Run-Based Semantics for RPQs

Claire David, **Nadime Francis**, **Victor Marsault**
Université Gustave Eiffel, CNRS, LIGM, France
{claire.david, nadime.francis, victor.marsault}@univ-eiffel.fr

Abstract

RPQs (regular path queries) are an important building block of most query languages for graph databases. They are generally evaluated under homomorphism semantics; in particular only the endpoints of the matched walks are returned.

However, practical applications often need the full matched walks to compute aggregate values. In those cases, homomorphism semantics are not suitable since the number of matched walks can be infinite. Hence, graph-database engines adapt the semantics of RPQs, often neglecting theoretical red flags. For instance, the popular query language Cypher uses trail semantics, which ensures the result to be finite at the cost of making computational problems intractable.

We propose a new kind of semantics for RPQs, including in particular simple-run and binding-trail semantics, as a candidate to reconcile theoretical considerations with practical aspirations. Both ensure the output to be finite in a way that is compatible with homomorphism semantics: projection on endpoints coincides with homomorphism semantics. Hence, testing the emptiness of result is tractable, and known methods readily apply. Moreover, simple-run and binding-trail semantics support bag semantics, and enumeration of the bag of results is tractable.

1 Introduction

When querying data graphs, users are not only interested in retrieving data, but also in *how* these pieces of data relate to each other. This is why most languages for querying data graphs, both in theory and in practice, are *navigational* languages. Informally, the querying process starts at some vertex and then *walks* through the graph: it follows edges from vertex to vertex, retrieving and testing data along the way, until the walk ends in some final vertex.

In database theory, this process is usually abstracted as Regular Path Queries (RPQs, Cruz, Mendelzon, and Wood 1987). An RPQ is defined by a regular expression R and is traditionally evaluated under *walk semantics* (also known as *homomorphism semantics*, Angles et al. 2017). In that case, it returns all pairs of vertices in the graph that are linked by a walk whose label conforms to R. This formalism enjoys many nice properties and has become an important building block of most query languages over graph databases.

However, RPQs do not entirely meet the needs of real-life graph database systems. Indeed, limiting the output of the query to the endpoints of the walk is not enough for many real-life applications, which might also require the number of matching walks (e.g. to rank answers or evaluate connectivity), or even the walks themselves (e.g. for route planning) (Robinson, Webber, and Eifrem 2015). Under walk semantics, the *space of matches* is infinite: there are infinitely many matching walks when the graph contains cycles, which renders these questions meaningless. Most graph database management systems have their own way of addressing this issue, with none of them being entirely satisfactory. We briefly describe the most common approaches below, as well as their shortcomings; we use the database given in Figure 1 and the following queries to illustrate them.

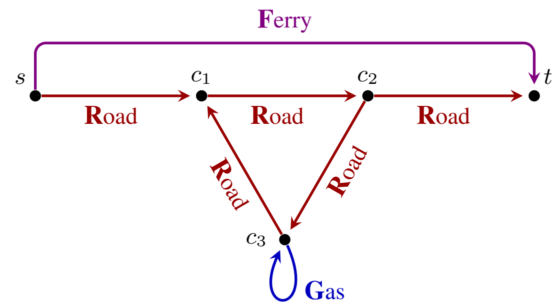


Figure 1: Our running graph database D

$$Q_1 = (\mathbf{R}\text{oad} + \mathbf{F}\text{erry})^*$$
$$Q_2 = (\mathbf{R}\text{oad} + \mathbf{F}\text{erry})^* \ \mathbf{G}\text{as} \ (\mathbf{R}\text{oad} + \mathbf{F}\text{erry})^*$$

Topological restriction This solution roots out unboundedness by forbidding cycles. Walks are only returned if no vertex (*simple-walk semantics*) or no edge (*trail semantics*) is visited twice. For instance, the language Cypher uses trail semantics (Francis et al. 2018). Moreover, the new query language GQL[1](Deutsch et al. 2022; Francis et al. 2023), currently designed by ISO[2] to be the first standard language for property graphs, implements several topological restrictions.

The output walks are in some sense representative of the possibilities in the space of matches. For instance, Q_1

[1]https://www.iso.org/standard/76120.html
[2]ISO stands for International Standards Organisation.

returns $s \to t$ and $s \to c_1 \to c_2 \to t$ under trail semantics, two unrelated possibilities in the space of matches. This is a crucial feature in real systems, in which pattern matching is usually just a first step before further processing. For instance, one could evaluate the connectivity between s and t by counting the number of walks from s to t matching Q_1.

The main weakness of this approach is that computational problems are intractable even for very simple queries. For instance, deciding whether two vertices are linked by a walk conforming to Q_2 is NP-complete in the size of the database under both trail (Martens, Niewerth, and Trautner 2020) and simple-walk semantics (Bagan, Bonifati, and Groz 2020). These semantics are also error-prone in that desirable results might be discarded unintentionally. For instance, Q_2 returns no walk from s to t under trail semantics; and in a bigger, more realistic, graph database the walk $s \to c_1 \to c_2 \to c_3 \to c_1 \to c_2 \to t$ would not be considered for further processing. These kinds of unwanted behaviours happen beyond theoretical settings: e.g., in (Robinson, Webber, and Eifrem 2015, p.132), the authors propose a query to solve a real-life scenario; the query does not work as intended due to trail semantics.

Witness selection Another approach consists in choosing a metric (length, cost, etc.), and then selecting only a few best-ranking walks in the space of matches. For instance, the semantics of GSQL (TigerGraph Team 2021; Deutsch et al. 2019) and G-Core (Angles et al. 2018) only return the shortest walks matching the query; GQL allows returning the k shortest walks (Deutsch et al. 2022). However, the length of the path is an arbitrary metric that may not fit every application: here, Q_1 would return the ferry route $v = s \to t$ over the road route $w = s \to c_1 \to c_2 \to t$ but it does not necessarily mean that v represents a faster or shorter route than w in reality. To circumvent this issue, GQL mentions other metrics, such as k-cheapest, as possible extensions to investigate. On the other hand, witness selection generally makes counting and aggregating meaningless: it counts or aggregates over something that is not representative of the space of matches.

Reducing expressivity Some systems disallow queries or operations that may lead to infinite outputs or ill-defined behaviours. Kleene stars in GQL queries are only allowed if they appear under some form of topological restriction or witness selection. In SPARQL[3], counting the number of walks matched by a property path is only allowed when the underlying regular language is finite. Otherwise, the returned number collapses to 0 (no walk matches the query) or 1 (at least one walk matches the query). Similarly, SPARQL equivalents of queries Q_1 or Q_2 only return the endpoints of matched walks. Note also that switching silently from bag to set semantics depending on the query is error-prone.

In this article, we propose another approach called *run-based*. We present two run-based semantics: *simple-run semantics*, whose input query is given as a finite automaton and provide sound theoretical foundations; and *binding-trail semantics* which operate directly on a regular expression in order to be closer to practical use. Akin to topological restriction, we aim at producing a finite output that faithfully represents the space of matches, and we do so by discarding cyclic results. Intuitively, run-based semantics discard a result only if a cycle in the walk coincides with a cycle in the computation of the query. For instance, binding-trail semantics filter out walks in which one edge is matched twice to the same atom of the regular expression. Indeed, the walk $w = s \to c_1 \to c_2 \to c_3 \to c_1 \to c_2 \to t$ is **not** in the output of Q_1: the edge $c_1 \to c_2$ is matched twice to the same Road atom. On the other hand, w is kept in the output of Q_2, because the two occurrences of the edge $c_1 \to c_2$ are matched to two different Road atoms. In general, the output under run-based semantics depends on the *syntax* of the query. This seeming drawback also provides a finer control of the output; see Remarks 14 and 30.

The paper is organised as follows. Section 2 covers necessary preliminaries and Section 3 gives the definition of simple-run semantics. In Section 4, we revisit classical computational problems and show that simple-run semantics enjoy efficient PTIME or polynomial-delay algorithms for emptiness, tuple membership and walk enumeration. Counting answers remains #P-Complete. Section 5 defines binding-trail semantics as an adaptation of simple-run semantics to queries given as regular expressions. As a side result, we show that any regular expression (in fact, its Glushkov automaton) may encode the same behaviour and topology as any arbitrary automaton, which means that all complexity lower and upper bounds translate from one setting to the other. Finally, we conclude this document in Section 6 by discussing possible extensions of our semantics.

2 Preliminaries

2.1 Graph Databases

In this document, we model graph databases as directed, multi-labeled, multi-edge graphs, and simply refer to them as *databases* for short. We will use the database shown in Figure 1, page 1, as our running example. Databases are formally defined as follows.

Definition 1. *A (graph)* database D *is a tuple* $(\Sigma, V, E, \textsc{Src}, \textsc{Tgt}, \textsc{Lbl})$ *where:* Σ *is a finite set of symbols, or* labels*;* V *is a finite set of* vertices*;* E *is a finite set of* edges*;* $\textsc{Src} : E \to V$ *is the* source *function;* $\textsc{Tgt} : E \to V$ *is the* target *function; and* $\textsc{Lbl} : E \to 2^{\Sigma}$ *is the* labelling *function.*

Definition 2. *A (directed)* walk w *in* D *is a non-empty finite sequence of alternating vertices and edges of the form* $w = (n_0, e_0, n_1, \dots, e_{k-1}, n_k)$ *where* $k \geq 0$, $n_0, \dots, n_k \in V$, $e_0, \dots, e_{k-1} \in E$, *such that:*

$$\forall i, 0 \leq i < k, \quad \textsc{Src}(e_i) = n_i \quad \text{and} \quad \textsc{Tgt}(e_i) = n_{i+1}$$

For ease of notation, we use $\to$ *to avoid naming the edge that connects two nodes when it is unique, as in* $w = n_0 \to n_1 \to \cdots \to n_k$.

[3]https://www.w3.org/TR/sparql11-query/#propertypaths

We call k the length *of w and denote it by* $\textsc{Len}(w)$. *We extend the functions* $\textsc{Src}$, $\textsc{Tgt}$ *and* $\textsc{Lbl}$ *to the walks in D as follows. For each walk $w = (n_0, e_0, n_1, \ldots, e_{k-1}, n_k)$ in D,* $\textsc{Src}(w) = n_0$, $\textsc{Tgt}(w) = n_k$, *and*

$$\textsc{Endpoints}(w) = \big(\textsc{Src}(w), \textsc{Tgt}(w)\big)$$

$$\textsc{Lbl}(w) = \big\{\, u_0 u_1 \cdots u_{k-1} \;\big|\; \forall i, 0 \le i < k,\; u_i \in \textsc{Lbl}(e_i) \,\big\}\ .$$

Finally, $s \xrightarrow{w} t$ means that $\textsc{Endpoints}(w) = (s,t)$ *and, for a word $u \in \Sigma^*$, we write $s \xrightarrow{u} t$ if there exists a walk w in D such that $s \xrightarrow{w} t$ and $u \in \textsc{Lbl}(w)$.*

We say that two walks w, w' concatenate *if* $\textsc{Tgt}(w) = \textsc{Src}(w')$, *in which case we define their* concatenation *as usual, and denote it by $w \cdot w'$, or simply ww' for short.*

Definition 3. *A* trail *is a walk with no repeated edge. A* simple walk *is a walk with no repeated vertex. We let* $\textsc{Trail}$ *(resp.* $\textsc{Simple}$*) denote the bag-to-bag function that takes as input a bag of walks B and returns the bag of the trails (resp. simple walks) in B.*

2.2 Regular Path Queries, Automata, Expressions

An RPQ is defined by a regular language (given as either an automaton or a regular expression). RPQs may be evaluated under various semantics. Several classical semantics, along with the novel *run-based semantics*, are defined in Section 3.

A (nondeterministic) automaton is a 5-tuple $\mathcal{A} = \langle \Sigma, Q, \Delta, I, F \rangle$ where Σ is a finite set of symbols, Q is a finite set of *states*, $I \subseteq Q$ is called the set of *initial* states, $\Delta \subseteq Q \times \Sigma \times Q$ is the set of *transitions* and $F \subseteq Q$ is the set of *final* states. As usual, we extend Δ into a relation over $Q \times \Sigma^* \times Q$ as follows: for every $q \in Q$, $(q, \varepsilon, q) \in \Delta$; and for every $q, q', q'' \in Q$ and every $u, v \in \Sigma^*$, if $(q, u, q') \in \Delta$ and $(q', u, q'') \in \Delta$ then $(q, uv, q'') \in \Delta$. We denote by $\mathrm{L}(\mathcal{A})$ the *language* of $\mathcal{A}$, defined as follows.

$$\mathrm{L}(\mathcal{A}) = \big\{\, u \in \Sigma^* \;\big|\; \exists i \in I,\ \exists f \in F,\ (i, u, f) \in \Delta \,\big\} \quad (1)$$

A *computation* in $\mathcal{A}$ is an alternating sequence of states and transitions that is defined similarly to walks in databases. We extend $\textsc{Src}$, $\textsc{Lbl}$, $\textsc{Tgt}$ and $\textsc{Endpoints}$ over computations. A computation is *successful* if it starts in an initial state and ends in a final state.

A regular expression R over an alphabet Σ is a formula obtained inductively from the letters in Σ, one unary function *, and two binary functions $+$ and $\cdot$, according to the following grammar.

$$R ::= \varepsilon \mid a \mid R^* \mid R \cdot R \mid R + R \qquad \text{where } a \in \Sigma \quad (2)$$

We usually omit the $\cdot$ operator and we let $L(R)$ denote the subset of Σ^* described by R.

3 Run-Based Query Evaluation

3.1 Run Database

In Section 3.1, we fix an automaton $\mathcal{A} = \langle \Sigma, Q, \Delta, I, F \rangle$ and a graph database $D = (\Sigma, V, E, \textsc{Src}, \textsc{Tgt}, \textsc{Lbl})$.

Definition 4. *The* run database $D \times \mathcal{A}$ *is the database* $D \times \mathcal{A} = (\Sigma, V', E', \textsc{Src}', \textsc{Tgt}', \textsc{Lbl}')$ *where*

$$V' = V \times Q$$

$$E' = \{\, (e, (q, a, q')) \in E \times \Delta \mid a \in \textsc{Lbl}(e) \,\}$$

and, for each $e' = (e, (q, a, q')) \in E'$,

$$\textsc{Src}'(e') = \big(\textsc{Src}(e), q\big) \quad \textsc{Tgt}'(e') = \big(\textsc{Tgt}(e), q'\big)$$

$$\textsc{Lbl}'(e') = \{\, a \,\} \quad .$$

We denote the projection from $D \times \mathcal{A}$ to D by π_D: for each $(n, q) \in V'$, $\pi_D((n, q)) = n$; for each $(e, t) \in E'$, $\pi_D((e, t)) = e$; and for each walk $w = (n_0, e_0, \ldots, n_k)$, $\pi_D(w) = (\pi_D(n_0), \pi_D(e_0), \ldots, \pi_D(n_k))$.

The run database is essentially a product of the automaton with the database. See Figure 2 for an example. In the figure, elements that do not contribute to any run are dashed.

Definition 5. *A walk w in $D \times \mathcal{A}$ is called* a run *if* $\textsc{Src}(w) \in V \times I$ *and* $\textsc{Tgt}(w) \in V \times F$. *We let* $\textsc{Match}_{\mathcal{A}}(D)$ *denote the bag*[4] *of all runs in $D \times \mathcal{A}$.*

A simple verification yields the following property.

Property 6. *For every walk w in D, there exists a run r in $D \times \mathcal{A}$ such that $\pi_D(r) = w$ if and only if* $\textsc{Lbl}(w) \cap \mathrm{L}(\mathcal{A}) \neq \emptyset$.

Remark 7. *Note that the run database $D \times \mathcal{A}$ depends on the structure of the automaton $\mathcal{A}$, and not only on $\mathrm{L}(\mathcal{A})$. Hence, when considering run databases, we cannot assume that $\mathcal{A}$ is deterministic, minimal, or of any particular shape.*

The run database allows rephrasing the most common semantics, as in Definitions 8, 9 and 10.

Definition 8. *Under* walk semantics, *RPQs return all walks of the input database whose label conforms to the query. It is defined as* $[\![\mathcal{A}]\!]_W(D) = \pi_D \circ \textsc{Match}_{\mathcal{A}}(D)$.

We will sometimes refer to the bag $[\![\mathcal{A}]\!]_W(D)$ as the *space of matches*, as it contains all walks that intuitively match the query. Note that it can be infinite, and thus cannot be returned as is. The following two semantics circumvent this issues by restricting $[\![\mathcal{A}]\!]_W(D)$ to a finite bag.

Definition 9. Trail semantics *return only the trails matching the query :* $[\![\mathcal{A}]\!]_T(D) = \textsc{Trail} \circ \pi_D \circ \textsc{Match}_{\mathcal{A}}(D)$.

Definition 10. Simple-walk semantics *return only the matching walks that are simple:* $[\![\mathcal{A}]\!]_{SW}(D) = \textsc{Simple} \circ \pi_D \circ \textsc{Match}_{\mathcal{A}}(D)$.

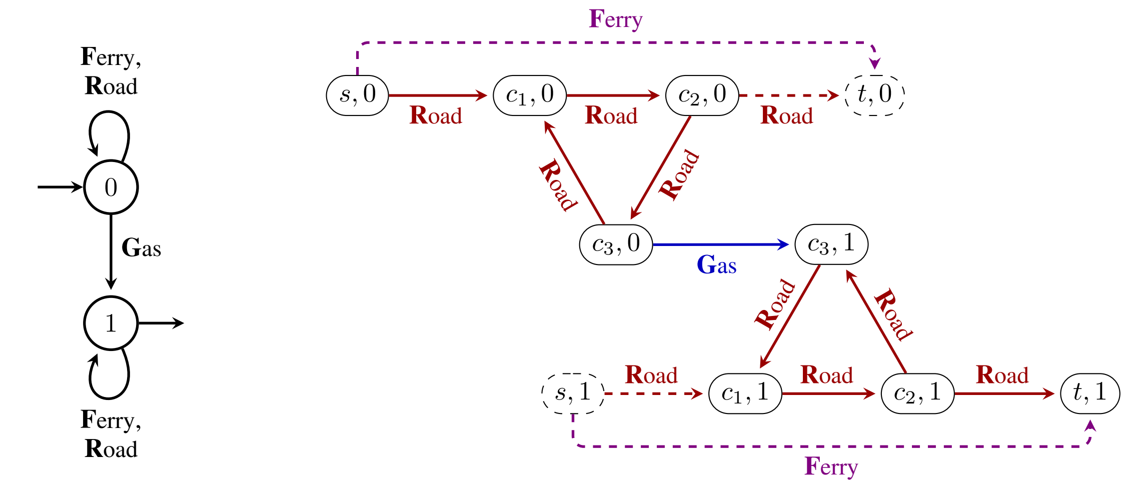
3.2 Simple-Run Semantics

In line with simple-walk and trail semantics, *simple-run semantics* keeps the output finite by filtering out *redundant* results. The difference amounts to the definition of *redundant*. Classical semantics filter based on redundancy in the computed walk (repeated edge, repeated vertex), hence filtering is done *after* projecting the runs to D. In the semantics we propose here, filtering is based on redundancy in the run, hence filtering is done *before* projecting to D.

Definition 11. *The* simple-run semantics *of an automaton $\mathcal{A}$, denoted by $[\![\mathcal{A}]\!]_{SR}$, is the mapping that associates, to each database D, the following bag of answers.*

$$[\![\mathcal{A}]\!]_{SR}(D) = \pi_D \circ \textsc{Simple} \circ \textsc{Match}_{\mathcal{A}}(D) \quad (3)$$

[4] Although the multiplicity of each element in $\textsc{Match}_{\mathcal{A}}(D)$ is one, we would rather not use the term *set* to avoid confusion when we apply bag-to-bag functions later on.



(a) An automaton $\mathcal{A}$ (b) The run database $D \times \mathcal{A}$

Figure 2: A run database constructed from D (Figure 1) and $\mathcal{A}$ (Figure 2a).

Example 12. *A run in the database from Figure 2b is a walk that goes from the top part to the bottom part. For instance, the walk* $r_1 = (s,0) \to (c_1,0) \to (c_2,0) \to (c_3,0) \to (c_3,1) \to (c_1,1) \to (c_2,1) \to (t,1)$ *is a run, which moreover is simple. Hence its projection* $w_1 = \pi_D(r_1) = s \to c_1 \to c_2 \to c_3 \to c_3 \to c_1 \to c_2 \to t$ *belongs to* $[\![\mathcal{A}]\!]_{SR}(D)$*. On the other hand,* w_1 *is neither a trail nor a simple walk, hence* $w_1 \notin [\![\mathcal{A}]\!]_T(D)$ *and* $w_1 \notin [\![\mathcal{A}]\!]_{SW}(D)$*. In fact,* $[\![\mathcal{A}]\!]_T(D)$ *and* $[\![\mathcal{A}]\!]_{SW}(D)$ *contain no walk going from* s *to* t*.*

One of the main features of simple-run semantics is that it covers the space of matches, in a precise way (Lemma 13). Essentially, if a walk w matching the query is **not** returned, at least one *subwalk* w' of w is returned; moreover, w' is obtained from w by removing superfluous cycles. Note that semantics based on topological restriction do not enjoy the same property, as shown in Example 12.

Lemma 13. *Let* D *be a database,* $\mathcal{A}$ *be an automaton, and* w *be a walk in* $\text{MATCH}_{\mathcal{A}}(D)$*. Then, there exists a decomposition of* w *as* $w = u_1 v_1 u_2 \cdots v_n u_{n+1}$ *such that every* u_i *satisfies* $\text{SRC}(u_i) = \text{TGT}(u_i)$*, and* $v_1 \cdots v_n \in [\![\mathcal{A}]\!]_{SR}(D)$*.*

Proof. By induction on the length of w. The statement obviously holds if w is a single vertex since a walk of length 0 is always simple.

Let $w \in \text{MATCH}_{\mathcal{A}}(D)$. Let r be a run in $D \times \mathcal{A}$ such that $\pi_D(r) = w$. If $w \in [\![\mathcal{A}]\!]_{SR}(D)$, there is nothing to prove: fix $n = 1$, u_1, u_2 as single vertices and $v_1 = w$. Otherwise, it means that r is not simple, that is there is a decomposition of r as $r = r_1 r_2 r_3$ such that $\text{TGT}(r_1) = \text{SRC}(r_2) = \text{TGT}(r_2) = \text{SRC}(r_3)$ and $\text{LEN}(r_2) \neq 0$. Hence, $r_1 r_3$ is a run in $D \times \mathcal{A}$ and the walk $w' = \pi_D(r_1 r_3)$ belongs to $\text{MATCH}_{\mathcal{A}}(D)$. Then, we conclude by induction on w' and reconstruct the decomposition of w. □

Remark 14. *Recall that the run database depends on the automaton itself (Remark 7). This dependence carries over to simple-run semantics:* $[\![\mathcal{A}]\!]_{SR}(D)$ *and* $[\![\mathcal{B}]\!]_{SR}(D)$ *might be different even if* $\text{L}(\mathcal{A}) = \text{L}(\mathcal{B})$*. Choosing* $\mathcal{A}$ *or* $\mathcal{B}$ *governs which representatives of the space of matches are returned, in the sense of Lemma 13.*

Remark 15. *Akin to simple-run semantics, one could define* trail-run semantics *that would return the* trails *of the run database. While trail-run semantics would generally enjoy the same properties as simple-run semantics, the meaning of a trail in the run database is much harder to grasp. Indeed, transitions of the automaton usually have no intrinsic meaning, whereas states encode the content of the memory.*

4 Computational Problems

In Section 4, we restate common computational problems related to query answering. We recall known results for the usual semantics, and give both lower and upper complexity bounds for simple-run semantics.

4.1 Existence of a Matching Walk

The problem TUPLE MEMBERSHIP consists in deciding whether there is a walk matching the query between two given endpoints. Under walk semantics, this problem corresponds to what is called *homomorphism semantics* in most theoretical contexts, hence it is unsurprisingly tractable in that case (Theorem 16). On the other hand, TUPLE MEMBERSHIP is intractable under trail or simple-walk semantics (Theorem 17). We show in Theorem 18 that it is tractable under simple-run semantics.

TUPLE MEMBERSHIP UNDER X SEMANTICS
• Data: A database D, and a pair (s,t) of vertices in D.
• Query: An automaton $\mathcal{A}$.
• Question: Does there exist a walk $w \in [\![\mathcal{A}]\!]_X(D)$ such that $\text{ENDPOINTS}(w) = (s,t)$?

Theorem 16 (Mendelzon and Wood 1995)**.** TUPLE MEMBERSHIP *is NL-complete under walk semantics.*

Theorem 17 (Martens, Niewerth, and Trautner 2020; Bagan, Bonifati, and Groz 2020)**.** TUPLE MEMBERSHIP *is NP-complete under trail or simple-walk semantics. It is already NP-hard for a fixed query in both cases.*

The typical query for which TUPLE MEMBERSHIP is hard under trail semantics is a^*ba^*. Indeed, one has to record which edges are matched by the left a^*, in order not to be matched by the right a^*. Under simple-run semantics it is not necessary to keep that record, which makes TUPLE MEMBERSHIP tractable as stated by Theorem 18.

Theorem 18. TUPLE MEMBERSHIP *is NL-complete under simple-run semantics.*

Theorem 18 is a corollary of Proposition 19 below, which is itself a direct consequence of Lemma 13.

Proposition 19. *Let D be a database, $\mathcal{A}$ be an automaton, and s, t be two vertices in D. We let $P_{s,t}$ denote the set $P_{s,t} = \{ w \in [\![\mathcal{A}]\!]_W(D) \mid \text{ENDPOINTS}(w) = (s,t) \}$. Each walk with minimal length in $P_{s,t}$ belongs to $[\![\mathcal{A}]\!]_{SR}(D)$.*

Proposition 19 implies that simple-run semantics and walk semantics are equivalent for TUPLE MEMBERSHIP. Hence known techniques for computing TUPLE MEMBERSHIP efficiently under walk semantics readily apply to simple-run semantics; and shortest-walk algorithms can be used to produce witnesses for TUPLE MEMBERSHIP.

4.2 Enumeration of Matching Walks

The problem QUERY EVALUATION consists in enumerating the walks returned by the query. It is perhaps the most important computational problem regarding query answering since it is close to what database engines do in practice. QUERY EVALUATION is ill-defined under walk semantics since $[\![\mathcal{A}]\!]_W(D)$ might be infinite[5]. Under trail or simple-walk semantics, QUERY EVALUATION is well-defined but it is intractable (Theorem 20). By using Yen's algorithm, we show that it is tractable under simple-run semantics.

QUERY EVALUATION UNDER X SEMANTICS
• Data: A database D.
• Query: An automaton $\mathcal{A}$.
• Output: All walks in $[\![\mathcal{A}]\!]_X(D)$.

Theorem 20. *Unless P = NP,* QUERY EVALUATION *under trail or simple-walk semantics cannot be enumerated with polynomial-time preprocessing.*

Theorem 20 follows easily from Theorem 17.

Theorem 21. QUERY EVALUATION *under simple-run semantics can be enumerated with polynomial delay and preprocessing.*

Sketch of proof. Computing $[\![\mathcal{A}]\!]_{SR}(D)$ amounts to computing all simple walks from (s,i) to (t,f) in the run database $D \times \mathcal{A}$, for each vertices s and t of D and each initial and final states i and f. This can be done for each (s,i) and (t,f) by using classical algorithms for simple-walk enumeration, such as Yen's algorithm (see Yen 1971; or Martens, Niewerth, and Trautner 2020 for a modern statement). ☐

[5]It might be of interest to define (non-terminating) enumeration procedure for this infinite bag. It is beyond the scope of this paper.

QUERY EVALUATION enumerates the walks in $[\![\mathcal{A}]\!]_X(D)$, which is a bag. Hence a walk in $[\![\mathcal{A}]\!]_X(D)$ with multiplicity m will be output m times. We call DEDUPLICATED QUERY EVALUATION the problem that enumerates the **distinct** matching walks.

DEDUPLICATED QUERY EVAL. UNDER X SEM.
• Data: A database D.
• Query: An automaton $\mathcal{A}$.
• Output: All walks in $[\![\mathcal{A}]\!]_X(D)$, without duplicates.

Note that Theorem 20 also holds for DEDUPLICATED QUERY EVALUATION for similar reasons. We leave its complexity under simple-run semantics as an open problem.

4.3 Counting Matching Walks

Counting the number of matching walks between two vertices, or TUPLE MULTIPLICITY, is also used in practice, for instance to evaluate the connectivity between two vertices. TUPLE MULTIPLICITY behaves differently under walk semantics, as some tuples might have infinite multiplicity. Under the variants based on witness selection (e.g. shortest walk semantics, as explained in the introduction), the problem takes a different meaning and no longer reflects the level of connectivity between vertices. Under trail or simple-walk semantics, this problem is known to be intractable; the same technique shows that it is also intractable under simple-run semantics (Theorem 22).

TUPLE MULTIPLICITY UNDER X SEMANTICS
• Data: A database D, and a pair (s,t) of vertices in D.
• Query: An automaton $\mathcal{A}$.
• Output: The total multiplicity of all walks $w \in [\![\mathcal{A}]\!]_X(D)$ such that $\text{ENDPOINTS}(w) = (s,t)$.

Theorem 22. TUPLE MULTIPLICITY *is #P-complete under trail, simple-walk and simple-run semantics. It is already #P-hard in data complexity: there exists a fixed automaton $\mathcal{A}$ for which the problem is #P-hard.*

In all cases, the upper bound comes from counting the successful computations of a nondeterministic polynomial-time machine that guesses a trail (resp. simple walk, resp. simple run) going from s to t and checks that it is accepted by $\mathcal{A}$. The hardness proof consists in a reduction from counting trails (or simple walks) in unlabelled graphs, two problems known to be #P-complete (Valiant 1979). One simply has to fix $\mathcal{A}$ as the one-state automaton accepting a^*. For simple-run semantics, one also has to note that for that particular $\mathcal{A}$ it holds $[\![\mathcal{A}]\!]_{SR}(D) = [\![\mathcal{A}]\!]_{SW}(D)$.

4.4 Walk Membership

The last problem we consider here is WALK MEMBERSHIP, which consists in deciding whether a given walk is returned.

This problem is usually considered whenever TUPLE MEMBERSHIP and QUERY EVALUATION are intractable; and as a matter of fact, it is known to be tractable for all usual semantics (Theorem 23). Surprisingly, it is intractable under simple-run semantics (Theorem 24).

WALK MEMBERSHIP UNDER X SEMANTICS
• Data: A database D and a walk w.
• Query: An automaton $\mathcal{A}$.
• Question: $w \in [\![\mathcal{A}]\!]_X(D)$?

Theorem 23. WALK MEMBERSHIP *is NL-complete under walk, trail or simple-walk semantics.*

Sketch of proof. For walk semantics, the problem amounts to checking acceptance of a word in a nondeterministic finite automaton. For trail (resp. simple-walk) semantics, one has to additionally check that the input walk is a trail (resp. a simple walk). Hardness comes from an easy reduction from ST-connectivity. ☐

Theorem 24. WALK MEMBERSHIP *is NP-complete under simple-run semantics. It is already NP-hard in data complexity: there exists a fixed automaton $\mathcal{A}$ for which the problem is NP-hard.*

Sketch of proof. The hardness proof is done by a direct reduction from 3-SAT. The fixed automaton $\mathcal{A}$ uses the alphabet $\{\mathsf{Var}, \mathsf{Keep}, \mathsf{Invert}, \mathsf{Eval}, \mathsf{Check}\}$, has three states $\{0, 1, \top\}$, $\top$ is the unique initial and final state, and its transition table is given below.

Keep:	0	$\to$	0
	1	$\to$	1
	$\top$	$\to$	$\top$
Reset:	$\{0,1,\top\}$	$\to$	$\top$
Check:	$\{0,\top\}$	$\to$	$\top$

Var:	$\{0,1,\top\}$	$\to$	$\{0,1\}$
Invert:	0	$\to$	1
	1	$\to$	0
Eval:	1	$\to$	$\{0,1\}$
	$\{0,\top\}$	$\to$	$\top$

Figure 3 gives an example of how the database D is built from a specific 3-SAT instance, and for this example, the input walk $w = w_V w_H$ goes through all edges of D, starting with vertical edges (w_V) and then horizontal edges (w_H):

$$w_V = \text{Start} \to x_1 \to \cdots \bar{x}_1 \to \cdots x_2 \to \cdots \bar{x}_4 \to \text{Mid}$$

$$w_H = \text{Mid} \to C_0 \to \bar{x}_1^1 \to \cdots C_1 \to x_1^2 \to \cdots$$
$$C_2 \to x_1^3 \to \cdots C_3 \to \text{End}$$

Each valuation of the variables corresponds in a one-to-one manner to one run in $D \times \mathcal{A}$ for the vertical part. For instance, valuation $x_1 \mapsto 1$, $x_2 \mapsto 1$, $x_3 \mapsto 0$, $x_4 \mapsto 0$ corresponds to the run r_V:

In D :	Start	x_1	x_1^2	x_1^3	$\bar{x}_1$	$\bar{x}_1^1$	x_2	$\cdots$	x_3	$\cdots$	x_4	$\cdots$	Mid
In $\mathcal{A}$:	$\top$	1	1	1	0	0	1	$\cdots$	0	$\cdots$	0	$\cdots$	$\top$

The horizontal walk w_H has three parts ($C_0 \to C_1$, $C_1 \to C_2$ and $C_2 \to C_3$); each $C_{i-1} \to C_i$ checks whether the valuation makes the clause C_i true. Let us take C_1 for instance. There is exactly one run r_1 for the part $C_0 \to C_1$ in order for $r_V r_1$ to be simple, given below.

In D :	C_0	$\bar{x}_1^1$	x_3^1	$\bar{x}_4^1$	C_1
In $\mathcal{A}$:	$\top$	1	1	0	$\top$

Indeed, the state reached at $\bar{x}_1^1$ is necessarily 1, otherwise the full run would not be simple: the vertex $(\bar{x}_1^1, 0)$ of the run database was already visited in the vertical part. One may see that the state reached at vertex C_1 is $\top$, which means that the valuation satisfies C_1. If the valuation did not make C_1 true, there would be no run r_1 such that $r_V r_1$ is simple. ☐

5 Query Given as a Regular Expression

This section aims at applying simple-run semantics to practical settings. In real life, users generally input RPQs as a regular expression and not as an automaton. The natural idea would consist in translating the expression and then applying simple-run semantics to the resulting automaton. In Section 5.1, we explain why this approach leads to unwanted behaviour. Section 5.2 introduces binding-trail semantics as an adhoc adaptation of simple-run semantics to the case where the query is given as a regular expressions. Then, we show that binding-trail semantics enjoy the same computational properties as simple-run semantics.

5.1 Expression to Automaton

Given a regular language, there are many known approaches for producing automata which accept the same language (see for instance Sakarovitch 2021). Then, all algorithms from Section 4 immediately apply. However, the crux of the matter lies in Remark 7: semantics do not only depend on the language accepted by the automaton, but on the automaton itself. Thus, *how* we choose to translate the expression into an equivalent automaton matters, and it seems that each translation algorithm features undesirable quirks. We give two compelling examples, and leave a complete account of the many translation algorithms for future work.

First, one could translate the expression into a minimal DFA, but this choice makes the semantics non-compositional. Consider an expression $R = R_1 + R_2$, one would expect R to return more results than R_1 or R_2; but it is not always the case. For instance, if $R_1 = b^*(ab^*ab^*)^*$ and $R_2 = (a+b)(a+b)^*$, then R is equivalent to $(a+b)^*$. The minimal DFA $\mathcal{A}_1$ associated with R_1 has two states, while the minimal DFA $\mathcal{A}$ associated with R has only one state. Hence one can easily find a database D such that $[\![\mathcal{A}]\!]_{SR}(D) \not\supseteq [\![\mathcal{A}_1]\!]_{SR}(D)$. Similar quirks can be found for concatenation and star. This example illustrates that the translated automaton must not only represent the *regular language* but also the *regular expression as written*.

Second, one could use Glushkov construction (recalled in Definition 25, below), which famously produces an automaton that stays close to the regular expression. However, this choice introduces a left-to-right bias. Typically, $\mathcal{A} = Gl(a^*b^*)$ is not the mirrored automaton of $Gl(b^*a^*)$, and if one considers the following database D, the walk `S` $\to$ `S` $\to$ `T` belongs to $[\![\mathcal{A}]\!]_{SR}(D)$ but not the walk `S` $\to$ `T` $\to$ `T`.

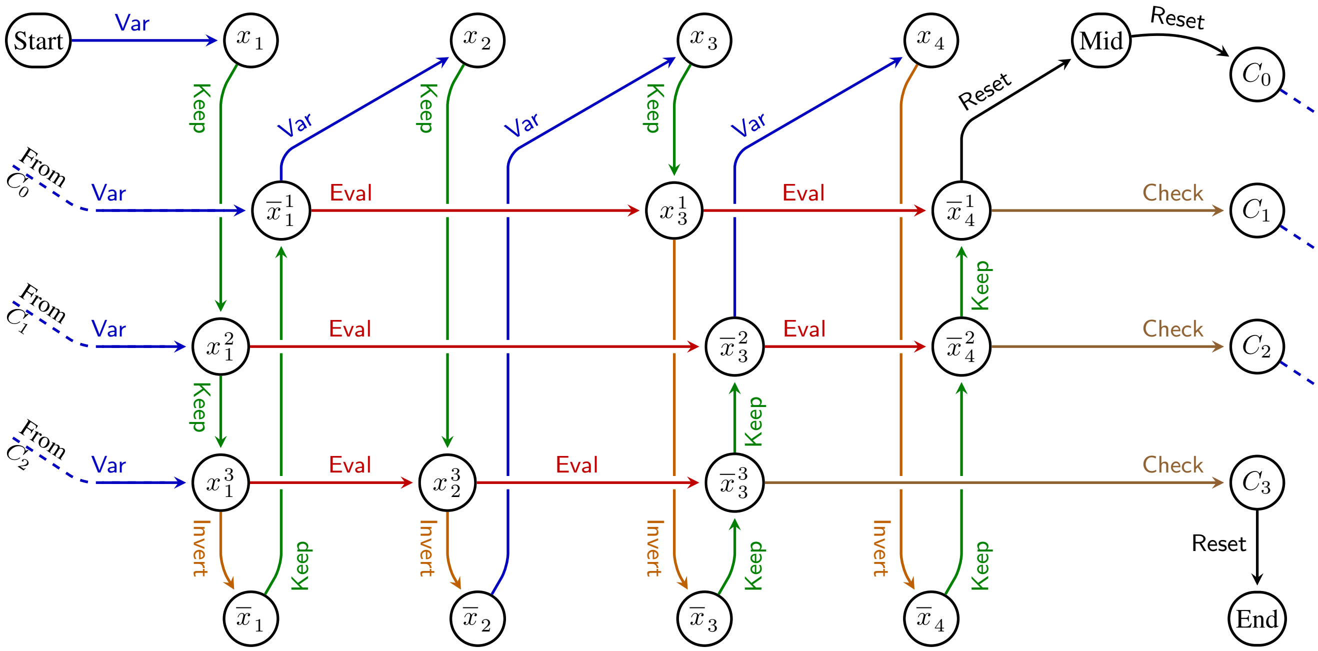
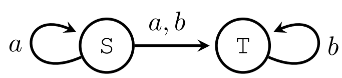


Figure 3: Graph encoding 3-SAT instance $C_1 \wedge C_2 \wedge C_3$ with $C_1 = \neg x_1 \vee x_3 \vee \neg x_4$, $C_2 = x_1 \vee \neg x_3 \vee \neg x_4$ and $C_3 = x_1 \vee x_2 \vee \neg x_3$



Definition 25. *Let R be a regular expression over Σ. A* linearisation *of R is a copy R' of R in which each atom in Σ is replaced by a different symbol in a new alphabet Γ, called the* positions *of R. Given $\alpha \in \Gamma$, we denote as $\overline{\alpha}$ the label of its antecedent in R.*

The Glushkov's automaton *of R, $Gl(R)$, is the automaton $\langle \Sigma, \{ i \} \uplus \Gamma, \Delta, \{ i \}, F \rangle$ defined as follows:*

$$\Delta = \{ (\alpha, \overline{\beta}, \beta) \mid \exists u, v \in \Gamma^*, u\alpha\beta v \in L(R') \} \cup \{ (i, \overline{\alpha}, \alpha) \mid \exists u \in \Gamma^*, \alpha u \in L(R') \}$$

$$F = \{ \alpha \mid \exists u \in \Gamma^*, u\alpha \in L(R') \} \cup \{ i \} \text{ if } \varepsilon \in L(R')$$

5.2 Binding-Trail Semantics

This section defines *binding-trail semantics*, a counterpart to simple-run semantics that operates directly on a given regular expression R, without translating R into an automaton.

Definition 26. *Let $D = (\Sigma, V, E, \textsc{Src}, \textsc{Tgt}, \textsc{Lbl})$ be a database, and R a regular expression. Let R' be a linearisation of R and Γ be the corresponding positions of R. A* binding trail *of D matching R is a sequence $(e_1, \alpha_1) \ldots (e_n, \alpha_n)$ of pairs in $E \times \Gamma$ such that:*

- *$e_1 \ldots e_n$ describes a walk of D and $\overline{\alpha_1 \cdots \alpha_n} \in \textsc{Lbl}(e_1 \ldots e_n)$;*
- *$\alpha_1 \cdots \alpha_n$ belongs to $L(R')$;*
- *All (e_i, α_i) are pairwise distinct.*

Binding-trail semantics are then defined by:

$$[\![R]\!]_{BT}(D) = \pi_D \left(\{ t \mid t \text{ is a binding trail of } D \text{ matching } R \} \right)$$

In other words, given a walk w in D, w conforms to binding-trail semantics if w matches R in such a way that the same edge of w cannot be used twice at the same position in R. The following lemma shows how binding-trail semantics relate to the run database.

Lemma 27. *For every regular expression R and database D,*

$$[\![R]\!]_{BT}(D) = \pi_D \circ \textsc{BindingTrail} \circ \textsc{Match}_{\mathcal{A}}(D) \,,$$

where $\mathcal{A} = Gl(R)$ and BindingTrail *is the run-bag filter that keeps only the runs*

$$(n_0, q_0)(e_0, \delta_0)(n_1, q_1) \cdots (e_{k-1}, \delta_{k-1})(n_k, q_k)$$

such that all (e_i, q_{i+1})'s, $0 \le i < k$, are pairwise distinct.

As is the case for simple-run semantics, binding-trail semantics also coincide with walk semantics for Tuple Membership, and produce the same shortest witnesses. Indeed, the proof of Proposition 19 can easily be adapted to prove the following:

Proposition 28. *Let D be a database and R be an expression. Let $\mathcal{A}$ be any automaton such that $L(R) = L(\mathcal{A})$. Let s, t be two vertices in D, and we denote by P the set of walks in $[\![\mathcal{A}]\!]_W(D)$ that go from s to t. Let w be a walk in P with minimal length. Then, w belongs to $[\![R]\!]_{BT}(D)$.*

Lemma 27 hints at the fact that the upper bounds of Section 4 for simple-run semantics immediately apply to binding-trail semantics, due to standard graph reduction techniques translating vertex-disjoint walks to edge-disjoint walks and back. The same does not necessarily hold true for lower bounds. Glushkov automata have additional properties: only one initial state, all incoming edges to a given state have the same label, and so on (Caron and Ziadi 2000). We show that these properties cannot be used to design more efficient algorithms.

Proposition 29. TUPLE MEMBERSHIP, TUPLE MULTIPLICITY, QUERY EVALUATION *and* WALK MEMBERSHIP *are computationally equivalent under binding-trail semantics and under simple-run semantics.*

Proposition 29 is actually a consequence of a much deeper result stating essentially that, given any automaton $\mathcal{A}$, there exists a regular expression R that encodes the topology of $\mathcal{A}$ in the sense that there is a strong connection between the computations of $\mathcal{A}$ and those of $Gl(R)$. Hence, any problem that takes automata as input will likely have hard instances that are of the form $Gl(R)$ for some expression R. Due to space constraints, we only sketch the main idea of the encoding.

Sketch of proof. Let $\mathcal{A} = \langle \Sigma, Q, \Delta, I, F \rangle$. Let $m = \text{CARD}(\Delta)$ and G denote any bijection $G : \Delta \to \{1, \ldots, m\}$. Let H be the only bijection $H : \Delta \to \{1, \ldots, m\}$ that meets: $\forall e \in \Delta,\ G(e) + H(e) = m + 1$. Finally, let σ be a fresh symbol that is not in Σ. We define the expression R over the alphabet $\Sigma \uplus \{\sigma\}$ as follows:

$$\left(\sum_{q \in Q} \left[\Big(\overbrace{\varepsilon +}^{\text{if } q \in I} \sum_{\substack{s \in Q,\ a \in \Sigma \\ e=(s,a,q) \in \Delta}} a^{G(e)} \Big) \sigma \Big(\overbrace{\varepsilon +}^{\text{if } q \in F} \sum_{\substack{a \in \Sigma,\ t \in Q \\ e=(q,a,t) \in \Delta}} a^{H(e)} \Big) \right] \right)^*$$

Note that there are exactly $\text{CARD}(Q)$ occurrences of the letter σ in R. We associate each state $s \in Q$ with the occurrence of σ appearing in the term of the external sum when $q = s$. Similarly, each transition $e = (s, a, t)$ with label a in $\mathcal{A}$ is encoded by the word $\sigma a^{m+1} \sigma$ that will be matched by the concatenation of the σ corresponding to s, the subexpression $a^{H(e)}$ on its right, the $a^{G(e)}$ on the left of σ corresponding to t followed by this σ.

We conclude by showing that a successful computation in $\mathcal{A}$ over a word $a_0 \cdots a_n \in \Sigma^*$ is encoded by matching the word $\sigma {a_0}^{m+1} \sigma \cdots {a_n}^{m+1} \sigma$ in R. Moreover, this encoding preserves relevant topological properties. For instance, a computation of $\mathcal{A}$ reuses a state if and only if its encoding reuses a position labelled by σ. ☐

Remark 30. *Similarly to simple-run semantics (see Remark 14), binding-trail semantics depend on the given regular expression and not only on the corresponding language. This allows for a finer control on which repetitions are permitted by the query. For instance, the three following expressions have different meanings under binding-trail semantics:*

- $[\![a^*]\!]_{BT}$ *returns all trails labelled by* a.
- $[\![a^* \cdot a^*]\!]_{BT}$ *returns the concatenations of two trails.*
- $[\![(a + a)^*]\!]_{BT}$ *returns all walks where edges are repeated at most twice.*

6 Perspectives

Run-based semantics are an attempt at addressing real-life concerns while maintaining good theoretical foundations. As such, our medium-term goal is to make sure that our work is indeed applicable to query languages used in practice. GQL offers a very plausible opportunity for integration, as it is in active development and already supports several semantics. This section aims at closing some of the gaps between theory and practice by discussing how run-based semantics adapt to commonly seen extensions or limitations.

6.1 Syntax Restrictions Used in Practice

Real query languages, such as GQL or Cypher, impose syntax restrictions on the regular expression given as input. We discuss here whether the lower bound complexity results change by imposing those restrictions. Note that no reasonable syntax restriction can change the complexity of TUPLE MULTIPLICITY since the lower bound already holds for the fixed expression a^*.

We consider three syntax restrictions and their respective impact on the complexity of WALK MEMBERSHIP. A regular expression has *star-height 1* if it has no nested Kleene stars. A regular expression is said to have *no union under star* (resp. *no concatenation under star*) if no union (resp. no concatenation) operator occurs in any subexpression nested under a Kleene star.

These restrictions match syntax rules commonly seen in practice: GQL only allows expressions with star-height 1 and Cypher queries cannot express concatenations under star. Moreover, users rarely use the full expressive power at their disposal. In (Bonifati, Martens, and Timm 2020), the authors make an analytical study of over 240,000 SPARQL queries: in the collected data set, every single RPQ has star-height 1, all queries but one have no concatenation under star, and only about 40% of queries use a union under a star.

The expression shown in Section 5.1 to simulate the computations of any arbitrary automaton has no nested stars. Hence all complexity lower bounds hold even when expressions are restricted to star-height 1. That expression has unions under star, but one can do without. Indeed, let $\mathcal{A} = \langle \Sigma, Q, \Delta, I, F \rangle$ be an automaton; for simplicity we assume that $\Sigma = \{0, 1, \ldots, k-1\}$ and that $Q = \{0, 1, \ldots, n-1\}$. Consider the following expression over the alphabet $\{a, b, c, \sigma\}$.

$$\sum_{i \in I} c^{i+1} \left(\prod_{i \in Q} \left(c^{n-i} \sigma a^{i+1}\right)^* \cdot \prod_{(i,x,j) \in \Delta} \left(a^{n-i} b^x c^{j+1}\right)^* \right)^* \sum_{i \in F} a^{n-i} \quad (4)$$

As in Section 5.2, one can show that this new expression[6] encodes the behaviour of $\mathcal{A}$. The key arguments are as follows: each letter $x \in \Sigma$ is encoded by $\lambda(x) = a^{n+1} b^x c^{n+1} \sigma$; each word $x_0 \cdots x_n$ is encoded by $c^{n+1} \lambda(x_1) \cdots \lambda(x_n) a^{n+1}$; the states of $\mathcal{A}$ are simulated by the positions with label σ.

[6] Strictly speaking, Equation (4) denotes a *family* of expressions, as concatenation (Π) is noncommutative. However, the reduction works for any of those expressions.

	Trail	Run-based	Homomorphism	Shortest-walk
TUPLE MEMBERSHIP	Intractable	Tractable	Tractable	Tractable
TUPLE MULTIPLICITY	Intractable	Intractable	$*$	$*$
QUERY EVALUATION	Intractable	Tractable	$*$	Tractable
DEDUPLICATED QUERY EVAL.	Intractable	Open	$*$	Tractable
WALK MEMBERSHIP	Tractable	Intractable	Tractable	Tractable

Table 1: Summary of computational complexity

Remark 31. *The internal Kleene stars in Equation (4) are used at most once, so these stars might be replaced by an* optional *operator, sometimes denoted by a "?". In that case, the expression would also be of star-height 1.*

When expressions have no concatenation under star, WALK MEMBERSHIP becomes tractable in combined complexity, as stated below.

Theorem 32. WALK MEMBERSHIP *is in PTIME under binding-trail semantics when restricted to expressions with no concatenation under star. The same holds under simple-run semantics when queries are restricted to the Glushkov automata of such expressions.*

Under binding-trail semantics, matching a starred subexpression with no concatenation in a walk amounts to counting that the number of repetitions of each edge in the walk is less than the number of compatible atoms of the expression. Under simple-run semantics, the proof relies on a reduction to matchings in bipartite graphs (Cormen et al. 2009, Section 26.3).

6.2 Extensions of Regular Expressions

Let us discuss how simple-run and binding-trail semantics behave with respect to some common extensions of RPQs.

One-or-more repetitions Many formalisms allow writing R^+ as a shorthand for $R \cdot R^*$. While expanding the notation is not suitable in our setting (the two R subexpressions in $R \cdot R^*$ would then be matched independently, resulting in a different behaviour), treating R^+ as a new operator poses no particular problem: the definitions of both binding-trail semantics and Glushkov automaton extend naturally over $+$.

Arbitrary repetitions Some formalisms allow an operator '$\{n,m\}$' with $n \in \mathbb{N}$ and $m \in \mathbb{N} \cup \{\infty\}$: $R^{\{n,m\}}$ means that R may be repeated between n and m times. Once again, expanding the notation would change the result. On the other hand, allowing this new operator would make Lemma 13 false, e.g., querying the database from Fig. 1 with $R^{\{6,\infty\}}$ would yield tuple (s, t) under walk semantics but not under binding-trail semantics. The impact of this operator on complexity results is left for future work.

Backward atoms Allowing backward atoms $\overleftarrow{a}$ in expressions, as for instance in 2RPQs (Angles et al. 2017), poses no particular problem: transitions of the automaton that are labelled with a backward atom are simply paired with reversed edges of the database in the run database.

Any-directed atoms Cypher and GQL allow any-directed atoms $\overleftrightarrow{a}$, that match edges labelled by a forward or backward. Allowing them would make Lemma 13 (and Prop. 28) false. For instance, let us query the database from Figure 1 with $R = \Big((\overrightarrow{\mathbf{R}} + \overrightarrow{\mathbf{G}}) \cdot \overleftrightarrow{\mathbf{R}} \cdot (\overrightarrow{\mathbf{R}} + \overrightarrow{\mathbf{G}})\Big)^*$. The walk $w = c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_3 \leftarrow c_2 \rightarrow t$ contradicts Lemma 13. This issue can be circumvented by expanding $\overleftrightarrow{a}$ into $\overrightarrow{a} + \overleftarrow{a}$ rather than treating it as a new operator, once again with some effect on the semantics.

7 Conclusion

Table 1, below, presents a summary of the computational complexity of popular semantics and compares them with run-based semantics. We also emphasize the following comparison points that do not appear in the table.

- Table 1 paints a negative picture of trail semantics, which seems at odds with its popularity in practice. We believe that one strength of trail semantics is that the output provides some kind of coverage of the space of matches, which enables rich aggregation and post-processing. Run-based semantics improves on this property by giving some guarantees on the coverage (Lemma 13).
- WALK MEMBERSHIP is a theoretical counterpart to QUERY EVALUATION: only the latter is implemented in real systems. Thus, we believe that run-based semantics offer a reasonable compromise by having tractable TUPLE MEMBERSHIP, tractable QUERY EVALUATION and intractable WALK MEMBERSHIP. It is in our view better than the other way around, as in trail semantics.
- Under homomorphism or shortest-walk semantics, some problems are marked with a $*$. Although tractable, complexity comparison with other semantics makes little sense since they behave differently. For instance, counting *all* matching walks leads to unboundedness under homomorphism semantics, and returns an information of dubious value under shortest-walk: 0 or 1 most of the time, and the number of uncomparable minimal runs otherwise.

In conclusion, simple-run and binding-trail semantics provide good computational properties overall, supports bag semantics and rich aggregation. However, while the RPQ formalism is a good model of the navigational part of most query languages over graph databases, it does not capture their ability to collect and compare data values along tested walks. Extending our proposed framework to handle data values lying in the edges and vertices of the database constitutes our main challenge going forward.

References

Angles, R.; Arenas, M.; Barceló, P.; Hogan, A.; Reutter, J. L.; and Vrgoč, D. 2017. Foundations of modern query languages for graph databases. *ACM Comput. Surv.* 50(5).

Angles, R.; Arenas, M.; Barceló, P.; Boncz, P. A.; Fletcher, G. H. L.; Gutierrez, C.; Lindaaker, T.; Paradies, M.; Plantikow, S.; Sequeda, J. F.; van Rest, O.; and Voigt, H. 2018. G-CORE: A core for future graph query languages. In *SIGMOD*, 1421–1432. ACM.

Bagan, G.; Bonifati, A.; and Groz, B. 2020. A trichotomy for regular simple path queries on graphs. *J. Comput. Syst. Sci.* 108:29–48.

Bonifati, A.; Martens, W.; and Timm, T. 2020. An analytical study of large SPARQL query logs. *VLDB J.* 29(2-3):655–679.

Brügemann-Klein, A. 1993. Regular expressions into finite automata. *Theoretical Computer Science* 120:197–213.

Caron, P., and Ziadi, D. 2000. Characterization of glushkov automata. *Theor. Comput. Sci.* 233(1-2):75–90.

Cormen, T.; Leiserson, C.; Rivest, R.; and Stein, C. 2009. *Introduction to Algorithms, third edition*. MIT Press.

Cruz, I. F.; Mendelzon, A. O.; and Wood, P. T. 1987. A graphical query language supporting recursion. In Dayal, U., and Traiger, I. L., eds., *SIGMOD'87*, 323–330. ACM.

Deutsch, A.; Xu, Y.; Wu, M.; and Lee, V. E. 2019. Tigergraph: A native MPP graph database. Preprint arXiv:1901.08248.

Deutsch, A.; Francis, N.; Green, A.; Hare, K.; Li, B.; Libkin, L.; Lindaaker, T.; Marsault, V.; Martens, W.; Michels, J.; Murlak, F.; Plantikow, S.; Selmer, P.; Voigt, H.; van Rest, O.; Vrgoč, D.; Wu, M.; and Zemke, F. 2022. Graph pattern matching in GQL and SQL/PGQ. In *SIGMOD'22*. ACM.

Diestel, R. 2012. *Graph Theory, 4th Edition*, volume 173 of *Graduate texts in mathematics*. Springer.

Francis, N.; Green, A.; Guagliardo, P.; Libkin, L.; Lindaaker, T.; Marsault, V.; Plantikow, S.; Rydberg, M.; Selmer, P.; and Taylor, A. 2018. Cypher: An evolving query language for property graphs. In *SIGMOD'18*. ACM.

Francis, N.; Gheerbrant, A.; Guagliardo, P.; Libkin, L.; Marsault, V.; Martens, W.; Murlak, F.; Peterfreund, L.; Rogova, A.; and Vrgoč, D. 2023. A Researcher's Digest of GQL. In Geerts, F., and Vandevoort, B., eds., *ICDT'23*, volume 255 of *LIPIcs*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Hopcroft, J. E., and Karp, R. M. 1973. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing* 2(4):225–231.

International Organization for Standardization. GQL. https://www.iso.org/standard/76120.html. Standard under development ISO/IEC CD 39075. Expected to be published in 2023.

Martens, W.; Niewerth, M.; and Trautner, T. 2020. A trichotomy for regular trail queries. In Paul, C., and Bläser, M., eds., *STACS'20*, volume 154 of *LIPIcs*, 7:1–7:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Mendelzon, A. O., and Wood, P. T. 1995. Finding regular simple paths in graph databases. *SIAM J. Comput.* 24(6):1235–1258.

Pin, J.-É., ed. 2021. *Handbook of Automata Theory*, volume 1. EMS.

Robinson, I.; Webber, J.; and Eifrem, E. 2015. *Graph databases, second edition*. O'Reilly.

Sakarovitch, J. 2021. Automata and rational expressions. Chapter 2 of (Pin 2021).

TigerGraph Team. 2021. TigerGraph Documentation – version 3.1.

Valiant, L. G. 1979. The complexity of enumeration and reliability problems. *SIAM J. Comput.* 8(3).

World Wide Web Consortium. 2013. SPARQL 1.1 query language, section 9: Property paths. https://www.w3.org/TR/sparql11-query/#propertypaths.

Yen, J. Y. 1971. Finding the k shortest loopless paths in a network. *Management Science* 17(11):712–716.

Acknowledgments

The authors would like to warmly thank Amélie Gheerbrant and Alexandra Rogova for pointing out the example in *Graph databases* (Robinson, Webber, and Eifrem 2015) that fails due to trail semantics. The authors are also very grateful to Antoine Meyer for the many fruitful conversations about the content of this paper. The authors acknowledge the support of *Agence Nationale de la Recherche* under grant *QUID*, code ANR-18-CE40-0031.

Appendix A: Recap figure

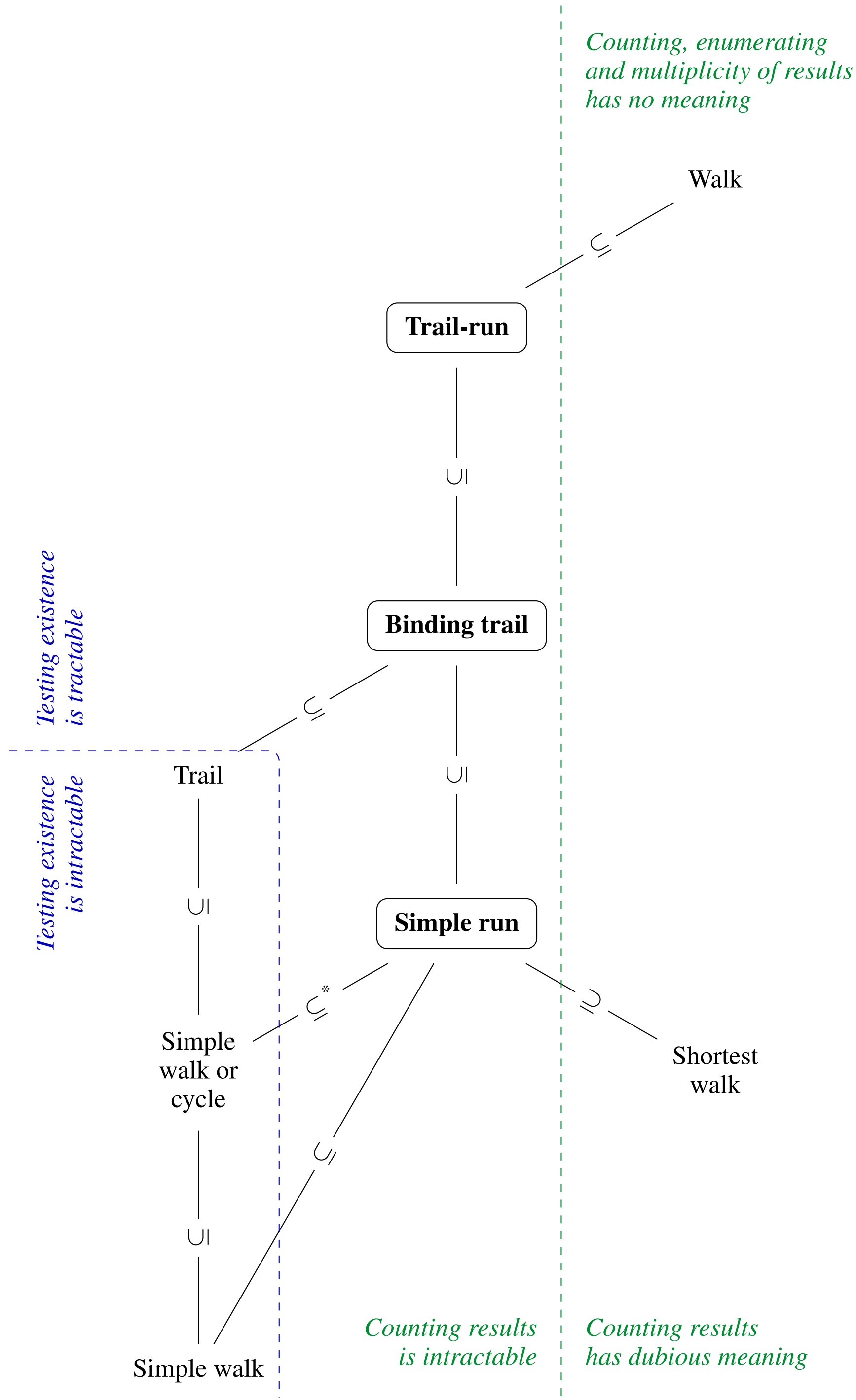


Figure 4: Summary of the different semantics mentioned in the introduction. (The inclusion marked with * holds only if the automaton is standard.)

Appendix B: Proof of Theorem 24

Theorem 24. WALK MEMBERSHIP *is NP-complete under simple-run semantics. It is already NP-hard in data complexity: there exists a fixed automaton $\mathcal{A}$ for which the problem is NP-hard.*

The problem is obviously in NP as one can guess a run r in $D \times \mathcal{A}$ and check that $\pi_D(r) = w$. The proof of hardness is done by reduction from 3-SAT. We define a fixed automaton $\mathcal{A}$ such that for any 3-SAT instance I we can build a polynomial size database D_I and walk p_I such that I is satisfiable iff $p_I \in [\![\mathcal{A}]\!]_{SR}(D_I)$.

Preliminary warning. Let us warn the reader that the sketch of proof as well as the Figure 3, page 7 give a simplified version of the construction presented here. The database D_I defined in this formal proof has many more *useless* vertices and edges in order to make the definition and statements easier to formulate. We omit those in the body of the paper to simplify Figure 3 and give better intuition in the sketch of proof. Figure 5 gives the database corresponding to the example of Figure 3 for the encoding defined below.

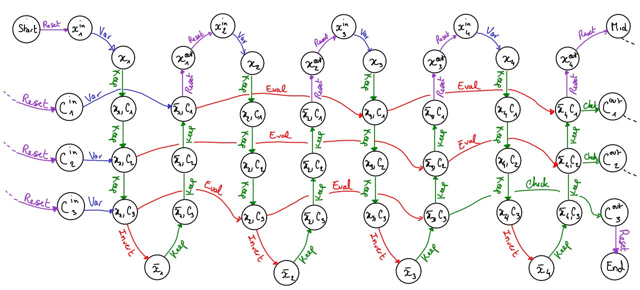


Figure 5: Drawing of the database D_I encoding the instance $I = C_1 \wedge C_2 \wedge C_3$ with $C_1 = \neg x_1 \vee x_3 \vee \neg x_4$, $C_2 = x_1 \vee \neg x_3 \vee \neg x_4$ and $C_3 = x_1 \vee x_2 \vee \neg x_3$

B1 The automaton $\mathcal{A}$

Let $\mathcal{A}$ be the automaton defined as follow:

- the alphabet is $\{$ Check, Eval, Invert, Keep, Reset, Var $\}$;
- $\mathcal{A}$ has three states 0, 1 and $\top$;
- $\top$ is both the only intial state and the only final state;
- its transition table is given below. (See Figure 6 for a graphical presentation.)

Keep:	0	$\to$	0
	1	$\to$	1
	$\top$	$\to$	$\top$
Reset:	$\{0,1,\top\}$	$\to$	$\top$
Check:	$\{0,\top\}$	$\to$	$\top$

Var:	$\{0,1,\top\}$	$\to$	$\{0,1\}$
Invert:	0	$\to$	1
	1	$\to$	0
Eval:	1	$\to$	$\{0,1\}$
	$\{0,\top\}$	$\to$	$\top$

B2 SAT instance I

Given a 3-SAT instance $I = C_1 \wedge \cdots \wedge C_\gamma$, we denote by X the set of the distinct variables $x_1, \ldots, x_n$ appearing in I and by $\bar{X}$ the set of the corresponding negated version of these variables: $\bar{x}_1, \ldots, \bar{x}_n$.

In the following, we will generally use the notation x, $\bar{x}$ and $\tilde{x}$ for elements in X, $\bar{X}$ and $X \cup \bar{X}$, respectively. Typically we write $C_i = \tilde{x}_\ell \vee \tilde{x}_k \vee \tilde{x}_m$.

B3 The Database D_I

Let I be a 3-SAT instance. We can now build a database D_I from I. An example is given in Figure 5.

Vertices

- D_I contains three special vertices: Start, Mid, and End
- For each x in X, D_I contains three vertices: x^{in}, x, and x^{out}.
- For each $\bar{x}$ in $\bar{X}$, D contains one vertex: $\bar{x}$.
- For each element $\tilde{x}$ in $X \cup \bar{X}$ and each clause C_i, D_I contains one vertex $(\tilde{x}, C_i)$
- For each clause C_i we create two vertices: C_i^{in} and C_i^{out}.

Edges

For each variable $x \in X$ we add the edges from the following walk p_x to D_I.

$$p_x = x^{\text{in}} \xrightarrow{\text{Var}} x \xrightarrow{\text{Keep}} (x, C_0) \xrightarrow{\text{Keep}} \cdots$$
$$\cdots \xrightarrow{\text{Keep}} (x, C_\gamma) \xrightarrow{\text{Invert}} \bar{x} \xrightarrow{\text{Keep}} (\bar{x}, C_\gamma) \xrightarrow{\text{Keep}} \cdots$$
$$\cdots \xrightarrow{\text{Keep}} (\bar{x}, C_0) \xrightarrow{\text{Reset}} x^{\text{out}}$$

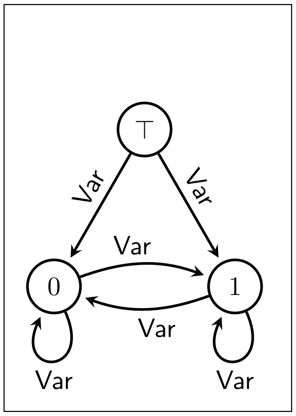
For each clause $C_i = \tilde{x}_k \vee \tilde{x}_\ell \vee \tilde{x}_m$, we add edges from the following walk p_{C_i} to D_I.

$$p_{C_i} = C_i^{\text{in}} \xrightarrow{\text{Var}} (\tilde{x}_k, C_i) \xrightarrow{\text{Eval}} (\tilde{x}_\ell, C_i)$$
$$\xrightarrow{\text{Eval}} (\tilde{x}_m, C_i) \xrightarrow{\text{Check}} C_i^{\text{out}}$$

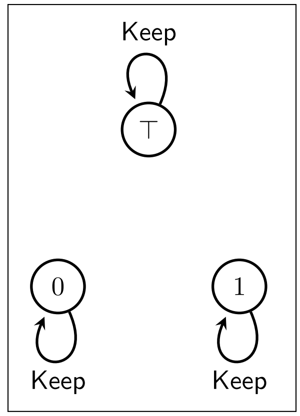
Then we connect these walks by Reset-labelled edges as follows.

$$\text{Start} \xrightarrow{\text{Reset}} x_1^{\text{in}}$$
$$\forall i,\ 1 \leq i < n \quad x_i^{\text{out}} \xrightarrow{\text{Reset}} x_{i+1}^{\text{in}}$$
$$x_n^{\text{out}} \xrightarrow{\text{Reset}} \text{Mid}$$

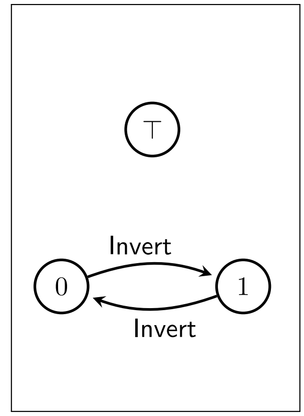
$$\text{Mid} \xrightarrow{\text{Reset}} C_1^{\text{in}}$$
$$\forall i, 1 \leq i < \gamma \quad C_i^{\text{out}} \xrightarrow{\text{Reset}} C_{i+1}^{\text{in}}$$
$$C_\gamma \xrightarrow{\text{Reset}} \text{End}$$



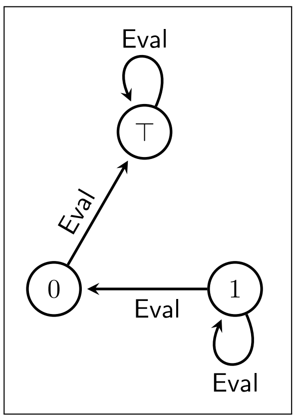
(a) Transitions labelled by Var



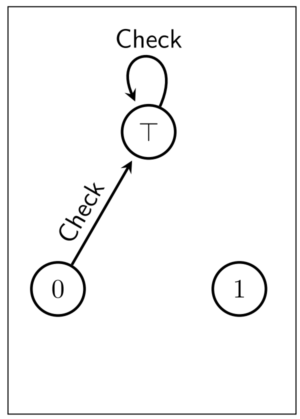
(b) Transitions labelled by Var



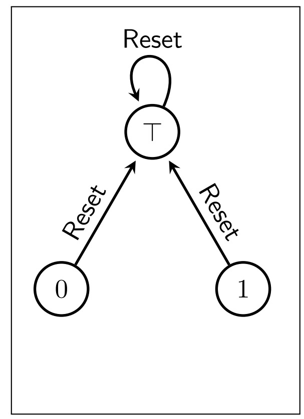
(c) Transitions labelled by Var



(d) Transitions labelled by Var



(e) Transitions labelled by Var



(f) Transitions labelled by Var

Figure 6: Transitions of the automaton $\mathcal{A}$ ($\top$ is initial and final)

B4 The walk p_I, and its components p_{setval} and p_{checksat}

The walk p_I consists in traversal from Start to End going through every single edge of the database D_I. Intuitively, a first walk p_{setval} going through every p_x will be used to define a valuation of the variables. A second walk p_{checksat} going through every p_{C_i} will be used to check that this valuation makes the instance true. Formally, we define the walk p_I as $p_{\text{setval}} \cdot p_{\text{checksat}}$ where p_{setval} and p_{checksat} are as follows :

$$p_{\text{setval}} = \text{Start} \xrightarrow{\textsf{Reset}} p_{x_1} \xrightarrow{\textsf{Reset}} p_{x_2} \xrightarrow{\textsf{Reset}} \cdots \\ \cdots \xrightarrow{\textsf{Reset}} p_{x_n} \xrightarrow{\textsf{Reset}} \text{Mid}$$

$$p_{\text{checksat}} = \text{Mid} \xrightarrow{\textsf{Reset}} p_{C_1} \xrightarrow{\textsf{Reset}} p_{C_2} \xrightarrow{\textsf{Reset}} \cdots \\ \cdots \xrightarrow{\textsf{Reset}} p_{C_\gamma} \xrightarrow{\textsf{Reset}} \text{End}$$

Note that the equation above makes a slight abuse of notation: the p_{x_i}'s and p_{C_i}'s are walks instead of vertices: $p \xrightarrow{\textsf{Reset}} p'$ means that we connect the last vertex of p with the first vertex of p'.

B5 Main statement

We call *valuation* any (total) function $\mu : X \cup \bar{X} \to \{0, 1\}$ such that for every $x \in X$, $\mu(\bar{x}) = 1 - \mu(x)$. The purpose of the remainder of this section is to show the following proposition, of which Theorem 24 is a direct consequence.

Proposition B33. *There are exactly N simple runs r in $D_I \times \mathcal{A}$ such that $\pi_D(r) = p_I$, where N is the number of distinct valuations that make I true.*

B6 Setting a valuation of the variables

Let us show that there is a bijection between the valuations and the runs in $D_I \times \mathcal{A}$ for the part p_{setval}.

Lemma B34. (a) *For every valuation μ there exists a simple run r_μ in $D_I \times \mathcal{A}$ such that: for every $x \in X$, the run r_μ passes through the vertex $(x, \mu(x))$.*

(b) *For every valuation μ, the run r_μ passes through the vertices $((\tilde{x}, C_i), \mu(\tilde{x}))$, for each $x \in X \cup \bar{X}$ and $1 \leq i \leq \gamma$.*

(c) *Let r be any run in $D_I \times \mathcal{A}$ such that $\pi_D(r) = p_{\text{setval}}$, then $r = r_\mu$ for some valuation μ.*

Proof. (a) For each $x \in X$, we denote by $r_{\mu,x}$ the following run of $D_I \times \mathcal{A}$.

$$\begin{aligned} r_{\mu,x} = (x^{\text{in}}, \top) \xrightarrow{\mathsf{Var}} (x, \mu(x)) &\xrightarrow{\mathsf{Keep}} ((x, C_0), \mu(x)) \\ &\xrightarrow{\mathsf{Keep}} \cdots \\ &\xrightarrow{\mathsf{Keep}} ((x, C_\gamma), \mu(x)) \\ \xrightarrow{\mathsf{Invert}} (\bar{x}, \mu(\bar{x})) &\xrightarrow{\mathsf{Keep}} ((\bar{x}, C_\gamma), \mu(\bar{x})) \\ &\xrightarrow{\mathsf{Keep}} \cdots \\ &\xrightarrow{\mathsf{Keep}} ((\bar{x}, C_0), \mu(\bar{x})) \\ &\qquad \xrightarrow{\mathsf{Reset}} (x^{\text{out}}, \top) \end{aligned}$$

Then we define r_μ as follows.

$$(\text{Start}, \top) \xrightarrow{\mathsf{Reset}} r_{\mu,x_1} \xrightarrow{\mathsf{Reset}} \cdots \xrightarrow{\mathsf{Reset}} r_{\mu,x_n} \xrightarrow{\mathsf{Reset}} (\text{Mid}, \top)$$

By construction, $r_{\mu,x}$ is a simple run of $D_I \times \mathcal{A}$.

(b) Follows from the definition of r_μ in item (a).

(c) For each variable $x \in X$, the vertex x appears exactly once in π_{setval} hence there is exactly one occurrence in r of a vertex of the form (x, s), for some $s \in \{0, 1, \top\}$. Note moreover that the edge coming into x in p_{setval} is labelled by Var hence that $s \in \{0, 1\}$ from the definition of $\mathcal{A}$ (cf. Figure 6a). We let μ denote the valuation that maps each $x \in X$ to the unique s such that (x, s).

From the definition of $\mathcal{A}$ (cf. Figure 6), the letters Reset, Keep and Invert are *deterministic*, in the sense that there is at most one transition going out from every state and labelled by one of these letters. It follows that $r = r_\mu$ since both runs coincide on all states coming after an edge labelled by Var. □

B7 Checking the clauses

Lemma B35. *For each valuation μ,*

(a) *if μ makes I true, then there exists a unique run r'_μ such that both $\pi(r'_\mu) = p_{checksat}$, and $r_\mu \cdot r'_\mu$ is simple;*

(b) *if μ makes I false, then there is no run r' such that $\pi_D(r') = p_{checksat}$ and $r_\mu \cdot r'$ is simple.*

We say that two runs r and r' are *mutually simple* if r and r' have no vertex in common.

Proof of (a). Let $C_i = \tilde{x}_k \vee \tilde{x}_\ell \vee \tilde{x}_m$ be a clause of I. By hypothesis, one of the atom is made true by μ. Let r'_{μ,C_i} be the run in $D_I \times \mathcal{A}$ defined as follows, depending on which among the atoms $\tilde{x}_k$, $\tilde{x}_\ell$ and $\tilde{x}_m$ is made true by μ

If $\mu(\tilde{x}_k) = 1$:

$$\begin{aligned} r'_{\mu,C_i} = (C_i^{\text{in}}, \top) &\xrightarrow{\mathsf{Var}} ((\tilde{x}_k, C_i), \quad 0) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_\ell, C_i), \quad \top) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_m, C_i), \top) \xrightarrow{\mathsf{Check}} (C_i^{\text{out}}, \top) \end{aligned}$$

If $\mu(\tilde{x}_k) = 0$ and $\mu(\tilde{x}_\ell) = 1$:

$$\begin{aligned} r'_{\mu,C_i} = (C_i^{\text{in}}, \top) &\xrightarrow{\mathsf{Var}} ((\tilde{x}_k, C_i), \quad 1) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_\ell, C_i), \quad 0) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_m, C_i), \top) \xrightarrow{\mathsf{Check}} (C_i^{\text{out}}, \top) \end{aligned}$$

If $\mu(\tilde{x}_k) = \mu(\tilde{x}_\ell) = 0$ and $\mu(\tilde{x}_m) = 1$:

$$\begin{aligned} r'_{\mu,C_i} = (C_i^{\text{in}}, \top) &\xrightarrow{\mathsf{Var}} ((\tilde{x}_k, C_i), \quad 1) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_\ell, C_i), \quad 1) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_m, C_i), \ 0) \xrightarrow{\mathsf{Check}} (C_i^{\text{out}}, \top) \end{aligned}$$

One may check that in each case, the vertex $((\tilde{x}_k, C_i), s)$ in r_{μ,C_i} is such that $s = 1 - \mu(\tilde{x_k})$; and that the same is true for x_ℓ and x_m. It follows that r_μ and r'_{μ,C_i} are mutually simple.

By definitions, r_{μ,C_i} and r_{μ,C_j} are mutually simple if $i \neq j$. Hence, the run $r_\mu r'_\mu$ is simple where r'_μ is defined as follows.

$$\begin{aligned} r'_\mu = (\text{Mid}, \top) \xrightarrow{\mathsf{Reset}} r_{\mu,C_1} &\xrightarrow{\mathsf{Reset}} \cdots \\ \cdots \xrightarrow{\mathsf{Reset}} r_{\mu,C_\gamma} &\xrightarrow{\mathsf{Reset}} \text{End} \end{aligned}$$

It may be verified that r'_μ is unique. The only letters that are nondeterministic in $\mathcal{A}$ are Var and Eval, and the choice is between state 0 and 1. More precisely, these letters always bring up a choice between between vertex $((\tilde{x}, C_i), 0)$ or $((\tilde{x}, C_i), 1)$ in $D_I \times \mathcal{A}$ for some $i \in \{1, \ldots, \gamma\}$ and some variable $\tilde{x} \in X \cup \bar{X}$. One of those two vertices necessarily appears in r_μ (as $((\tilde{x}_k, C_i), \mu(\tilde{x}))$), hence only the other one may appear in r'_μ. □

Proof of (b). For the sake of contradiction, let us assume that there exists a run r' such that $\pi_D(r') = \pi_{\text{checksat}}$ and such that $r_\mu r'$ is simple. Since μ makes I false, it makes $C_i = \tilde{x}_k \vee \tilde{x}_\ell \vee \tilde{x}_m$ false for some i. Let r'_{C_i} be the subwalk of r' such that $\pi_D(r'_{C_i}) = p_{C_i}$; it may we written as follows.

$$\begin{aligned} r'_{C_i} = (C_i^{\text{in}}, s^{\text{in}}) &\xrightarrow{\mathsf{Var}} ((\tilde{x}_k, C_i), \quad s_k) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_\ell, C_i), \quad s_\ell) \\ &\xrightarrow{\mathsf{Eval}} ((\tilde{x}_m, C_i), s_m) \xrightarrow{\mathsf{Check}} (C_i^{\text{out}}, s^{\text{out}}) \end{aligned}$$

for some $s^{\text{in}}, s_k, s_\ell, s_m, s^{\text{out}} \in \{0, 1, \top\}$. Since μ makes I false, it follows that $\mu(\tilde{x}_k) = 0$, hence from Lemma B34(b) that r_μ contains the vertex $((\tilde{x}_k, C_i), 0)$. Moreover, since $r_\mu r'$ is simple, it follows that $s_k \neq 0$. A similar reasoning yields that $s_\ell \neq 0$ and $s_m \neq 0$ Regardless of the value of s^{in}, it follows from the transitions of $\mathcal{A}$ for letter Var (cf. Figure 6a) that $s_k \in \{0, 1\}$, hence $s_k = 1$; it then follows from the transitions of $\mathcal{A}$ for letter Eval (cf. Figure 6d) that $s_\ell = 1$, and with the same argument that $s_m = 1$. Since $\mathcal{A}$ has no transition labelled by Check and going out from state 1 (cf. Figure 6e), this leads to a contradiction. □

B8 Proof of Proposition B33

Let N be the number of valuations that make I true. Let μ be a valuation that makes I true. Lemma B34(a) yields a run r_μ and Lemma B35(a) yields a run r'_μ such that $r_\mu \cdot r'_\mu$ is simple and $\pi_D(r_\mu \cdot r'_\mu) = p_I$.

Note that if μ, μ' denote two valuations that makes I true, $r_\mu = r'_\mu$ implies that for every $x \in X$, $\mu(x) = \mu'(x)$

(due to the condition in Lemma B34 (a)), hence that $\mu = \mu'$. Thus, the previous paragraph defines N distinct simple runs in $D_I \times \mathcal{A}$.

Let $r \cdot r'$ be any simple run in $D_I \times \mathcal{A}$ such that $\pi_D(r) = p_{\text{setval}}$ and $\pi_D(r') = p_{\text{checksat}}$. Hence from Lemma B34 (c), there exists μ such that $r = r_\mu$. It is impossible that μ makes I false since the existence of r' would be in contradiction with Lemma B35 (b). Hence, μ makes I true and the unicity in Lemma B35 (a) implies that $r' = r'_\mu$. □

Remark 36. *The size of the database D_I built in Section B3 may be up to quadratic in the size of the 3-SAT instance I, namely in $O(\gamma n)$. This is due to the fact that we create a lot of "useless" vertices. In particular, every vertex $(\tilde{x}, C_i)$ can be omitted if $\tilde{x}$ does not appear in clause C_i. By omitting those vertices, the database D_I would be of a size in $O(\gamma + n)$, at the cost of making the definition more involved. Figure 3, page 7, gives an example of that simplified D.*

Appendix C: Glushkov construction provides hard instance for WALK MEMBERSHIP

The purpose of appendix C is to prove the proof of Theorem C37, below, which is the center piece behind Proposition 29. The proof uses the novel notion of *topological coding* of an automaton, which we plan to flesh-out in a future independent document; we leave a preview of it here in order for this preprint to be self-contained.

Theorem C37. *There exists a fixed expression R such that* WALK MEMBERSHIP *is NP-hard for $\mathcal{A} = Gl(R)$ under simple-run semantics.*

Section C1 introduces a different, equivalent definition for the Glushkov automaton of an expression. It allows for more effective notations, whereas the initial definition was only stated in a declarative way. Section C2 defines *topological codings* and Section C3 gives a precise meaning to the intuition that a topological coding somehow simulates another automaton. Section C4 states and shows Theorem C43, the main result of appendix C: any automaton can be encoded into a Glushkov automaton. Section C5 applies Theorem C43 to the proof of Theorem C37. Finally Section C6 briefly explains how a counterpart to Theorem C37 can be proved under binding-trail semantics.

C1 Glushkov automaton

A linearisation of an expression R over an alphabet Γ is a pair $\langle \Gamma, R' \rangle$ where

- Γ is a finite set of *annotations*;
- R' is an expression over $\Sigma \times \Gamma$ such that
  - every letter in $\Sigma \times \Gamma$ appears at most once in R'
  - $f(R') = R$ where f is the projection $\Sigma \times \Gamma \to \Sigma$ on the first component, lifted to regular expressions.

We denote the letter $(a, i) \in \Sigma \times \Gamma$ as $\left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right]$. Classically, one linearises R using $\Gamma = \{1, \ldots, n\}$, where n is the number of atoms in R, annotating the i-th leftmost atom in R with i. For instance, the linearisation of $b^*(ab^*ab^*)^*$ would be $\langle R', \Gamma \rangle$ with:

$$R' = \left[\begin{smallmatrix} b \\ 1 \end{smallmatrix}\right]^* \left(\left[\begin{smallmatrix} a \\ 2 \end{smallmatrix}\right] \left[\begin{smallmatrix} b \\ 3 \end{smallmatrix}\right]^* \left[\begin{smallmatrix} a \\ 4 \end{smallmatrix}\right] \left[\begin{smallmatrix} b \\ 5 \end{smallmatrix}\right]^* \right)^*$$
$$\Gamma = \{1, 2, 3, 4, 5\}.$$

Given a regular expression R over an alphabet Σ, the *Glushkov automaton* associated with R, denoted by $Gl(R) = \langle \Sigma, Q, \Delta, I, F \rangle$, is defined as follows from any linearisation $\langle \Gamma, R' \rangle$ of R.

$$Q = \{\mathsf{init}\} \cup \Sigma \times \Gamma \tag{5}$$

$$\Delta = \left\{ \left(\left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right], b, \left[\begin{smallmatrix} b \\ k \end{smallmatrix}\right]\right) \,\middle|\, \begin{array}{l} \left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right], \left[\begin{smallmatrix} b \\ k \end{smallmatrix}\right] \in \Sigma \times \Gamma \\ \exists w, w' \text{ such that } w \left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right] \left[\begin{smallmatrix} b \\ k \end{smallmatrix}\right] w' \in \mathrm{L}(R')) \end{array} \right\}$$
$$\bigcup \left\{ \left(\mathsf{init}, a, \left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right]\right) \,\middle|\, \begin{array}{l} \left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right] \in \Sigma \times \Gamma \\ \exists w \text{ such that } \left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right] w \in \mathrm{L}(R') \end{array} \right\} \tag{6}$$

$$I = \{\mathsf{init}\} \tag{7}$$

$$F = \left\{ \left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right] \,\middle|\, \exists w \text{ such that } w \left[\begin{smallmatrix} a \\ i \end{smallmatrix}\right] \in \mathrm{L}(R') \right\} \tag{8}$$

C2 Topological coding of an automaton

The notion of topological coding is an adaptation to automata of the notion of *topological minor* for directed graphs (Diestel 2012). Topological codings are defined formally in Definition C38. We give a first intuitive definition below.

Intuitively, $\mathcal{B}$ is a topological coding of $\mathcal{A}$ if $\mathcal{B}$ may be built from $\mathcal{A}$ by the following process:

- for each letter a in $\Sigma_{\mathcal{A}}$, the alphabet of $\mathcal{A}$, choose a nonempty word $\lambda(a)$ over $\Sigma_{\mathcal{B}}$, the alphabet of $\mathcal{B}$;
- replace each transition (s, q, t) in $\mathcal{A}$ by a fresh walk labelled by $\lambda(a)$ that starts at s and ends at t;
- optionally, choose a word u_i over $\Sigma_{\mathcal{B}}$, add a fresh initial state $init$, and for each initial state s of $\mathcal{A}$, remove its initial status and add a fresh walk going from $init$ to s
- optionally, proceed similarly for final states with a word u_f and a fresh final state $final$;
- then, one may add states and transitions, as long as it does not create a walk labelled by a word in $\mathrm{IM}(\lambda)$, or a walk starting from an initial (resp. ending in a final) state labelled by u_i (resp. u_f).

Given an automaton $\mathcal{A}$, we let $\mathsf{Comp}(\mathcal{A})$ denote the set of computations in $\mathcal{A}$.

Definition C38. *We say that $\mathcal{B} = \langle \Sigma_{\mathcal{B}}, Q_{\mathcal{B}}, \Delta_{\mathcal{B}}, I_{\mathcal{B}}, F_{\mathcal{B}} \rangle$ is a* topological out-coding *of $\mathcal{A} = \langle \Sigma_{\mathcal{A}}, Q_{\mathcal{A}}, \Delta_{\mathcal{A}}, I_{\mathcal{A}}, F_{\mathcal{A}} \rangle$, or simply a* topological coding[7] *of $\mathcal{A}$, if there exist:*

(a) *two words $u_i, u_f \in \Sigma_{\mathcal{B}}^*$;*

(b) *an injective function $\lambda : \Sigma_{\mathcal{A}} \to \Sigma_{\mathcal{B}}^+$*

(c) *an injective function $\nu : Q_{\mathcal{A}} \to Q_{\mathcal{B}}$.*

---

[7] The *out-* comes from the fact that the definitions of the $W_{\mathcal{B}}^{\text{something}}$'s are not symmetric: $W_{\mathcal{B}}$ contains the walks going **out** of a state...

and, denoting

$$W_{\mathcal{B}}^{\mathit{initial}} = \left\{ w \in \mathit{Comp}(\mathcal{B}) \,\middle|\, \begin{array}{l} \mathrm{SRC}(w) \in I_{\mathcal{B}} \\ \mathrm{LBL}(w) = u_i \end{array} \right\} \quad (9)$$

$$W_{\mathcal{B}}^{\mathit{transition}} = \left\{ w \in \mathit{Comp}(\mathcal{B}) \,\middle|\, \begin{array}{l} \mathrm{SRC}(w) \in \mathrm{IM}(\nu) \\ \mathrm{LBL}(w) \in \mathrm{IM}(\lambda) \end{array} \right\} \quad (10)$$

$$W_{\mathcal{B}}^{\mathit{final}} = \left\{ w \in \mathit{Comp}(\mathcal{B}) \,\middle|\, \begin{array}{l} \mathrm{SRC}(w) \in \mathrm{IM}(\nu) \\ \mathrm{LBL}(w) = u_f \end{array} \right\} \quad (11)$$

there exist:

(d) *a bijection $\eta_i : I_{\mathcal{A}} \to W_{\mathcal{B}}^{\mathit{initial}}$ such that*
- *for every $s \in I_{\mathcal{A}}$, the walk $\eta_i(s)$ ends in $\nu(s)$*
- *for every $w \in \mathrm{IM}(\eta_i)$ and every state s that appears in w at a position that is not the last one, then $s \notin \mathrm{IM}(\nu)$;*

(e) *a bijection $\eta : \Delta_{\mathcal{A}} \to W_{\mathcal{B}}^{\mathit{transition}}$ such that for every (s, a, t) in $\Delta_{\mathcal{A}}$, the walk $w = \eta((s, a, t))$ satisfies:*
- *$\eta((s, a, t))$ starts in $\nu(s)$*
- *$\eta((s, a, t))$ is labelled by $\lambda(a)$*
- *$\eta((s, a, t))$ ends in $\nu(t)$*
- *every internal state[8] in $\eta((s, a, t))$ is not in $\mathrm{IM}(\nu)$;*

(f) *a bijection $\eta_f : F_{\mathcal{A}} \to W_{\mathcal{B}}^{\mathit{final}}$ such that*
- *for every $s \in F_{\mathcal{A}}$, the walk $\eta_f(s)$ ends in $F_{\mathcal{B}}$*
- *for every $w \in \mathrm{IM}(\eta_f)$ and every state s that appears in w at a position that is not the first one, then $s \notin \mathrm{IM}(\nu)$.*

that collectively satisfy:

(g) *for every $w_1, w_2 \in \big(\mathrm{IM}(\eta_i) \cup \mathrm{IM}(\eta) \cup \mathrm{IM}(\eta_f)\big)$, if w_1 and w_2 have a transition or an internal state in common, then $w_1 = w_2$.*

C3 Properties of a topological coding

In this section, we fix two automata $\mathcal{A} = \langle \Sigma_{\mathcal{A}}, Q_{\mathcal{A}}, \Delta_{\mathcal{A}}, I_{\mathcal{A}}, F_{\mathcal{A}} \rangle$ and $\mathcal{B} = \langle \Sigma_{\mathcal{B}}, Q_{\mathcal{B}}, \Delta_{\mathcal{B}}, I_{\mathcal{B}}, F_{\mathcal{B}} \rangle$ such that $\mathcal{B}$ is a *topological out-coding* of $\mathcal{A}$ and we reuse the notations of Definition C38.

Remark 39. *The properties below follow from the definition of a topological coding.*

(a) *The conditions C38*(d) *and C38*(e) *imply that $W_{\mathcal{B}}^{\mathit{transition}}$ and $W_{\mathcal{B}}^{\mathit{initial}}$ are disjoint. Indeed, if $u_i \neq \varepsilon$, the first state in each walk in $W_{\mathcal{B}}^{\mathit{transition}}$ is in $\mathrm{IM}(\nu)$, while the first state in each walk in $W_{\mathcal{B}}^{\mathit{initial}}$ is not; and if $u_i = \varepsilon$, all walks in $W_{\mathcal{B}}^{\mathit{initial}}$ has length 0 while all walks in $W_{\mathcal{B}}^{\mathit{transition}}$ have a positive length since $\mathrm{LEN}(\lambda(a)) > 0$ for every $a \in \Sigma_{\mathcal{A}}$.*

(b) *Similarly, $W_{\mathcal{B}}^{\mathit{transition}}$ and $W_{\mathcal{B}}^{\mathit{final}}$ are disjoint.*

(c) *Similarly, $W_{\mathcal{B}}^{\mathit{initial}}$ and $W_{\mathcal{B}}^{\mathit{final}}$ are disjoint unless $u_i = u_f = \varepsilon$ and $I_{\mathcal{A}} \cap F_{\mathcal{A}} \neq \emptyset$.*

As expected, there is a strong correspondence between computations in an automaton and its topological coding.

[8] A state in a computation is *internal* if it appears at a position that is not the first or the last one.

Lemma C40. *Let $\mathcal{A}$ and $\mathcal{B}$ be two automata such that $\mathcal{B}$ is a topological coding of $\mathcal{A}$.*

(a) *Let $s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} \cdots \xrightarrow{a_k} s_k$ be a computation in $\mathcal{A}$. Then, $\nu(s_0) \xrightarrow{\lambda(a_1)} \nu(s_1) \xrightarrow{\lambda(a_2)} \cdots \xrightarrow{\lambda(a_k)} \nu(s_k)$ is a computation in $\mathcal{B}$.*

(b) *Conversely, let $\pi_{\mathcal{B}}$ be a computation in $\mathcal{B}$ of the form*

$$\pi_{\mathcal{B}} = \nu(s_0) \xrightarrow{\lambda(a_1)} \nu(s_1) \xrightarrow{\lambda(a_2)} \cdots \xrightarrow{\lambda(a_k)} \nu(s_k)$$

then $s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} \cdots \xrightarrow{a_k} s_k$ is a computation in $\mathcal{A}$.

Proof. Item (a) follows from condition C38(e) applied to each transition of $\Delta_{\mathcal{A}}$. Item (b) follows from the fact that each walk $\nu(s_i) \xrightarrow{\lambda(a_{i+1})} \nu(s_{i+1})$ belongs to $W_{\mathcal{B}}^{\mathrm{transition}}$, which allows to apply C38(e) and concludes the proof. ☐

Definition C41. *Let $\mathcal{B}$ be a topological coding of $\mathcal{A}$. Given a successful computation $\pi_{\mathcal{A}} = s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} \cdots \xrightarrow{a_k} s_k$ in $\mathcal{A}$, we call* corresponding computation *in $\mathcal{B}$, the computation $\pi_{\mathcal{B}}$ defined as*

$$\overbrace{x \xrightarrow{u_i} \nu(s_0)}^{\eta_i(s_0)} \underbrace{\xrightarrow{\lambda(a_1)} \nu(s_1)}_{\eta((s_0,a_1,s_1))} \xrightarrow{\lambda(a_2)} \cdots$$

$$\cdots \xrightarrow{\lambda(a_{k-1})} \underbrace{\nu(s_{k-1}) \xrightarrow{\lambda(a_k)} \nu(s_k)}_{\eta((s_{k-1},a_1,s_k))} \overbrace{\xrightarrow{u_f} y}^{\eta_f(s_k)}$$

where $x = \mathrm{SRC}(\eta_i(s_0))$ and $y = \mathrm{TGT}(\eta_f(s_k))$.

We show that there is a bijection between successful computations in $\mathcal{A}$ and successful computations in $\mathcal{B}$ of a specific shape.

Proposition C42. *Let $\mathcal{B}$ be a topological coding of $\mathcal{A}$.*

(a) *Let $\pi_{\mathcal{A}}$ be a successful computation in $\mathcal{A}$, then the corresponding computation $\pi_{\mathcal{B}}$ in $\mathcal{B}$ is successful.*

(b) *Let $\pi_{\mathcal{B}}$ be a successful computation in $\mathcal{B}$ such that $\mathrm{LBL}(\pi_{\mathcal{B}}) = u_i\,\lambda(u)\,u_f$ for some $u \in \Sigma_{\mathcal{A}}^*$. Then there exists a successful computation in $\mathcal{A}$, of which $\pi_{\mathcal{B}}$ is the corresponding computation.*

(c) *For every word $u \in \Sigma_{\mathcal{A}}^*$, u is accepted by $\mathcal{A}$ if and only if $u_i\,\lambda(u)\,u_f$ is accepted by $\mathcal{B}$.*

Proof. Item (a) follows from Definition C38(d) which ensures that $\mathrm{SRC}(\pi_{\mathcal{B}}) \in I_{\mathcal{B}}$ and Definition C38(f) which ensures that $\mathrm{TGT}(\pi_{\mathcal{B}}) \in F_{\mathcal{B}}$.

Item (b). We write $u = a_1 \cdots a_k$. The computation $\pi_{\mathcal{B}}$ may be factorised as $\pi_{\mathcal{B}} = \pi_0 \pi_1 \cdots \pi_k \pi_{k+1}$ with $\mathrm{LBL}(\pi_0) = u_i$, $\mathrm{LBL}(\pi_{k+1}) = u_f$, and for every i, $0 < i \leq k$, $\mathrm{LBL}(\pi_i) = \lambda(a_i)$. Since $\pi_{\mathcal{B}}$ is successful, π_0 is such that

$$\mathrm{SRC}(\pi_0) \in I_{\mathcal{B}} \quad \text{and} \quad \mathrm{LBL}(\pi_0) = u_i$$

which implies that $\pi_0 \in W_{\mathcal{B}}^{\mathrm{initial}}$ by definition, and $\mathrm{TGT}(\pi_0) \in \mathrm{IM}(\nu)$ from the first item of Condition C38(d).

Then a simple induction using the third item of Condition C38(e) yields that for every i, $0 < i \leq k$, $\pi_i \in W_{\mathcal{B}}^{\text{transition}}$ and $\textsc{Tgt}(\pi_i) \in \textsc{Im}(\nu)$. Finally since π_{k+1} is such that

$$\textsc{Src}(\pi_{k+1}) \in \textsc{Im}(\nu) \quad \text{and} \quad \textsc{Lbl}(\pi_{k+1}) = u_f$$

it holds $\pi_{k+1} \in W_{\mathcal{B}}^{\text{final}}$. We let $\pi_{\mathcal{A}}$ denote the following computation.

$$\pi_{\mathcal{A}} = \eta_i^{-1}(\pi_0) \cdot \eta^{-1}(\pi_1) \cdots \eta^{-1}(\pi_k) \cdot \eta_f^{-1}(\pi_{k+1})$$

Conditions C38(d), C38(e) and C38(f) ensure that $\pi_{\mathcal{A}}$ is a well-defined computation in $\mathcal{A}$ and that it is successful.

Item (c) follows directly from (a) and (b). □

C4 Main statement

Theorem C43. *Let $\mathcal{A} = \langle \Sigma, Q, \Delta, I, F\rangle$ be a trim automaton. There exists a regular expression R such that $Gl(R)$ is a topological coding of $\mathcal{A}$.*

Moreover, let $m = \textsc{Card}(\Delta)$. The size of R is in $O(m^2)$, there is exactly one Kleene star in R, and the number of transitions in $Gl(R)$ is in $O(m^2)$.

The remainder of section C4 is dedicated to the proof of Theorem C43.

We write $\mathcal{A} = \langle \Sigma_{\mathcal{A}}, Q_{\mathcal{A}}, \Delta_{\mathcal{A}}, I_{\mathcal{A}}, F_{\mathcal{A}}\rangle$ and $m = \textsc{Card}(\Delta_{\mathcal{A}})$. We let G denote any bijection $G : \Delta_{\mathcal{A}} \to \{1, \ldots, m\}$. We let H denote the only bijection $H : \Delta_{\mathcal{A}} \to \{1, \ldots, m\}$ that meets the following.

$$\forall e \in \Delta_{\mathcal{A}}, \quad G(e) + H(e) = m + 1 \tag{12}$$

For each state $q \in Q$ we let R_q^{left} and R_q^{right} denote the following expressions.

$$R_q^{\text{left}} = \overbrace{\varepsilon \ +}^{\text{if } q \in I_{\mathcal{A}}} \sum_{\substack{s \in Q,\ a \in \Sigma \\ e=(s,a,q) \in E}} \overbrace{a \cdots a}^{G(e) \text{ times}} \tag{13}$$

$$R_q^{\text{right}} = \overbrace{\varepsilon \ +}^{\text{if } q \in F_{\mathcal{A}}} \sum_{\substack{a \in \Sigma,\ t \in Q \\ e=(q,a,t) \in E}} \overbrace{a \cdots a}^{H(e) \text{ times}} \tag{14}$$

Finally, given a fresh letter $\sigma \notin \Sigma_{\mathcal{A}}$, we define the expression R over $\Sigma_{\mathcal{A}} \uplus \{\sigma\}$ as follows:

$$R = \left(\sum_{q \in Q} R_q^{\text{left}} \cdot \sigma \cdot R_q^{\text{right}} \right)^* \tag{15}$$

Note that for R to be well defined, we need the automaton $\mathcal{A}$ to be trim. Indeed, an automaton that is not accessible might feature a state q that is not initial and that has no incoming transition. In that case, the subexpression R_q^{left} would be an empty sum, and the neutral element for the sum of regular expression is not allowed in our formalism for regular expressions. A similar phenomenon occurs would $\mathcal{A}$ not be coaccessible.

Now, let us show that $Gl(R)$ is a topological coding of $\mathcal{A}$. We define a particular linearisation $\bar{R}$ of R in order to be able to explicitly state the elements u_i, u_f, λ, ν, η_i, η and η_f that realise the the topological coding.

Let $\langle \bar{R}, \Gamma\rangle$ be the linearisation of R with $\bar{R}$ defined as follows and Γ defined implicitly .

$$\bar{R} = \left(\sum_{q \in Q} \bar{R}_q^{\text{left}} \cdot \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \cdot \bar{R}_q^{\text{right}} \right)^* \tag{16}$$

where:

$$\bar{R}_q^{\text{right}} = \overbrace{\varepsilon \ +}^{\text{if } q \in F_{\mathcal{A}}} \sum_{\substack{a \in \Sigma,\ t \in Q \\ e=(q,a,t) \in E}} \left[\begin{smallmatrix}a\\ e,0\end{smallmatrix}\right] \cdots \left[\begin{smallmatrix}a\\ e,H(e)-1\end{smallmatrix}\right] \tag{17}$$

$$\bar{R}_q^{\text{left}} = \overbrace{\varepsilon \ +}^{\text{if } q \in I_{\mathcal{A}}} \sum_{\substack{s \in Q,\ a \in \Sigma \\ e=(s,a,q) \in E}} \left[\begin{smallmatrix}a\\ e,H(e)\end{smallmatrix}\right] \cdots \left[\begin{smallmatrix}a\\ e,m\end{smallmatrix}\right] \tag{18}$$

Notice that, in $\bar{R}_q^{\text{left}}$ (18), there are indeed $G(e)$ concatenated atom in the member of the sum corresponding to transition e since $G(e) + H(e) = m + 1$.

In the following, $\mathcal{B} = \langle \Sigma_{\mathcal{B}}, Q_{\mathcal{B}}, \Delta_{\mathcal{B}}, I_{\mathcal{B}}, F_{\mathcal{B}}\rangle$ denotes the Glushov automaton built from the linearisation $\langle \bar{R}, \Gamma\rangle$. Intuitively, each state q of $\mathcal{A}$ is encoded by the state $\left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right]$ of $\mathcal{B}$.

Using notations from Definition C38, we now prove that $\mathcal{B}$ is a topological coding of $\mathcal{A}$. We define the words u_i, u_f as $u_i = \sigma$ and $u_f = \varepsilon$ and the function $\lambda, \nu, \eta_i, \eta, \eta_f$ as follows.

$$\begin{array}{rccl} \lambda : & \Sigma_{\mathcal{A}} & \to & \Sigma_{\mathcal{A}} \uplus \{\sigma\} \\ & a & \mapsto & a^{m+1}\sigma \end{array} \tag{19}$$

$$\begin{array}{rccl} \nu : & Q_{\mathcal{A}} & \to & Q_{\mathcal{B}} \\ & q & \mapsto & \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \end{array} \tag{20}$$

$$\begin{array}{rccl} \eta_i : & I_{\mathcal{A}} & \to & \mathsf{Comp}(\mathcal{B}) \\ & q & \mapsto & \mathsf{init} \xrightarrow{\sigma} \left[\begin{smallmatrix}q\\ \sigma\end{smallmatrix}\right] \end{array} \tag{21}$$

$$\begin{array}{rccl} \eta : & \Delta_{\mathcal{A}} & \to & \mathsf{Comp}(\mathcal{B}) \\ & e & \mapsto & \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \xrightarrow{a} \left[\begin{smallmatrix}a\\ e,0\end{smallmatrix}\right] \xrightarrow{a} \left[\begin{smallmatrix}a\\ e,1\end{smallmatrix}\right] \cdots \xrightarrow{a} \left[\begin{smallmatrix}a\\ e,m\end{smallmatrix}\right] \xrightarrow{\sigma} \left[\begin{smallmatrix}\sigma\\ q'\end{smallmatrix}\right] \end{array} \tag{22}$$

where $e = (q, a, q')$

$$\begin{array}{rccl} \eta_f : & F_{\mathcal{A}} & \to & \mathsf{Comp}(\mathcal{B}) \\ & q & \mapsto & \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \end{array} \tag{23}$$

Notice that the image of η_f are computations of length 0, hence reduced to a single state.

It remains to show that conditions of Definition C38 are satisfied.

Conditions C38(a) **to C38**(c) Condition C38 (a) is trivially satisfied, and it is easy to see that λ and ν are injective; hence Conditions C38 (b) and C38 (c) hold.

Condition C38(d) Since $I_{\mathcal{B}} = \{\,\mathsf{init}\,\}$ and $u_i = \sigma$, it follows from (9) and (6):

$$\begin{aligned} W_{\mathcal{B}}^{\text{initial}} &= \{\, w \in \mathsf{Comp}(\mathcal{B}) \mid \textsc{Src}(w) \in I_{\mathcal{B}} \text{ and } \textsc{Lbl}(w) = u_i \,\} \\ &= \left\{ \mathsf{init} \xrightarrow{\sigma} \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \;\middle|\; \begin{array}{l} q \in Q_{\mathcal{A}} \\ \varepsilon \in L(\bar{R}_q^{\text{left}}) \end{array} \right\} \\ &= \left\{ \mathsf{init} \xrightarrow{\sigma} \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \;\middle|\; q \in I_{\mathcal{A}} \right\} \\ &= \textsc{Im}(\eta_i) \end{aligned}$$

Hence η_i is a bijection $I_{\mathcal{A}} \to W_{\mathcal{B}}^{\text{initial}}$. Then, it may be verified that Condition C38 (d) is met by definition.

Condition C38(e) Showing this condition amounts to showing the following.

$$W_{\mathcal{B}}^{\text{transition}} = \textsc{Im}(\eta) \tag{24}$$

Indeed, all other requirements follow from the definition of η. It is clear that $\textsc{Im}(\eta) \subseteq W_{\mathcal{B}}^{\text{transition}}$ so let us show the other direction. Let $w \in W_{\mathcal{B}}^{\text{transition}}$, hence

- there exists $q_1 \in Q_{\mathcal{A}}$ such that $\textsc{Src}(w) = \left[\begin{smallmatrix}\sigma\\ q_1\end{smallmatrix}\right]$
- there exists $a \in \Sigma_{\mathcal{A}}$ such that $\textsc{Lbl}(w) = a^{m+1}\sigma$
- since $\textsc{Lbl}(w)$ ends with the letter σ, there exists $q_2 \in Q_{\mathcal{A}}$ such that $\textsc{Tgt}(w) = \left[\begin{smallmatrix}\sigma\\ q_2\end{smallmatrix}\right]$.

Thus, the computation w is of the form $w = w_1 w_2 w_3$ with $\textsc{Lbl}(w_1) \in L(\bar{R}_{q_1}^{\text{right}})$, $\textsc{Lbl}(w_2) \in L(\bar{R}_{q_2}^{\text{left}})$ and $\textsc{Lbl}(w_3) = \sigma$. Let $\ell_1, \ell_2 \in \mathbb{N}$ such that $\textsc{Lbl}(w_1) = a^{\ell_1}$ and $\textsc{Lbl}(w_2) = a^{\ell_2}$.

If $\ell_1 > 0$ and $\ell_2 > 0$, there exist $e_1 = (q_1, a, t) \in \Delta_{\mathcal{A}}$ and $e_2 = (s, a, q_2) \in \Delta_{\mathcal{A}}$ that satisfy:

$$\begin{aligned} w = \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \to \left[\begin{smallmatrix}a\\ e_1,0\end{smallmatrix}\right] \xrightarrow{a} \left[\begin{smallmatrix}a\\ e_1,1\end{smallmatrix}\right] \xrightarrow{a} \cdots \\ \cdots \xrightarrow{a} \left[\begin{smallmatrix}a\\ e_1,H(e_1)-1\end{smallmatrix}\right] \xrightarrow{a} \left[\begin{smallmatrix}a\\ e_2,H(e_2)\end{smallmatrix}\right] \xrightarrow{a} \left[\begin{smallmatrix}a\\ e_2,H(e_2)+1\end{smallmatrix}\right] \xrightarrow{a} \cdots \\ \cdots \xrightarrow{a} \left[\begin{smallmatrix}a\\ e_2,m\end{smallmatrix}\right] \xrightarrow{\sigma} \left[\begin{smallmatrix}\sigma\\ q'\end{smallmatrix}\right] \end{aligned}$$

The only way for the label of w to be $a^{m+1}\sigma$ is if $H(e_1) = H(e_2)$, that is if $e_1 = e_2$. It follows that $(q_1, a, q_2) \in \Delta_{\mathcal{A}}$ and one may verify that $w = \eta((q_1, a, q_2))$.

Otherwise, if $\ell_1 = 0$, then $\ell_2 = m + 1$. This is a contradiction with $\textsc{Lbl}(w_2) \in L(\bar{R}_{q_2}^{\text{left}})$ because all words in $L(\bar{R}_{q_2}^{\text{left}})$ are of length at most m. The case where $\ell_2 = 0$ is impossible for similar reasons.

Condition C38(f) Since $u_f = \varepsilon$, Condition C38 (f) amounts to showing that $\{\, \left[\begin{smallmatrix}\sigma\\ q\end{smallmatrix}\right] \mid q \in F_{\mathcal{A}} \,\} \subseteq F_{\mathcal{B}}$. It is true from the definition of $\bar{R}$: indeed $\varepsilon \in L(\bar{R}_q^{\text{right}})$ if and only if $q \in F_{\mathcal{A}}$.

Condition C38(g) Let $w_1, w_2 \in \big(\textsc{Im}(\eta_i) \cup \textsc{Im}(\eta) \cup \textsc{Im}(\eta_f)\big)$ such that w_1 and w_2 have a transition or an internal state in common. The walks in $\textsc{Im}(\eta_f)$ have no transitions nor internal states, hence $w_1, w_2 \in \big(\textsc{Im}(\eta_i) \cup \textsc{Im}(\eta)\big)$. Since the walks in $\textsc{Im}(\eta_i)$ consists of a single transition and that transition is never used by any walk in $\textsc{Im}(\eta)$, then either $w_1, w_2 \in \textsc{Im}(\eta_i)$ or $w_1, w_2 \in \textsc{Im}(\eta)$. If $w_1, w_2 \in \textsc{Im}(\eta_i)$, by hypothesis w_1 and w_2 have a transition in common (since they don't have internal states) hence $w_1 = w_2$. Let us now treat the case where $w_1, w_2 \in \textsc{Im}(\eta)$. If w_1 and w_2 have an internal state in common $\left[\begin{smallmatrix}a\\ e,i\end{smallmatrix}\right]$ for some $a \in \Sigma_{\mathcal{A}}$, $e \in \Delta_{\mathcal{A}}$ and $i \in \{\,0, \ldots, m\,\}$, which implies that $w_1 = w_2 = \eta(e)$. Otherwise, w_1 and w_2 have a transition in common, and since they are both of length $m + 2$ it also means that they have an internal state in common and we may apply the previous case.

This concludes the proof of the main statement of Theorem C43. We now show the second part.

Lemma C44. *In R there are exactly $(\textsc{Card}(Q) + m(m+1))$ atoms and $(\textsc{Card}(I) + \textsc{Card}(F))$ occurrences of ε.*

Proof. Let $e = (s, a, t) \in \Delta_{\mathcal{A}}$. It gives rise to two subexpressions in $R_{\mathcal{A}}$: $a^{H(e)}$ in R_s^{right} and $a^{G(e)}$ in R_t^{left}. In total, $H(e) + G(e) = m + 1$ atoms. Moreover, there are exactly $\textsc{Card}(Q)$ occurrences of σ and $(\textsc{Card}(I) + \textsc{Card}(F))$ occurrences of ε. ☐

Since $\mathcal{A}$ is trim, $\textsc{Card}(Q) \leq m$ hence Lemma C44 yields that the size of R is in $O(m^2)$.

Lemma C45. *The only states in $Gl(R)$ that have more than one outgoing edges are in*

$$\textsc{Im}(\nu) \cup \{\, \left[\begin{smallmatrix}a\\ e,H(e)-1\end{smallmatrix}\right] \mid e \in \Delta_{\mathcal{A}} \text{ and } a = \textsc{Lbl}(e) \,\} \tag{25}$$

The only states in $Gl(R)$ that have more than one incoming edges are in

$$\textsc{Im}(\nu) \cup \{\, \left[\begin{smallmatrix}a\\ e,H(e)\end{smallmatrix}\right] \mid e \in \Delta_{\mathcal{A}} \text{ and } a = \textsc{Lbl}(e) \,\} \tag{26}$$

Corollary C46. *The number of transitions in $Gl(R)$ is in $O(m^2)$.*

C5 Application of Theorem C43 to the proof of Theorem C37

Theorem C37 follows from Theorems 24 and C43, together with the next proposition. Note that the database in the proof of Theorem 23 is simply-labelled: We say that a database $D = (\Sigma, V, E, \textsc{Src}, \textsc{Tgt}, \textsc{Lbl})$ is *simply-labelled* if $\textsc{Card}(\textsc{Lbl}(v)) = 1$ for every $v \in V$.

Proposition C47. *Let $\mathcal{A}$ and $\mathcal{B}$ be two automata such that $\mathcal{B}$ is a topological coding of $\mathcal{A}$. Let D be a simply-labelled database and w a walk in D. There exists a simply-labelled database D' and a walk w' such that* WALK MEMBERSHIP *returns true on $D, w, \mathcal{A}$ if and only if* WALK MEMBERSHIP *returns true on $D', w', \mathcal{B}$.*

Proof. We reuse notation from Definition C38 for $u_i, u_f, \lambda, \nu, \eta_i, \eta, \eta_f$. The database $D' = (\Sigma_{\mathcal{B}}, V', E', \textsc{Src}', \textsc{Tgt}', \textsc{Lbl}')$ is built from the database $D = (\Sigma_{\mathcal{A}}, V, E, \textsc{Src}, \textsc{Tgt}, \textsc{Lbl})$ as follows:

- V' contains V
- each edge $e \in E$ is replaced in D' by a walk w_e with label $\lambda(\textsc{Lbl}(e))$: $\textsc{Len}(\lambda(i))$ fresh edges are added to E' and $\textsc{Len}(\lambda(i)) - 1$ fresh nodes are added to V' for each $e \in E$.
- we add in D' one walk w_i from a fresh node S to $\textsc{Src}(w)$ and one walk w_f from $\textsc{Tgt}(w)$ to a fresh node T:

$$w_i = \mathsf{S} \xrightarrow{u_i} src(w)$$
$$w_f = \textsc{Tgt}(w) \xrightarrow{u_f} \mathsf{T}$$

That is $(\textsc{Len}(u_i) + \textsc{Len}(u_f))$ fresh nodes and edges.

The walk w' is built from w as follows: we let $e_1, e_2, \ldots, e_k$ be the edges in w and:

$$w' = w_i \cdot w_{e_1} \cdot w_{e_2} \cdots w_{e_k} \cdot w_f$$

Assume that there is $r \in [\![\mathcal{A}]\!]_{SR}(D)$ such that $\pi_D(r) = w$. In the following, we denote the vertices in the run database in column to improve readability; typically (n, s) is written $\left\langle \begin{matrix} n \\ s \end{matrix} \right\rangle$. We denote r and w as follows.

$$r = \left\langle \begin{matrix} n_0 \\ s_0 \end{matrix} \right\rangle \xrightarrow{a_1} \left\langle \begin{matrix} n_1 \\ s_1 \end{matrix} \right\rangle \xrightarrow{a_2} \cdots \xrightarrow{a_k} \left\langle \begin{matrix} n_k \\ s_k \end{matrix} \right\rangle \tag{27}$$

$$w = (n_0, e_1, n_1, \cdots, e_k, n_k) \tag{28}$$

with $s_0 \in I_{\mathcal{A}}$ and $s_0 \in F_{\mathcal{A}}$. Hence $\pi_{\mathcal{A}}$, below, is a successful computation in $\mathcal{A}$.

$$\pi_{\mathcal{A}} = s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} \cdots \xrightarrow{a_k} s_k \tag{29}$$

We denote:

$$\forall i, 0 < i \leq k, \quad \delta_i = (n_{i-1}, a_i, n_i) \tag{30}$$

Hence the corresponding computation (Definition C41), denoted by $\pi_{\mathcal{B}}$ and given below, is successful in $\mathcal{B}$ (from Lemma C42(a)).

$$\overbrace{x \xrightarrow{u_i} \nu(s_0)}^{\eta_i(s_0)} \underbrace{\xrightarrow{\lambda(a_1)} \nu(s_1)}_{\eta(\delta_1)} \xrightarrow{\lambda(a_2)} \cdots$$

$$\cdots \xrightarrow{\lambda(a_{k-1})} \underbrace{\nu(s_{k-1}) \xrightarrow{\lambda(a_k)} \nu(s_k)}_{\eta(\delta_k)} \overbrace{\xrightarrow{u_f} y}^{\eta_f(s_k)}$$

where $x = \textsc{Src}(i(s_0))$ and $y = \textsc{Tgt}(f(s_k))$. Hence r', defined below, is a run in $D' \times \mathcal{B}$.

$$r' = \left\langle \begin{matrix} \mathsf{S} \\ x \end{matrix} \right\rangle \xrightarrow[\eta_i(s_0)]{w_i} \left\langle \begin{matrix} n_0 \\ \nu(s_0) \end{matrix} \right\rangle \xrightarrow[\eta(\delta_1)]{w_{e_1}} \left\langle \begin{matrix} n_1 \\ \nu(s_1) \end{matrix} \right\rangle$$
$$\cdots \xrightarrow[\eta(\delta_k)]{w_{e_k}} \left\langle \begin{matrix} n_k \\ \nu(s_k) \end{matrix} \right\rangle \xrightarrow[\eta_f(s_k)]{w_f} \left\langle \begin{matrix} \mathsf{T} \\ y \end{matrix} \right\rangle$$

It remains to show that r' is simple. We assume that it is not simple for the sake of contradiction; let N, M be two vertices in r' such that $N = M$.

(1) If $N = \left\langle \begin{matrix} S \\ x \end{matrix} \right\rangle$ then no vertex M can be equal to it since S is a fresh vertex in D': it does not occur in w_i or w_f, nor in any w_{e_j}.

(2) The three following cases are treated in the same way: (2a) $N = \left\langle \begin{matrix} T \\ y \end{matrix} \right\rangle$; (2b) N is an internal node of $\xrightarrow[\eta_i(s_0)]{w_i}$; and (2c) N is an internal node of $\xrightarrow[\eta_f(s_k)]{w_f}$.

(3) Case where $N = \left\langle \begin{matrix} n_i \\ \nu(s_i) \end{matrix} \right\rangle$ and $M = \left\langle \begin{matrix} n_j \\ \nu(s_j) \end{matrix} \right\rangle$ for some i, j. Since ν is a bijection, $N = M$ implies $\left\langle \begin{matrix} n_i \\ s_i \end{matrix} \right\rangle = \left\langle \begin{matrix} n_j \\ s_j \end{matrix} \right\rangle$, hence $i = j$ since r is simple.

(4) Case where N is an internal vertex in $\xrightarrow[\eta(\delta_i)]{w_{e_i}}$ for some i and M is an internal vertex in $\xrightarrow[\eta(\delta_j)]{w_{e_j}}$ for some j. It implies that $\eta(\delta_i)$ and $\eta(\delta_j)$ have an internal state in common, hence that $\eta(\delta_i) = \eta(\delta_j)$ from Definition C38(g), hence that $\nu(s_i) = \nu(s_j)$. Similarly, the internal vertices in w_{e_i} and w_{e_j} were created fresh, hence $w_{e_i} = w_{e_j}$. It follows that $n_i = n_j$.

Finally, we have $\left\langle \begin{matrix} n_i \\ \nu(s_i) \end{matrix} \right\rangle = \left\langle \begin{matrix} n_j \\ \nu(s_j) \end{matrix} \right\rangle$ and we apply case (3).

(5) The last case is where, for some i, j, N is an internal vertex in $\xrightarrow[\eta(\delta_i)]{w_{e_i}}$ and $M = \left\langle \begin{matrix} n_j \\ \nu(s_j) \end{matrix} \right\rangle$. It would implies that an internal node in w_{e_i}, which is a fresh node in D', is equal to n_j, which was already in D, a contradiction.

It remains to show the converse: the existence of a simple run r in $D' \times \mathcal{B}$ implies the existence of a simple run r' in $D \times \mathcal{A}$. We use Lemma C42(b) to build a successful computation in $\mathcal{A}$ from the successful computation in $\mathcal{B}$ underlying r, and then we build a run $r' \in D \times \mathcal{A}$. In that direction, showing that r' is simple is directly implied by the fact that r is simple. □

C6 About Proposition 29

The technique developed earlier in appendix C allows to prove Proposition 29, recalled below.

Proposition 29. Tuple Membership, Tuple Multiplicity, Query Evaluation *and* Walk Membership *are computationally equivalent under binding-trail semantics and under simple-run semantics.*

Indeed, binding trail semantics is closely linked to the Glushkov automaton, and we show next how to use Theorem C37 to show one of reductions required for Proposition 29. Other reductions require similar classical graph techniques.

Proposition C48. *There exists a fixed expression R such that* Walk Membership *is NP-hard for binding-trail semantics.*

Proof. Let R be the expression and $D = (\Sigma, V, E, \textsc{Src}, \textsc{Tgt}, \textsc{Lbl})$ the database given by Theorem C37. Let D' be the database constructed from D by splitting vertices in order for every vertex in D' to have at most one incoming transition. More precisely $D' = (\Sigma, V' \cup V, E' \cup E, \textsc{Src}', \textsc{Tgt}', \textsc{Lbl}')$ where

- $V' = \{ (e, v) \in E \times V \mid \textsc{Tgt}(e) = v \}$;
- $E' = \{ ((e, v), e') \in V' \times E \mid \textsc{Src}(e') = v \}$;
- $\textsc{Src}'(((e, v), e')) = (e, v)$; $\quad \textsc{Src}'(e) = \textsc{Src}(e)$;
- $\textsc{Tgt}'(((e, v), e')) = (e', \textsc{Tgt}(e'))$; $\quad \textsc{Tgt}'(e) = (e, \textsc{Tgt}(e))$;
- $\textsc{Lbl}'(((e, v), e')) = \textsc{Lbl}(e')$; $\quad \textsc{Src}'(e) = \textsc{Src}(e)$.

The simple runs in $D \times Gl(R)$ are in bijection with the binding trails in D matching R that start with an edge in E. Indeed, the simple-run $r = ((v_0, s_0), (e_1, t_1), (v_1, s_1), \ldots, (e_n, t_n), (v_n, s_n))$ in $D \times Gl(R)$ is associated with: $(f_1, \alpha_1) \cdots (f_n, \alpha_n)$, where

- $f_1 = e_1$, $f_i = ((e_{i-1})$ for every i, $1 < i \leq n$; and
- α_n is the letter labelling transition t_n for every i, $0 < i \leq n$.

One may verify that tfae

- $(f_i, \alpha_i) = (f_j, \alpha_j)$
- $(v_i, s_i) = (v_j, s_j)$

Note also that $(v_0, s_0) = (v_i, s_i)$ implies $i = 0$: the state s_0 is necessarily the special initial state from the Glushkov Construction and thus has no incoming transition. $\square$

Appendix D: Proof of Theorem 32

The purpose of Appendix D is to show Theorem 32, restated below.

Theorem 32. WALK MEMBERSHIP *is in PTIME under binding-trail semantics when restricted to expressions with no concatenation under star. The same holds under simple-run semantics when queries are restricted to the Glushkov automata of such expressions.*

We only show the statement for Glushkov automaton; the proof is similar for binding-trail semantics.

First, Lemma D49 states that as soon as we don't allow concatenation under star, the expression may be simplified syntactically.

Lemma D49. *Let R be an expression with no concatenation under star. Then,*

- *deleting all stars that are nested inside another star,*
- *deleting every occurrence of ε that appears inside a star*

yields an expression R' such that $Gl(R) = Gl(R')$.

Lemma D49 describes a much simplified version of the algorithm to put an expression in *star-normal form* (Brügemann-Klein 1993; Sakarovitch 2021).

Proposition D50. WALK MEMBERSHIP *under simple-run semantics is in P-time if the input is $\mathcal{A} = Gl(R)$ where R is an expression with no concatenation under star.*

Proof. Let R be an expression with no concatenation under star. From Lemma D49, we may assume that R has no star nor occurrence of ε inside a star. Let $w = (n_0, e_1, n_1, \ldots, e_m, n_m)$ be a walk in the database $D = (\Sigma, V, E, \text{SRC}, \text{TGT}, \text{LBL})$.

The general strategy is to compute inductively the set S_X given below, for each subexpression X of R.

$$S_X = \left\{ (i, j) \,\middle|\, \begin{array}{l} 0 \leq i \leq j \leq m \\ (n_i, e_{i+1}, \ldots, n_j) \in [\![Gl(X)]\!]_{SR}(D) \end{array} \right\}$$

First, $S_{X \cdot Y}$ and S_{X+Y} are easy to compute in polynomial time from S_X and S_Y. Second, $S_\varepsilon = \{ (i,i) \mid 0 \leq i \leq m \}$ and, for each $a \in \Sigma$ $S_a = \{ (i, i+1) \mid 0 \leq i < m \text{ and } a \in \text{LBL}(e_{i+1}) \}$ is built in linear time. The remainder of the proof is about the last case, that is where $X = (a_1 + \cdots + a_n)^*$, for some atoms $a_1, \ldots, a_n$. Note that it is possible that $a_i = a_j$ for some $i \neq j$.

We build S_X by testing whether $(\ell, k) \in S_X$ for each ℓ, k such that $0 \leq \ell \leq k \leq m$. We now describe a polynomial time algorithm to test whether $(\ell, k) \in S_X$. For each vertex v in w, we let I_v denote the set $I_v = \big\{ i \in \{\ell, \ldots, k-1\} \,\big|\, \text{TGT}(e_i) = v \big\}$. Consider the following undirected graph $H_v = (V, U)$

- V contains $\text{CARD}(I_v) + n$ vertices, one vertex P_i for each position i in I_v, plus one vertex A_j for each atom a_j:

$$V = \{ P_i \mid i \in I_v \} \cup \{ A_j \mid 0 < j \leq n \}$$

- U contains an edge between P_i and A_j if and only if a_j is a label of e_i:

$$U = \left\{ (P_i, A_j) \,\middle|\, \begin{array}{l} i \in I_v \\ 0 < j \leq n \\ a_j \in \text{LBL}(e_i) \end{array} \right\}$$

Note that H_v is a bipartite graph. We may then use a classical algorithm to compute the maximal matching M_v of H_v in polynomial time (Cormen et al. 2009, Section 26.3)(Hopcroft and Karp 1973).

Then, one may use the different M_v's to test whether $(\ell, k) \in S_X$, as stated below.

Claim D50.1. *The following are equivalent.*

(a) *For each i, $\ell < i \leq k$, $\text{CARD}(M_{n_i}) = \text{CARD}(I_{n_i})$;*

(b) $(\ell, k) \in S_X$.

Proof of Claim D50.1. (a) $\Rightarrow$ (b). Let us consider the automaton $Gl(X)$. It has $(n+1)$ states: one initial state q_0, plus one state q_j for each atom a_j, $1 \leq j \leq n$. All states are final. It has $(n+1)n$ transitions: (q_i, a_j, q_j) for each i, j such that $0 \leq i \leq n$ and $0 < j \leq n$.

Let us construct a run r such that $\pi_D(r) = (n_\ell, e_{\ell+1}, \ldots, n_k)$. It is defined by

$$r = \left\langle \begin{matrix} n_\ell \\ s_\ell \end{matrix} \right\rangle \xrightarrow[b_{\ell+1}]{e_{\ell+1}} \left\langle \begin{matrix} n_{\ell+1} \\ s_{\ell+1} \end{matrix} \right\rangle \xrightarrow[b_{\ell+2}]{e_{\ell+2}} \cdots \xrightarrow[b_k]{e_k} \left\langle \begin{matrix} n_k \\ s_k \end{matrix} \right\rangle$$

where $s_\ell = q_0$ and for every i, $\ell < i \leq k$, we set $s_i = q_j$, and $b_i = a_j$, where j is the index of the vertex A_j matched to vertex P_i in M_{n_i}. By construction $s_\ell \xrightarrow{b_\ell+1} \cdots \xrightarrow{b_k} s_k$ is an (accepting) computation in $Gl(X)$. For the sake of contradiction, assume that r is not simple. There exists i, i', j such that $n_i = n_{i'}$ and $s_i = s_{i'} = q_j$ hence by definition, A_j is matched to both P_i and P_i' in the matching M_{n_i}, a contradiction.

The proof of (b) $\Rightarrow$ (a) is similar: the construction of the run from the matchings is actually bijective. (The matchings built from the run are necessarily maximal since they use all P_i's.)

In order to compute S_X, we use the algorithm above for each pair (k, ℓ), $0 \leq k < \ell \leq m$; which results in a polynomial time algorithm overall. The number of subexpressions of R are in polynomial number so computing all S_X's may be done in polynomial time, and then one simply has to check that S_R contains $(0, m)$. ☐