Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Document spanners: from expressive power to decision problems

PLEASE CITE THE PUBLISHED VERSION

http://dx.doi.org/10.4230/LIPIcs.ICDT.2016.17

PUBLISHER

Schloss Dagstuhl – Leibniz Center for Informatics

VERSION

VoR (Version of Record)

PUBLISHER STATEMENT

This work is made available according to the conditions of the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Full details of this licence are available at: http://creativecommons.org/licenses/ by/4.0/

LICENCE

CC BY-NC-ND 4.0

REPOSITORY RECORD

Freydenberger, Dominik D., and Mario Holldack. 2019. “Document Spanners: From Expressive Power to Decision Problems”. figshare. https://hdl.handle.net/2134/24360.

Document Spanners: From Expressive Power to Decision Problems

Dominik D. Freydenberger[*1] **and Mario Holldack**[2]

1 **University of Bayreuth, Bayreuth, Germany**
2 **Goethe University, Frankfurt am Main, Germany**

Abstract

We examine *document spanners*, a formal framework for information extraction that was introduced by Fagin et al. (PODS 2013). A document spanner is a function that maps an input string to a relation over *spans* (intervals of positions of the string). We focus on document spanners that are defined by *regex formulas*, which are basically regular expressions that map matched subexpressions to corresponding spans, and on *core spanners*, which extend the former by standard algebraic operators and string equality selection.

First, we compare the expressive power of core spanners to three models – namely, *patterns*, *word equations*, and a rich and natural subclass of *extended regular expressions* (regular expressions with a repetition operator). These results are then used to analyze the complexity of query evaluation and various aspects of static analysis of core spanners. Finally, we examine the relative succinctness of different kinds of representations of core spanners and relate this to the simplification of core spanners that are extended with difference operators.

1998 ACM Subject Classification H.2.1 Data models, H.2.4 Textual databases, Relational databases, Rule-based databases, F.4.3 Classes defined by grammars or automata, Decision problems, F.1.1 Relations between models

Keywords and phrases Information extraction, document spanners, regular expressions, regex, patterns, word equations, decision problems, descriptional complexity

Digital Object Identifier 10.4230/LIPIcs.ICDT.2016.17

1 Introduction

Information Extraction (IE) is the task of automatically extracting structured information from texts. This paper examines *document spanners*, a formalization of the IE query language AQL, which is used in IBM's SystemT. Document spanners were introduced by Fagin et al. [7] in order to allow the theoretical examination of AQL, and were also used in [6].

A *span* is an interval on positions of a string w, and a *spanner* is a function that maps w to a relation over spans of w. A central topic of [7] and of the present paper are *core spanners*. The primitive building blocks of core spanners are *regex formulas*, which are regular expressions with variables. Each of these variables corresponds to a subexpression, and whenever a regex formula α matches a string w, each variable is mapped to the span in w that matches that subexpression. Hence, each match of α on w determines a tuple of spans; and as there can be multiple matches of a regex formula to a string, this process creates a relation over spans of w. Core spanners are then defined by extending regex formulas with the relational operations projection, union, natural join, and string equality selection.

* Supported by Deutsche Forschungsgemeinschaft (DFG) under grant FR 3551/1-1.



© Dominik D. Freydenberger and Mario Holldack;
licensed under Creative Commons License CC-BY



One of the two main topics of the present paper is the examination of decision problems for core spanners, in particular evaluation and static analysis. These results are mostly derived from the other main topic, the examination of the expressive power of core spanners in relation to three other models that use repetition operators, which act similar to the spanners' string equality selection.

The first of these models are *patterns*. A pattern is word that consists of variables and terminals, and generates the language of all words that can be obtained by substitution of the variables with arbitrary terminal words. For example, the pattern $\alpha = xx\mathtt{ab}y$ (where x and y are variables, and $\mathtt{a}$ and $\mathtt{b}$ are terminals) generates the language of all words that have a prefix that consists of a square, followed by the word $\mathtt{ab}$. Although pattern languages have a simple definition, various decision problems for them are surprisingly hard. For example, their membership problem is NP-complete (cf. Jiang et al. [19]), and their inclusion problem is undecidable (cf. Bremer and Freydenberger [3]). As we show that core spanners can recognize pattern languages, this allows us to conclude that evaluation of core spanners is NP-hard, and that spanner containment is undecidable.

The second model we consider are *word equations*, which are equations of the form $\alpha = \beta$, where α and β are patterns, which can be used to define word relations. We show that word equations with regular constraints can express all relations that are expressible with core spanners. By using an improved version of Makanin's algorithm (cf. Diekert [5]), this allows us to show that satisfiability and hierarchicality for core spanners can be decided in PSPACE. Moreover, using coding techniques from word equations, we show that two common relations from combinatorics on words can be selected with core spanners.

The third model are *regexes* (also called *extended regular expressions* in literature). These are regular expressions that can use a repetition operator, that is available in most modern implementations for regular expressions (see, e. g., Friedl [14]) and that allows the definition of non-regular languages. For example, the regex $x\{\Sigma^*\}\&x\&x$ generates all words www with $w \in \Sigma^*$, as $x\{\Sigma^*\}$ generates some word w which is stored in the variable x, and each occurrence of $\&x$ repeats that w. As a consequence of this increase in expressive power, many decision problems are harder for regexes than for their "classical" counterparts. In particular, various problems of static analysis are undecidable (Freydenberger [11]).

But as shown by Fagin et al. [7], document spanners cannot define all languages that are definable by regexes. Intuitively, the reason for this is that regexes can use their repetition operators inside a Kleene star, which allows them to repeat an arbitrary word an unbounded number of times, while core spanners have to express repetitions with variables and string equality selections. Inspired by this observation, we introduce *variable-star free (or vstar-free) regexes* as those regexes that neither define nor use variables inside a Kleene star. We show that every vstar-free regex can be converted into an equivalent core spanner. Since all undecidability results by Freydenberger [11] also apply to vstar-free regexes, these undecidability results carry over to core spanners. This also has various consequences to the minimization and the relative succinctness of classes of spanner representations, and to the simplification of core spanners with difference operators. As a further contribution, we also develop tools to prove inexpressibility for vstar-free regular expressions and for core spanners.

As we shall see, many of the observed lower bounds hold even for comparatively restricted classes of core spanners (in particular, most of the results hold for spanners that do not use join). Hence, the authors consider it reasonable to expect that these results can be easily adapted to other information extraction languages that combine regular expressions with capture variables and a string equality operator.

In addition to regex formulas, Fagin et al. [7] also consider two types of automata as basic building blocks of spanner representations. While the present paper does not discuss these

in detail, most of the results on spanner representations that are based on regex formulas can be directly converted to the respective class of spanner representations that are based on automata.

Related work. For an overview of related models, we refer to Fagin et al. [7]. In addition to this, we highlight connections to models with similar properties. In [7], Fagin et al. showed that there is a language that can be defined by regexes, but not by core spanners. Furthermore, they compared the expressive power of core spanners and a variant of conjunctive regular path queries (CRPQs), a graph querying language. Barceló et al. [1] introduced extended CRPQs (ECRPQs), which can compare paths in the graph with regular relations. While there is no direct connection between ECRPQs and core spanners, both models share the basic idea of combining regular languages with a comparison operator that can express string equality. As shown by Freydenberger and Schweikardt [13], ECRPQs have undecidability results that are comparable to those in the present paper, and to those for regexes (cf. Freydenberger [11]). Furthermore, Barceló and Muñoz [2] have used word equations with regular constraints for variants of CRPQs.

Structure of the paper. In Section 2, we give definitions of regexes and of core spanners. Section 3 compares the expressive power of core spanners to patterns, word equations, and vstar-free regular expressions. The results from this section are then used in Section 4 to examine the complexity of evaluation and static analysis of spanners. We also examine the consequences of these results to the relative succinctness of different spanner representations. Section 5 concludes the paper. Due to space reasons, all proofs were moved to an appendix that is contained in the full version of the paper.

2 Preliminaries

Let $\mathbb{N}$ and $\mathbb{N}_{>0}$ be the sets of non-negative and positive integers, respectively. Let Σ be a fixed finite alphabet of *(terminal) symbols*. Except when stated otherwise, we assume $|\Sigma| \geq 2$. We use ε to denote the *empty word*. For every word $w \in \Sigma^*$ and every $a \in \Sigma$, let $|w|$ denote the length of w, and $|w|_a$ the number of occurrences of a in w. A word $x \in \Sigma^*$ is a *subword* of a word $y \in \Sigma^*$ if there exist $u, v \in \Sigma^*$ with $y = uxv$. A word $x \in \Sigma^*$ is a *prefix* of a word $y \in \Sigma^*$ if there exists a $v \in \Sigma^*$ with $y = xv$, and a *proper prefix* if it is a prefix and $x \neq y$. For every $n \in \mathbb{N}$, an *n-ary word relation (over Σ)* is a subset of $(\Sigma^*)^n$.

2.1 Regexes (Extended Regular Expressions)

This section introduces the syntax and semantics of regexes, which we shall also use for spanners in Section 2.2. We begin with the syntax, which follows the definition from [7].

▶ **Definition 1.** We fix an infinite set X of *variables* and define the set M of *meta symbols* as $M := \{\varepsilon, \emptyset, (,), \{, \}, \cdot, \vee, ^*, \&\}$. Let Σ, X, and M be pairwise disjoint. The set of *regexes (extended regular expressions)* is defined as follows:

1. The symbols $\emptyset$, ε, and every $a \in \Sigma$ are regexes.
2. If α_1 and α_2 are regex, then $(\alpha_1 \cdot \alpha_2)$ *(concatenation)*, $(\alpha_1 \vee \alpha_2)$ *(disjunction)*, and (α_1^*) *(Kleene star)* are regexes.
3. For every $x \in X$ and every regex α that contains neither $x\{\cdots\}$ nor $\&x$ as a subword, $x\{\alpha\}$ is a regex *(variable binding)*.
4. For every $x \in X$, $\&x$ is a regex *(variable reference)*.

If a subword β of a regex α is a regex itself, we call β a *subexpression (of α)*. The set of all subexpressions of α is denoted by $\mathsf{Sub}(\alpha)$, and the set of variables occurring in variable bindings in a regex α is denoted by $\mathsf{Vars}(\alpha)$. If a regex α contains neither variable references, nor variable bindings, we call α a *proper regular expression*.

In other words, we use the term "proper" to distinguish those expressions that are usually just called "regular expressions" from the more general extended regular expressions. We use the notation α^+ as a shorthand for $\alpha \cdot \alpha^*$. Parentheses can be added freely. We may also omit parentheses and the concatenation operator, where we assume $*$ and $+$ are taking precedence over concatenation, and concatenation precedes disjunction. Furthermore, we use Σ as a shorthand for the regular expression $\bigvee_{a \in \Sigma} a$.

Before introducing the semantics of regexes formally, we give an intuitive explanation. An expression of the form $\alpha = x\{\beta\}$ matches the same strings as β, but α additionally stores the matched string in the variable x. Using a variable reference $\&x$, this string can then be repeated. For example, let $\alpha := (x\{\Sigma^*\} \cdot \&x)$. The subexpression $x\{\Sigma^*\}$ matches any string $w \in \Sigma^*$ and stores this match in x. The following variable reference $\&x$ repeats the stored w. Thus, α defines the (non-regular) *copy-language* $\{ww \mid w \in \Sigma^*\}$.

The following definition of the semantics of regexes is based on the semantics by Freydenberger [11], which is an adaption of the semantics from Câmpeanu et al. [4] (the former uses variables, the latter backreferences). In comparison to [11], the case for Kleene star has been changed, in order to make the definition compatible with the parse trees from Fagin et al. [7].

▶ **Definition 2.** Let γ be a regex over Σ and X. A *γ-parse tree* is a finite, directed, and ordered tree T_γ. Its nodes are labeled with tuples of the form $(w, \gamma') \in (\Sigma^* \times \mathsf{Sub}(\gamma))$. The root of every γ-parse tree T_γ is labeled with (w, γ), $w \in \Sigma^*$; and the following rules must hold for each node v of T_γ:

1. If v is labeled (w, a) with $a \in (\Sigma \cup \{\varepsilon\})$, then v is a leaf, and $w = a$.
2. If v is labeled $(w, (\beta_1 \cdot \beta_2))$, then v has exactly one left child v_1 and exactly one right child v_2 with respective labels (w_1, β_1) and (w_2, β_2), and $w = w_1 w_2$.
3. If v is labeled $(w, (\beta_1 \vee \beta_2))$, then v has a single child, labeled (w, β_1) or (w, β_2).
4. If v is labeled (w, β^*), then one of the following cases holds:
   - (a) $w = \varepsilon$, and v is a leaf, or
   - (b) $w = w_1 w_2 \dots w_k$ for words $w_1, \dots, w_k \in \Sigma^+$ (with $k \geq 1$), and v has k children $v_1, \dots, v_k$ (ordered from left to right) that are labeled $(w_1, \beta), \dots, (w_k, \beta)$.
3. If v is labeled $(w, x\{\beta\})$, then v has a single child, labeled (w, β).
4. If v is labeled $(w, \&x)$, let $\prec$ denote the post-order of the nodes of T_γ (that results from a left-to-right, depth-first traversal). Then one of the following cases applies:
   - (a) If there is no node v' with $v' \prec v$ that is labeled $(w', x\{\beta'\}) \in \Sigma^* \times \mathsf{Sub}(\gamma)$, then v is a leaf, and $w = \varepsilon$.
   - (b) Otherwise, let v' be the node with $v' \prec v$ that is $\prec$-maximal among nodes labeled $(w', x\{\beta'\})$. Then v is a leaf, and $w = w'$.

If the root of a γ-parse tree T_γ is labeled (w, γ), we call T_γ a *γ-parse tree for w*. If the context is clear, we omit γ and call T_γ a parse tree.

There is no parse tree for $\emptyset$, and references to unbound variables (i. e., variables that were not assigned a value with a variable binding operator) default to ε. For an example of a parse tree, see Figure 1.

We use parse trees to define the semantics of regexes:

▶ **Definition 3.** A regex γ recognizes the language $\mathcal{L}(\gamma)$ of all $w \in \Sigma^*$ for which there exists a γ-parse tree T_γ with (w, γ) as root label.

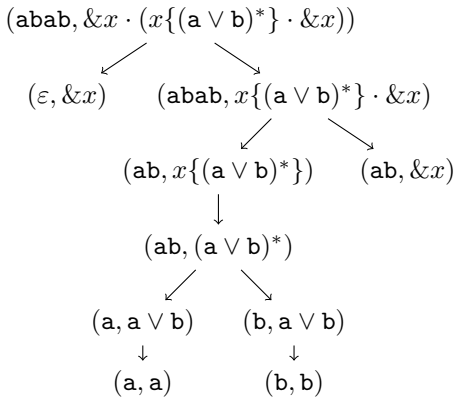


Figure 1 The α-parse tree for w, where $\alpha := \&x \cdot (x\{(\mathtt{a} \vee \mathtt{b})^*\} \cdot \&x)$ and $w := \mathtt{abab}$.

▶ **Example 4.** Let $\alpha := x\{\Sigma^+\} \cdot (\&x)^+$. Then $\mathcal{L}(\alpha) = \{w^n \mid w \in \Sigma^+, n \geq 2\}$. Furthermore, let $\beta := x\{\Sigma^+\} \cdot \&x \cdot x\{\Sigma^+\} \cdot \&x$. Then $\mathcal{L}(\beta) = \{x_1x_1x_2x_2 \mid x_1, x_2 \in \Sigma^+\}$. Finally, for some $a \in \Sigma$, let $\gamma := x\{aa^+\} \cdot (\&x)^+$. Then $\mathcal{L}(\gamma) = \{a^n \mid n \geq 2, n \text{ is not prime}\}$.

2.2 Document Spanners

Let $w := a_1a_2 \cdots a_n$ be a word over Σ, with $n \in \mathbb{N}$ and $a_1, \ldots, a_n \in \Sigma$. A *span of* w is an interval $[i, j\rangle$ with $1 \leq i \leq j \leq n + 1$ and $i, j \in \mathbb{N}$. For each span $[i, j\rangle$ of w, we define a subword $w_{[i,j\rangle} := a_i \cdots a_{j-1}$. In other words, each span describes a subword of w by its bounding indices. Two spans $[i, j\rangle$ and $[i', j'\rangle$ of w are equal if and only if $i = i'$ and $j = j'$. These spans *overlap* if $i \leq i' < j$ or $i' \leq i < j'$, and are *disjoint*, otherwise. The span $[i, j\rangle$ *contains* the span $[i', j'\rangle$ if $i \leq i' \leq j' \leq j$. The *set of all spans of* w is denoted by $\mathsf{Spans}(w)$.

▶ **Example 5.** Let $w := \mathtt{aabbcabaa}$. As $|w| = 9$, both $[3, 3\rangle$ and $[5, 5\rangle$ are spans of w, but $[10, 11\rangle$ is not. As $3 \neq 5$, the first two spans are not equal, even though $w_{[3,3\rangle} = w_{[5,5\rangle} = \varepsilon$. The whole word w is described by the span $[1, 10\rangle$.

▶ **Definition 6.** Let SVars be a fixed, infinite set of *span variables*, where Σ and SVars are disjoint. Let $V \subset \mathsf{SVars}$ be a finite subset of SVars, and let $w \in \Sigma^*$. A (V, w)-*tuple* is a function $\mu \colon V \to \mathsf{Spans}(w)$, that maps each variable in V to a span of w. If context allows, we write w-tuple instead of (V, w)-tuple. A set of (V, w)-tuples is called a (V, w)-*relation*.

As V and $\mathsf{Spans}(w)$ are finite, every (V, w)-relation is finite by definition. Our next step is the definition of spanners, which map words w to (V, w)-relations:

▶ **Definition 7.** Let V and Σ be alphabets of variables and symbols, respectively. A *(document) spanner* is a function P that maps every word $w \in \Sigma^*$ to a (V, w)-relation $P(w)$. Let V be denoted by $\mathsf{SVars}(P)$. A spanner P is n-*ary* if $|\mathsf{SVars}(P)| = n$, and *Boolean* if $\mathsf{SVars}(P) = \emptyset$. For all $w \in \Sigma^*$, we say $P(w) = \mathtt{True}$ and $P(w) = \mathtt{False}$ instead of $P(w) = \{()\}$ and $P(w) = \emptyset$, respectively. Let P be a spanner and $w \in \Sigma^*$. A w-tuple $\mu \in P(w)$ is *hierarchical* if for all $x, y \in \mathsf{SVars}(P)$ at least one of the following holds:

1. The span $\mu(x)$ contains $\mu(y)$,
2. the span $\mu(y)$ contains $\mu(x)$, or
3. the spans $\mu(x)$ and $\mu(y)$ are disjoint.

A spanner P is *hierarchical* if, for every $w \in \Sigma^*$, every $\mu \in P(w)$ is hierarchical.

A spanner P is *total on* w if $P(w)$ contains all w-tuples over $\mathsf{SVars}(P)$. Let $Y \subset \mathsf{SVars}$ be a finite set of variables. The *universal spanner over* Y is denoted by Υ_Y. It is the unique spanner P' such that $\mathsf{SVars}(P') = Y$ and P' is total on every $w \in \Sigma^*$. Furthermore, a spanner P is *hierarchical total on* w if $P(w)$ is exactly the set of all hierarchical w-tuples over $\mathsf{SVars}(P)$; and the *universal hierarchical spanner* over a set Y is the unique spanner $\Upsilon_Y^{\mathbf{H}}$ that is hierarchical total on every $w \in \Sigma^*$.

For two spanners P_1 and P_2, we write $P_1 \subseteq P_2$ if $P_1(w) \subseteq P_2(w)$ for every $w \in \Sigma^*$, and $P_1 = P_2$ if $P_1(w) = P_2(w)$ for every $w \in \Sigma^*$.

Hence, a spanner can be understood as a function that maps a word w to a set of functions, each of which assigns spans of w to the variables of the spanner. As Boolean spanners are functions that map words to truth values, they can be interpreted as characteristic functions of languages. For every Boolean spanner P, we define the *language recognized by* P as $\mathcal{L}(P) := \{w \in \Sigma^* \mid P(w) = \texttt{True}\}$. We extend this to arbitrary spanners P by $\mathcal{L}(P) := \{w \in \Sigma^* \mid P(w) \neq \emptyset\}$.

▶ **Definition 8.** A *regex formula* is a regex α over Σ and $X := \mathsf{SVars}$ such that α does not contain any variable references, and for every $\beta \in \mathsf{Sub}(\alpha)$ with $\beta = \gamma^*$, no subexpression of γ may be a variable binding.

In other words, a regex formula is a proper regular expression that is extended with variable binding operators, but these operators may not occur inside a Kleene star. We define $\mathsf{SVars}(\gamma) := \mathsf{Vars}(\gamma)$ for all regex formulas γ.

To define the semantics of regex formulas, we use the definition of parse trees for regexes, see Definition 2. Intuitively, the goal of this definition is that, each occurrence of a variable x in a γ-parse tree is matched to the corresponding span. Here, two problems can arise. Firstly, a variable might not occur in the parse tree; for example, when matching the regex formula $(x\{\mathtt{a}\} \vee \mathtt{bb})$ to the word $\mathtt{bb}$. Secondly, a variable might be defined too often, as e. g. in the regex formula $x\{\Sigma^+\} \cdot x\{\Sigma^+\}$. In order to avoid such problems, we introduce the notion of a functional regex formula.

▶ **Definition 9.** Let γ be a regex formula. We call γ *functional* if for every $w \in \Sigma^*$ and every γ-parse tree T_γ for w, each variable in $\mathsf{SVars}(\gamma)$ occurs in exactly one node label of T_γ. The class of all functional regex formulas is denoted by RGX.

As shown in Proposition 3.5 in Fagin et al. [7], functionality has a straightforward syntactic characterization: Basically, variables may not be redeclared, variables may not be used inside of Kleene stars, and if variables are used in a disjunction, each side of a disjunction has to contain exactly the same variables. Consider the following example:

▶ **Example 10.** The regex formula $\gamma_1 := (x\{\mathtt{a}\} \vee x\{\mathtt{b}\})$ is functional even though it contains two occurrences of variable definitions for x. There are just two γ_1-parse trees, both of which only contain one node labeled $(c, x\{c\})$, where $c \in \{\mathtt{a}, \mathtt{b}\}$. As a trivial case, even $\gamma_2 := x\{\emptyset\}$ is functional (as no γ_2-parse tree exists). Furthermore, the regex formulas $\gamma_3 := x\{(\mathtt{a} \vee \mathtt{b})^*\} \cdot x\{\mathtt{b}^+\}$ and $\gamma_4 := \mathtt{a}^* \vee x\{\mathtt{b}\}$ are not functional. Finally, $\gamma_5 := x\{\mathtt{a}\}^*$ is not a regex formula at all.

For functional regex formulas, we use parse trees to define the semantics:

▶ **Definition 11.** Let γ be a functional regex formula and let T be a γ-parse tree for a word $w \in \Sigma^*$. For every node v of T, the subtree that is rooted at v naturally maps to a span $p(v)$

of w. As γ is functional, for every $x \in \mathsf{SVars}(\gamma)$, exactly one node v_x of T has a label that contains x. We define $\mu^T \colon \mathsf{SVars}(\gamma) \to \mathsf{Spans}(w)$ by $\mu^T(x) := p(v_x)$. Each $\gamma \in \mathsf{RGX}$ defines a spanner $[\![\gamma]\!]$ by $[\![\gamma]\!](w) := \{\mu^T \mid T \text{ is a } \gamma\text{-parse tree for } w\}$ for each $w \in \Sigma^*$.

▶ **Example 12.** Assume that $\mathtt{a}, \mathtt{b} \in \Sigma$. We define the regex formula

$$\alpha := \Sigma^* \cdot x\big\{\mathtt{a} \cdot y\{\Sigma^*\} \cdot (z\{\mathtt{a}\} \vee z\{\mathtt{b}\})\big\} \cdot \Sigma^*.$$

Let $w := \mathtt{baaba}$. Then $[\![\alpha]\!](w)$ consists of the tuples $([2,4\rangle, [3,3\rangle, [3,4\rangle)$, $([2,5\rangle, [3,4\rangle, [4,5\rangle)$, $([2,6\rangle, [3,5\rangle, [5,6\rangle)$, $([3,5\rangle, [4,4\rangle, [4,5\rangle)$, $([3,6\rangle, [4,5\rangle, [5,6\rangle)$.

For every $w \in \Sigma^*$, a spanner P defines a (V, w)-relation $P(w)$. In order to construct more sophisticated spanners, we introduce spanner operators.

▶ **Definition 13.** Let P, P_1, P_2 be spanners and let $w \in \Sigma^*$. The algebraic operators *union*, *projection*, *natural join* and *selection* are defined as follows.

Union Two spanners P_1 and P_2 are *union compatible* if $\mathsf{SVars}(P_1) = \mathsf{SVars}(P_2)$, and their *union* $(P_1 \cup P_2)$ is defined by $\mathsf{SVars}(P_1 \cup P_2) := \mathsf{SVars}(P_1) \cup \mathsf{SVars}(P_2)$ and $(P_1 \cup P_2)(w) := P_1(w) \cup P_2(w)$ for every $w \in \Sigma^*$.

Projection Let $Y \subseteq \mathsf{SVars}(P)$. The *projection* $\pi_Y P$ is defined by $\mathsf{SVars}(\pi_Y P) := Y$ and $\pi_Y P(w) := P|_Y(w)$ for all $w \in \Sigma^*$, where $P|_Y(w)$ is the restriction of all w-tuples in $P(w)$ to Y.

Natural join Let $V_i := \mathsf{SVars}(P_i)$ for $i \in \{1, 2\}$. The *(natural) join* $(P_1 \bowtie P_2)$ of P_1 and P_2 is defined by $\mathsf{SVars}(P_1 \bowtie P_2) := \mathsf{SVars}(P_1) \cup \mathsf{SVars}(P_2)$ and, for all $w \in \Sigma^*$, $(P_1 \bowtie P_2)(w)$ is the set of all $(V_1 \cup V_2, w)$-tuples μ for which there exist (V_i, w)-tuples μ_i $(i \in \{1, 2\})$ with $\mu|_{V_1}(w) = \mu_1(w)$ and $\mu|_{V_2}(w) = \mu_2(w)$.

Selection Let $R \in (\Sigma^*)^k$ be a k-ary relation over Σ^*. The *selection operator* ζ^R is parameterized by k variables $x_1, \dots, x_k \in \mathsf{SVars}(P)$, written as $\zeta^R_{x_1,\dots,x_k}$. The *selection* $\zeta^R_{x_1,\dots,x_k} P$ is defined by $\mathsf{SVars}\big(\zeta^R_{x_1,\dots,x_k} P\big) := \mathsf{SVars}(P)$ and, for all $w \in \Sigma^*$, $\zeta^R_{x_1,\dots,x_k} P(w)$ is the set of all $\mu \in P(w)$ for which $\big(w_{\mu(x_1)}, \dots, w_{\mu(x_k)}\big) \in R$.

Like [7], we mostly consider the string equality selection operator $\zeta^=$. Hence, unless otherwise noted, the term "selection" refers to selection by the n-ary string equality relation. Note that unlike selection (which compares strings), join requires that the spans are identical.

Regarding the join of two spanners P_1 and P_2, $P_1 \bowtie P_2$ is equivalent to the intersection $P_1 \cap P_2$ if $\mathsf{SVars}(P_1) = \mathsf{SVars}(P_2)$, and to the Cartesian Product $P_1 \times P_2$ if $\mathsf{SVars}(P_1)$ and $\mathsf{SVars}(P_2)$ are disjoint. Hence, if applicable, we write $\cap$ and $\times$ instead of $\bowtie$.

For convenience, we may add and omit parentheses. We assume there is an order of precedence with projection and selection ranking over join ranking over union, e.g. we may write $\pi_Y \zeta^=_{x,y} P_1 \cup P_2 \bowtie P_3$ instead of $(\pi_Y \zeta^=_{x,y} P_1 \cup (P_2 \bowtie P_3))$, where projection and selection are applied to P_1, and the result is united with the join of P_2 and P_3.

▶ **Example 14.** Let $P_1 := \zeta^=_{x,y}[\![x\{\Sigma^*\} \cdot y\{\Sigma^*\}]\!]$ and $P_2 := \zeta^=_{x,y,z}[\![x\{\Sigma^*\} \cdot y\{\Sigma^*\} \cdot z\{\Sigma^*\}]\!]$. Then $\mathcal{L}(P_1) = \{ww \mid w \in \Sigma^*\}$, and the variables x and y always refer to the span of the first and second occurrence of w, respectively. Analogously, $\mathcal{L}(P_2) = \{www \mid w \in \Sigma^*\}$ (and z refers to the third occurrence of w). Assume that we want to construct a spanner for the language $\{w^n \mid w \in \Sigma^*, n \in \{2, 3\}\}$. As P_1 and P_2 are not union compatible, we cannot simply define $P_1 \cup P_2$. Union compatibility can be achieved by projecting P_2 to the set of common variables; i. e., $\pi_{\{x,y\}} P_2$.

▶ **Definition 15.** A *spanner algebra* is a finite set of spanner operators. If O is a spanner algebra, then $\mathsf{RGX}^{\mathsf{O}}$ denotes the set of all *spanner representations* that can be constructed

by (repeated) combination of the symbols for the operators from O with regex formulas from RGX. For each spanner representation of the form $o\rho$ (or $\rho_1\, o\, \rho_2$), where $o \in \mathsf{O}$, we define $[\![o\rho]\!] = o[\![\rho]\!]$ (and $[\![\rho_1\, o\, \rho_2]\!] = [\![\rho_1]\!]\, o\, [\![\rho_2]\!]$). Furthermore, $[\![\mathsf{RGX}^{\mathsf{O}}]\!]$ is the closure of $[\![\mathsf{RGX}]\!]$ under the spanner operators in O.

We define $\mathcal{L}(\rho) := \mathcal{L}([\![\rho]\!])$ for every spanner representation ρ. Fagin et al. [7] refer to $[\![\mathsf{RGX}]\!]$ as the class of *hierarchical regular spanners* and to $[\![\mathsf{RGX}^{\{\pi,\cup,\bowtie\}}]\!]$ as the class of *regular spanners*. In addition to (hierarchical) regular spanners, Fagin et al. also introduced the so-called *core spanners*, which are obtained by combining regex formulas with the four algebraic operators projection, selection, union, and join – in other words, the class of core spanners is the class $[\![\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}]\!]$. Analogously, $\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ is the class of *core spanner representations*.

3 Expressibility Results

3.1 Pattern Languages

We begin our examination of the expressive power of core spanners by comparing them to one of the simplest mechanisms with repetition operators:

▶ **Definition 16.** Let X be an infinite variable alphabet that is disjoint from Σ. A *pattern* is a word $\alpha \in (\Sigma \cup X)^+$ that generates the language $\mathcal{L}(\alpha) := \{\sigma(\alpha) \mid \sigma \text{ is a pattern substitution}\}$, where a *pattern substitution* is a homomorphism $\sigma\colon (\Sigma \cup X)^* \to \Sigma^*$ with $\sigma(a) = a$ for all $a \in \Sigma$. We denote the set of all variables in α by $\mathsf{Vars}(\alpha)$.

Intuitively, a pattern α generates exactly those words that can be obtained by replacing the variables in α with terminal words homomorphically (i. e., multiple occurrences of the same variable have to be replaced in the same way). This type of pattern languages is also called *erasing pattern language* (cf. Jiang et al. [19]).

▶ **Example 17.** Let $x, y \in X$ and $\mathtt{a}, \mathtt{b} \in \Sigma$. The patterns $\alpha := xx$ and $\beta := x\mathtt{a}y\mathtt{b}x$ generate the languages $\mathcal{L}(\alpha) = \{ww \mid w \in \Sigma^*\}$ and $\mathcal{L}(\beta) = \{v\mathtt{a}w\mathtt{b}v \mid v, w \in \Sigma^*\}$.

From every pattern α, we can straightforwardly construct a regex for $\mathcal{L}(\alpha)$. A similar observation holds for core spanners:

▶ **Theorem 18.** *Given a pattern α, we can compute in polynomial time a $\rho_\alpha \in \mathsf{RGX}^{\{\zeta^=\}}$ such that $\mathcal{L}(\rho_\alpha) = \mathcal{L}(\alpha)$.*

▶ **Example 19.** Let $x, y, z \in X$, $\mathtt{a}, \mathtt{b} \in \Sigma$, and define the pattern $\alpha := x\mathtt{a}yy\mathtt{b}xzx$. The construction in the proof of Theorem 18 leads to the spanner representation $\zeta^=_{x_1,x_2,x_3}\zeta^=_{y_1,y_2}\gamma$, where $\gamma = x_1\{\Sigma^*\} \cdot \mathtt{a} \cdot y_1\{\Sigma^*\} \cdot y_2\{\Sigma^*\} \cdot \mathtt{b} \cdot x_2\{\Sigma^*\} \cdot z_1\{\Sigma^*\} \cdot x_3\{\Sigma^*\}$.

While the construction in the proof of Theorem 18 is so easy that it might not seem noteworthy, it will prove quite useful: In contrast to their simple definition, many canonical decision problems for them are surprisingly hard. Via Theorem 18, the corresponding lower bounds also apply to spanners, as we discuss in Sections 4.1 and 4.2.

3.2 Word Equations and Existential Concatenation Formulas

In this section, we introduce word equations, which are equations of patterns (cf. Definition 16) and can be used to define languages and relations, cf. Karhumäki et al. [20]:

▶ **Definition 20.** A *word equation* is a pair $\eta = (\eta_L, \eta_R)$ of patterns η_L and η_R. A pattern substitution σ is a *solution* of η if $\sigma(\eta_L) = \sigma(\eta_R)$. We define $\mathsf{Vars}(\eta) := \mathsf{Vars}(\eta_L) \cup \mathsf{Vars}(\eta_R)$. For $k \geq 1$, a relation $R \subseteq (\Sigma^*)^k$ is defined by a word equation $\eta = (\eta_L, \eta_R)$ if there exist variables $x_1, \ldots, x_k \in \mathsf{Vars}(\eta)$ such that $R = \{(\sigma(x_1), \ldots, \sigma(x_k)) \mid \sigma \text{ is a solution of } \eta\}$.

We also write (η_L, η_R) as $\eta_L = \eta_R$. The following relations are well known examples of relations that are definable by word equations:

▶ **Definition 21.** Over Σ^*, we define relations $R_{\text{com}} := \{(x, y) \mid x, y \in \{u\}^* \text{ for some } u \in \Sigma^*\}$ and $R_{\text{cyc}} := \{(x, y) \mid x \text{ is a cyclic permutation of } y\}$.

As shown in Lothaire [22], the relation R_{com} is defined by the equation $xy = yx$, and R_{cyc} is defined by the equation $xz = zy$.

Let R be a k-ary string relation, and let C be a class of spanners. We say that R is *selectable* by C, if for every spanner $P \in C$ and every sequence of variables $\vec{x} = (x_1, \ldots, x_k)$ with $x_1, \ldots, x_k \in \mathsf{SVars}(P)$, the spanner $\zeta^R_{\vec{x}} P$ is also in C.

▶ **Proposition 22.** *The relations R_{com} and R_{cyc} are selectable by core spanners.*

In particular, this means that we can add $\zeta^{R_{\text{com}}}$ and $\zeta^{R_{\text{cyc}}}$ to core spanner representations, without leaving the class $[\![\mathsf{RGX}^{\{\pi, \zeta^=, \cup, \bowtie\}}]\!]$.

▶ **Example 23.** Define $L_{\text{imp}} := \{w^n \mid w \in \Sigma^+, n \geq 2\}$ and $\rho := \zeta^{R_{\text{com}}}_{x,y}(x\{\Sigma^+\}y\{\Sigma^+\})$. Then $\mathcal{L}(\rho) = L_{\text{imp}}$.

This does not imply that R_{com} can be used to select relations like $R_{\text{pow}} := \{(x, x^n) \mid n \geq 0\}$. For example, if $x := \mathtt{abab}$, $(x, y) \in R_{\text{com}}$ holds for all $y \in \{\mathtt{ab}\}^*$. The authors conjecture that R_{pow} is not selectable by core spanners.

Furthermore, the spanner that is constructed for R_{com} in the proof of Proposition 22 is more complicated than the corresponding word equation $xy = yx$. In fact, we constructed both spanners not from the equations, but from a characterization of the solutions. This appears to be necessary, due the fact that spanners need to relate their variables to an input w, while word equations use their variables without such constrictions. We shall see in Theorem 28 further down that, if this restriction is kept in mind, core spanners can be used to simulate word equations.

Before we consider this topic further, we examine how word equations can simulate spanners, as this shall provide useful insights on some question of static analysis in Section 4.2. One drawback of word equations is that they are unable to express many comparatively simple regular languages; like A^* for any non-empty $A \subset \Sigma^*$ (cf. Karhumäki et al. [20]). In order to overcome this problem, we consider the following extension:

▶ **Definition 24.** Let $\eta = (\eta_L, \eta_R)$ be a word equation. A *regular constraints function*[1] is a function Cstr that maps each $x \in \mathsf{Vars}(\eta)$ to a regular language $\mathsf{Cstr}(x)$, where each of these languages is defined by a nondeterministic finite automaton. A solution σ of η is a *solution of η under constraints* Cstr if $\sigma(x) \in \mathsf{Cstr}(x)$ holds for every $x \in \mathsf{Vars}(\eta)$.

Hence, regular constraints restrict the possible substitutions of a variable x to a regular language $\mathsf{Cstr}(x)$.

A syntactic extension of word equations are *existential concatenation formulas*, which are obtained by extending word equations with $\vee$, $\wedge$, and existential quantification over

[1] While most existing literature uses the term *rational constraints*, we follow the terminology of [2].

variables. For example, R_{cyc} is expressed by the formula $\varphi_{\text{cyc}}(x, y) := \exists z\colon (xz = zy)$. Using appropriate coding techniques, one can transform every existential concatenation formula into an equivalent word equation (see Diekert [5]). In particular, this transformation is possible in polynomial time.

Like word equations, these formulas can be further extended by adding regular constraints. For each variable x and each nondeterministic finite automaton (NFA) A, the *(regular) constraint* $L_A(x)$ is satisfied for a solution σ if $\sigma(x) \in \mathcal{L}(A)$. We call the resulting formulas *existential concatenation formulas with regular constraints*, or $\mathsf{EC}^{\mathsf{reg}}$*-formulas*.

▶ **Example 25.** Let A be an NFA with $\mathcal{L}(A) = \{\mathtt{ab}^i\mathtt{a} \mid i \geq 1\}$, and define the $\mathsf{EC}^{\mathsf{reg}}$-formula $\varphi(x, y) := \exists z\colon (L_A(z) \wedge (\exists z_1, z_2\colon x = z_1 z z_2) \wedge (\exists z_1, z_2\colon y = z_1 z z_2))$.

Then φ expresses the relation of all (x, y) that have a common subword z from $\mathcal{L}(A)$.

Using the same techniques as for formulas without constraints, $\mathsf{EC}^{\mathsf{reg}}$-formulas can be transformed into equivalent word equations with regular constraints, and this construction is possible in polynomial time as well (cf. Diekert [5]). In order to express core spanners with $\mathsf{EC}^{\mathsf{reg}}$-formulas, we introduce the following definition:

▶ **Definition 26.** Let P be a core spanner with $\mathsf{SVars}(P) = \{x_1, \dots, x_n\}$, $n \geq 0$, and let $\varphi(x_w, x_1^P, x_1^C, \dots x_n^P, x_n^C)$ be an $\mathsf{EC}^{\mathsf{reg}}$-formula. We say that φ *realizes* P if, for all $w, w_1^P, w_1^C, \dots, w_n^P, w_n^C \in \Sigma^*$, $\varphi(w, w_1^P, w_1^C, \dots, w_n^P, w_n^C) = \mathtt{True}$ holds if and only if there is a $\mu \in P(w)$ with $w_k^P = w_{[1,i_k\rangle}$ and $w_k^C = w_{[i_k,j_k\rangle}$ for each $1 \leq k \leq n$, where $[i_k, j_k\rangle = \mu(x_k)$.

This definition uses the fact that spans are always defined in relation to a word w. Note that every span $[i, j\rangle \in \mathsf{Spans}(w)$ is characterized by the words $w_{[1,i\rangle}$ and $w_{[i,j\rangle}$. Hence, if $\mu \in [\![\rho]\!](w)$, the $\mathsf{EC}^{\mathsf{reg}}$-formula models $\mu(x_k) = [i_k, j_k\rangle$ by mapping x_w to w, x_k^P to $w_{[1,i_k\rangle}$, and x_k^C to $w_{[i_k,j_k\rangle}$. In the naming of the variables, C stands for *content*, and P for *prefix*. This allows us to model spanners in $\mathsf{EC}^{\mathsf{reg}}$-formulas:

▶ **Theorem 27.** *Given $\rho \in \mathsf{RGX}^{\{\pi, \zeta^=, \cup, \bowtie\}}$ with $\mathsf{SVars}(\rho) = \{x_1, \dots, x_n\}$, $n \geq 0$, we can compute in polynomial time an $\mathsf{EC}^{\mathsf{reg}}$-formula $\varphi_\rho(x_w, x_1^P, x_1^C, \dots x_n^P, x_n^C)$ that realizes $[\![\rho]\!]$.*

As we shall see in Section 4.2, this result allows us to find upper bounds on two problems from the static analysis of spanners. We now examine how spanners can simulate word equations (and, thereby, also $\mathsf{EC}^{\mathsf{reg}}$-formulas). As discussed above, spanners need to relate their variables to an input word. Hence, we only state the following result, which is a weaker form of simulation than for the other direction:

▶ **Theorem 28.** *Every word equation $\eta = (\eta_L, \eta_R)$ with regular constraints Cstr can be converted computably into a $\rho \in \mathsf{RGX}^{\{\zeta^=, \bowtie\}}$ with $\mathsf{SVars}(\rho) \subseteq \mathsf{Vars}(\eta)$ such that for all $w \in \Sigma^*$, there is a solution σ of η under constraints Cstr with $w = \sigma(\eta_L) = \sigma(\eta_R)$ if and only if there is a $\mu \in [\![\rho]\!](w)$ with $\sigma(x) = w_{\mu(x)}$ for all $x \in \mathsf{Vars}(\eta)$.*

While this form of simulation is weaker (as w has to be present), it still shows that the constructed spanner is satisfiable if and only if the word equation (with constraints) is satisfiable; and computed (V, w)-relation encodes solutions of the equation.

▶ **Example 29.** Let $\mathtt{a}, \mathtt{b} \in \Sigma$ and define $\eta := (xy, yx)$ with $\mathsf{Cstr}(x) = \mathcal{L}(\mathtt{aab}^+)$, $\mathsf{Cstr}(y) = \Sigma^+$. The construction from the proof of Theorem 28 results in $\rho := \zeta^=_{x,x_2}\zeta^=_{y,y_2}(\hat{\eta}_L \times \hat{\eta}_R)$, where $\hat{\eta}_L := x\{\mathtt{aab}^+\} \cdot y\{\Sigma^+\}$ and $\hat{\eta}_R := y_2\{\Sigma^+\} \cdot x_2\{\mathtt{aab}^+\}$.

The only reason that this construction is not necessarily possible in polynomial time is that regular constraints are specified with NFAs, while core spanners use regular expressions, which can lead to an exponential increase in the size.

There is a similar construction that does not use the join operator: By adding new variables z_1, z_2, we can construct $\hat{\rho} := \zeta^{=}_{x,x_2}\zeta^{=}_{y,y_2}\zeta^{=}_{z_1,z_2}(z_1\{\hat{\eta}_L\}z_2\{\hat{\eta}_R\})$, which behaves almost like ρ; the only difference that the solution is encoded in $w = \sigma(\eta_L \cdot \eta_R)$, instead of $\sigma(\eta_L)$.

3.3 Regexes

As shown by Fagin et al. [7], there are languages that are recognized by regexes, but not by core spanners. In order to prove this, [7] introduced the so-called "uniform-0-chunk"-language L_{uzc}: Assuming $0, 1 \in \Sigma$, L_{uzc} is defined as the language of all $w = s_1 \cdot t \cdot s_2 \cdot t \cdots s_{n-1} \cdot t \cdot s_n$, where $n > 0$, $s_1, \ldots, s_n \in \{1\}^+$, and $t \in \{0\}^+$. Then $\mathcal{L}(\alpha_{\text{uzc}}) = L_{\text{uzc}}$ holds for the regex $\alpha_{\text{uzc}} := 1^+ \cdot x\{0^*\} \cdot (1^+ \cdot \&x)^* \cdot 1^+$, but no core spanner recognizes L_{uzc}.

Considering that the syntax of regex formulas does not allow the use of variables inside a Kleene star (or plus), this inexpressibility result might be considered expected, as α_{uzc} uses variable references inside of a Kleene plus. This raises the question whether regexes that restrict variables in a similar manner can still recognize languages that core spanners cannot. In order to examine this question, we define the following subclass of regexes:

▶ **Definition 30.** A regex α is *variable star-free* (short: *vstar-free*) if, for every $\beta \in \mathsf{Sub}(\alpha)$ with $\beta = \gamma^*$, no subexpression of γ is a variable binding or a variable reference. We denote the class of all vstar-free regexes by vsfRX.

As we shall see in Theorem 36 below, every language that is recognized by a vstar-free regex is also recognized by a core spanner. While this observation might be considered not very surprising, its proof needs to deal with some technicalities. In particular, one needs to deal with expressions like $\alpha := x\{\Sigma^*\} \cdot (\&x \vee \&x\&x)$. A conversion in the spirit of Theorem 18 would need to replace the $\&x$ with distinct variables and ensure equality with selections; but as the disjunction contains subexpressions with distinct numbers of occurrences of $\&x$, we would not be able to ensure functionality of the resulting regex formula. We avoid these problems by working with the following syntactically restricted class of vstar-free regexes:

▶ **Definition 31.** An $\alpha \in \mathsf{vsfRX}$ is a *regex path* if, for every $\beta \in \mathsf{Sub}(\alpha)$ with $\beta = (\gamma_1 \vee \gamma_2)$, no subexpression of γ_1 or γ_2 is a variable binding or a variable reference. We denote the class of all regex paths by RXP.

Intuitively, a regex path $\alpha \in \mathsf{RXP}$ can be understood as a concatenation $\alpha = \alpha_1 \cdots \alpha_n$, where each α_i is either a proper regular expression, a variable reference, or a variable binding of the form $\alpha_i = x\{\hat{\alpha}\}$, where $\hat{\alpha}$ is also a regex path. By "multiplying out" disjunctions that contain variables, we can convert every vstar-free regex into a disjunction of regex paths.

▶ **Lemma 32.** *Given $\alpha \in \mathsf{vsfRX}$, we can compute $\alpha_1, \ldots, \alpha_n \in \mathsf{RXP}$ with $\mathcal{L}(\alpha) = \bigcup_{i=1}^{n} \mathcal{L}(\alpha_i)$.*

▶ **Example 33.** Let $\alpha := x\{\Sigma^*\} \cdot (\&x \vee y\{\Sigma^*\}) \cdot (\&x \vee \&y)$. Multiplying out the disjunctions, we obtain regex paths $\alpha_1 = x\{\Sigma^*\} \cdot \&x \cdot \&x$, $\alpha_2 = x\{\Sigma^*\} \cdot y\{\Sigma^*\} \cdot \&x$, $\alpha_3 = x\{\Sigma^*\} \cdot \&x \cdot \&y$, and $\alpha_4 = x\{\Sigma^*\} \cdot y\{\Sigma^*\} \cdot \&y$. Then $\mathcal{L}(\alpha) = \bigcup_{i=1}^{4} \mathcal{L}(\alpha_i)$.

This transformation process might result in an exponential number of regex paths; but as efficiency is not of concern right now, this is not a problem. Each of these regex paths is then transformed into a functional regex formula:

▶ **Lemma 34.** *Given $\alpha \in \mathsf{RXP}$, we can be compute a $\rho \in \mathsf{RGX}^{\{\pi,\zeta^{=}\}}$ with $\mathcal{L}(\rho) = \mathcal{L}(\alpha)$.*

▶ **Example 35.** Consider the regex path $\alpha := \&x \cdot x\{\Sigma^* \cdot y\{\Sigma^*\}\} \cdot \&x \cdot \&y \cdot y\{\Sigma^*\} \cdot \&x \cdot \&y$. The construction from the proof of Lemma 34 leads to the equivalent regex path $\gamma := \varepsilon \cdot x\{\Sigma^* \cdot y\{\Sigma^*\}\} \cdot \&x \cdot \&y \cdot \hat{y}\{\Sigma^*\} \cdot \&x \cdot \&\hat{y}$, from which we derive the functional regex formula

$$\delta := x\,\{\Sigma^* y\{\Sigma^*\}\}\, x_1\{\Sigma^*\} y_1\{\Sigma^*\} \hat{y}\{\Sigma^*\} x_2\{\Sigma^*\} \hat{y}_1\{\Sigma^*\},$$

which we use in the spanner representation $\rho := \pi_\emptyset \zeta^=_{x,x_1,x_2} \zeta^=_{y,y_1} \zeta^=_{\hat{y},\hat{y}_1} \delta$. Then $\mathcal{L}(\alpha) = \mathcal{L}(\rho)$.

As these spanner representations are Boolean, they are also union compatible. Hence, we can now combine Lemma 32 and Lemma 34 to observe the following.

▶ **Theorem 36.** *Given $\alpha \in$ vsfRX, we can compute a $\rho \in \mathsf{RGX}^{\{\pi,\zeta^=,\cup\}}$ with $\mathcal{L}(\rho) = \mathcal{L}(\alpha)$.*

In Section 4.2, we use Theorem 36 together with the undecidability results from [11] to obtain multiple lower bounds for static analysis problems. Theorem 36 also raises the question whether every language that is recognized by a core spanner is also recognized by a vstar-free regular expression. As we have already seen in Example 23, it is possible to express the language $L_{\text{imp}} := \{w^n \mid w \in \Sigma^+, n \geq 2\}$ with core spanners. Hence, under certain conditions, core spanners can simulate constructions like $(\&x)^*$.

While L_{imp} might seem to be an obvious witness that separates the classes of languages that are recognized by core spanners and by vstar-free regexes, proving this appears to be quite involved. Instead, we consider a related language, which allows us to use the following tool:

▶ **Definition 37.** Let $k \in \mathbb{N}_{>0}$. We call a set $A \subseteq \mathbb{N}^k$ *linear* if there exist an $r \geq 0$ and $m_0, \ldots, m_r \in \mathbb{N}^k$ with $A = \{m_0 + m_1 i_1 + m_2 i_2 + \cdots + m_r i_r \mid i_1, i_2, \ldots, i_r \in \mathbb{N}\}$. A set $A \subseteq \mathbb{N}^k$ is *semi-linear* if it is a finite union of linear sets. Assume Σ is ordered; i. e., $\Sigma = \{a_1, a_2 \ldots, a_k\}$. The *Parikh map* $\Psi\colon \Sigma^* \to \mathbb{N}^k$ is defined by $\Psi(w) := (|w|_{a_1}, |w|_{a_2}, \ldots, |w|_{a_k})$, and is extended to languages by $\Psi(L) := \{\Psi(w) \mid w \in L\}$. We call L *semi-linear* if $\Psi(L)$ is semi-linear.

According to Parikh's Theorem [24], every context-free language is semi-linear. Moreover, as shown by Ginsburg and Spanier [15], a set is semi-linear if and only if it is definable in Presburger arithmetic. Building on this, we state the following:

▶ **Theorem 38.** *For every $\alpha \in$ vsfRX, the language $\mathcal{L}(\alpha)$ is semi-linear.*

We use Theorem 38 to separate the classes of languages that are recognized by core spanners and by vstar-free regexes:

▶ **Lemma 39.** *Let $L_{\text{nsl}} := \{(\mathtt{ab}^m)^n \mid m, n \geq 2\}$ and $\rho := \zeta^{R_{\text{com}}}_{x,y}(x\{\mathtt{abb}^+\}y\{\Sigma^+\})$ for $\Sigma := \{\mathtt{a}, \mathtt{b}\}$. Then $L_{\text{nsl}} = \mathcal{L}(\rho)$, but there is no $\alpha \in$ vsfRX with $\mathcal{L}(\alpha) = L_{\text{nsl}}$.*

We do not need the join operator to define non-semi-linear languages: Consider the core spanner representation ρ from Example 29 with $\mathcal{L}(\rho) = L_{\text{nsl}}$. If we construct $\hat{\rho}$ as explained below that example, we obtain $\mathcal{L}(\hat{\rho}) = \{ww \mid w \in L_{\text{nsl}}\}$, which is also not semi-linear.

It is worth pointing out Lemma 39 does not resolve the open question from [7] whether there is a language that is recognized by a core spanner, but not by a regex, as Theorem 38 only applies to vstar-free regexes. We have already seen languages that are not semi-linear, but are recognized by regexes: The language L_{nsl} is recognized by $\alpha_{\text{nsl}} := x\{\mathtt{abb}^+\}\&x^+$; and a similar approach is used for the following language (which we already met in Example 4):

▶ **Example 40.** Let $\Sigma := \{\mathtt{a}\}$, and define the language $L_{\text{npr}} := \{\mathtt{a}^{mn} \mid m, n \geq 2\}$. In other words, L_{npr} is the language of all words $\mathtt{a}^i$ with $i \geq 4$ such that i is not a prime number. Let $\alpha_{\text{npr}} := x\{\mathtt{aa}^+\} \cdot (\&x)^+$. Then $\mathcal{L}(\alpha_{\text{npr}}) = L_{\text{npr}}$.

While L_{nsl} and L_{npr} are defined by very similar regexes, the latter cannot be recognized by core spanners. In order to show this with a semi-linearity argument, we observe:

▶ **Theorem 41.** *Let $|\Sigma| = 1$ and let P be a core spanner over Σ. Then $\mathcal{L}(P)$ is semi-linear.*

Apart from the observation that L_{npr} from Example 40 is not recognized by core spanners, Theorem 41 also allows us to conclude that on unary alphabets, core spanners recognize exactly the class of regular languages (which, on unary alphabets, is identical to the class of context-free languages).

4 Decision Problems

4.1 Spanner Evaluation

We first examine the *combined complexity* of the evaluation problem for core spanners, the problem CSp-Eval: Given a $\rho \in \mathsf{RGX}^{\{\pi,\zeta^{=},\cup,\bowtie\}}$, a $w \in \Sigma^*$, and a $(\mathsf{SVars}(\rho), w)$-tuple μ, is $\mu \in [\![\rho]\!](w)$? In order to prove lower bounds for this problem, we consider the membership problem for pattern languages: Given a pattern α and a word w, decide whether $w \in \mathcal{L}(\alpha)$. As shown by Jiang et al. [19], this problem is NP-complete. Due to Theorem 18, we observe the following (the proof of NP-membership is straightforward).

▶ **Theorem 42.** *CSp-Eval is NP-complete, even if restricted to $\mathsf{RGX}^{\{\pi,\zeta^{=}\}}$.*

The question arises whether there are natural restrictions to CSp-Eval that make this problem tractable. It appears that any subclass of the core spanners that extends regular spanners in a meaningful way while having a tractable evaluation problem can not be allowed to recognize the full class of pattern languages.

For pattern languages, it was shown by Ibarra et al. [18] that bounding the number of variables in the pattern leads to an algorithm for the membership problem with a running time that is polynomial, although in $\mathcal{O}(n^k)$ (where n is the length of the word w, and k the number of variables). From a parameterized complexity point of view, this is usually not considered satisfactory. In fact, it was first observed by Stephan et al. [26] that the membership problem for pattern languages is $W[1]$-complete if the number of variable occurrences (not of variables) is used as a parameter. As the number of variable occurrences in a pattern corresponds to the number of variables in an equivalent spanner, this implies that using the number of variables in a spanner as parameter leads to $W[1]$-hardness for this parameter of CSp-Eval.

Fernau and Schmid [9] and Fernau et al. [10] discuss various other potential restrictions to pattern languages that still do not lead to tractability (among these a bound on the length of the replacement of each variable, which corresponds to a bound on the length of spans). On the other hand, Reidenbach and Schmid [25] and Fernau et al. [8] examine parameters for patterns that make the membership problem tractable. While this does not directly translate to spanners, the authors consider these directions promising for further research.

We also consider the *data complexity* of the evaluation problem for core spanners. For every core spanner representation ρ over Σ, we define the decision problem CSp-Eval(ρ): Given a $w \in \Sigma^*$ and a w-tuple μ, is $\mu \in [\![\rho]\!](w)$? Using a slight variation of the proof of Theorem 42, we obtain the following.

▶ **Theorem 43.** *For every $\rho \in \mathsf{RGX}^{\{\pi,\zeta^{=},\cup,\bowtie\}}$, CSp-Eval($\rho$) is in NL.*

4.2 Static Analysis

We consider the following common decision problems for core spanner representations, where the input is $\rho \in \mathsf{RGX}^{\{\pi,\zeta^{=},\cup,\bowtie\}}$ or $\rho_1, \rho_2 \in \mathsf{RGX}^{\{\pi,\zeta^{=},\cup,\bowtie\}}$:

1. CSp-Sat: Is $[\![\rho]\!](w) \neq \emptyset$ for some $w \in \Sigma^*$?
2. CSp-Hierarchicality: Is $[\![\rho]\!]$ hierarchical?
3. CSp-Universality: Is $[\![\rho]\!] = \Upsilon_{\mathsf{SVars}(\rho)}$?
4. CSp-Equivalence: Is $[\![\rho_1]\!] = [\![\rho_2]\!]$?
5. CSp-Containment: Is $[\![\rho_1]\!] \subseteq [\![\rho_2]\!]$?
6. CSp-Regularity: Is $[\![\rho]\!] \in [\![\mathsf{RGX}^{\{\pi,\cup,\bowtie\}}]\!]$?

We approach the first two of these problems by using Theorem 27 to convert core spanner representations to $\mathsf{EC}^{\mathsf{reg}}$-formulas, for which satisfiability is in PSPACE (cf. Diekert [5]). Hence, we observe:

▶ **Theorem 44.** *CSp-Sat is* PSPACE*-complete, even if restricted to* $\mathsf{RGX}^{\{\zeta^=\}}$.

For the lower bound, the proof of Theorem 44 uses the PSPACE-hardness of the intersection emptiness problem for regular expressions. But even if the variables in the regex formulas were only bound to Σ^*, it follows from Theorem 28 that this problem would still be at least as hard as the satisfiability problem for word equations without constraints (cf. Diekert [5]).

Furthermore, it is possible to use $\mathsf{EC}^{\mathsf{reg}}$-formulas to express a violation of the criteria for hierarchicality. This allows us to state the following result:

▶ **Theorem 45.** *CSp-Hierarchicality is* PSPACE*-complete, even if restricted to* $\mathsf{RGX}^{\{\zeta^=,\bowtie\}}$.

For the remaining problems, we use Theorem 36, and the fact that the undecidability results from Freydenberger [11] also hold for vstar-free regexes:

▶ **Theorem 46.** *CSp-Universality and CSp-Equivalence are not semi-decidable, and CSp-Regularity is neither semi-decidable, nor co-semi-decidable. This holds even if the input is restricted to* $\mathsf{RGX}^{\{\pi,\zeta^=,\cup\}}$.

As the proof of Theorem 46 relies only on Boolean spanners, the decidability status of CSp-Regularity does not change if the problem asks for hierarchical regularity (i. e., membership in $[\![\mathsf{RGX}]\!]$) instead of regularity, as the two classes coincide for Boolean spanners. Likewise, CSp-Universality remains not semi-decidable if one replaces $\Upsilon_{\mathsf{SVars}(\rho)}$ with $\Upsilon^{\mathbf{H}}_{\mathsf{SVars}(\rho)}$.

In the construction from this proof, variables are only bound to a language a^+. Hence, the same undecidability results hold for spanners that use selections by equal length relation, instead of the string equality relation. While the proof builds on regexes $\alpha_{\mathcal{X}}$ that use only a single variable x, the resulting core spanners use an unbounded amount variables, as every occurrence of a variable reference $\&x$ in a regex path is converted to a spanner variable x_i. But undecidability remains even if we bound the number of variables in the spanners, as the $\alpha_{\mathcal{X}}$ can be modified to use only a bounded number of variable references (see Section 4.1 in [11]). Theorem 46 also implies that CSp-Containment is not semi-decidable. This holds even for a more restricted class of spanners:

▶ **Theorem 47.** *CSp-Containment is not semi-decidable, even if restricted to* $\mathsf{RGX}^{\{\pi,\zeta^=\}}$.

As shown by Bremer and Freydenberger [3], the inclusion problem for pattern languages remains undecidable if the number of variables in the patterns is bounded. In fact, that proof constructs patterns where even the number of variable occurrences is bounded. Therefore, CSp-Containment is not semi-decidable even if restricted to representations from $\mathsf{RGX}^{\{\pi,\zeta^=\}}$ with a bounded number of variables. It is a hard open question whether the equivalence problem for pattern languages is decidable (cf. Ohlebusch and Ukkonen [23], Freydenberger and Reidenbach [12]). Undecidability of this problem would imply undecidability of CSp-Equivalence, even if restricted to representations from $\mathsf{RGX}^{\{\pi,\zeta^=\}}$.

4.2.1 Minimization and Relative Succinctness

In order to address the minimization of spanner representations, we first formalize the notion of the size or complexity of a spanner representation. Even for proper regular expressions, there are various different definitions of size, see e. g. Holzer and Kutrib [17], and there might be convincing reasons to add additional weight to the number of variables or other parameters. As we shall see, these distinctions do not affect the negative results that we prove further down. Hence, instead of defining a single fixed notion of size, we use the following general definition of complexity measures from Kutrib [21]:

▶ **Definition 48.** Let SR be a class of spanner representations. A *complexity measure* for SR is a recursive function $c : \mathsf{SR} \to \mathbb{N}$ such that for each Σ, the set of all $\rho \in \mathsf{SR}$ that represent spanners over Σ can be computably enumerated in order of increasing $c(\rho)$, and does not contain infinitely many $\rho \in \mathsf{SR}$ with the same value $c(\rho)$.

By *recursive*, we mean a function that is total and computable. Definition 48 is general enough to include all notions of complexity that take into account that descriptions are commonly encoded with a finite number of distinct symbols, and that it should be decidable if a word over these symbols is a valid encoding from SR. Regardless of the chosen complexity measure, computable minimization of core spanners is impossible:

▶ **Theorem 49.** *Let c be a complexity measure for $\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$. There is no algorithm that, given a $\rho \in \mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$, computes an equivalent $\hat{\rho} \in \mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ that is c-minimal.*

In addition to regex formulas, Fagin et al. [7] also define spanner representations that are based on so-called vset- and vstk-automata (denoted by $\mathsf{VA}_{\mathsf{set}}$ and $\mathsf{VA}_{\mathsf{stk}}$) and extended with the same spanner operators; and they compare the expressive power of these spanner representations to $\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ and its subclasses. Without going into details, we note that their equivalence proofs use computable conversions between the models. Hence, Theorem 49 also applies to those spanner representations from [7] that can express core spanners, like ${\mathsf{VA}_{\mathsf{stk}}}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ and ${\mathsf{VA}_{\mathsf{set}}}^{\{\pi,\zeta^=,\cup,\bowtie\}}$, and it implies that an algorithm that converts from one of these classes of representations to another cannot guarantee that its result is minimal.

Using a technique by Hartmanis [16], we can use the fact that CSp-Regularity is not co-semi-decidable to compare the relative succinctness of regular and core spanner representations:

▶ **Theorem 50.** *Let c_1, c_2 be complexity measures for $\mathsf{RGX}^{\{\pi,\cup,\bowtie\}}$ and $\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$, respectively. For every recursive function $f : \mathbb{N} \to \mathbb{N}$, there exists a $\rho \in \mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ such that $[\![\rho]\!] \in [\![\mathsf{RGX}^{\{\pi,\cup,\bowtie\}}]\!]$, but $c_1(\hat{\rho}) > f(c_2(\rho))$ holds for every $\hat{\rho} \in \mathsf{RGX}^{\{\pi,\cup,\bowtie\}}$ with $[\![\hat{\rho}]\!] = [\![\rho]\!]$.*

Hence, the blowup from $\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ to $\mathsf{RGX}^{\{\pi,\cup,\bowtie\}}$ is not bounded by a recursive function. As above, we can replace each of this classes with a class with the same expressive power; for example, we can replace $\mathsf{RGX}^{\{\pi,\cup,\bowtie\}}$ with ${\mathsf{VA}_{\mathsf{stk}}}^{\{\pi,\cup,\bowtie\}}$, $\mathsf{VA}_{\mathsf{set}}$, or ${\mathsf{VA}_{\mathsf{set}}}^{\{\pi,\cup,\bowtie\}}$ (or, as the proof uses Boolean spanners, RGX or $\mathsf{VA}_{\mathsf{stk}}$, or any class between those).

We also consider the relative succinctness of representations of core spanners and representations of their complements. For every spanner P, we define its *complement* $\mathsf{C}(P) := \Upsilon_{\mathsf{SVars}(P)} \setminus P$, and its *hierarchical complement* $\mathsf{C}^{\mathbf{H}}(P) := \Upsilon^{\mathbf{H}}_{\mathsf{SVars}(P)} \setminus P$.

▶ **Theorem 51.** *Let c be a complexity measure for $\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$. For every recursive function $f : \mathbb{N} \to \mathbb{N}$, there exists a $\rho \in \mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ such that $\mathsf{C}([\![\rho]\!]) \in [\![\mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}]\!]$, but $c(\rho) > f(c(\hat{\rho}))$ holds for every $\hat{\rho} \in \mathsf{RGX}^{\{\pi,\zeta^=,\cup,\bowtie\}}$ with $[\![\hat{\rho}]\!] = \mathsf{C}([\![\rho]\!])$. This also holds if we consider $\mathsf{C}^{\mathbf{H}}$ instead of C.*

In other words, there are core spanners where the (hierarchical) complement is also a core spanner, but the blowup between their representations is not bounded by any recursive function. Again, this holds for the other classes of representations as well.

This result has consequences to an open question of Fagin et al. One of the central tools in [7] is the core-simplification-lemma, which states that every core spanner is definable by an expression of the form $\pi_V SA$, where A is a vset-automaton, $V \subseteq \mathsf{SVars}(A)$, and S is a sequence of selections $\zeta^{=}_{x,y}$ for $x, y \in \mathsf{SVars}(A)$.

In addition to core spanners, Fagin et al. also discuss adding a set difference operator $\setminus$, and ask "whether we can find a simple form, in the spirit of the core-simplification lemma, when adding difference to the representation of core spanners". It is a direct consequence of Theorem 51 that such a simple representation, if it exists, cannot be obtained computably, as reducing the number of difference operators can lead to a non-recursive blowup. While this observation does not prove that such a simple form does not exist, it suggests that any proof of its existence should be expected to be non-constructive.

5 Conclusions and Further Work

In Section 3, we have seen that core spanners can express languages that are defined by patterns or by vstar-free regexes. We used this in Section 4 to derive various lower bounds on decision problems, even for subclasses of core spanner representations. Note that in most of the cases, these lower bounds do not require the join operator, and mostly rely on the string equality selection. This can be interpreted as a sign that string equality (or repetition) is an expensive operator, in particular as similar results have been observed for related models (e. g., [1, 11, 13]).

On a more positive note, there is reason to hope that these connections can be beneficial for spanners: There is recent work on restricted classes of pattern languages with an efficient membership problem (e. g., [9, 25]), which could lead to subclasses of spanners that can be evaluated more efficiently. Furthermore, as Theorems 27 and 28 show, core spanners and word equations with regular constraints are closely related. Recent work on word equations has also considered tasks like enumerating all solutions of an equation. The employed compression techniques (cf. [5]) might also be used to improve the evaluation of core spanners. In particular, the $\mathsf{EC}^{\mathsf{reg}}$-formulas that are constructed in the proof of Theorem 27 have the special property that there is a variable x_w (for w), and for every solution σ and every variable x, $\sigma(x)$ is a subword of $\sigma(x_w)$. It remains to be seen whether this restriction leads to favorable lower bounds.

Also note that conversion from vstar-free regular expressions to core spanner representations that is used for Theorem 36 can lead to an exponential increase in size. If this size blowup cannot be avoided, allowing vstar-free regexes as primitive spanner representations might be useful as syntactic sugar.

Finally, while we only mentioned this explicitly in Section 4.2.1, note that most of the other results in this paper can also be directly converted to the appropriate spanner representations that use vset- and vstk-automata from [7].

Acknowledgements. We thank Florin Manea for his suggestion to use word equations with regular constraints, and Thomas Zeume for reporting a list of typos. We also thank the anonymous reviewers for their feedback.

References

1 P. Barceló, L. Libkin, A. W. Lin, and P. T. Wood. Expressive languages for path queries over graph-structured data. *ACM T. Database Syst.*, 37(4):31, 2012.

2 P. Barceló and P. Muñoz. Graph logics with rational relations: the role of word combinatorics. In *Proc. CSL-LICS 2014*, 2014.

3 J. Bremer and D. D. Freydenberger. Inclusion problems for patterns with a bounded number of variables. *Inform. Comput.*, 220–221:15–43, 2012.

4 C. Câmpeanu, K. Salomaa, and S. Yu. A formal study of practical regular expressions. *Int. J. Found. Comput. Sci.*, 14:1007–1018, 2003.

5 V. Diekert. Makanin's Algorithm. In M. Lothaire, editor, *Algebraic Combinatorics on Words*, chapter 12, pages 387–442. Cambridge University Press, 2002.

6 R. Fagin, B. Kimelfeld, F. Reiss, and S. Vansummeren. Cleaning inconsistencies in information extraction via prioritized repairs. In *Proc. PODS 2014*, 2014.

7 R. Fagin, B. Kimelfeld, F. Reiss, and S. Vansummeren. Document spanners: A formal approach to information extraction. *J. ACM*, 62(2):12, 2015.

8 H. Fernau, F. Manea, R. Mercas, and M. L. Schmid. Pattern matching with variables: Fast algorithms and new hardness results. In *Proc. STACS 2015*, 2015.

9 H. Fernau and M. L. Schmid. Pattern matching with variables: A multivariate complexity analysis. *Inf. Comput.*, 242:287–305, 2015.

10 H. Fernau, M. L. Schmid, and Y. Villanger. On the parameterised complexity of string morphism problems. *Theory Comput. Sys.*, 2015.

11 D. D. Freydenberger. Extended regular expressions: Succinctness and decidability. *Theory Comput. Sys.*, 53(2):159–193, 2013.

12 D. D. Freydenberger and D. Reidenbach. Bad news on decision problems for patterns. *Inform. Comput.*, 208(1):83–96, 2010.

13 D. D. Freydenberger and N. Schweikardt. Expressiveness and static analysis of extended conjunctive regular path queries. *J. Comput. Syst. Sci.*, 79(6):892–909, 2013.

14 J. E. F. Friedl. *Mastering Regular Expressions*. O'Reilly Media, 3rd edition, 2006.

15 S. Ginsburg and E. Spanier. Semigroups, presburger formulas, and languages. *Pac. J. Math.*, 16(2):285–296, 1966.

16 J. Hartmanis. On Gödel speed-up and succinctness of language representations. *Theor. Comput. Sci.*, 26(3):335–342, 1983.

17 M. Holzer and M. Kutrib. Descriptional complexity–an introductory survey. *Scientific Applications of Language Methods*, 2:1–58, 2010.

18 O. H. Ibarra, T.-C. Pong, and S. M. Sohn. A note on parsing pattern languages. *Pattern Recogn. Lett.*, 16(2):179–182, 1995.

19 T. Jiang, E. Kinber, A. Salomaa, K. Salomaa, and S. Yu. Pattern languages with and without erasing. *Int. J. Comput. Math.*, 50:147–163, 1994.

20 J. Karhumäki, F. Mignosi, and W. Plandowski. The expressibility of languages and relations by word equations. *J. ACM*, 47(3):483–505, 2000.

21 M. Kutrib. The phenomenon of non-recursive trade-offs. *Int. J. Found. Comput. Sci.*, 16(5):957–973, 2005.

22 M. Lothaire. *Combinatorics on Words*. Cambridge University Press, 1997.

23 E. Ohlebusch and E. Ukkonen. On the equivalence problem for E-pattern languages. *Theor. Comput. Sci.*, 186:231–248, 1997.

24 R. J. Parikh. On context-free languages. *J. ACM*, 13(4):570–581, 1966.

25 D. Reidenbach and M. L. Schmid. Patterns with bounded treewidth. *Inform. Comput.*, 239:87–99, 2014.

26 F. Stephan, R. Yoshinaka, and T. Zeugmann. On the parameterised complexity of learning patterns. In *Proc. ISCIS 2011*, 2011.