent if, and only if, they become identical after deleting from each of

them the monomials that contain complementary tokens. Hence, the congruence classes in $\mathbb{N}[X, \bar{X}]$ are in one-to-one correspondence with the polynomials in $\mathbb{N}[X \cup \bar{X}]$ such that none of their monomials contains complementary tokens.

Note that the semirings $\mathbb{N}[X \cup \bar{X}]$ are +-positive and root-integral, but not positive, since they obviously admit divisors of 0.

The semirings $\mathbb{N}[X, \bar{X}]$ turn out to be adequate for a general provenance analysis of full first-order logic (with negation) [9], and hence also for full relational algebra (not just its positive fragment). This extends to fragments of first order logic such as modal and guarded logics [4] and (as we propose in this paper) description logics. However for logics with fixed points or with mechanisms of unbounded iteration, polynomial semirings are not sufficient. Even for a formalism as simple as datalog (avoiding all complications arising from universal quantification and negation) one has to impose additional conditions on the semirings to guarantee the existence of least fixed points [5]. Of particular importance are ω-continuous semirings. Many application semirings are ω-continuous, but $\mathbb{N}$, and the polynomial semirings $\mathbb{N}[X]$ and $\mathbb{N}[X, \bar{X}]$ are not. The ω-continuous completion of $\mathbb{N}$ is $\mathbb{N}^\infty := \mathbb{N} \cup \{\infty\}$ (with $a + \infty = a \cdot \infty = \infty$), but the completion of $\mathbb{N}[X]$ is $\mathbb{N}^\infty[\![X]\!]$ which is not a semiring of polynomials, but of formal power series (possibly infinite sums of monomials), with coefficients in $\mathbb{N}^\infty$ and indeterminates in X, with addition and multiplication defined in the standard way. We combine this with our approach for dealing with negation by taking quotients.

Definition 3. The semiring $\mathbb{N}^\infty[\![X, \bar{X}]\!]$ is the quotient of the semiring of power series $\mathbb{N}^\infty[\![X \cup \bar{X}]\!]$ by the congruence generated by the equalities $p \cdot \bar{p} = 0$ for all $p \in X$. The congruence classes in $\mathbb{N}^\infty[\![X, \bar{X}]\!]$ are in one-to-one correspondence with the power series in $\mathbb{N}^\infty[\![X \cup \bar{X}]\!]$ such that none of their monomials contain complementary tokens. We call these *dual-indeterminate power series*.

Every function $f : X \cup \bar{X} \to K$ into an ω-continuous semiring K with the property that $f(p) \cdot f(\bar{p}) = 0$ for all $p \in X$ extends uniquely to an ω-continuous semiring homomorphism $h : \mathbb{N}^\infty[\![X, \bar{X}]\!] \to K$ that coincides with f on $X \cup \bar{X}$.

3 Provenance for Model Checking Problems

Provenance analysis has been developed for query evaluation and, more generally, model checking problems in logic, in particular first-order logic and its fragments.

Let τ be a vocabulary, which in the case of description logics contains only unary predicates (concept names) and binary predicates (role names), and fix a finite universe Δ. We denote by $\mathrm{Atoms}_\Delta(\tau)$ the set of all atoms $R\bar{a}$ with $R \in \tau$ and $\bar{a} \in \Delta^k$. Further, let $\mathrm{NegAtoms}_\Delta(\tau)$ be the set of all negated atoms $\neg R\bar{a}$ where $R\bar{a} \in \mathrm{Atoms}_\Delta(\tau)$, and consider the set of all τ-literals on A, $\mathrm{Lit}_\Delta(\tau) := \mathrm{Atoms}_\Delta(\tau) \cup \mathrm{NegAtoms}_\Delta(\tau) \cup \{a \text{ op } b : a, b \in A\}$, where op stands for $=$ or $\neq$.

Definition 4. Given any commutative semiring K, a *K-interpretation* (for τ and Δ) is a function $\pi : \mathrm{Lit}_\Delta(\tau) \to K$ that maps equalities and inequalities

to their truth values 0 or 1. A K-interpretation is *sound for negation* if $\pi[\alpha] \cdot \pi[\neg\alpha] = 0$ for every atom $\alpha \in \mathrm{Atoms}_\Delta(\tau)$. In this paper, all K-interpretations are assumed to be sound for negation.

The equality and inequality atoms are interpreted in K as 0 or 1, i.e., their provenance is not tracked. One could give a similar treatment to other relations with a fixed meaning, e.g., assuming a linear order on A. However, we do not pursue this in this paper.

We have defined in [9] how a semiring interpretation extends to a full valuation $\pi : \mathrm{FO}(\tau) \to K$ mapping any fully instantiated formula $\psi(\bar{a})$ to a value $\pi[\![\psi]\!]$, by setting

$$\pi[\![\psi \vee \varphi]\!] := \pi[\![\psi]\!] + \pi[\![\varphi)]\!] \qquad \pi[\![\psi \wedge \varphi]\!] := \pi[\![\psi]\!] \cdot \pi[\![\varphi]\!]$$
$$\pi[\![\exists x\varphi(x)]\!] := \sum_{a\in\Delta} \pi[\![\varphi(a)]\!] \qquad \pi[\![\forall x\varphi(x)]\!] := \prod_{a\in\Delta} \pi[\![\varphi(a)]\!].$$

For negation, we set $\pi[\![\neg\varphi]\!] := \pi[\![\mathrm{nnf}(\neg\varphi)]\!]$ where $\mathrm{nnf}(\varphi)$ is the negation normal form of φ.

As shown in [9], for positive semirings, and also for the interpretations in semirings of dual indeterminate polynomials that we are interested in, the soundness for negation extends from atoms to arbitrary first-order formulae and implies that $\pi[\![\varphi]\!] \cdot \pi[\![\neg\varphi]\!] = 0$ for all $\varphi \in \mathrm{FO}$. However, since we admit semirings with divisors of 0, soundness for negation does not necessarily imply that one of $\pi[\![\varphi]\!]$ and $\pi[\![\neg\varphi]\!]$ must be 0.

For modal and guarded logic similar definitions of provenance interpretations have been given and analysed in [4]. It is not difficult to adapt these definitions for description logics. Here is one for $\mathcal{ALC}$. For simplicity of notation we identify individual names with elements of the universe. Further, all concept assertions of form $a\!:\!C$ where C is a concept name or the negation of a concept name, and all role assertions $(a,b)\!:\!r$ for a role name r are viewed as literals in some set $\mathrm{Lit}_\Delta(\tau)$.

Definition 5. Let $\pi : \mathrm{Lit}_\Delta(\tau) \to K$ be a K-interpretation for a finite universe Δ and a vocabulary τ of concept names and role names. Given a role name r and an element $a \in \Delta$, let $r(a) := \{b : \pi((a,b)\!:\!r) \neq 0\}$. For shortness we define $\pi(rab) := \pi((a,b)\!:\!r)$. We extend π to *concept assertions* $a\!:\!C$ consisting of an $\mathcal{ALC}$ concept description C, assumed to be given in negation normal form, and an element $a \in \Delta$ by

$$\pi[\![a\!:\!\bot]\!] := 0 \qquad \pi[\![a\!:\!\top]\!] := 1$$
$$\pi[\![a\!:\!C \sqcup D]\!] := \pi[\![a\!:\!C]\!] + \pi[\![a\!:\!D]\!] \qquad \pi[\![a\!:\!C \sqcap D]\!] := \pi[\![a\!:\!C]\!] \cdot \pi[\![a\!:\!D]\!]$$
$$\pi[\![a\!:\!\exists r.C]\!] := \sum_{b\in r(a)} (\pi(rab) \cdot \pi[\![b\!:\!C]\!]) \qquad \pi[\![a\!:\!\forall r.C]\!] := \prod_{b\in r(a)} (\pi(rab) \cdot \pi[\![b\!:\!C]\!]).$$

The close relationship between description logics and modal logics admits to carry over the complexity results for computing provenance values of modal formulae [4] to this setting.

Proposition 1. *Let K be an arbitrary semiring. Given a concept description C in $\mathcal{ALC}$, a K-interpretation $\pi : \mathrm{Lit}_\Delta(\tau) \to K$, and an element $a \in \Delta$, the provenance value $\pi[\![a:C]\!]$ can be computed with $O(|C| \cdot |\pi|)$ semiring operations.*

Notice that for concept descriptions in full first-order logic rather than $\mathcal{ALC}$, the number of semiring operations needed to compute provenance values may be much higher. Indeed, the straightforward approach requires an exponential number of operations with respect to the length of a first-order concept description, and since even in the Boolean case, the model checking problem for first-order logic is PSPACE-complete, it is unlikely that polynomial bounds are possible.

Nevertheless, despite the relatively small *number* of semiring operations that are needed to compute provenance values for $\mathcal{ALC}$, the *complexity* of such computations may, depending on the *costs* of representing elements in the given semiring and the costs of addition and multiplication, still be rather high, in fact doubly exponential in the length of the concept description. See [4] for a detailed complexity analysis for the case of modal and guarded logic.

4 Provenance Semantics for ABoxes and TBoxes

We have described basic observations about the definition and computation of provenance values for concept assertions in $\mathcal{ALC}$. However the important reasoning tasks associated with description logics are not so much the evaluation of a concept assertion in a given interpretation. Description logics are used as knowledge representation languages. A *knowledge base* typically consists of a TBox $\mathcal{T}$ which is a finite set of *general concept inclusions* $C \sqsubseteq D$, describing conceptual knowledge about the domain of application, and an ABox $\mathcal{A}$, which is a finite set of *concept assertions* $a:C$ and *role assertions* $(a,b):r$, describing specific data. Relevant questions, given an $\mathcal{ALC}$ knowledge base $(\mathcal{A},\mathcal{T})$, concern for instance the subsumption and equivalence of two given concepts in all models of $\mathcal{T}$, the consistency of the knowledge base, or the question whether a given concept assertion $a:C$ is entailed by the knowledge base.

Can semiring provenance provide any additional insights for knowledge representation by description logics? To discuss such questions, we first discuss what provenance semantics might mean for ABoxes and TBoxes.

Provenance Semantics for an ABox. Since an ABox defines a set of statements that are asserted to be true, a natural possibility to define its provenance semantics could be to assign to every assertion in the ABox a non-zero value in the semiring, defining its precise 'shade of truth'. However, we propose a definition that is a little more general, which gives us also the possibility to declare that $a:C$ just has *some* shade of truth $\geq k$ or $> k$ without a commitment to a precise value.

Definition 6. A *K-valued ABox* is a finite set of statements of form $\pi[\![\alpha]\!]$ op k where α is a concept assertion or role assertion, k is an element of the semiring K, and op is $=, \geq$, or $>$.

In DL one sometimes restricts attention to *simple ABoxes* admitting only concept assertions $a:C$ where C is a concept name, and simple K-valued ABoxes are defined analogously. This comes with no loss of expressive power since one can replace each assertion $a:C$ by $a:A_C$, where A_C is a new concept name, and then add an equivalence $A_C \equiv C$ to the TBox.

Provenance Semantics for a TBox. For a given TBox $\mathcal{T}$, let τ be a vocabulary containing all concept names and role names appearing in $\mathcal{T}$, let Δ be a finite universe and K a commutative semiring. We want to discuss what it means that a K-interpretation $\pi : \mathrm{Lit}_\Delta(\tau) \to A$ is *consistent* with $\mathcal{T}$.

There are two main possibilities. For the stronger one we assume, without loss of generality, that $\mathcal{T}$ is given as a finite set of concept inclusions $C \sqsubseteq D$.

Definition 7. A K-interpretation $\pi : \mathrm{Lit}_\Delta(\tau) \to K$ is *strongly consistent* with $\mathcal{T}$, if for every concept inclusion $C \sqsubseteq D$ in $\mathcal{T}$ and every $a \in \Delta$, we have that $\pi[\![a:C]\!] \leq \pi[\![a:D]\!]$.

Recall that the natural order in a semiring K is defined by $x \leq y :\Longleftrightarrow \exists z(x + z = y)$. The requirement that our semirings are naturally ordered means that $\leq$ is antisymmetric (i.e. $x \leq y \wedge y \leq x$ only for $x = y$). Hence, if $\mathcal{T}$ contains both $C \sqsubseteq D$ and $D \sqsubseteq C$, and thus imposes an equivalence $C \equiv D$, strong consistency means that $\pi[\![a:C]\!] = \pi[\![a:D]\!]$ for all a.

This strong notion of consistency is rather restrictive. In many applications it may not be adequate to require that a subsumption between two concepts translates in this precise way into an ordering between their truth values. A less restrictive possibility is to view a concept inclusion $C \sqsubseteq D$ as a requirement that whenever $a : C$ has a positive 'shade of truth' then so has $a : D$. On the other side, this does not seem right in the case of concept definitions $A \equiv C$, where A is a concept name; in this case we should, of course, require that all provenance values of A and C are the same.

For the weaker notion of consistency that we have in mind we therefore rewrite a TBox as a disjoint union $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1$ where $\mathcal{T}_0$ is an *acyclic* TBox, consisting of concept definitions $A \equiv C$, without cyclic dependencies among them, and $\mathcal{T}_1$ is written as a finite set of equations $C \sqcap D = \bot$. Notice that in the Boolean case, this is just an equivalent rewriting because any concept inclusion $C \sqsubseteq D$ is equivalent to $C \sqcap \neg D = \bot$.

Definition 8. A K-interpretation $\pi : \mathrm{Lit}_\Delta(\tau) \to K$ is *weakly consistent* with a TBox $\mathcal{T}_0 \cup \mathcal{T}_1$, if

(1) for every concept definition $A \equiv C$ in $\mathcal{T}_0$ and every $a \in \Delta$, we have that $\pi[\![a:A]\!] = \pi[\![a:C]\!]$, and
(2) for every equation $C \sqcap D = \bot$ in $\mathcal{T}_1$ we have that

$$\sum_{a \in \Delta} \pi[\![a:C]\!] \cdot \pi[\![a:D]\!] = 0.$$

As a sanity check for these definitions, we prove

Proposition 2. *If π is strongly consistent with a TBox $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1$, then it is also weakly consistent with $\mathcal{T}$.*

Proof. We just have to show that, for every equation $C \sqcap D = \bot$ in $\mathcal{T}_1$ and every $a \in \Delta$, we have that $\pi[\![a : C]\!] \leq \pi[\![a : \neg D]\!]$ implies $\pi[\![a : C]\!] \cdot \pi[\![a : D]\!] = 0$. But $\pi[\![a : C]\!] \leq \pi[\![a : \neg D]\!]$ implies that also $\pi[\![a : C]\!] \cdot \pi[\![a : D]\!] \leq \pi[\![a : \neg D]\!] \cdot \pi[\![a : D]\!] = 0$ by distributivity and soundness for negation. Further, since the semiring is assumed to be $+$-positive, it follows that $\pi[\![a : C]\!] \cdot \pi[\![a : D]\!] = 0$. □

Definition 9. A *provenance knowledge base* consists of a K-valued ABox $\mathcal{A}$ and a TBox $\mathcal{T}$. We say that a K-interpretation $\pi : \mathrm{Lit}_\Delta(\tau) \to K$ is strongly (or weakly) consistent with $(\mathcal{A}, \mathcal{T})$ if τ contains all role names and concept names occuring in $\mathcal{A}$ and $\mathcal{T}$ and $\mathcal{T}$, if Δ contains all individual names occurring in $\mathcal{A}$ and

(1) π satisfies all assertions occurring in $\mathcal{A}$,
(2) π is strongly (or weakly) consistent with $\mathcal{T}$.

Such a K-interpretation is also called a K-model (or a weak K-model) of $(\mathcal{A}, \mathcal{T})$.

5 Reasoning Problems for Provenance Knowledge Bases

The distinction between strong and weak consistency corresponds with a distinction between strong and weak subsumption between two concept descriptions. We say that C is *strongly subsumed* by D, in a K-interpretation π, in symbols $C \sqsubseteq_\pi D$, if $\pi[\![a : C]\!] \leq \pi[\![a : D]\!]$ for all elements a of π. Similarly C is *weakly subsumed* by D in π, in symbols $C \sqsubseteq^w_\pi D$ if $\pi[\![a : C]\!] \cdot \pi[\![a : \neg D]\!] = 0$ for all a. This also implies two notions of strong and weak equivalence between two concept descriptions, denoted $C \equiv_\pi D$, and $C \equiv^w_\pi D$. Further, we write $C \sqsubseteq_\mathcal{T} D$ and $C \sqsubseteq_{(\mathcal{A},\mathcal{T})} D$ to denote that such a subsumption holds in all models of a TBox $\mathcal{T}$ or in all models of a provenance knowledge base $(\mathcal{A}, \mathcal{T})$, and analogously for the other subsumption and equivalence properties.

In analogy to and generalisation of the standard reasoning problems in DL we propose the following problems, for a given provenance knowledge base $(\mathcal{A}, \mathcal{T})$.

Subsumption. What kind of subsumption and equivalence properties hold between concept descriptions in K-models of $\mathcal{T}$? In particular, describe the subsumption hierarchy and the weak subsumption hierarchy entailed by $\mathcal{T}$.

Consistency. Do there exist K-interpretations that are (strongly or weakly) consistent with $(\mathcal{A}, \mathcal{T})$?

Provenance values. Given a concept assertion $a : C$, what are the possible provenance values $\pi[\![a : C]\!]$ in (weak) models of $(\mathcal{A}, \mathcal{T})$? In particular is there a possible provenance value $\pi[\![a : C]\!] \neq 0$ in some such model; this generalizes the satisfiability problem.

Query answering. Given a (Boolean) query q, formulated in some appropriate query language, what are the possible provenance values $\pi[\![q]\!]$ in models of $(\mathcal{A}, \mathcal{T})$? In particular, is $\pi[\![q]\!] \neq 0$ in all such models?

Depending on the choice of the semiring, this permits to answer questions about issues such as cost, confidences, or required clearance levels for statements that we derive from the knowledge base. Here are a few examples:

(1) Consider a provenance knowledge base $(\mathcal{A}, \mathcal{T})$ with interpretations in the tropical semiring $\mathbb{T} = (\mathbb{R}_+^\infty, \min, +, \infty, 0)$. We view $\pi[\![a : A]\!]$ as the cost of using the assertion $a : A$. If $(\mathcal{A}, \mathcal{T})$ entails a strong subsumption $C \sqsubseteq D$ then this means that for all a, it is less expensive to establish the assertion $a : C$ than $a : D$. If $(\mathcal{A}, \mathcal{T})$ entail such a subsumption only in the weak sense, then this means that whenever $a : D$ can be established for free (with cost 0), then this is also the case for $a : C$.
(2) Given a TBox $\mathcal{T}$ and an $\mathbb{A}$-valued ABox $\mathcal{A}$ (i.e. with valuations in the access control semiring), the consistency of the provenance knowledge base $(\mathcal{A}, \mathcal{T})$ means that the clearance levels required by the $\mathcal{A}$ are compatible with the hierarchy of access restrictions as imposed by the TBox. For instance, $(\mathcal{A}, \mathcal{T})$ would be inconsistent if the TBox imposes a subsumption $C \sqsubseteq D$, but $\mathcal{A}$ declares $a : C$ to be top secret and $a : D$ only confidential.
(3) Given a provenance knowledge base $(\mathcal{A}, \mathcal{T})$ with interpretations in the Viterbi semiring of confidence scores, the maximal provenance value $\pi[\![q]\!]$ of a Boolean query q in models π of $(\mathcal{A}, \mathcal{T})$ describes the confidence we can have that q holds in *some* model of $(\mathcal{A}, \mathcal{T})$.

The question arises to what extent, with what algorithmic and complexity theoretic consequences, the common reasoning techniques, such as tableaux, automata based methods, query rewriting, and so on extend to the semiring provenance setting.

6 Tableaux Rules for Provenance Knowledge Bases

A standard approach in description logics for checking the consistency of a knowledge base or an ABox is based on tableaux. A tableaux algorithm uses a system of rules to extend a given ABox by more and more assertions; for instance if an ABox $\mathcal{A}$ contains the assertion $a : C \sqcap D$, but not both $a : C$ and $a : D$, then one extends $\mathcal{A}$ to $\mathcal{A}' = \mathcal{A} \cup \{a : C, a : D\}$. This process of adding new assertions is iterated until one can either read off a model from the incremented ABox, or it contains a clash of the form $a : C$ and $a : \neg C$, so that that one can conclude that the original ABox is inconsistent. See for instance [3] for a full description of a tableaux algorithm for $\mathcal{ALC}$.

The question arises whether the tableaux approach also works for provenance knowledge bases. We show that this is indeed the case if we restrict ourselves to the class of absorptive semirings for which the natural order is a linear order. For this class, we can present a tableaux algorithm which correctly determines

whether the given provenance knowledge base is consistent, provided the ABox does not contain equality statements. Moreover, for the subclass of *max-min-semirings* our tablaux rules do not only check consistency but also produce more detailed descriptions of the K-models. Additionally for max-min semirings we can allow equality statements in the ABox.

We call a K-valued ABox $\mathcal{A}$ *normalized* if each assertion in $\mathcal{A}$ is in negation normal form and for each $\alpha = a:C$ or $\alpha = (a,b):r$ there is at most one statement about the K-value of α in $\mathcal{A}$. Additionally we disallow trivial statements $\pi[\![\alpha]\!] \geq 0$. We can normalize any K-valued ABox $\mathcal{A}$ by simply deleting all assertions $\pi[\![\alpha]\!] \geq k$ and $\pi[\![\alpha]\!] > k$ for which k is not maximal and by deleting $\pi[\![\alpha]\!] \geq k$ if $\pi[\![\alpha]\!] > j \in \mathcal{A}$ for some $j \geq k$ or if $k = 0$.

Tableaux Rules for K-valued ABox Consistency. For simplicity we will not define rules for assertions of the form $\pi[\![\alpha]\!] > k$ because they are easy adaptations of the rules for assertions of the form $\pi[\![\alpha]\!] \geq k$. However we have to exclude assertions of the form $\pi[\![\alpha]\!] = k$, because for most semirings we cannot guarantee to satisfy for instance $p \cdot q = k$ by requirements on p and q that do not depend on the value of the respective other factor. Though this is a significant restriction, it is fair to assume that in many cases it suffices to require that a concept or role assertion has 'at least truth value k' instead of requiring the provenance value to be an exact $k \in K$.

So let $\mathcal{A}$ be a K-valued ABox consisting of assertions of the form $\pi[\![a:C]\!] \geq k$ or $\pi[\![(a,b):r]\!] \geq k$, where C is not necessarily atomic and k is a value from a provenance semiring K, which we assume to be absorptive and totally ordered by its natural order. In particular this implies that addition in K is max and that multiplication in K is deflationary in both arguments with respect to the natural order, i.e. $a \cdot c = c \cdot a \leq a$ for any $a, c \in K$. The reason for this requirement is that we would like to be able to deduce form $\pi[\![a:C \sqcup D]\!] \geq k$ that one of the assertions $\pi[\![a:C]\!] \geq k$ and $\pi[\![a:D]\!] \geq k$ also has to hold, and from $\pi[\![a:C \sqcap D]\!] \geq k$ that both of them are true. In a general semiring, this is not necessarily the case and in fact we might not get any useful information about $\pi[\![a:C]\!]$ and $\pi[\![a:D]\!]$ from $\pi[\![a:C \sqcap (\sqcup)D]\!] \geq k$. With these restrictions we are able to define tableaux rules for consistency checking of K-valued ABoxes:

$\sqcap$-rule: if
1. $\pi[\![a:C \sqcap D]\!] \geq k \in \mathcal{A}$, and
2. $\{\pi[\![a:C]\!] \geq i, \pi[\![a:D]\!] \geq j\} \nsubseteq \mathcal{A}$ for all $i, j \geq k$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![a:C]\!] \geq k, \pi[\![a:D]\!] \geq k\}$

$\sqcup$-rule: if
1. $\pi[\![a:C \sqcup D]\!] \geq k \in \mathcal{A}$, and
2. $\{\pi[\![a:C]\!] \geq j, \pi[\![a:D]\!] \geq j\} \cap \mathcal{A} = \emptyset$ for all $j \geq k$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![a:X]\!] \geq k\}$ for some $X \in \{C, D\}$

$\exists$-rule: if
1. $\pi[\![a:\exists r.C]\!] \geq k \in \mathcal{A}$, and
2. there is no b and no $i, j \geq k$ such that $\{\pi[\![(a,b):r]\!] \geq i, \pi[\![b:C]\!] \geq j\} \subseteq \mathcal{A}$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![(a,d):r]\!] \geq k, \pi[\![d:C]\!] \geq k\}$, where d is new in $\mathcal{A}$

$\forall$-rule: if

1. $\{\pi[\![a\!:\!\forall r.C]\!] \geq k, \pi[\![(a,b)\!:\!r]\!] \geq \ell\} \subseteq \mathcal{A}$ for some $\ell \in K, \ell > 0$, and
2. there are no $i, j \geq k$ such that $\{\pi[\![(a,b)\!:\!r]\!] \geq i, \pi[\![b\!:\!C]\!] \geq j\} \subseteq \mathcal{A}$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![(a,b)\!:\!r]\!] \geq k, \pi[\![b\!:\!C]\!] \geq k\}$

The tableaux rules are then applied in an algorithm which works as follows. It receives a normalized K-valued ABox as input and chooses one applicable rule. Then it applies that rule, creating an extended ABox which is then transformed into a normalized one. This continues until either a clash occurs, i.e. $\mathcal{A}$ contains assertions $\pi[\![a\!:\!C]\!] \geq j$ and $\pi[\![a\!:\!\neg C]\!] \geq k$ for $j, k > 0$, or no more tableaux rules are applicable. If the algorithm registers a clash, it returns 'inconsistent', and if it does not and no more rules are applicable, it returns 'consistent' and the ABox that has been constructed.

Similarily to the algorithm for a classical (non-provenance) ABox described in [3] this algorithm is non-deterministic in two ways. Firstly, it does not specify in which order the rules are applied. This is not a problem, since these choices do not affect the outcome of the algorithm, nor the ABox that is returned. The other form of non-determinism lies in choosing the concept X in the $\sqcup$-rule. This is a relevant choice but one can determinize the algorithm by simultaneously tracking all ABoxes one could construct at once and checking that not all of them contain a clash.

The tableaux rules are based on the implications that in K if $p \cdot q \geq k$, then $p \geq k$ and $q \geq k$ and if $p + q \geq k$, then $p \geq k$ or $q \geq k$, which hold in absorptive semirings with linear natural order. Thus it is easy to check that if the algorithm observes a clash, then the original ABox was already inconsistent. However the implication for multiplication is not an equivalence. As a consequence, not every K-model of the set of atomic assertions in the final ABox $\mathcal{A}$ will be a K-model of the original ABox. Still we can construct a K-model from these atomic assertions by setting $\pi[\![\alpha]\!] = 1$ and $\pi[\![\overline{\alpha}]\!] = 0$ if $\pi[\![\alpha]\!] \geq k \in \mathcal{A}$ for some k. Here, $\overline{\alpha}$ describes the complementary statement (in negation normal form) to α, for instance $\overline{a\!:\!\neg C} = a : C$. If there is no k such that either $\pi[\![\alpha]\!] \geq k \in \mathcal{A}$ or $\pi[\![\overline{\alpha}]\!] \geq k \in \mathcal{A}$, we assign 0 to non-negated statements α and 1 to negated ones. It is important to note that in an absorptive semiring, 1 is always the maximal element with respect to the natural order. And since $1 + 1 = 1$ and $1 \cdot 1 = 1$ in these semirings, all nonatomic β which occur in assertions in $\mathcal{A}$ will also have K-value 1 and thus satisfy their respective assertions. Thus the tableaux algorithm is sound and complete and it terminates because each step simplifies the formulae which can only be done a finite number of times.

Notice that if the algorithm returns 'consistent', we also return $\mathcal{A}$. This is because while not every K-model of the atomic assertions in $\mathcal{A}$ is a K-model of the ABox, they still are necessary conditions for satisfying the ABox. Thus the new K-valued ABox gives us some information about the K-models of the original one. This is of course not new information as the new ABox has exactly the same K-models as the old one, but it gives us some requirements for the K-values of the atomic statements.

In max-min-semirings this information on the atomic statements is even more useful. In these semirings $p \cdot q \geq k$ is equivalent to ($p \geq k$ and $q \geq k$) and

$p+q \geq k$ is equivalent to ($p \geq k$ or $q \geq k$). Hence we do not lose any information by applying the tableaux rules and discarding the initial assertion while keeping the added ones. It follows that if our tableaux algorithm returns 'consistent', any K-model of the atomic assertions in the newly constructed ABox that sets all K-values of positive statements not in the ABox to 0 and negative statements to 1 will also be a K-model of the assertions from the original ABox. Thus we do not only get the one model where every relevant fact is set to 1, but possibly many more K-models. Additionally for max-min semirings we can define rules for assertions of the form $\pi[\![\alpha]\!] = k$ by introducing assertions using $\leq k$ for which we can in turn define rules. This is because any addition and multiplication will always take the value of one of its summands or factors. For example $p+q = k$ is equivalent to ($p = k$ and $q \leq k$) or ($q = k$ and $p \leq k$). We will not write down the resulting rules here, but they can be easily constructed from such equivalences.

Tableaux Rules for Provenance Knowledge Base Consistency. A more general problem than K-valued ABox consistency, is the consistency of a given provenance knowledge base $(\mathcal{A}, \mathcal{T})$. For an acyclic TBox $\mathcal{T}$ we can do this by adding a rule for $\sqsubseteq$. This rule depends on wether we require weak or strong consistency with $\mathcal{T}$ with respect to $\sqsubseteq$. For strong consistency this looks as follows. Again we require $\mathcal{A}$ to be normalized and K to be absorptive and to have a linear natural order.

strong $\sqsubseteq$-rule: if
1. $\pi[\![a\!:\!C]\!] \geq k \in \mathcal{A}$, $C \sqsubseteq D \in \mathcal{T}$, and
2. $\pi[\![a\!:\!D]\!] \geq j \notin \mathcal{A}$ for all $j \geq k$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![a\!:\!D]\!] \geq k\}$

If we now adjust the tableaux algorithm to check a provenance knowledge base instead of an ABox and add the strong $\sqsubseteq$-rule to the tableaux rules, we get an algorithm that checks consistency for acyclic knowledge bases.

If we consider weak consistency, we first need an equivalence rule.

$\equiv$-rule: if
1. $\pi[\![a\!:\!C]\!] \geq k \in \mathcal{A}$, $\{C \equiv D, D \equiv C\} \cap \mathcal{T} \neq \emptyset$, and
2. $\pi[\![a\!:\!D]\!] \geq j \notin \mathcal{A}$ for all $j \geq k$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![a\!:\!D]\!] \geq k\}$

For the weak $\sqsubseteq$-rule we encounter a small issue, which has to do with the fact that we restricted ourselves to assertions of the form $\pi[\![\alpha]\!] \geq k$ instead of also allowing $> k$. As mentioned, this restriction is not necessary and it is easy to define the corresponding rules for $>$ for all tableaux rules defined so far. So if we allow assertions $\pi[\![\alpha]\!] > k$, the $\sqsubseteq$-rule for weak consistency looks like this:

weak $\sqsubseteq$-rule: if
1. $\{\pi[\![a\!:\!C]\!] \geq k, \pi[\![a\!:\!C]\!] > k\} \cap \mathcal{A} \neq \emptyset$, $C \sqcap D = \bot \in \mathcal{T}$, and
2. $\{\pi[\![a\!:\!\neg D]\!]$ op $j, \pi[\![a\!:\!\neg D]\!] > 0 \mid \text{op} \in \{\geq, >\}\} \cap \mathcal{A} = \emptyset$ for all $j \in K$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![a\!:\!D]\!] > 0\}$

The problem with defining this rule with only $\geq$ is that we cannot express that some value is non-zero. While using $\geq$ might seem more intuitive at first glance, this is a reasonable argument for using $>$ if one wants to restrict to only one kind of comparison. It is still possible to define a weak $\sqsubseteq$-rule using only $\geq$ but this adds some additional non-determinism. This time, it does not lie in the choice of the concept, as in the $\sqcup$-rule, but in the choice of semiring value.

weak $\sqsubseteq$-rule, $\geq$-version: if
1. $\pi[\![a:C]\!] \geq k \in \mathcal{A}$, $C \sqcap D = \bot \in \mathcal{T}$, and
2. $\pi[\![a:\neg D]\!] \geq \varepsilon \notin \mathcal{A}$ for all $\varepsilon \in K$

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![a:D]\!] \geq \varepsilon\}$ for some $\varepsilon \in K$

With one of the weak $\sqsubseteq$-rules added to the algorithm in place of the strong $\sqsubseteq$-rule we again get a consistency checking algorithm for acyclic provenance knowledge bases. This time it checks weak consistency within the TBox. If we use the $\geq$-version of the rule however, this algorithm is not only non-deterministic but it can in general not be determinised in the same way as the algorithm containing only the $\sqcup$-rule. The reason is that unlike for the $\sqcup$-rule we might have infinitely many choices for ε in the weak $\sqsubseteq$-rule which we cannot track all at once. With the weak $\sqsubseteq$-rule allowing $>$ we do not run into this issue.

Lastly, we can consider general TBoxes, which are not necessarily acyclic. Here we encounter the same challenge as in the Boolean case that we have to guarantee termination. Consider for instance strong TBox consistency and assume we use the rules as they are defined right now. If $\mathcal{T}$ contains $C \sqsubseteq \exists r.C$ and $\pi[\![a:C]\!] \geq k \in \mathcal{A}$ then we will add $\pi[\![a:\exists r.C]\!] \geq k$ to $\mathcal{A}$. After that we will apply their $\exists$-rule and add $\pi[\![(a,d):r]\!] \geq k, \pi[\![d:C]\!] \geq k$ for a new symbol d and then we will repeat the same process with d. This will repeat over and over again and never terminate.

In order to avoid this issue, we need to introduce an additional termination condition for the $\exists$-rule. In the Boolean case this is done by the concept of a blocked individual name (see for instance [3]). We call a an *ancestor* of b if there is a sequence of relations $r_1, \ldots, r_l$ and of individual names $c_1, \ldots, c_{l-1}$ such that $(a,c_1):r_1 \in \mathcal{A}, (c_1,c_2):r_2 \in \mathcal{A}, \ldots, (c_{l-1},b):r_l \in \mathcal{A}$. An individual name b is called *blocked* by a if a is an ancestor of b and $\{C \mid b:C \in \mathcal{A}\} \subseteq \{C \mid a:C \in \mathcal{A}\}$. To put it less technically this means that b can be reached from a via some relation assertions in $\mathcal{A}$ and a has to satisfy any concept assertion that b has to satisfy. If we think of constructing a model, this means that if we reach such a point with the Boolean tableaux rules, we can set $b = a$ and form a loop at that point. A detailed explanation on why this is possible can be found in [3].

Now we need to adapt this termination condition to the provenance setting. We call a a *K-ancestor* of b if there is a sequence of relations $r_1, \ldots, r_l$ and of individual names $c_1, \ldots, c_{l-1}$ such that $\pi[\![(a,c_1):r_1]\!] \geq k_1 \in \mathcal{A}, \pi[\![(c_1,c_2):r_2]\!] \geq k_2 \in \mathcal{A}, \ldots, \pi[\![(c_{l-1},b):r_l]\!] \geq k_l \in \mathcal{A}$ for some $k_1, \ldots, k_l > 0$. We define an individual name b to be *K-blocked* by a if a is a K-ancestor of b and for each C such that $\pi[\![b:C]\!] \geq k \in \mathcal{A}$ we have $\pi[\![a:C]\!] \geq j \in \mathcal{A}$ for some $j \geq k$. Again the intuition is that a has to satisfy all constraints on b, also taking into account the lower bound on the K-value. We say that b is *K-blocked* if b is K-blocked by

some a. Again if b is blocked by a this makes it possible to form a loop. Hence we can define the new $\exists$-rule as follows.

$\exists$-rule: if
1. $\pi[\![a:\exists r.C]\!] \geq k \in \mathcal{A}$, and
2. there is no b and no $i, j \geq k$ such that $\{\pi[\![(a,b):r]\!] \geq i, \pi[\![b:C]\!] \geq j\} \subseteq \mathcal{A}$, and
3. a is not K-blocked

then $\mathcal{A} \longrightarrow \mathcal{A} \cup \{\pi[\![(a,d):r]\!] \geq k, \pi[\![d:C]\!] \geq k\}$, where d is new in $\mathcal{A}$

In order to ensure termination with the help of this rule, we need to check that none of the rules will again and again increase the lower bounds that occur in $\mathcal{A}$. This would avoid the blocking condition as it would further and further restrict the conditions on the individual names that are introduced. But almost all rules do not introduce a bound which is larger than the bound of the original assertion. Only the weak $\sqsubseteq$-rule in the $\geq$-version has to be adapted slightly. Intuitively the ε which can be chosen as lower bound in that rule should be as small as possible but theoretically it may be set to a value larger than k. We can simply fix this issue by requiring that $\varepsilon \leq k$ since the only information we want to reflect is that the value is larger than 0. In this new version, the algorithm is guaranteed to terminate both for strong and weak consistency because as in the Boolean case, there will be only finitely many assertions about non-blocked names if the values from the semiring, which are introduced, do not grow. Soundness and completeness can also be proved similarly to the Boolean case for which a proof can be found in [3].

7 Provenance-Tracking Interpretations

Also for classical reasoning problems in DL, for purely Boolean knowledge bases $\mathcal{K} = (\mathcal{A}, \mathcal{T})$ a provenance approach might be helpful, at least over a fixed universe. Provenance interpretations in polynomial semirings can track precisely which combinations of atomic facts are responsible for the truth and falsity of a statement, and thus may help to 'repair' an interpretation that is inconsistent with some requirement.

Definition 10. *An* $\mathbb{N}[X, \bar{X}]$*-interpretation is* provenance-tracking *if it is induced by a mapping* $\pi : \mathrm{Lit}_{\Delta}(\tau) \to X \cup \bar{X} \cup \{0,1\}$ *such that* $\pi(\mathrm{Atoms}_{\Delta}(\tau)) \subseteq X \cup \{0,1\}$ *and* $\pi(\mathrm{NegAtoms}_{\Delta}(\tau)) \subseteq \bar{X} \cup \{0,1\}$*. Further,* π *maps equalities and inequalities to their truth values 0 or 1.*

The idea is that if π annotates a positive or negative atom with a token, then we wish to track that literal through the model-checking computation. On the other hand annotating with 0 or 1 is done when we do not track the literal, yet we need to recall whether it holds or not in the model. See [9] for more details and potential applications of provenance-tracking interpretations.

Consider now a simple ABox $\mathcal{A}$ and some fixed, but sufficiently large, universe Δ that in particular contains all individual constants appearing in $\mathcal{A}$. Any

concept or role assertion in $\mathcal{A}$ is identified with an atom $\alpha \in \mathrm{Atoms}_\Delta(\tau)$ for an appropriate vocabulary τ. Further, let X be the set of provenance tokens p_α, for $\alpha \in \mathrm{Atoms}_\Delta(\tau)$, and let $\bar{X}$ be the corresponding set of negative tokens $\bar{p}_\alpha$. We say that a knowledge base $\mathcal{K} = (\mathcal{A}, \mathcal{T})$ is consistent over Δ if it has a model with universe Δ.

We define the provenance tracking interpretation $\pi_\mathcal{A} : \mathrm{Lit}_\Delta(\tau) \to \mathbb{N}[X, \bar{X}]$ by

$$\pi_\mathcal{A}(\alpha) := \begin{cases} 1 & \text{if } \alpha \in \mathcal{A} \\ p_\alpha & \text{otherwise} \end{cases}$$

$$\pi_\mathcal{A}(\neg\alpha) := \begin{cases} 0 & \text{if } \alpha \in \mathcal{A} \\ \bar{p}_\alpha & \text{otherwise} \end{cases}$$

Notice that for each assertion $a : C$ the provenance value $\pi_\mathcal{A}[\![a : C]\!]$ is a polynomial in $\mathbb{N}[X, \bar{X}]$ with indeterminates p_α and $\bar{p}_\alpha$ for $\alpha \notin \mathcal{A}$. An equation system in $\mathbb{N}[X, \bar{X}]$ is a set E of equations of form $f = 0$ with $f \in \mathbb{N}[X, \bar{X}]$. A solution of E in a semiring K is a function $h : X \cup \bar{X} \to K$, making all equations in E true, such that for each token $p \in X$ we have that $h(p) = 0$ if, and only if, $h(\bar{p}) \neq 0$. In particular, such a solution is a model-defining K-interpretation [9], defining the unique structure over Δ making precisely those atoms $\alpha \in \mathrm{Atoms}_\Delta(\tau)$ true for which $h(p_\alpha) \neq 0$.

Definition 11. We associate with every knowledge base $\mathcal{K} = (\mathcal{A}, \mathcal{T})$ and every universe Δ the equation system $E^\Delta_\mathcal{K}$ consisting of the equations

$$\pi_\mathcal{A}[\![a : C]\!] \cdot \pi_\mathcal{A}[\![a : \neg D]\!] = 0$$

for all concept inclusions $C \sqsubseteq D \in \mathcal{T}$ and all $a \in \Delta$.

Proposition 3. *A knowledge base $\mathcal{K} = (\mathcal{A}, \mathcal{T})$ is consistent over Δ if, and only if, the equation system $E^\Delta_\mathcal{K}$ has a solution (in any semiring K).*

Due to the assumption that our semirings are +-positive, we can expand the equation system $E^\Delta_\mathcal{K}$ into a single polynomial

$$f^\Delta_\mathcal{K}(X, \bar{X}) := \sum_{C \sqsubseteq D \in \mathcal{T}} \sum_{a \in \Delta} \pi_\mathcal{A}[\![a : C]\!] \cdot \pi_\mathcal{A}[\![a : \neg D]\!]$$

and we have that the solutions of the equation $f^\Delta_\mathcal{K}(X, \bar{X}) = 0$ are in correspondence with the models of the knowledge base $\mathcal{K}$ on the universe Δ. Notice that for just finding the zeros of $f^\Delta_\mathcal{K}(X, \bar{X})$, it makes no difference whether we write it as a polynomial in $\mathbb{N}[X, \bar{X}]$, or in a simpler semiring such as $\mathbb{B}[X, \bar{X}]$, $\mathbb{W}[X, \bar{X}]$, $\mathbb{S}[X, \bar{X}]$, or even the semiring of positive Boolean functions. Notice further, that the problem whether such zeros exist is NP-complete.

However, provenance polynomials allow us to do more. We can compare solutions, and we can use this approach to find solutions that describe models

that are close to a given interpretation. Assume for instance that we have an interpretation $\mathcal{I}$ that is a model of a given knowledge base, but then, after adding further facts to the to ABox and/or making changes to the TBox, it happens that $\mathcal{I}$ is no longer consistent with $(\mathcal{A}, \mathcal{T})$. We may want to get back a model by a set of changes that has minimal costs in some sense. This approach is related to work in [15] on missing query answers and integrity repairs for databases.

By dualizing $f_{\mathcal{K}}^{\Delta}(X, \bar{X})$, we obtain the polynomial

$$g_{\mathcal{K}}^{\Delta}(X, \bar{X}) := \prod_{C \sqsubseteq D \in \mathcal{T}} \prod_{a \in \Delta} (\pi_{\mathcal{A}}[\![a : \neg C]\!] + \pi_{\mathcal{A}}[\![a : D]\!])$$

and we have that $g_{\mathcal{K}}^{\Delta}(X, \bar{X}) = 0$ (as a polynomial in $\mathbb{N}[X, \bar{X}]$) if, and only if, $\mathcal{K}$ is inconsistent on Δ. More interestingly, if this is not the case, then by writing out $g_{\mathcal{K}}^{\Delta}(X, \bar{X})$ as a sum of monomials $p_1^{e_1} \dots p_k^{e_k}$, we see that, for each such monomial, every interpretation that makes all those literals true that are associated with the tokens $p_1, \dots p_k$ is a model of $\mathcal{K}$. In general, such a monomial does not define a specific model, but a whole class of models, because those literals α for which neither p_α nor $\bar{p}_\alpha$ occur in the monomial can be interpreted in any way. Choices between different classes of models can then be made on the basis of any (partial) order between monomials in $\mathbb{N}[X, \bar{X}]$, and this can then be refined on the basis of selection criteria between different interpretations that make the same monomial true.

Coming back to the example of defining a model that is close to a given interpretation $\mathcal{I}$ (that itself is not anymore consistent with $\mathcal{K}$) we may for instance define a *cost interpretation* $\rho : \mathrm{Lit}_{\Delta}(\tau) \to \mathbb{T}$ into the tropical semiring $\mathbb{T} = (\mathbb{R}_+^{\infty}, \min, +, \infty, 0)$ that associates with the addition of a fact to $\mathcal{I}$ a cost $c \in \mathbb{R}$, and with the deletion of a fact a cost $d \in \mathbb{R}$. More precisely, for each atom $\alpha \in \mathrm{Atoms}_{\Delta}(\tau)$, we would put $\rho(\alpha) = 0$ and $\rho(\neg\alpha) = d$ if $\mathcal{I} \models \alpha$, and $\rho(\alpha) = c$ and $\rho(\neg\alpha) = 0$ if $\mathcal{I} \models \neg\alpha$. By setting $\hat{\rho}(p_\alpha) := \rho(\alpha)$ and $\hat{\rho}(\bar{p}_\alpha) := \rho(\neg\alpha)$, we obtain a semiring homomorphism $\hat{\rho} : \mathbb{N}[X, \bar{X}] \to \mathbb{T}$. We would then select the monomial m in $g_{\mathcal{K}}^{\Delta}(X, \bar{X})$ with minimal value $\hat{\rho}[\![m]\!]$; notice that this coincides with the provenance value $\hat{\rho}[\![g_{\mathcal{K}}^{\Delta}(X, \bar{X})]\!]$. Given the original interpretation $\mathcal{I}$ and the monomial m, we can then define a new interpretation $\mathcal{I}(m)$ with $\mathcal{I}(m) \models \alpha$ whenever p_α occurs in m, $\mathcal{I}(m) \models \neg\alpha$ whenever $\bar{p}_\alpha$ occurs in m, and $\mathcal{I}(m) \models \alpha \iff \mathcal{I} \models \alpha$ for all other atoms $\alpha \in \mathrm{Atoms}_{\Delta}(\tau)$.

We can view $\mathcal{I}(m)$ as a model of $(\mathcal{A}, \mathcal{T})$ which, among all interpretations with universe Δ, is obtained from $\mathcal{I}$ by a set of additions and deletions of facts that leads to minimal costs for establishing the consistency with $(\mathcal{A}, \mathcal{T})$. Notice in this context, that in case $\mathcal{K} = (\mathcal{A}, \mathcal{T})$ is inconsistent, and hence $g_{\mathcal{K}}^{\Delta}(X, \bar{X})$ is the zero polynomial, then $\hat{\rho}[\![g_{\mathcal{K}}^{\Delta}(X, \bar{X})]\!] = \infty$.

Instead of such a cost based choice, by means of an interpretation in the tropical semiring, the semiring framework permits also choices by other criteria, for instance by maximizing consistency scores, using an interpretation into the Viterbi semiring $\mathbb{V}$, or by minimizing the required clearance level, by an interpretation into the access control semiring $\mathbb{A}$.

Notice that all this is algorithmically nontrivial. First of all it assumes that we have determined a universe Δ on which we evaluate the provenance polynomials. This is a separate, nontrivial, problem, but for most description logics, we can determine bounds on the size of minimal models without too much effort, so this seems not infeasible. Second, neither the problem of finding zeros, nor the computation of a provenance polynomial in standard form, as a sum of monomials, are computationally easy, in general. However, it is a fact, that at least for reasonably expressive description logics, the common reasoning problems do have a rather high complexity anyway. It is thus not at all the case that provenance analysis makes easy problems complicated. To the contrary, we hope that it actually may help to provide a more principled approach to a number of interesting questions.

8 Conclusion and Outlook

We have reported on an algebraic framework for the provenance analysis of logics with negation that we believe to be suitable and interesting also for applications in description logics. As a first step, we have seen that provenance values of concept assertions from $\mathcal{ALC}$ on a fixed interpretation can be computed with a moderate number of semiring operations. We have then discussed which variations of the traditional reasoning problems for description logics may be interesting when we evaluate concept and role assertions in a commutative semiring, and what kind of new questions might be investigated with such an approach. We have further discussed the issue of extending the familiar tableaux based algorithmic methods to provenance knowledge basis, and we have illustrated this for certain specific cases. Finally we have investigated how provenance tracking interpretations in semirings of dual-indeterminate polynomials may also help to give a new approach to traditional (purely Boolean) reasoning problems such as the consistency of a knowledge base, by means of provenance polynomials that describe multiple models, and allow us to repair inconsistencies and to make choices between different models on a principled basis. Of course, this work so far is rather preliminary, and proposes more definitions and questions than that it provides answers.

An interesting area that we have left largely untouched so far is query rewriting. This is the problem of rewriting a (say, conjunctive or first-order) query q for a given TBox $\mathcal{T}$ as a new query $q_{\mathcal{T}}$ that evaluated on any given ABox $\mathcal{A}$ should provide the same answers as the (certain) answers of the original query q on (models of) the knowledge base $(\mathcal{A}, \mathcal{T})$. First-order rewritings are only possible for rather inexpressive description logics, but for certain somewhat more expressive ones, rewritings in Datalog are possible (see [3, Chap. 7]). A provenance approach to this problem has recently been explored in [16], but it is rather different from our methods and does not make use of dual-indeterminate polynomials. It should be interesting to combine these methods with ours, taking also into account the semirings of dual-indeterminate formal power series that provide the algebraic framework for a provenance analysis of languages that include both recursion and negation.

References

1. Amsterdamer, Y., Deutch, D., Tannen, V.: On the limitations of provenance for queries with difference. In: 3rd Workshop on the Theory and Practice of Provenance, TaPP 2011 (2011). CoRR abs/1105.2255
2. Amsterdamer, Y., Deutch, D., Tannen, V.: Provenance for aggregate queries. In: Principles of Database Systems, PODS, pp. 153–164 (2011). CoRR abs/1101.1110
3. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press, Cambridge (2017)
4. Dannert, K., Grädel, E.: Semiring Provenance for Guarded Logics (submitted for publication)
5. Deutch, D., Milo, T., Roy, S., Tannen, V.: Circuits for datalog provenance. In: Proceedings of 17th International Conference on Database Theory ICDT, pp. 201–212 (2014)
6. Foster, J., Green, T., Tannen, V.: Annotated XML: queries and provenance. In: Principles of Database Systems, PODS, pp. 271–280 (2008)
7. Geerts, F., Poggi, A.: On database query languages for K-relations. J. Appl. Logic **8**(2), 173–185 (2010)
8. Geerts, F., Unger, T., Karvounarakis, G., Fundulaki, I., Christophides, V.: Algebraic structures for capturing the provenance of SPARQL queries. J. ACM **63**(1), 7:1–7:63 (2016)
9. Grädel, E., Tannen, V.: Semiring provenance for first-order model checking. arXiv:1712.01980 [cs.LO] (2017)
10. Grädel, E., Tannen, V.: Provenance analysis for logic and games (2019, submitted for publication)
11. Green, T.: Containment of conjunctive queries on annotated relations. Theory Comput. Syst. **49**(2), 429–459 (2011)
12. Green, T., Ives, Z., Tannen, V.: Reconcilable differences. In: Database Theory - ICDT 2009, pp. 212–224 (2009)
13. Green, T., Karvounarakis, G., Tannen, V.: Provenance semirings. In: Principles of Database Systems PODS, pp. 31–40 (2007)
14. Green, T., Tannen, V.: The semiring framework for database provenance. In: Proceedings of PODS, pp. 93–99 (2017)
15. Xu, J., Zhang, W., Alawini, A., Tannen, V.: Provenance analysis for missing answers and integrity repairs. IEEE Data Eng. Bull. **41**(1), 39–50 (2018)
16. Ozaki, A., Penaloza, R.: Provenance in ontology-based data access. In: Proceedings of the 31st International Workshop on Description Logics (2018)
17. Tannen, V.: Provenance propagation in complex queries. In: Tannen, V., Wong, L., Libkin, L., Fan, W., Tan, W.-C., Fourman, M. (eds.) In Search of Elegance in the Theory and Practice of Computation. LNCS, vol. 8000, pp. 483–493. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41660-6_26