Verification of Evolving Graph-structured Data under Expressive Path Constraints

Diego Calvanese¹, Magdalena Ortiz², and Mantas Šimkus²

1 Free University of Bozen-Bolzano, Bozen, Italy
2 TU Wien, Viena, Austria

Abstract

Integrity constraints play a central role in databases and, among other applications, are fundamental for preserving data integrity when databases evolve as a result of operations manipulating the data. In this context, an important task is that of static verification, which consists in deciding whether a given set of constraints is preserved after the execution of a given sequence of operations, for every possible database satisfying the initial constraints. In this paper, we consider constraints over graph-structured data formulated in an expressive Description Logic (DL) that allows for regular expressions over binary relations and their inverses, generalizing many of the well-known path constraint languages proposed for semi-structured data in the last two decades. In this setting, we study the problem of static verification, for operations expressed in a simple yet flexible language built from additions and deletions of complex DL expressions. We establish undecidability of the general setting, and identify suitable restricted fragments for which we obtain tight complexity results, building on techniques developed in our previous work for simpler DLs. As a by-product, we obtain new (un)decidability results for the implication problem of path constraints, and improve previous upper bounds on the complexity of the problem.

1998 ACM Subject Classification H.2 [Database Management] General

Keywords and phrases Path constraints, Description Logics, Graph databases, Static verification

1 Introduction

Integrity constraints play a central role in databases and, among many other applications, are fundamental for preserving data integrity when databases evolve as a result of operations manipulating the data [1, 23, 6]. A fundamental problem in this context is static verification: given a set of integrity constraints, and a sequence of operations that describe changes on databases (over the same schema), the goal is to verify whether the constraints are preserved by the operations, that is, they are satisfied after their application, for every database that initially satisfies the constraints. This allows one to establish the acceptability of sequences of operations, which guarantees that applications maintain data integrity at runtime, independently of the specific database states that may be reached. However, static verification is very hard, and identifying sufficiently expressive languages for integrity constraints and data operations that allow for decidable verification is challenging.

In this paper, we consider graph-structured data (GSD), that is, relational data that contains unary and binary relations only, and thus admits a natural representation as a labeled graph. This data model is well suited for those settings where the data does not comply to a fixed schema, and the topology of the data relations is central. The GSD model became important already two decades ago due to the close relationship with semi-structured data [2, 13]. In the last decade it has gained renewed interest due to its relevance in the
Semantic Web and a diverse range of areas, including social networks, life-sciences, and program analysis, see e.g., [5, 36] and references therein. The study of query languages for GSD has been the focus of extensive research efforts in the database community over the last decade, based on the common consensus that GSD requires navigational query languages that allow, as minimal required functionality, to extract nodes that are connected by paths complying with a given regular language [36, 30]. The exploration of constraints for this data model is somehow more limited, partly due to the fact that even simple formalisms for constraining relations between regular paths result in undecidability of basic inference problems. Indeed, early proposals for path constraint languages are still viewed as adequate [3, 14, 16, 26, 17], but their wider adoption is hindered by the fact their implication problem is only known to be decidable under very strong restrictions. Recently there has been much interest in the study of containment for expressive query languages for GSD [27, 20]. In this basic form of static analysis, queries can be seen as expressing constraints over GSD, but also here inference turns out to be undecidable unless severe restrictions are imposed. For example, containment of Graph-XPath queries, a variation of XPath advocated for querying GSD, is undecidable in general, and has been shown decidable only when restricted to the so-called path-positive fragment [27]. In this paper we show undecidability for a path constraint language that is significantly more restricted than Graph-XPath, and even than the path constraints in [14], and improve the previous undecidability results that required to express paths that return to the initial point [14, 27].

We advocate an expressive Description Logic (DL) as constraint language for GSD. DLs are a family of languages tailored for representing structured knowledge, and for supporting inference over it [8]. They formalize domain knowledge by describing complex classes of objects, called concepts, and binary relations between them, called roles. Most DLs can be seen as (syntactic variants of) decidable fragments of classical first-order (FO) logic, or its extension with transitive closure. Different DLs provide different expressive means to describe knowledge, with the computational complexity of inference varying accordingly. DLs are the basis of state-of-the-art ontology languages for sharing domain conceptualizations [9], and they are naturally suited for describing data sources. They have been applied for the static analysis of traditional data models, such as UML class diagrams [10] and Entity Relationship schemata [7]. In the paradigm of ontology based data access [34, 28], which has gained great importance in the last decade, DL ontologies are used to describe possibly heterogeneous data sources, facilitating their management, and leveraging domain knowledge to improve access to them. Query answering and containment in this setting has been extensively studied, for a range of DLs and query languages [11, 15, 33].

In this paper we show that DLs are adequate also as constraint languages for GSD. We focus on the expressive DL $ZOI$, also known as $\mathcal{ALCOI}^{\text{Set}}_{\text{reg}}$ [18], which features regular expressions over binary relations and their inverses, and allows for using them to impose complex relations between concepts and roles. $ZOI$ can express the full path-positive fragment of Graph-XPath, and supports additional features that allow it to express even richer constraints on GSD. We also show that well-known path constraint languages proposed for GSD in the past [3, 14] can be naturally expressed in (variations of) $ZOI$. Moreover, we can leverage results from the DL community to generalize and improve previous upper bounds on the complexity of the implication problem for decidable fragments of path constraints. This requires, however, to lift existing algorithms and complexity results for reasoning in $ZOI$ to finite structures, since in the setting of verification of evolving GSD we are interested in finite data instances, and so far this DL had been studied over unrestricted, possibly infinite, models only [18]. This transfer of results to the finite setting is fortunately possible.
since \( \mathcal{ZOI} \) enjoys finite model property (FMP), as we are able to prove. This is a crucial stepping stone for our results, and an interesting contribution on its own right. Indeed, while the FMP has been long known for the closely related converse PDL [24, 21], \( \mathcal{ZOI} \) has several further features. In particular, it allows for Boolean combinations of roles that make standard filtration techniques not directly applicable, and call for more subtle arguments that borrow ideas from FMP proofs for the two-variable fragment of first-order logic [32, 25].

For expressing operations on GSD we use the action language proposed in [4], that allows for the (possibly conditional) composition of basic operations that add or delete from a predicate the objects selected by a complex DL concept or role. The undecidability of general path constraint implication implies that static verification becomes undecidable when if complex roles are allowed in actions. However, we regain decidability by restricting the roles in the action to be simple roles that allow for union, intersection, and difference of possibly inverse roles, but disallow composition and the Kleene star. Under these restrictions, we can rely on the techniques of [4] to reduce static verification in the presence of \( \mathcal{ZOI} \) constraints to satisfiability of \( \mathcal{ZOI} \) KBs, obtaining a tight ExpTime upper bound for the former.

The paper is organized as follows. In Section 2 we introduce the DL \( \mathcal{ZOI} \), and establish the finite model property for it. Section 3 is devoted to path constraint languages. We tighten the undecidability of path constraint implication shown in [14]. We also generalize the decidability in [3] to a richer class, and improve the upper bound from 2ExpSpace to ExpTime by reducing the problem to reasoning in \( \mathcal{ZOI} \). Section 4 studies the static verification of \( \mathcal{ZOI} \) constraints over GSD, for actions expressed in the language proposed in [4]. We show that the problem is undecidable if arbitrary \( \mathcal{ZOI} \) roles occur in actions, and impose suitable restrictions to obtain ExpTime decidability using the techniques of [4].

2 Expressive DLs for Expressing Constraints over GSD

In this paper, we formalize GSD as relational structures, which we call instances, over a unary and binary relational signature. We propose to use the rich DL \( \mathcal{ZOI} \), also known as \( \mathcal{ALCOI}^{reg} \) [18], to express constraints over graph structured data. This is natural as, like other DLs, \( \mathcal{ZOI} \) is defined over a relational vocabulary that contains unary and binary predicates only (respectively called concept names and role names in DL jargon) and the structures over which is interpreted are precisely GSD instances. A distinguishing feature of \( \mathcal{ZOI} \), which makes it especially adequate for describing GSD, is that it can express relations between objects by allowing for complex roles defined using regular expressions.

**Definition 1 (\( \mathcal{ZOI} \) syntax).** We consider fixed, countably infinite sets \( \mathbb{N}_C \) of concept names, \( \mathbb{N}_R \) of role names, and \( \mathbb{N}_I \) of individual names. We assume that the set \( \mathbb{N}_C \) contains the special concepts \( \top \) (top) and \( \bot \) (bottom), while \( \mathbb{N}_R \) contains the top (universal) role \( T \) and the bottom (empty) role \( B \). We define (\( \mathcal{ZOI} \)) atomic concepts \( B \), concepts \( C, C' \), atomic roles \( P \), simple roles \( S, S' \), and roles \( R, R' \), where \( a, b \in \mathbb{N}_I \), \( A \in \mathbb{N}_C \), \( r \in \mathbb{N}_R \), and \( r \neq T \), according to the following syntax:

\[
\begin{align*}
B & \rightarrow A \mid \{a\} \\
C, C' & \rightarrow B \mid \neg C \mid C \cap C' \mid C \cup C' \mid \forall R.C \mid \exists R.C \mid \exists S.\text{Self} \\
P & \rightarrow r \mid r^{-} \mid \{a, b\} \\
S, S' & \rightarrow P \mid S \cap S' \mid S \cup S' \mid S \setminus S' \\
R, R' & \rightarrow T \mid \varepsilon \mid \text{id}(C) \mid S \mid R \cup R' \mid R \circ R' \mid R'
\end{align*}
\]

We call \( \mathcal{ZOI} \) expressions \( \{a\} \) nominal concepts, and \( \{a, b\} \) nominal roles.

\[1\] In our setting, updates are performed data that is viewed as complete, and DL constraints are not used to infer new knowledge. Thus we do not run in the expressiveness issues considered e.g., in [31, 19].
We use $\mathcal{ZOI}$ concepts and roles to define a general form of knowledge bases, in which we allow for Boolean combinations of intensional and extensional level statements.

**Definition 2 ($\mathcal{ZOI}$ knowledge bases).** A concept inclusion is an expression of the form $C \subseteq C'$, where $C, C'$ are arbitrary concepts, and a role inclusion is an expression of the form $S \subseteq S'$, where $S, S'$ are simple roles. An assertion is an expression of the form $C(a), S(a, a')$, or $a \neq a'$, where $C$ is a concept, $S$ a simple role, and $\{a, a'\} \subseteq \mathbb{N}_i$. Then, ($\mathcal{ZOI}$) knowledge bases (KBs) are defined inductively as follows:

(i) every inclusion and every assertion is a KB;

(ii) if $\mathcal{K}, \mathcal{K}'$ are KBs, so are $\mathcal{K} \land \mathcal{K}'$, $\mathcal{K} \lor \mathcal{K}'$, and $\neg \mathcal{K}$.

The semantics of $\mathcal{ZOI}$ is based on standard relational structures. An instance (or interpretation) $I = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ consists of a non-empty domain $\Delta^\mathcal{I}$ and an interpretation function $\cdot^\mathcal{I}$ that maps each individual $a \in \mathbb{N}_i$ to an element $a^\mathcal{I} \in \Delta^\mathcal{I}$, each concept name $A \in \mathbb{N}_C$ to a set $A^\mathcal{I} \subseteq \Delta^\mathcal{I}$, and each role name $r \in \mathbb{N}_R$ to a set $r^\mathcal{I} \subseteq \Delta^\mathcal{I} \times \Delta^\mathcal{I}$, in such a way that $\top^\mathcal{I} = \Delta^\mathcal{I}$, $\bot^\mathcal{I} = \emptyset$, $\sqsubseteq^\mathcal{I} = \Delta^\mathcal{I} \times \Delta^\mathcal{I}$, and $\emptyset^\mathcal{I} = \emptyset$. The function $\cdot^\mathcal{I}$ is inductively extended to all $\mathcal{ZOI}$ concepts and roles as follows:

$$
\{a\}^\mathcal{I} = \{a^\mathcal{I}\} \\
(\neg C)^\mathcal{I} = \Delta^\mathcal{I} \setminus C^\mathcal{I} \\
(C \land C')^\mathcal{I} = C^\mathcal{I} \cap C'^\mathcal{I} \\
(C \lor C')^\mathcal{I} = C^\mathcal{I} \cup C'^\mathcal{I} \\
(\forall R.C)^\mathcal{I} = \{x \mid \forall y. (x, y) \in R^\mathcal{I} \rightarrow y \in C^\mathcal{I}\} \\
(\exists R.C)^\mathcal{I} = \{x \mid \exists y. (x, y) \in R^\mathcal{I} \land y \in C^\mathcal{I}\} \\
(\exists S.\text{id})^\mathcal{I} = \{x \mid (x, x) \in S^\mathcal{I}\}
$$

where $\cap, \cup,$ and $\setminus$ are overloaded to denote also the standard set-theoretic operations, $\circ$ to denote composition, and $\cdot^*$ to denote the reflexive transitive closure of a binary relation.

$I$ satisfies the inclusion $E \subseteq E'$ if $E^\mathcal{I} \subseteq E'^\mathcal{I}$, the assertions $A(a)$ if $a^\mathcal{I} \in A^\mathcal{I}$, $S(a, b)$ if $(a^\mathcal{I}, b^\mathcal{I}) \in S^\mathcal{I}$, and $a \neq b$ if $a^\mathcal{I} \neq b^\mathcal{I}$. Satisfaction is extended in the usual way to KBs, which are Boolean combinations of inclusions and assertions. When $I$ satisfies $\mathcal{K}$, we also say that $I$ is a model of $\mathcal{K}$, and denote it with $I \models \mathcal{K}$.

As basic reasoning task we consider KB satisfiability, which consists in deciding, given a KB $\mathcal{K}$, whether $\mathcal{K}$ admits a model. Other standard reasoning tasks, like concept (resp., role) satisfiability, that is, deciding whether there exists an interpretation where the extension of a given concept (resp., role) is not empty, can be reduced to KB satisfiability.

**Remark.** The roles $\varepsilon$ and $\{(a, b)\}$, which are not usually included in $\mathcal{ZOI}$, are just syntactic sugar. Indeed, $\varepsilon$ has the same meaning as $\text{id}(\top)$ and captures the identity relation. Nominal roles $\{(a, b)\}$ can be easily simulated in $\mathcal{K}$ by replacing each occurrence of $\{(a, b)\}$ by a fresh role name $r_{ab}$, and conjunctively adding to $\mathcal{K}$ the KB $r_{ab}(a, b) \land (\forall r_{ab}.(\top \sqsubseteq \{a\}) \land (\top \sqsubseteq \forall r_{ab}.\{b\})$. This ensures that $r_{ab}^\mathcal{I} = \{(a^\mathcal{I}, b^\mathcal{I})\}$ in every model of the modified KB.

**Example 3.** As a running example, we consider the following self-explanatory instance $I_{unf}$. For simplicity, in the examples we interpret individuals as themselves (i.e., we make the standard name assumption).
Consider the KB $\mathcal{K}_{\text{Uni}}$ defined as the conjunction of the following ZOI constraints: $\phi_1$ says that the domain of ‘offers’ are the departments, and $\phi_2$ says that its range is the union of programs and courses. Similarly, $\phi_3$ and $\phi_4$ restrict the domain of ‘requires’ to programs, and its range to courses other entities that comprise courses, like modules. Finally $\phi_5$ says that every course that is required (directly, or because it is part of a required module) must be offered.

$$\phi_1 = \exists \text{offers}. \top \sqsubseteq \text{Dept}$$
$$\phi_2 = \top \sqsubseteq \forall \text{offers}. (\text{Program} \sqcup \text{Course})$$
$$\phi_3 = \exists \text{requires}. \top \sqsubseteq \text{Program}$$
$$\phi_4 = \top \sqsubseteq \forall \text{requires}. (\exists \text{partOf}^- \cdot \text{Course})$$
$$\phi_5 = \text{Course} \sqcap \exists (\text{partOf}^* \circ \text{requires}^-). \top \sqsubseteq \exists \text{offers}^-. \top$$

Note that all these constraints are satisfied by our instance, that is, $\mathcal{I}_{\text{Uni}} \models \mathcal{K}_{\text{Uni}}$.

We note that ZOI is closely related to path-positive Graph-XPath (abbreviated GXPath$^\text{path-pos}$) introduced in [30]. By viewing arc labels as role names, node formulas in GXPath$^\text{path-pos}$ can be written as ZOI concepts, and GXPath$^\text{path-pos}$ path formulas as ZOI roles. Additionally ZOI extends GXPath$^\text{path-pos}$ with other features, such as Boolean combinations of node and path labels (i.e., Boolean concepts and roles), nominals, and concepts of the form $\exists S$.Self.

In [18], a tree-automata based algorithm for checking satisfiability of ZOI concepts is provided, and by using a variant of internalization [35], this is exploited to check satisfiability of ZOI KBs constituted by a conjunction of (positive) assertions and inclusions. It is easy to extend internalization also to ZOI KBs of the more general form considered here, and thus reduce satisfiability of a ZOI KB to satisfiability of a ZOI concept. The proof is given in the extended version of the paper.

▼ Theorem 4. Given a ZOI KB $\mathcal{K}$, one can construct in linear time a ZOI concept $C_{\mathcal{K}}$ such that $\mathcal{K}$ is satisfiable if and only if $C_{\mathcal{K}}$ is so.

From this result and the ExpTime upper bound for concept satisfiability given in [18], it follows immediately that satisfiability of ZOI KBs is decidable in single exponential time. This is worst-case optimal, since the problem is ExpTime-hard even for significantly simpler DLs like $\mathcal{ALC}$ [8].

▼ Theorem 5 ([18]). Checking satisfiability of ZOI KBs is an ExpTime-complete problem.
Finite Model Reasoning in \(ZOI\). In the setting of GSD we are usually interested in finite instances. In the DL literature, however, finite model reasoning has received significantly less attention than reasoning with respect to unrestricted models. To our knowledge, finite model reasoning for \(ZOI\) has not been addressed so far. However, as we show in the following, \(ZOI\) enjoys the finite model property, which states that every satisfiable KB admits a model whose domain is finite. In line with what has been done for other logics that cannot express functionality, keys, or number restrictions, we can show this through a filtration argument [24, 21]. However, due to the presence of both transitive closure over roles and role intersection and difference, the proof is more involved than for logics that involve none or only one of the two kinds of constructs.

We say a KB \(K\) is finitely satisfiable if it admits a finite model, i.e., a model with a finite domain. The proof of the following result is given in the extended version of the paper.

\[\text{Theorem 6. Let } K \text{ be a } ZOI \text{ KB. Then } K \text{ is satisfiable if and only if } K \text{ is finitely satisfiable.}\]

3 Path Constraints

We define a language for path constraints inspired by [3, 14, 26] and closely related to \(ZOI\).

\[\text{Definition 7 (Path constraints). A path constraint } \varphi \text{ has the form } [R_p](R_\ell \subseteq R_r), \text{ where } R_p, R_\ell, \text{ and } R_r \text{ are arbitrary } ZOI \text{ roles. The role } R_p \text{ is called prefix of } \varphi, \text{ while the roles } R_\ell \text{ and } R_r \text{ are respectively called the left tail and the right tail of } \varphi. \text{ If } R_p = \varepsilon, \text{ we call } \varphi \text{ a prefix-empty constraint}\]², and write it simply as \(R_\ell \subseteq R_r\).

This definition generalizes the well-known path constraint languages from [3, 14]. A complex role \(R\) built from the symbols in \(N_R \cup \{\varepsilon\}\) using \(\circ, \cup, \text{ and } *\) is called a (one-way) regular path role, and if additionally it does not contain \(\cup\) or \(*\) then it is called a (one-way) word role. A one-way regular path constraint (called simply path constraint in [3]) is a prefix-empty constraint where \(R_\ell\) and \(R_r\) are one-way regular path roles. If, in addition, \(R_\ell\) and \(R_r\) are one-way word roles, the path constraint is called a word constraint in [3]. The language of path constraints in [14] allows for non-empty prefixes, but restricts the left tail to be a one-way word role, and the right tail to be either a one-way word role (in the so-called forward constraints), or an inverted one-way word role (in backward constraints), which is a sequence of concatenated inverses of role names (that is, \(r_1^{-1} \circ \cdots \circ r_n^{-1}\) with \(n \geq 0\)).³

Now we define the semantics of path constraints and their fundamental reasoning problem, namely implication of path constraints, both in its finite and in its unrestricted variants.

In the semantics of early path constraint languages [3, 14], every instance has a distinguished root object at which the constraints are enforced. We introduce a minor variation of this semantics, which we call pointed semantics, where rather than a fixed name for the root node, we allow for any individual name to be used as its identifier. Later works advocated what we call the global semantics [22, 26], in which constraints are enforced at every point in the model, rather than at just one. We note that the global semantics is in general computationally more costly, and causes undecidability of the implication problem for some fragments that are decidable under the pointed semantics [3, 14]. We discuss below how both

² Prefix-empty constraints were called simple in [14]. We use a different name to avoid confusion with the simple roles of Definition 1.

³ We note that [14] uses a different syntax with explicit variables.
semantics can be naturally captured in DLs, and provide decidability and undecidability results for both of them.

Definition 8 (Pointed and rooted semantics, implication problem). Let \( \varphi = [R_p](R_t \subseteq R_r) \) be a path constraint. For an interpretation \( \mathcal{I} \), we let \( \varphi^\mathcal{I} \) be the set of objects \( d \in \Delta^\mathcal{I} \) such that for each \( d', d'' \in \Delta^\mathcal{I} \), if \( (d, d') \in R^\mathcal{I}_p \) and \( (d', d'') \in R^\mathcal{I}_r \), then \( (d', d'') \in R^\mathcal{I}_r \).

A pointed instance is a pair \( \mathcal{I}, a \) of an instance \( \mathcal{I} \) and an individual \( a \). We call \( \mathcal{I}, a \) a pointed model of \( \varphi \), and write \( \mathcal{I}, a \models \varphi \) if \( \varphi^\mathcal{I} \) is non-empty. Similarly, we write \( \mathcal{I}, a \models \Gamma \) for a set \( \Gamma \) of constraints, if \( \mathcal{I}, a \models \varphi \) for each \( \varphi \in \Gamma \). We write \( \mathcal{I}, a \models \varphi \) if \( \mathcal{I}, a \models \varphi \) for every pointed model \( \mathcal{I}, a \) of \( \Gamma \), and write \( \Gamma, a \models \varphi \) if \( \mathcal{I}, a \models \varphi \) for every finite pointed model \( \mathcal{I}, a \) of \( \Gamma \).

The (finite) pointed implication problem consists in deciding, given an individual \( a \), a set \( \Gamma \) of path constraints, and a path constraint \( \varphi \), whether \( \Gamma, a \models \varphi \).

Let \( \varphi = [R_p](R_t \subseteq R_r) \) be a path constraint. We call \( \mathcal{I} \) a global model of \( \varphi \), and write \( \mathcal{I} \models \varphi \) if \( \varphi^\mathcal{I} \) is non-empty. We write \( \mathcal{I} \models \Gamma \) for a set \( \Gamma \) of constraints, if \( \mathcal{I} \models \varphi \) for each \( \varphi \in \Gamma \). We write \( \mathcal{I} \models \varphi \) if \( \mathcal{I} \models \varphi \) for every \( \mathcal{I} \) with \( \mathcal{I} \models \Gamma \), and write \( \Gamma \models \varphi \) if \( \mathcal{I} \models \varphi \) for every finite \( \mathcal{I} \) with \( \mathcal{I} \models \Gamma \).

The (finite) global implication problem consists in deciding, given a set \( \Gamma \) of path constraints and a path constraint \( \varphi \), whether \( \Gamma \models \varphi \).

Example 9. Consider the constraint \( \varphi_1 = R_1 \subseteq R_2 \), where

\[
R_1 = \text{id(Dept)} \circ \text{partOf}^{-1} \circ \text{offers} \circ \text{requires} \circ \text{partOf}^{-*} \\
R_2 = \text{id(Dept)} \circ \text{partOf}^{-1} \circ \text{offers}
\]

Intuitively, (a node interpreting) a department satisfies \( \varphi_1 \) if every course required by a program offered by (a suborganization of) the department is offered by (a suborganization of) the same department. With the rooted semantics, we can enforce the constraint for some specific departments. For example, we may require it for computer science, and our example instance satisfies it: \( \mathcal{I}_{\text{uni}}, \text{CS_Dept} \models \varphi_1 \). With the global semantics, the constraint would apply to all departments, but we can easily modify it so that it applies only to the desired departments, e.g., \( (\text{id(CS_dept)} \cdot R_1) \subseteq R_2 \).

To illustrate the use of prefixes, suppose that the policy expressed by \( \varphi_1 \) is enforced not at the department level, but at the higher faculty/school level. For example, suppose that the School of Science and Engineering requires that all departments offer within their department every mandatory course in their programs (while other schools may allow for mandatory courses that are offered by different departments). This is captured by the constraint with non-empty prefix \( \varphi_1 = [\text{id(SciEng)} \circ \text{hasDepartment}(R_1 \subseteq R_2) \).

Expressing Path Constraints in ZOI with Role Difference. To express path constraints, we extend ZOI with difference \( R \setminus R' \) of arbitrary roles, resulting in the logic we call ZOI\!

Definition 10. ZOI\! roles are defined analogously to ZOI roles, except that for complex roles we have \( R, R' \rightarrow \mathcal{T} \mid \text{id(C)} \mid S \mid R \cup R' \mid R \circ R' \mid R \setminus R' \mid R^* \).

The syntax and semantics of ZOI\! concepts, assertions, axioms, and knowledge bases are defined as for ZOI, but allowing for ZOI\! roles in the place of ZOI roles.

Entailment of path constraints defined above can be reduced to reasoning in ZOI\!:

Lemma 11. For a path constraint \( \varphi = [R_p](R_t \subseteq R_r) \), let \( C_{\varphi} = \forall R_p, (\forall (R_t \setminus R_r), \bot) \). Then, for every instance \( \mathcal{I} \), we have that \( \varphi^\mathcal{I} = C_{\varphi}^\mathcal{I} \). Consequently, for a set \( \Gamma \) of path constraints, a
path constraint $\varphi$, and $a \in \mathbb{N}_i$, we have:

$$
\begin{align*}
\Gamma, a \models \varphi & \iff (\bigwedge_{\gamma \in \Gamma} (T \subseteq C_\gamma) \land \neg (T \subseteq C_\varphi)) \text{ is unsatisfiable}, \\
\Gamma, a \models_{\text{fin}} \varphi & \iff (\bigwedge_{\gamma \in \Gamma} (T \subseteq C_\gamma) \land \neg (T \subseteq C_\varphi)) \text{ is finitely unsatisfiable}, \\
\Gamma \models \varphi & \iff \bigwedge_{\gamma \in \Gamma} (T \subseteq C_\gamma) \land \neg (T \subseteq C_\varphi) \text{ is unsatisfiable}, \\
\Gamma \models_{\text{fin}} \varphi & \iff \bigwedge_{\gamma \in \Gamma} (T \subseteq C_\gamma) \land \neg (T \subseteq C_\varphi) \text{ is finitely unsatisfiable}.
\end{align*}
$$

We will see that, unfortunately, both implication of path constraints and reasoning in $\ZOI$ are undecidable. Before moving to these negative results, though, we point out that the lemma above implies that the upper bounds for reasoning in plain $\ZOI$ extend to the implication of path constraints $\gamma$ where only simple $\ZOI$ roles occur. Since in this case the resulting $C_\gamma$ is a $\ZOI$ concept, from Lemma 11 and theorem 5 we get:

**Corollary 12.** Let $\Gamma$ be a set of path constraints and $\varphi$ a path constraint such that, for each $\gamma = [R_p][R_t \subseteq R_c] \in \Gamma \cup \{\varphi\}$, $R_p$, $R_t$ and $R_c$ are all simple $\ZOI$ roles. Then $\Gamma, a \models \varphi$, $\Gamma, a \models_{\text{fin}} \varphi$, $\Gamma \models \varphi$, and $\Gamma \models_{\text{fin}} \varphi$ are all decidable in $\text{ExpTime}$.

**Undecidability of Path Constraint Implication.** Unfortunately, both the finite and the unrestricted implication problems are undecidable in rather restricted settings. The following result was established already several years ago:

**Theorem 13 ([14]).** Assume that every instance has a distinguished root element $o$, and that there is some $a_o \in \mathbb{N}_i$ such that $a_o^2 = o$ in every $I$. The problem of checking whether $\Gamma, a_o \models \varphi$ and whether $\Gamma, a_o \models_{\text{fin}} \varphi$ are undecidable, even when $\varphi$ and all constraints in $\Gamma$ satisfy one of the following two restrictions:

- All prefixes, left tails, and right tails are one-way word roles (that is, only forward constraints according to [14] are allowed).
- All prefixes and left tails are one-way word roles different from $\varepsilon$, and each right tail is a one-way word role or an inverted one-way word role, and is different from $\varepsilon$.

We strengthen this result, showing undecidability when both restrictions apply: only one-way word roles of length one or two are allowed. Our proof encodes a Turing rather than a two-register machine as in [14], and we believe some readers may find it simpler.

**Theorem 14.** The problems of checking whether $\Gamma, a \models \varphi$ and whether $\Gamma, a \models_{\text{fin}} \varphi$, given $\Gamma$, $a$, and $\varphi$ are undecidable. This holds even when $\varphi$ is of the form $r_1 \subseteq r_2$ and $\Gamma$ contains only constraints of the following forms, where $\{r, r_1, r_2, r_3\} \subseteq \mathbb{N}_R$:

$$
\begin{align*}
\forall r_1 & \subseteq r_3 \\
\forall r_1 \subseteq r_2 \circ r_3 & \\
[r]r_1 \circ r_2 \subseteq r_3 & \\
[r]r_1 \subseteq r_2 \circ r_3 &
\end{align*}
$$

**Proof.** As in [14] we employ the notion of **conservative reduction classes** to simultaneously deal with general and finite implication. In addition, we see deciding $\Gamma, a \models \varphi_M$ as checking unsatisfiability of the first order formula that corresponds to $\Gamma$ and the negation of $\varphi_M$. The same is true for finite implication. Let $\text{FO}$ denote the set of FO formulae, $X$ be a recursive subset of $\text{FO}$, and let $f : \text{FO} \rightarrow X$ be a recursive function such that:

- if $\beta \in \text{FO}$ is unsatisfiable, then $f(\beta)$ is unsatisfiable, and
- if $\beta \in \text{FO}$ has a finite model, then $f(\beta)$ has a finite model.

---

4 In fact, the authors of [14] show that implication is r.e. complete, and finite implication co-r.e. complete.
Then $X$ is a conservative reduction class and thus satisfiability of formulae in $X$ is co-r.e.-complete and finite satisfiability r.e.-complete [12]. It is well known that for a first-order formula $\beta$ we can build a procedure that takes no input and terminates iff $\beta$ is unsatisfiable or has a finite model. Thus to show the undecidability of general and finite implication, it suffices to show how the computation of such a procedure can be simulated using path constraints. We consider a deterministic Turing Machine (TM) $M$ with state set $Q$ and alphabet $\Sigma$. We assume that $M$ has two designated states $q_{in}$ and $q_{fin}$. We build $I$ and $\varphi_M$ such that the following conditions are satisfied:

1. If $M$ reaches $q_{in}$ starting from the empty string as input, then $\Gamma_M, a \models \varphi_M$.
2. If $M$ reaches $q_{fin}$ starting from the empty string as input, then there exists a finite $I$ such that $I, a \models \Gamma_M$ and $I, a \not\models \varphi_M$.

We assume that $M$ starts at the initial tape position, and never moves to the left of it. Moreover, the tape is initially empty, that is, it only contains the blank symbol $\sqcup$. The transition function of $M$ is of the form $\delta : Q \times \Sigma \to \Sigma \times Q \times \{R, L\}$, with the usual reading, where $R$ and $L$ stand for right and left, respectively. The initial state of $M$ is $q_{ini}$.

The constraint $\varphi_M$ takes the form $u_{\text{ini}} \subseteq u_{\text{halt}}$ where $u_{\text{ini}}$ and $u_{\text{halt}}$ are two role names. We define the set $\Gamma_M$ next. Intuitively, the idea of the reduction is the following. Assume an arbitrary pointed structure $I, a$ and let $a$ denote $a^I$. Whenever there is some $\sigma' \in \Delta_I$ such that $(o, \sigma') \in u_{\text{ini}}^I$, the constraints in $\Gamma_M$ will ensure that if $I, a \models \Gamma_M$, then $I$ contains a (possibly infinite) structure that represents the computation of $M$, and that $(o, \sigma') \in u_{\text{ini}}^I$ whenever the computation halts.

To provide an intuitive description of how this structure is enforced, we will use the term $r$-arc to refer to a pairs $(d, d') \in r^I$ for a role name $r$. Now we describe the constraints in $\Gamma_M$, which ensure that from the initial arc $u_{\text{ini}}$ we build a full computation of $M$. We use role names of the form $l_{q, \sigma}$ and $f_{q, \sigma}$ for each $q \in Q \cup \{\#\}$ and each $\sigma \in \Sigma \cup \{\}$. Each of these symbols stands for a tape position containing the symbol $\sigma$. The marker $\#$ indicates that the head of $M$ is not on the current position, while $q \in Q$ indicates that the head is on the current position and $M$ is in state $q$. The symbols $f_{q, \sigma}$ are used to distinguish the right-most tape position, while regular ‘inner’ positions are represented by symbols $l_{q, \sigma}$.

The first two constraints are prefix-empty and use the auxiliary role $u_{\text{aux}}$. They ensure that whenever there is an $u_{\text{ini}}$ arc, there exists also an arc labeled $f_{q_{ini}}$, indicating that $M$ is in state $q_{ini}$, the current tape position contains the symbol $\sqcup$, and the current tape position is the right-most one that has been visited so far:

$$u_{\text{ini}} \subseteq u_{\text{aux}} \circ u_{\text{out}}$$

(1)

$$u_{\text{aux}} \subseteq u_{\text{ini}} \circ f_{q_{ini}}$$

(2)

Note that there is an $u_{\text{ini}}$ arc from $o$ to the beginning of $f_{q_{ini}}$, and an $u_{\text{out}}$ arc from its end to $o'$. For the initial and final points of this $f_{q_{ini}}$ arc, and of all the arcs $r$ that our construction will generate below in order to simulate the runs of $M$, we want the role name $u_\text{in}$ to connect $o$ to the point, and $u_\text{out}$ to connect the point to $o'$. We ensure this by adding the following constraints:

$$[u_{\text{in}} \circ r \subseteq u_{\text{ini}}]$$

for every $r \in NR \setminus \{u_{\text{ini}}, u_{\text{halt}}, u_{\text{in}}, u_{\text{out}}, u_{\text{aux}}\}$ occurring in $\Gamma_M$

(3)

$$[u_{\text{in}} \circ u_{\text{out}} \subseteq u_{\text{out}}]$$

for every $r \in NR \setminus \{u_{\text{ini}}, u_{\text{halt}}, u_{\text{in}}, u_{\text{out}}, u_{\text{aux}}\}$ occurring in $\Gamma_M$

(4)

Note that there is a pair of these constraints for every role name occurring in the rest of the proof, except for the $u$ roles that do not have a direct correspondence with the configurations of $M$. With these axioms, the initial and final points of the $f_{q_{ini}}$ arc are both reachable.
Additionally to the contents to the left, which have been taken care of, we also need to propagate tape contents to the right of the current position. We use constraints analogous to (7) and (8):

\[\forall n \in Q \cup \{\#, \_\}, \sigma \in \Sigma \cup \{\_\} : \]
\[u_n(n_{\sigma q} \subseteq d_{\sigma q}) \quad \text{for each } q \in Q \cup \{\#\} \] (14)
\[u_n(d_{\sigma q} \subseteq n_{\sigma q} \circ t_{\sigma q}) \quad \text{for each } q \in Q \cup \{\#\} \] (15)

Figure 1 Model of \(\Gamma_M\) and \(u_{ini}(o, o')\) representing the initial configuration of \(M\).
When we propagate to the right, we need to ensure that the final position \( t_{q \sigma} \) is also copied to the next configuration:

\[
[u_{in}] \langle n_{-} \circ f_{q \sigma} \subseteq d_{q \sigma} \rangle \quad \text{for each } q \in Q \cup \{\#\}, \sigma \in \Sigma \cup \{\_\} \quad (16) \\
[u_{in}] \langle d_{q \sigma} \subseteq f_{q \sigma} \circ n_{+} \rangle \quad \text{for each } q \in Q \cup \{\#\}, \sigma \in \Sigma \cup \{\_\} \quad (17)
\]

Finally we handle transitions to the left from a non-final tape position:

\[
[u_{in}] \langle t_{q \sigma} \subseteq d_{-q} \circ n_{-} \circ n_{-} \rangle \quad \text{for each } q \in Q, \sigma \in \Sigma \cup \{\_\} \quad (18) \\
[u_{in}] \langle d_{-q} \circ n_{-} \subseteq n_{-} \circ t_{\#} \circ n_{-} \rangle \quad \text{for each } q \in Q, \sigma \in \Sigma \cup \{\_\} \quad (19) \\
[u_{in}] \langle t_{\#} \circ n_{+} \subseteq d_{-q} \rangle \quad \text{for each } q \in Q, \sigma \in \Sigma \cup \{\_\} \quad (20)
\]

Figure 2 illustrates how the constraints simulate the transitions of \( \mathcal{M} \). For readability, the \( u_{in} \) arcs to and \( u_{out} \) arcs from every node are omitted.

An interpretation that is a model of the transitions we have presented correctly executes a computation of \( \mathcal{M} \). It is only left to enforce the special role \( u_{halt} \) to hold between \( o \) and \( o' \) iff the computation halts in \( q_{fin}^1 \). We first ensure that if an arc \( t_{\#i, o} \) or \( f_{\#i, o} \) for some \( \sigma \in \Sigma \cup \{\_\} \) was generated (that is, \( q_{fin}^1 \) was reached in the computation), an \( u_{halt} \) arc is created; note that if \( \mathcal{M} \) halts in \( q_{fin}^2 \) nothing enforces \( u_{halt} \) and there will be a model of the constraints where \( \varphi_M \) does not hold. Then we ensure that if there is an \( u_{halt} \) arc anywhere, then there is such an arc between \( o \) and \( o' \). This is easy to do, exploiting the fact that all points are reachable via \( u_{in} \) from \( o \), and reach \( o' \) via \( u_{in} \):

\[
\begin{align*}
    u_{in} \circ t_{\#i, \sigma} &\subseteq u_{halt} &\text{for each } \sigma \in \Sigma \cup \{\_\} \quad (21) \\
    u_{in} \circ f_{\#i, \sigma} &\subseteq u_{halt} &\text{for each } \sigma \in \Sigma \cup \{\_\} \quad (22) \\
    u_{halt} \circ u_{\text{out}} &\subseteq u_{halt} \quad (23)
\end{align*}
\]

Let \( \Gamma_M \) contain the constraints (1) – (23). We show that the reduction is as desired:

1. \( \mathcal{M} \) reaches \( q_{fin}^1 \) iff \( \Gamma_M, a \models \varphi_M \).

Suppose \( \mathcal{M} \) reaches \( q_{fin}^1 \). Assume an arbitrary \( I, a \) such that \( I, a \models \Gamma_M \), and assume there is an arbitrary \( d \in \Delta^2 \) such that \( (a^2, d) \in u_{halt}^T \). Since \( I, a \models \Gamma_M \) and \( \mathcal{M} \) reaches \( q_{fin}^1 \), then \( (a^2, d) \in u_{halt}^T \) follows. This implies \( I, a \models \varphi_M \) and hence \( \Gamma, a \models \varphi_M \). For the converse, assume \( \Gamma, a \not\models \varphi_M \). That is, there exists some \( I, a \) such that \( I, a \models \Gamma_M \) but \( I, a \not\models \varphi_M \). That is, there is \( d \in \Delta^2 \) such that \( (a^2, d) \in u_{halt}^T \) but \( (a^2, d) \not\in u_{halt}^T \). Since \( I, a \models \Gamma_M \), by constraints (3), (4), (21), and (23) we have \( (a^2, d) \not\in u_{halt}^T \), and \( I \) cannot
contain any arcs of the form \( t_{a,b} \) or \( f_{a,b} \), which means that the computation of \( M \) does not reach the state \( q_{\text{fin}}^1 \).

2. If \( M \) reaches \( q_{\text{fin}}^2 \), then there exists a finite \( I \) such that \( I, a \models \Gamma_M \) and \( I, a \not\models \varphi_M \).

Suppose that there is a computation of \( M \) that reaches \( q_{\text{fin}}^2 \). Then this computation does not reach \( q_{\text{fin}}^1 \). Moreover, there is some \( I, a \) that simulates this computation, that is, such that \( (a^2, d) \in u^R_{\text{fin}} I_{\text{fin}} \) and \( I, a \models \Gamma_M \). Since \( M \) does not reach \( q_{\text{fin}}^1 \), we can assume there are no arcs \( t_{a,b} \) or \( f_{a,b} \) in \( I \), and since the only constraints that imply \( u_{\text{halt}} \) are (21) and (23), we can assume \( u^R_{\text{halt}} = \emptyset \). Hence we have \( I, a \models \Gamma_M \) with \( (a^2, d) \in u^R_{\text{fin}} \) but \( u^R_{\text{halt}} = \emptyset \) as desired.

This concludes the proof of Theorem 14.

\[\Box\]

The same proof that we gave for the pointed semantics of path constraints applies also to the global semantics, even if we further restrict the constraints to be prefix-empty.

**Theorem 15.** The problems of deciding \( \Gamma \models \varphi \) and \( \Gamma \models_{\text{fin}} \varphi \) given \( \Gamma \) and \( \varphi \) are undecidable, even when \( \varphi \) is of the form \( r_1 \subseteq r_2 \) and \( \Gamma \) contains only constraints of the forms \( r_1 \circ r_2 \subseteq r_3 \) and \( r_1 \subseteq r_2 \circ r_3 \), for \( \{r_1, r_2, r_3\} \subseteq N_R \).

**Proof.** It suffices to observe that in the proof of Theorem 14, all constraints \([R_p](R_f \subseteq R_e)\) with \( R_f \neq \epsilon \) have \( R_e = u_{\text{fin}} \), and the only role of the prefix is to ensure that \( R_f \subseteq R_e \) fires at all nodes, and not only at \( a \). Under global semantics, this prefix is unnecessary and we can replace each constraint \([R_p](R_f \subseteq R_e)\) simply by \( R_f \subseteq R_e \).

We formulated Theorems 14 and 15 for path constraint implication, but we could just as well formulate them for the problem of deciding whether a fact \( r_1(a, b) \) and a set of constraints \( \Gamma \) that are satisfied (at \( a \) with the pointed semantics for the prefixed version, or everywhere for the prefix-empty version) imply the existence of a pair of objects in \( r_2 \), which need not be \( a, b \) (note that in this case we don’t need constraint (23)). Our negative results also apply to other languages that can express \( \Gamma \). For instance, the prefix-empty version of \( \Gamma \) can be expressed in the following restricted class of *tuple generating dependencies*:

\[
\begin{align*}
    r_1(x, y), r_2(y, z) & \rightarrow r_3(x, z) & r_1(x, z) & \rightarrow \exists y. r_2(x, y), r_3(y, z)
\end{align*}
\]

Hence we obtain a proof of undecidability of (finite) entailment of a query \( \exists x, y, r'(x, y) \), from one single fact \( r(a, b) \) in the presence of dependencies in this restricted class.

We note that the undecidability of path constraint implication shows the undecidability of (finite) satisfiability in \( \ZO I \). In fact, it shows the undecidability of (finite) satisfiability of a complex \( \ZO I \) role \( id(\forall R, \perp) \circ (r_1 \setminus r_2) \), where \( R \) is a union of roles of the forms

\[
(r_1 \circ r_2) \setminus r_3 \quad r_1 \setminus (r_2 \circ r_3) \quad r \circ ((r_1 \circ r_2) \setminus r_3) \quad r \circ (r_1 \setminus (r_2 \circ r_3))
\]

Undecidability also applies to the (finite) entailment of \( r_2(a, b) \) from \( (id(\forall R, \perp) \circ r_1)(a, b) \). If we use a set \( \mathcal{K} \) of inclusions of the form \( \top \subseteq \forall R_e \setminus R_e, \perp \) where each \( R_e \) is the concatenation of at most two role names) to enforce prefix-empty constraints to hold everywhere, we get undecidability of testing whether the KB \( \mathcal{K} \cap \{r_1(a, b)\} \) (finitely) entails \( r_2(a, b) \).

We remark that the constraints in the proof above can also be written in \( \GXP a th_{r e g} \), the extension of \( \GXP a th_{r e g} \) that allows for negation (and hence, intersection and difference) of path formulas. Hence Theorem 14 implies the undecidability of the (finite) implication of two \( \GXP a th_{r e g} \) formulas \( \psi_1 \) and \( \psi_2 \), even when \( \psi_2 \) is restricted to a role name \( r_2 \) and \( \psi_1 \) is a formula of the form \( [-\langle \xi \rangle] \circ r_1 \) for \( r_1 \), a role name and \( \xi \) a union of roles of the forms in (24) above. Note that this result is tighter that the one in [27], whose proof heavily uses \( \varepsilon \).
Improving the Upper bound for Prefix-empty Constraints. We have seen that under global semantics, or with non-empty prefixes, there is no hope for decidability of path constraint implication. However, prefix-empty constraints under the pointed semantics are expressible in plain \( ZOI \) without using role difference, using a nominal to denote the point at which the constraint is evaluated. Hence we can reduce their (finite) implication problem to (finite) KB satisfiability in \( ZOI \). This implies decidability and an \( \text{ExpTime} \) upper bound, thus significantly improving the previous upper bound shown in [3] for the subclass of prefix-empty one-way regular path constraints.

\[ \text{Theorem 16. Let } \Gamma \text{ be a set of prefix-empty path constraints, } \varphi \text{ a prefix-empty path constraint, and } a \in N \text{. Then we can obtain a set of } ZOI \text{ of axioms } \mathcal{T}_I \text{ and a concept } C_\varphi \text{ such that } \Gamma \models \varphi \iff \{ \mathcal{T}_I \} \land \lnot C_\varphi \text{ is unsatisfiable. Moreover, } \mathcal{T}_I \text{ and } C_\varphi \text{ can be constructed in linear time and the size of } \mathcal{T}_I \text{ and } C_\varphi \text{ are both linear, in the combined sizes of } \Gamma \text{ and } \varphi. \]

Proof. For each \( \gamma = R_e \subseteq R_e \in \Gamma \cup \varphi \), let \( \mathcal{T}_I \) be the axiom \( \{ a \} \subseteq \forall R_e, \exists \text{inv}(R_e), \{ a \} \) where \( \text{inv}(R_e) \) is the inverted \( R_e \) defined as follows, where \( r \in N_R \) and \( R, R' \) denote arbitrary roles:

\[
\begin{align*}
\text{inv}(r) &= r^- \\
\text{inv}(id(C)) &= id(C) \\
\text{inv}(R^*) &= (\text{inv}(R))^* \\
\text{inv}(R \cap R') &= \text{inv}(R) \cap \text{inv}(R') \\
\text{inv}(R \cup R') &= \text{inv}(R) \cup \text{inv}(R') \\
\text{inv}(R \circ R') &= \text{inv}(R') \circ \text{inv}(R)
\end{align*}
\]

and \( \text{inv}(\varepsilon) = \varepsilon \). It is easy to see that \( (d, d') \in \Gamma \Leftrightarrow (d', d) \in \text{inv}(\Gamma) \) for every \( d, d' \in \Delta_I \).

It is then a straightforward consequence of the semantics of \( ZOI \) and of \( \gamma \), that for every pointed instance \( I, a \) we have \( I, a \models \gamma \) if and only if \( I \models \mathcal{T}_I \). Hence it easily follows that \( \Gamma \models \varphi \) if and only if \( (\bigcup_{\gamma \in \Gamma} \mathcal{T}_I) \land \lnot C_\varphi \) is unsatisfiable.

From this reduction and Theorem 5 we get:

\[ \text{Corollary 17. The unrestricted and the finite pointed implication problems for prefix-empty path constraints are decidable in } \text{ExpTime}. \]

We can combine this result and Corollary 12, obtaining that entailment is decidable in \( \text{ExpTime} \) whenever all constraints are prefix-empty or involve only simple roles. Although this is a strong restriction, it still allows for fairly non-trivial constraints. Apart from capturing these relevant decidable subclasses of path constraints, \( ZOI \) can also express many other natural constraints, some of which are not easily expressible as path constraints. For example, \( \phi_5 \) in Example 3 does not directly correspond to a path constraint. \( ZOI \) provides flexible means to express quite involved expressions, for example, that a course required in an undergraduate program must be taught by a faculty member that is a member of an institute, or a suborganization of it:

\[ \text{Course} \cap \exists \text{requires}^- \subseteq \exists \text{teaches}^- \cdot (\exists \text{memberOf o partOf}^+).\text{Inst}. \]

To conclude this section, we remark that both the pointed and the global semantics are directly supported within \( ZOI \); the logic has a global semantics, but we can use nominals to ensure that any assertion or inclusion only ‘fires’ at the interpretation of a given individual.

4 Verification of Evolving Graph Structured Data

We have advocated the use of \( ZOI \) as a constraint language for GSD. We define a language for manipulating GSD and show that it is possible to effectively reason about the preservation of \( ZOI \) constraints when data instances evolve as a result of operations in this language.
Updating Graph Structured Data. The basic operations in our manipulation language allow to insert or delete individuals from extensions of concepts, and pairs of individuals from extensions of roles. The candidates for additions and deletions are instances of complex concepts and roles in $ZOI$. The language allows, in particular, to select one object and add it to or remove it from a concept name, using a nominal \{a\}. It similarly allows to add or remove pairs of objects to/from role names using a nominal role \{(a,b)\}. Moreover, we can add to or delete from some role or concept the answers to query expressed as a complex concept or role (in the latter case, a so-called regular path query). We also allow for variables in the place of individuals, to have a more natural manipulation language with parameters that can be instantiated with different values. Finally, these basic actions can be combined into complex ones using composition and conditional actions.

The language we define next is like the one in [4], but there basic actions use concepts and roles in a different DL called $ALCHOIQbr$, instead of $ZOI$. $ALCHOIQbr$ does not allow for regular expressions as roles, hence it cannot express path queries and path constraints. On the other hand, it supports number restrictions, which are not allowed in $ZOI$. We note that the language in [4] allows roles of the form $R|_C$ (or $\text{inv}(R)|_C$), which stand for the pairs of objects in $R$ whose first (resp., second) component is an instance of $C$. Such roles are expressible in $ZOI$ using $R \circ \text{id}(C)$.

Definition 18 (Action language). In what follows, additionally to $N_i$, $N_C$ and $N_R$, we consider a countably infinite set $N_\gamma$ of variables, disjoint from the other sets. We use $ZOI_\gamma$ concepts, roles, and KBs, which are defined as for $ZOI$, but allowing for variables in the place of individuals, that is, atomic concepts take the forms $A$ and \{t\}, and atomic roles the forms $r$, $r^-$ and \{(t, t')\}, where $t, t' \in N_\gamma \cup N_i$, $A \in N_C$, $r \in N_R \setminus \{T\}$.

Basics action $\beta$ and (complex) actions $\alpha$ are defined by the following grammar:

$$
\beta \rightarrow (A \oplus C) \mid (A \odot C) \mid (r \oplus R) \mid (r \odot R) \quad \alpha \rightarrow \epsilon \mid \beta \cdot \alpha \mid (K ? \alpha[A]) \cdot \alpha
$$

with $A$ a concept name, $C$ an arbitrary $ZOI_\gamma$ concept, $r$ a role name, $R$ an arbitrary $ZOI_\gamma$ role, and $K$ an arbitrary $ZOI_\gamma$ KB. The symbol $\epsilon$ denotes the empty action.

We call a concept, role, KB or an action ground if it has no variables. A substitution is a function $\sigma$ from $N_\gamma$ to $N_i$. For a concept, role, KB or an action $\gamma$, we use $\sigma(\gamma)$ to denote the result of replacing in $\gamma$ every occurrence of a variable $x$ by the individual $\sigma(x)$. An action $\alpha'$ is called a ground instance of an action $\alpha$ if $\alpha' = \sigma(\alpha)$ for some substitution $\sigma$.

Intuitively, an application of an action $(A \odot C)$ on an instance $I$ is the addition of the content of $C^I \times A^T$. Similarly, $(A \oplus C)$ removes the content of $C^I$ from $A^T$. The two operations can also be performed on roles. Composition stands for successive action execution, and a conditional action $K ? \alpha_1[\alpha_2]$ expresses that $\alpha_1$ is executed if the instance is a model of $K$, and $\alpha_2$ is executed otherwise. If $\alpha_2 = \epsilon$ then we have an action with a simple pre-condition as in classical planning languages, and we write it as $K ? \alpha_1$, omitting $\alpha_2$.

To formally define the semantics of actions, we introduce the notion of instance update.

Definition 19 (Instance update, semantics of actions). Assume an instance $I$ and let $E$ be a concept or role name. If $E$ is a concept, let $W \subseteq \Delta^I$, otherwise, if $E$ is a role, let $W \subseteq \Delta^I \times \Delta^I$. Then, $I \oplus_E W$ (resp., $I \odot_E W$) denotes the instance $I'$ such that $\Delta^{I'} = \Delta^I \cup W$, $E^{I'} = E^I \cup W$ (resp., $E^{I'} = E^I \setminus W$), and $E_{1}^{I'} = E_{1}^{I}$, for all symbols $E_{1} \neq E$. Given a
ground action $\alpha$, we define a mapping $S_{\alpha}$ from instances to instances as follows:

$$S_{\alpha}^I = I \quad \text{if } I \models K, \quad S_{\alpha}^I = \alpha(I) \quad \text{if } I \not\models K.$$  

Note that we have not defined the semantics of actions with variables, i.e., for non-ground actions. In our approach, all variables of an action are seen as parameters whose values are given before execution with a substitution with actual individuals, i.e., by grounding.

**Example 20.** Consider the next action with a free variable $x$, and its ground instance $\alpha_{dl}$:

RemoveCourse($x$) = (Course $\circ \{x\}$) $\cdot$ (requires $\circ$ (requires $\circ$ id($\{x\}$)))$\cdot$ (offers $\circ$ (offers $\circ$ id($\{x\}$)))$\cdot$ (partOf $\circ$ (partOf $\circ$ (id($\{x\}$) $\circ$ partOf)))

In $S_{\alpha_{dl}}^I$ we have the following changes, and the rest of the instance remains unchanged:

Course$^I$ = {DataStruct:CS202, FoundDBs:CS327}
requires$^I$ = {(BSc_CSci, DataStruct:CS202), (MSc_CompLogic, FoundDBs:CS327), (MSc_CompLogic, mod_KR)}
offers$^I$ = {(CS_Dept, BSc_CSci), (CS_Dept, MSc_CompLogic), (CS_Dept, MSc_Bioinformatics)}{(CS_Dept, DataStruct:CS202), (CS_Dept, FoundDBs:CS327)}
partOf$^I$ = {}

**The Static Verification Problem.** We consider now the scenario where DL KBs are used to impose integrity constraints on GSD. A basic reasoning problem for analyzing the effects of actions in the presence of integrity constraints is static verification, which consists in checking whether the execution of an action $\alpha$ always preserves the satisfaction of integrity constraints given by a KB $K$.

**Definition 21 (The static verification problem).** Let $K$ be a KB. We say that an action $\alpha$ is $K$-preserving if for every ground instance $\alpha'$ of $\alpha$ and every finite interpretation $I$, we have that $I \models K$ implies $S_{\alpha'}^I$ $\models$ $K$. The static verification problem consists on deciding, given an action $\alpha$ and a KB $K$, whether $\alpha$ is $K$-preserving.

**Example 22.** Recall the constraints $K_{\text{UAI}}$ from Ex. 3, and the action $\alpha_{\text{dl}}$ from Ex. 20. Note that $\alpha_{\text{dl}}$ is not $K_{\text{UAI}}$-preserving. In fact, this is witnessed by our instance $I_{\text{UAI}}$. We saw that in $S_{\alpha_{\text{dl}}}^I$ we have (MSc_CompLogic, mod_KR) $\in$ requires$^I$, but mod_KR $\not\in$ ($\exists$partOf$^I$.Course)$^I$, that is, the mandatory KR module does not contain any courses, violating $\phi_4 = \top \subseteq \forall \text{requires}.(\exists$partOf$^I$$.\text{Course})$.

Our technique for static verification relies on a transformation $\text{TR}_{\alpha}(K)$ that rewrites $K$ incorporating the effects of an action $\alpha$. The technique is similar in spirit to regression in reasoning about actions [29], and it can be seen as a way to compute the weakest precondition of $\alpha$ and $K$. Intuitively, the models of $\text{TR}_{\alpha}(K)$ are exactly the interpretations $I$ such that applying $\alpha$ on $I$ leads to a model of $K$. In this way, we can effectively reduce reasoning about changes in any database that satisfies a given $K$, to reasoning about a single KB.

**Definition 23.** Given a ZOT KB $K$, we use $K_{E\leftarrow E'}$ to denote the KB that is obtained from $K$ by replacing every name $E$ by the (possibly more complex) expression $E'$. Given a
KB $\mathcal{K}$ and a ground action $\alpha$, we define $\text{TR}_\alpha(\mathcal{K})$ as follows.

$$
\begin{align*}
\text{TR}_\alpha(\mathcal{K}) &= \mathcal{K} \\
\text{TR}_{(r@R)}(\mathcal{K}) &= (\text{TR}_\alpha(\mathcal{K}))_{r\rightarrow r\cup R} \\
\text{TR}_{(A@C)}(\mathcal{K}) &= (\text{TR}_\alpha(\mathcal{K}))_{A\leftarrow A\cup C} \\
\text{TR}_{(A@C)}(\mathcal{K}) &= (\text{TR}_\alpha(\mathcal{K}))_{A\leftarrow A\cap C} \\
\end{align*}
$$

Assume a ground action $\alpha$ and a $\mathcal{ZOI}$ KB $\mathcal{K}$. Note that the transformation $\text{TR}_\alpha(\mathcal{K})$ may introduce role differences involving complex roles from $\alpha$. Hence $\text{TR}_\alpha(\mathcal{K})$ is a $\mathcal{ZOI}$ and need not be a $\mathcal{ZOI}$ KB. Note also that the size of $\text{TR}_\alpha(\mathcal{K})$ might be exponential in the size of $\alpha$. By employing the same argument as [4], we see that the transformation correctly captures the effects of complex actions. In particular, for every interpretation $I$, we have $S_\alpha(I) \models \mathcal{K}$ iff $I = \text{TR}_\alpha(\mathcal{K})$. With the transformation $\text{TR}_\alpha(\mathcal{K})$ above we have a reduction from static verification to finite (un)satisfiability of $\mathcal{ZOI}$ KBs: an action $\alpha$ is not $\mathcal{K}$-preserving iff some finite model of $\mathcal{K}$ does not satisfy $\text{TR}_\alpha(\mathcal{K})$, where $\alpha^*$ is a ‘canonical’ grounding of $\alpha$. Formally, we have:

**Theorem 24 ([4]).** For a (complex) action $\alpha$ and a KB $\mathcal{K}$, the following are equivalent:

(i) $\text{The action } \alpha \text{ is not } \mathcal{K}\text{-preserving.}$

(ii) $\mathcal{K} \land \neg \text{TR}_\alpha(\mathcal{K})$ is finitely satisfiable, where $\alpha^*$ is obtained from $\alpha$ by replacing each variable with a fresh individual name not occurring in $\alpha$ and $\mathcal{K}$.

**Undecidability of unrestricted static verification.** The first and foremost consequence of this reduction is that for the action language we have defined, the static verification problem is undecidable, even if the input $\mathcal{K}$ is trivial, and the actions are quite restricted:

**Theorem 25.** Deciding whether $\alpha$ is $\mathcal{K}$-preserving is undecidable, even when $\mathcal{K}$ is a trivial KB of the form $r(a, b)$, and $\alpha$ is just a sequence of basic actions of the forms $(r \oplus R)$ and $(r \odot R)$, with $R$ a sequence of one or two concatenated role names.

**Proof.** We provide a reduction from deciding $\Gamma \models_{\text{fin}} r_1 \subseteq r_2$, where $\Gamma$ contains only constraints of the forms $r_1 \circ r_2 \subseteq r_3$ and $r_1 \subseteq r_2 \circ r_3$ for $\{r_1, r_2, r_3\} \subseteq N_R$. We have seen above that this problem is undecidable. In particular, we construct an action $\alpha$ such that $\Gamma \models_{\text{fin}} r_1 \subseteq r_2$ iff $\alpha$ is $r_1(a, b)$-preserving. Let $R_1^1 \subseteq R_1^2, \ldots, R_2^1 \subseteq R_2^2$ be an enumeration of all constraints in $\Gamma$. For every $1 \leq i \leq n$, let $p_i^1$ and $p_i^2$ be fresh role names. Then $\alpha$ is the concatenation of the following actions in the given order:

- $(r_1 \odot r_1) \cdot (r_1 \odot r_2)$
- $(p_1^1 \oplus p_1^1) \cdot \cdots \cdot (p_i^1 \oplus p_i^1) \cdot (p_i^2 \oplus p_i^2) \cdot \cdots \cdot (p_n^2 \oplus p_n^2)$
- $(p_1^1 \oplus R_1^1) \cdot \cdots \cdot (p_i^1 \oplus R_i^1) \cdot (p_i^2 \oplus R_i^2) \cdot \cdots \cdot (p_n^2 \oplus R_n^2)$
- $(p_1^1 \oplus p_1^1) \cdot \cdots \cdot (p_i^1 \oplus p_i^2) \cdot (r_1 \odot p_1^1) \cdot \cdots \cdot (r_1 \odot p_n^1)$.

Recall that $\alpha$ is not $r_1(a, b)$-preserving iff $r_1(a, b) \land \neg \text{TR}_\alpha(r_1(a, b))$ is finitely satisfiable. It’s not hard to see that $\text{TR}_\alpha(r_1(a, b)) = R_F(a, b)$, where $R_F$ is equivalent to the role $r_2 \cup (R_1^1 \setminus R_2^1) \cup \cdots \cup (R_n^1 \setminus R_n^2)$. Thus $r_1(a, b) \land \neg \text{TR}_\alpha(r_1(a, b))$ is finitely satisfiable iff $\Gamma \not\models_{\text{fin}} r_1 \subseteq r_2$. 

Unfortunately we cannot allow for complex roles in our actions, not even of the form $r \circ r$, but we get positive results if we restrict actions to simple roles.

An undesired effect of disallowing complex roles in actions is that we cannot express $r|_C$ as $r \circ id(C)$, and as our examples illustrate, this construct is quite useful. For nominals we can, however, simulate $r|_{\{a\}}$ in $\mathcal{ZOI}$. We use a special role name $T|_{\{a\}}$ with the intended
semantics \((T|_{\{a\}})^T = \Delta^T \times \{a\}^T\), and write the simple role \(r \cap T|_{\{a\}}\) in the place of \(r|_{\{a\}}\). The intended meaning of \(T|_{\{a\}}\) is easily enforced by adding \(T \subseteq \exists T|_{\{a\}}.a\) to any \(ZOI\) KB.

We call an action \(\alpha\) role-restricted if in every basic action of the form \((r \oplus R)\) or \((r \ominus R)\), we have that \(R\) is a simple \(ZOI\) role that may use the special role names \(T|_{\{a\}}\).

Note \(\alpha_{dl}\) in Example 20 can be rewritten as a role-restricted action as follows:

\[
\alpha'_{dl} = \left(\text{Course} \ominus \{\text{DLs:CS451}\}\right) \cdot \left(\text{requires} \ominus \left(\text{requires} \cap T|_{\{\text{DLs:CS451}\}}\right)\right) \cdot \left(\text{offers} \ominus \left(\text{offers} \cap T|_{\{\text{DLs:CS451}\}}\right)\right) \cdot \left(\text{partOf} \ominus \left(\text{partOf} \cap (T|_{\{\text{DLs:CS451}\}})\right)\right).
\]

▶ Theorem 26. Deciding whether \(\alpha\) is \(K\)-preserving for a given \(ZOI\) KB \(K\) and a role-restricted \(\alpha\) is \(\text{ExpTime}\)-complete.

Proof sketch. Since the union and the difference of simple roles are simple roles, it is not hard to see that the result of iteratively replacing role names by simple roles involving union and difference in a \(ZOI\) role results in a \(ZOI\) role. Hence, for any \(ZOI\) KB \(K\) and a role-restricted action \(\alpha\), the KB \(T_R(\alpha)(K)\) is not only a \(ZOI\) KB but also a \(ZOI\) KB (i.e., difference is applied to simple roles only). Then from the decidability of (finite) satisfiability of \(ZOI\) it follows that checking whether \(\alpha\) is \(K\)-preserving is decidable. For the complexity upper bound, recall that the size of \(T_R(\alpha)(K)\) might be exponential in the size of \(\alpha\). However, as argued in [4], there are only exponentially many conjunctive clauses in disjunctive normal form of \(K \land \lnot T_R^*(\alpha)(K)\), each with size polynomial in the size of \(\alpha\) and \(K\). Thus from Theorems 5 and 24 we obtain the desired result.

5 Conclusions and Outlook

The main goal of this work was to advocate the use of the DL \(ZOI\) to specify constraints over graph structured data and to show the decidability of static verification in a rich language for manipulating such data. Along the way, we have shown several undecidability and complexity results that concern not only our setting, but also formalisms that were introduced in the 90s, as well as recently introduced query languages for GSD like GXPath\(_{reg}\). In our future work we aim at providing some support for identification constraints, which is clearly desirable but naturally requires equality reasoning. This is challenging, as e.g., the decidability of the extension of \(ZOI\) where some roles must be interpreted as partial functions is a long standing open problem.

Acknowledgements. This work was supported by the EU IP project Optique (Scalable End-user Access to Big Data), grant agreement n. FP7-318338, the Vienna Science and Technology Fund project ICT12-15, and the Austrian Science Fund projects P25207 and T515.

References

Verification of Evolving GSD under Expressive Path Constraints


