Symbolic Boolean Derivatives for Efficiently Solving Extended Regular Expression Constraints

Anonymous Author(s)

Abstract

The manipulation of raw string data is ubiquitous in security-critical software, and verification of such software relies on efficiently solving string and regular expression constraints via SMT. However, the typical case of Boolean combinations of regular expression constraints exposes blowup in existing techniques. To address solvability of such constraints, we propose a new theory of derivatives of symbolic extended regular expressions (extended meaning that complement and intersection are incorporated), and show how to apply this theory to obtain more efficient decision procedures. Our implementation of these ideas, built on top of Z3, matches or outperforms state-of-the-art solvers on standard and handwritten benchmarks, showing particular benefits on examples with Boolean combinations.

Our work is the first formalization of derivatives of regular expressions which both handles intersection and complement and works symbolically over an arbitrary character theory. It unifies existing approaches involving derivatives of extended regular expressions, alternating automata and Boolean automata by lifting them to a common symbolic platform. It relies on a parsimonious augmentation of regular expressions: a construct for symbolic conditionals is shown to be sufficient to obtain relevant closure properties for derivatives over extended regular expressions.

Keywords regex, SMT, automaton, string

1 Introduction

Regular expressions and finite automata play a fundamental role in many areas, ranging from applications in natural sciences [21] and NLP [33] to core problems in applied computer science, such as matching [19, 36, 39], model-checking [22], and solving of string constraints in SMT [23]. Recent years have seen a resurgence of interest in solvers for quantifier-free string and regular expression constraints, driven by software verification and security applications [4, 11]. However, there remains a gap between the theory of regular expressions (or regexes) and the constraints that arise in practice in such applications. We focus here on two aspects of this gap: (1) in typical applications, regexes exist over a symbolic potentially complex character theory rather than over a finite alphabet; and (2) in typical applications, multiple regex membership constraints may be combined using Boolean connectives. Modern SMT solvers thus need to efficiently solve Boolean combinations of regex constraints over a symbolic alphabet, rather than solving individual constraints in isolation over a finite one.

Although regexes are widely supported in most modern SMT string solvers [1, 5, 7, 14, 15, 20, 35, 49–51], no current state-of-the-art tool provides a satisfactory solution to both of these challenges simultaneously. With respect to (1), modern strings that arise in applications are generally written in Unicode, but as of today, no SMT solver supports even the Basic Multilingual Plane (*BMP* or also known as *Plane 0*), while most widely used regex standards, e.g., the .NET regex standard [31] are based on BMP. Additionally, regexes that arise in practice employ *character classes* such as `\w` which denotes a word character, i.e. the subset of the character space (e.g. Unicode) which includes the Latin alphabet `a-z` and other alphabetic symbols. With respect to (2), we follow existing work by defining *extended regexes* to be those that allow intersection and complement. As we will see shortly, an efficient treatment of extended regexes has eluded existing techniques.

We believe that Boolean combinations of constraints represent the norm, rather than the exception, in practice. To give one example: cloud policy languages, such as Amazon AWS policies [4] and Microsoft Azure resource manager policies [30] utilize regexes for lightweight pattern matching. For example, Figure 1 shows a combination of constraints used to match a *date format*: a string which appears like a date, such as `2020-Nov-25`. A sanity check here for SMT would be to make sure that the constraint is indeed satisfiable — for example, if we made a mistake and wrote `.*2019` and `.*2020` instead of `2019.*` and `2020.*`, then it would be unsatisfiable because the year was accidentally specified to be both at the beginning and at the end of the string. This would render this hypothetical audit policy useless (never activated) and would not match the user's intention. To combine the date constraints into a single *classical* regex (i.e., without any use of complement or intersection), is theoretically possible because regular languages are closed under Boolean operations. However, this not only might be less succinct, but interestingly, industrial policy languages actually restrict regex syntax in various ways, forcing users to write Boolean combinations. For example, both the Amazon AWS and Microsoft Azure languages, as of 2020, allow Kleene star in `.*` only (here `.*` is the regex matching any string). One rationale behind such language restrictions is to simplify the regex matcher engine implementation in order to avoid performance bottlenecks that could otherwise be exploited

PL'18, January 01–03, 2018, New York, NY, USA
2018.

```
{"if":{"allOf":[{"field":"date", "match":"####-???-##"},
                {"anyOf":[{"field":"date", "like":"2019*" },
                          {"field":"date", "like":"2020*"}]}]}
"then":{"effect":"audit"}}
```

$$\text{meaning}: \quad \begin{array}{l} \mathit{date} \in \texttt{\textbackslash d\{4\}-[a-zA-Z]\{3\}-\textbackslash d\{2\}} \;\wedge \\ (\mathit{date} \in \texttt{2019.*} \;\;\vee\;\; \mathit{date} \in \texttt{2020.*}). \end{array}$$

Figure 1. Example Boolean combination of regex constraints arising in practice: users of the Azure resource policy language [30] write a restricted form of regexes to control when a cloud resource should be audited. The semantics of the policy (top) is a Boolean combination of regex membership constraints (bottom), where # denotes a number (`\d`), ? denotes a letter (`[a-zA-Z]`), `*` denotes any sequence (`.*`), and we write `{n}` for n-fold iteration of a regex. Large Boolean combinations are either challenging or beyond reach for existing SMT string solvers (see Section 6).

for regex denial-of-service attacks (ReDoS). This makes the use of top level conjunction and complement, as used in this date example, a way to safely express more complex regular constraints, while at the same time raises the need to deal with Boolean combinations of regex contraints for analysis.

The way in which current state-of-the-art solvers deal with Boolean combinations (intersection and complement) can be summarized by two main approaches:

1. Convert a regular expression r into an automaton M_r and then propagate the logical connectives into corresponding Boolean operations over automata. Thus $(s \in r_1) \wedge (s \in r_2)$ can be converted into $s \in L(M_{r_1} \times M_{r_2})$ and $\neg(s \in r)$ can be converted into $s \in L(M_r^{\complement})$ [48].
2. Propagate the operations over regexes, by considering extended regexes, such as `(.*\d.*)&(.*[a-z].*)`, where & is intersection. Then, directly algebraically manipulate such extended regexes using *derivatives* [29].

While it is possible to extend classical automata algorithms to work modulo a character theory [18], the first approach has the following fundamental bottleneck. The construction of M_r is typically eager (the entire state space is constructed), and intersection and complement cause state space blowup for most automata models that are used. This means that constructing the state space for M_r is infeasible, such as for r = `~(.*a.{100})` (where `.*` matches any string, $\{n\}$ is n-fold repetition, and ~ is complement). This is a limitation because constructing M_r eagerly might not be needed in the first place: for example if checking satisfiability of r, it may be that an accepting state of M_r can be reached through exploration without constructing all states. On the other hand, if checking unsatisfiability of r, in product and complement constructions on automata, many more states are constructed than may actually be reachable (these can be eliminated through minimization of automata, but only after the fact). This suggests that we may be able to avoid constructing them in the first place.

On the other hand, the second approach addresses this state space blowup by leveraging *derivatives*, a syntactic way of exploring the state space of a regular expression without converting it to automata, pioneered by Brzozowski [9] and Antimorov [3]. The summary of the approach is that the derivatives of a regular expression correspond to the states of M_r, but they are constructed *lazily*. However, the second approach has another fundamental drawback: the lack of an appropriate formalism which both works symbolically and incorporates intersection and complement. As shown in [26], the classical theory of derivatives does not directly extend to the symbolic setting, because taking a symbolic derivative (derivative with respect to a character predicate denoting a set B of characters) of an extended *symbolic* regular expression r does in general not preserve its language semantics. It either results in an *over-approximation* or an *under-approximation* of the actual language, depending on whether the positive derivative $\Delta_B(r)$ or the negative derivative $\nabla_B(r)$ is taken [26, Lemma 3]. On the other hand, a classical generalization of Antimorov derivatives to extended regular expressions is possible (over a finite alphabet Σ) although challenging [12]; however, leveraging this work for the symbolic SMT setting would require explicitly enumerating (finitizing) the entire alphabet upfront (also known as *mintermization* in the literature [17, 18]), as we explain further in Section 2. Doing so may be infeasible or prohibitively expensive (e.g. for Unicode), requires considering all regex constraints in an SMT formula globally, and for general predicates may cause another exponential blowup [18]. Considering only intersection, and not complement, avoids some of this complexity and represents a state-of-the-art approach [29], but this loses the full generality of the Boolean operations.

In this work, we fill these gaps by proposing the first theory of derivatives of symbolic regexes which incorporates intersection and complement. Unlike previous work, our approach can be used to avoid the state-space blowup of automata-based solvers without assuming a finite alphabet and without under- and over-approximation. The *key new insight* that enables us to define derivatives of regexes directly, while allowing Boolean operations, is that we augment regexes with *conditionals* (if-then-else), and define the derivative of a regex to be a regex with conditionals, called a *transition regex*. We show that transition regexes allow for efficient algebraic manipulation rules for complementation and intersection: for example, given a regex which is a Boolean combination of classical regexes, we show that the number of derivatives is strictly linear (Theorem 7.3). We give a decision procedure based on our derivatives which integrates into a broader SMT context: a set of inference rules that incrementally unfolds regular expression constraints into symbolic constraints over the background character theory. Derivatives enable this lazy unfolding; the symbolic conditionals directly map to the underlying character theory;

and the succinct handling of Boolean combinations via extended regexes avoids the blowup in existing techniques. We also introduce an accompanying theory of symbolic Boolean finite automata (SBFAs): the deriatives of an extended regular expression correspond to the states in the SBFA. This is used to prove the succinctness theorem and to study the connection with classical approaches and other techniques.

We have implemented symbolic Boolean derivatives in a new regular expression solver, dZ3, which is built on top of Z3 and fully replaces the existing solver. We show that the lack of blowup shows the expected benefits in practice. Using a large benchmark suite and compared to an array of state-of-the-art solvers, we show that our decision procedure matches or outperforms other solvers in terms of number of benchmarks solved and average time per benchmark. It shows particular benefits on examples with Boolean combinations: although CVC4 and Ostrich are competitive on subsets of the benchmarks, no solver consistently shows good performance across benchmark sets involving Boolean combinations. For example, dZ3 is 1.54x faster than the next best solver (CVC4) on average for existing benchmarks with Boolean combinations, and solves 88% of handwritten examples such as the date example in Figure 1, compared to 57% for CVC4.

Contributions.

- We introduce a new theory of symbolic derivatives of extended regexes, which avoids blowup in existing techniques. It works via translation to *transition regexes* which augment extended regexes with a conditional construct. (Section 4)
- We propose a sound and conditionally complete decision procedure for solving extended regular expression constraints in an SMT context. (Section 5)
- We provide a proof-of-concept open-source implementation on top of Z3, called dZ3. Using existing benchmark sets, we show that our solver matches or outperforms state-of-the-art solvers for string constraints and shows particular performance and solvability improvements on Boolean combinations. (Section 6)
- To formally study the benefits of our approach, we introduce a theory of Symbolic Boolean Finite Automata (SBFAs) that generalizes the various classical approaches of alternating and Boolean automata to the symbolic setting. In particular, we use SBFAs to show that for a common subclass of extended regexes, the set of symbolic derivatives has linear size (**Theorem 7.3**). (Section 7)
- We provide an in-depth comparison of our theory of derivatives with the classical theory. (Section 8).

2 Motivating Running Example

We discuss here a motivating example that helps us highlight some of the main ideas behind *transition regexes*, the key to defining derivatives for symbolic extended regular expressions. The example also serves as a running example and is referenced in the later sections. It is similar in spirit to the date example in Figure 1 and is typical to many of the benchmarks used in Section 6.

Suppose we are given a membership constraint $s \in R$, where s is a string term over an alphabet type Σ, i.e., s has type Σ^*, and R is a concrete regex over Σ^*. Our goal is to solve the *satisfiability* problem for that membership constraint: does there exist a concrete instance of s in Σ^* such that R accepts that instance? Using the approach of derivatives, we plan to attack the problem by calculating the derivatives of R, by deducing the following case split: [1]

$$(|s| = 0 \land \mathit{IsNullable}(R)) \lor (|s| > 0 \land s_{1..} \in \delta(R)(s_0)),$$

where *IsNullable*(R) is true if R accepts the empty string, and $\delta(R)$ is a function of R called its *derivative*: it takes a regex (R) and a first character (s_0), and returns a regex for the language of *suffixes* w such that $s_0 w \in R$ holds.

> *However, the classical theory of derivatives does not directly apply here: the problem is that the string s may be uninterpreted (we don't know the first character s_0), and classical derivatives are only defined for a given input character. We could naively enumerate all possible characters $\bigvee_{a\in\Sigma}(s_{1..}{\in}D_a(R) \land s_0{=}a)$, but this does not scale.*

Our contribution is to address this by providing a closed definition of $\delta(R)$ above: in particular, we want to be able to evaluate $\delta(R)$ symbolically, before knowing s_0. We call this the *symbolic derivative*, and we call the resulting term a *transition regex*: it denotes a function from Σ to regexes.

More concretely, take R to be a typical *password constraint*:

$$(s \in \texttt{.*\textbackslash d.*}) \land \neg(s \in \texttt{.*01.*})$$

This constraint states that s contains at least one digit but not the subsequence `01`. Regular expressions such as this one are used in the generation and validation of password strings. In typical real-world cases, they may involve many more similar simultaneous constraints (cf. [37]), which can be encoded as large intersections (cf. [43]). The motivation for derivative-based approaches is that such constraints — in particular because they are also combined with bounded loops such as `.{8,128}` — cause an explosion of the state space when converted to automata [17]. By unfolding the derivatives of R, we will explore possible strings for s without constructing the state space up front.

We now show how to solve the constraint $s \in R$ for this example, using our approach, and following our implementation in dZ3. The negation is first converted into a regex complement and then the conjuction into an intersection:

```
s ∈ (.*\d.*) & ~(.*01.*)
```

[1] We write s_i for its i'th element and $s_{i..}$ for its suffix from i. Note that these can be purely symbolic expressions, s itself may be uninterpreted.

Let $R_1 = \mathtt{.*\backslash d.*}$, $R_2 = \mathtt{{\sim}(.*01.*)}$ and $R = R_1 \,\&\, R_2$. Since R is not nullable (does not accept the empty string), the case split we started from reduces to the assertion

$$|s| > 0 \land s_{1..} \in \delta(R)(s_0)$$

To calculate $\delta(R)$ as a transition regex, we need to deal with the problem that we do not know s_0. The solution is to *augment regexes with conditionals (if-then-else)*, and then allow conditionals in transition regexes. When taking the derivative of a regex such as `01`, we generate the term $\textbf{IF}(x = \mathtt{0}, \mathtt{1}, \bot)$, read as *if* $x = \mathtt{0}$ *then* $\mathtt{1}$ *else* $\bot$. This idea allows for the derivative of R to be computed using algebraic rules as follows. The $\equiv$ below also shows simplification steps using distributivity, DeMorgan's laws, and other properties. Below, φ_d is the predicate for `\d` (characters that are digits).

$$\begin{aligned}
\delta(R) &= \delta(R_1) \,\&\, \delta(R_2)\\
\delta(R_1) &= R_1 \mid \textbf{IF}(\varphi_d(x), \mathtt{.*}, \bot) \equiv \textbf{IF}(\varphi_d(x), \mathtt{.*}, R_1)\\
\delta(R_2) &= {\sim}(\delta(\mathtt{.*01.*})) = {\sim}(\mathtt{.*01.*} \mid \delta(\mathtt{01.*}))\\
&= {\sim}(\mathtt{.*01.*} \mid \textbf{IF}(x = \mathtt{0}, \mathtt{1.*}, \bot))\\
&\equiv {\sim}(\mathtt{.*01.*}) \,\&\, {\sim}(\textbf{IF}(x = \mathtt{0}, \mathtt{1.*}, \bot))\\
&\equiv R_2 \,\&\, \textbf{IF}(x = \mathtt{0}, {\sim}(\mathtt{1.*}), .*)\\
&\equiv \textbf{IF}(x = \mathtt{0}, R_2 \,\&\, {\sim}(\mathtt{1.*}), R_2)\\
\delta(R) &\equiv \textbf{IF}(\varphi_d(x), \mathtt{.*}, R_1) \,\&\, \textbf{IF}(x = \mathtt{0}, R_2 \,\&\, {\sim}(\mathtt{1.*}), R_2)\\
&\overset{\text{(i)}}{\equiv} \textbf{IF}(x = \mathtt{0}, R_2 \,\&\, {\sim}(\mathtt{1.*}), \textbf{IF}(\varphi_d(x), \mathtt{.*}, R_1) \,\&\, R_2)\\
&\equiv \textbf{IF}(x = \mathtt{0}, R_2 \,\&\, {\sim}(\mathtt{1.*}), \textbf{IF}(\varphi_d(x), R_2, R))
\end{aligned}$$

Observe that all conditional predicates are extracted from the regex itself: e.g. `0` in a conditional arises from `0` in the original regex. Step (i) uses (among other properties) that $\neg\varphi_d(x) \land x = \mathtt{0}$ is unsat. Note that ${\sim}\bot \equiv \mathtt{.*}$ and $\mathtt{.*} \mid \ldots \equiv \mathtt{.*}$.

There is no direct classical counterpart to the above derivation sequence, because classical regexes do not have *if-then-else*. In particular, there is no direct classical counterpart which handles complement. For example, consider the regular expression `01.*` above. Classically, we would take the derivative as $D_{\mathtt{0}}(\mathtt{01.*}) = \mathtt{1.*}$. But what if we want to now take the derivative of the complement of `01.*`? Then we need to know not just this derivative where the first character is `0` but also the derivative if the first character is *not* `0`, because while the latter case was impossible before it becomes relevant when considering the complement. Using conditionals solves this problem: we write the derivative as $\textbf{IF}(x = \mathtt{0}, \mathtt{1.*}, \bot)$, which has the case where the first character is not `0` present. Then when complementing this, we get $\textbf{IF}(x = \mathtt{0}, {\sim}(\mathtt{1.*}), \mathtt{.*})$. Thus, viewing the derivative as a conditional regex (transition regex) is what enables us to treat complement algebraically.

Having calculated the derivative $\delta(R)$, we then continue as follows. Let $R_3 = R_2 \,\&\, {\sim}(\mathtt{1.*})$. So $s_{1..} \in \delta(R)(s_0)$ reduces to

$$s_{1..} \in \textbf{IF}(s_0 = \mathtt{0}, R_3, \textbf{IF}(\varphi_d(s_0), R_2, R))$$

Going forward, this creates the further case split:

$$(s_0 = \mathtt{0} \land s_{1..} \in R_3) \lor (s_0 \neq \mathtt{0} \land s_{1..} \in \textbf{IF}(\varphi_d(s_0), R_2, R))$$

where $s_{1..} {\in} R_3$ splits further into two subcases:

$$(|s_{1..}| = 0 \land \mathit{IsNullable}(R_3)) \lor (|s_{1..}| > 0 \land s_{2..} \in \delta(R_3)(s_1))$$

where $(s_{1..})_{1..} = s_{2..}$ and $(s_{1..})_0 = s_1$, and the procedure repeats. Here R_3 is nullable so dZ3 can generate a model for $|s| > 0 \land |s_{1..}| = 0 \land s_0 = \mathtt{0}$ — provided that these constraints are consistent with other constraints on s in the context. For example if there was a constraint $s_0 > \mathtt{0}$, this case would be blocked and the search would backtrack to the other case.[2]

3 Preliminaries

Sequences. When working with sequences over a domain Σ we make the standard simplifying assumption that $\Sigma^{(1)} = \Sigma$, and let $\Sigma^{(0)} = \{\epsilon\}$, $\Sigma^{(k+1)} = \Sigma \cdot \Sigma^{(k)}$, for $k \geq 0$, and $\Sigma^* = \bigcup_{k \geq 0} \Sigma^{(k)}$, $\Sigma^+ = \bigcup_{k \geq 1} \Sigma^{(k)}$. Moreover, for $v \in \Sigma^{(k)}$, the length of v is k, $|v| = k$. In contrast, when Σ^* is implemented in an SMT solver the type Σ^* is *sequence over* Σ that is disjoint from Σ. For $X, Y \subseteq \Sigma^*$, define $X \cdot Y \subseteq \Sigma^*$ such that $X \cdot Y = \{x \cdot y \mid x \in X, y \in Y\}$ where concatenation $\cdot$ is associative and ϵ is the empty sequence. We write xy for $x \cdot y$ when it is clear from the context that juxtaposition stands for concatenation. Also, X^* stands for the closure of X under concatenation when it is clear from the context that $X \subseteq \Sigma^*$.

Boolean Algebras. Let D be a nonempty *universe*. A *Boolean algebra over* D is a tuple $\mathcal{A} = (D, \Psi, [\![\_]\!], \bot, \top, \lor, \land, \neg)$ where Ψ is a set of *predicates* closed under the Boolean connectives; $[\![\_]\!] : \Psi \to 2^D$ is a *denotation function*; $\bot, \top \in \Psi$; $[\![\bot]\!] = \emptyset$, $[\![\top]\!] = D$, and for all $\varphi, \psi \in \Psi$, $[\![\varphi \lor \psi]\!] = [\![\varphi]\!] \cup [\![\psi]\!]$, $[\![\varphi \land \psi]\!] = [\![\varphi]\!] \cap [\![\psi]\!]$, and $[\![\neg\varphi]\!] = D \setminus [\![\varphi]\!]$. For $\varphi, \psi \in \Psi$ we write $\varphi \equiv \psi$ (φ is *equivalent* to ψ) to mean that $[\![\varphi]\!] = [\![\psi]\!]$. In particular, if $\varphi \equiv \bot$ then φ is *unsatisfiable* and if $\varphi \equiv \top$ then φ is *valid*. $\mathcal{A}$ being *effective* means that all components of $\mathcal{A}$ are recursively enumerable, and satisfiability of $\varphi \in \Psi$ ($\varphi \not\equiv \bot$) is decidable.

Boolean Combinations. If Q is a set of basic elements or *atoms* then $\mathbb{B}(Q)$ denotes the *Boolean closure* over Q using | for disjunction, & for conjunction, and ~ for complement. $\mathbb{B}^+(Q)$ denotes the *positive* Boolean closure of Q (without use of ~). Both & and | are treated as idempotent, associative and commutative operators and lifted to finite nonempty subsets $S \subseteq Q$ through $\mathit{AND}(S)$ and $\mathit{OR}(S)$, respectively.

Symbolic Regexes. Let $\mathcal{A} = (\Sigma, \Psi, [\![\_]\!], \bot, ., \lor, \land, \neg)$ be a fixed effective Boolean algebra called an *alphabet theory*. Note that Σ may be infinite. We first recall the definitions of the two standard subclasses of regexes and extended regexes, where $\varphi \in \Psi$. We always work *modulo* $\mathcal{A}$ and we do not mention this explicitly every time.

$$\begin{aligned}
\mathit{RE} &::= \varphi \mid \varepsilon \mid \bot \mid \mathit{RE}_1 \cdot \mathit{RE}_2 \mid \mathit{RE}{*} \mid \mathit{RE}_1 | \mathit{RE}_2\\
\mathit{ERE} &::= \varphi \mid \varepsilon \mid \bot \mid \mathit{ERE}_1 \cdot \mathit{ERE}_2 \mid \mathit{ERE}{*} \mid \mathbb{B}(\mathit{ERE})
\end{aligned}$$

[2] The condition $s_0 > \mathtt{0}$ is possible because the underlying character theory (by default bitvectors in dZ3) is equipped with a total order.

The class *RE* corresponds to all standard regexes. The fragment $\mathbb{B}(RE) \subset ERE$ comprises all Boolean combinations over *RE* and covers *all* of our practical scenarios. The *language accepted by R* is $\mathbf{L}(R) \subseteq \Sigma^*$:

$$\begin{aligned}&\mathbf{L}(\varphi) = [\![\varphi]\!],\ \mathbf{L}(\varepsilon) = \{\epsilon\},\ \mathbf{L}(\bot) = \emptyset,\\ &\mathbf{L}(R_1 \cdot R_2) = \mathbf{L}(R_1) \cdot \mathbf{L}(R_2),\ \mathbf{L}(R*) = \mathbf{L}(R)^*,\\ &\mathbf{L}(R_1 \mid R_2) = \mathbf{L}(R_1) \cup \mathbf{L}(R_2),\ \mathbf{L}(R_1 \,\&\, R_2) = \mathbf{L}(R_1) \cap \mathbf{L}(R_2),\\ &\mathbf{L}({\sim}R) = \Sigma^* \setminus \mathbf{L}(R)\end{aligned}$$

A regex R is *nullable* ($\nu(R)$) iff $\epsilon \in \mathbf{L}(R)$: $\nu(\varphi)=\nu(\bot)=\mathit{false}$; $\nu(\varepsilon)=\nu(R*)=\mathit{true}$; $\nu(R_1 \cdot R_2) \Leftrightarrow \nu(R_1)\ \mathit{and}\ \nu(R_2)$; $\nu(R_1 \& R_2) \Leftrightarrow \nu(R_1)\ \mathit{and}\ \nu(R_2)$; $\nu(R_1 \mid R_2) \Leftrightarrow \nu(R_1)\ \mathit{or}\ \nu(R_2)$; $\nu({\sim}R) \Leftrightarrow \mathit{not}\ \nu(R)$.

4 Symbolic Derivatives

Here we formally introduce the concept of *transition regexes TR*, define *symbolic derivatives* for $R \in ERE$ in terms *TR*, and prove their correctness in Theorem 4.3. We also discuss some algebraic laws that hold in *TR* — used as simplification rules in dZ3 — as illustrated in Section 2.

Transition Regexes. In order to define symbolic derivatives we first introduce the key concept of *transition regexes TR* in which regexes are augmented with conditionals. The definition of *TR* depends on a parameter Q — referred to below as TR_Q — here $Q = ERE$. Let $\diamond \in \{\mid, \&\}$, $\bar{\&} = \mid$ and $\bar{\mid} = \&$.

$$TR ::= Q \ \mid\ \textbf{IF}(\varphi, TR_1, TR_2) \ \mid\ \mathbb{B}(TR)$$

We call $\textbf{IF}(\varphi, \tau_1, \tau_2)$ a *conditional regex*. A transition regex τ denotes the function $\tau : \Sigma \longrightarrow \mathbb{B}(Q)$ defined as follows.[3]

$$\begin{aligned} R(x) &= R \quad (\text{for } R \in Q)\\ \textbf{IF}(\varphi, \tau, \rho)(x) &= \begin{cases} \tau(x), & \text{if } x \in [\![\varphi]\!];\\ \rho(x), & \text{otherwise.}\end{cases}\\ \tau \diamond \rho(x) &= \tau(x) \diamond \rho(x)\\ {\sim}\tau(x) &= {\sim}(\tau(x))\end{aligned}$$

Transition regexes τ and ρ are *equivalent*, denoted $\tau \equiv \rho$, when $\forall x \in \Sigma$, $\tau(x) \equiv \rho(x)$. The concatenation operation of regexes is lifted to transition regexes τ in $\tau \cdot R$ for $R \in ERE$.

$$\begin{aligned} \textbf{IF}(\varphi, \tau, \rho) \cdot R &= \textbf{IF}(\varphi, \tau \cdot R, \rho \cdot R)\\ (\tau \mid \rho) \cdot R &= (\tau \cdot R) \mid (\rho \cdot R)\\ {\sim}\tau \cdot R &= \overline{\tau} \cdot R\\ (\tau \& \rho) \cdot R &= \mathit{lift}(\tau \& \rho) \cdot R\end{aligned}$$

Negation $\overline{\tau}$ of τ is defined as follows.

$$\overline{R} = {\sim}R,\ \overline{{\sim}\tau} = \tau,\ \overline{\tau \diamond \rho} = \overline{\tau}\ \bar{\diamond}\ \overline{\rho},\ \overline{\textbf{IF}(\varphi, \tau, \rho)} = \textbf{IF}(\varphi, \overline{\tau}, \overline{\rho})$$

The definition of $\mathit{lift}(\tau)$ is such that if $\tau \in Q$ then $\mathit{lift}(\tau) = \tau$ else τ is transformed into an equivalent conditional regex by lifting the character predicates to the top while pushing conjuction into the leaves.[4]

The following lemmas represent key semantic properties that are used in several contexts. Lemma 4.1 is used in the proof of Theorem 4.3 and Lemma 4.2 is correctness of negation that is for example exploited in normal forms. Both lemmas are proved by induction over τ using various algebraic laws of *TR*.

Lemma 4.1. $\mathbf{L}(\tau \cdot R(x)) = \mathbf{L}(\tau(x)) \cdot \mathbf{L}(R)$

Lemma 4.2. ${\sim}\tau \equiv \overline{\tau}$

The *symbolic derivative* $\delta(R)$ of a regex $R \in ERE$ is defined as the following transition regex, where $\varphi \in \Psi$.

$$\begin{aligned} \delta(\varepsilon) = \delta(\bot) &= \bot\\ \delta(\varphi) &= \textbf{IF}(\varphi, \varepsilon, \bot)\\ \delta(R \cdot R') &= \begin{cases} \delta(R) \cdot R' \mid \delta(R'); & \text{if } R \text{ is nullable,}\\ \delta(R) \cdot R'; & \text{otherwise.}\end{cases}\\ \delta(R*) &= \delta(R) \cdot R*\\ \delta(R \diamond R') &= \delta(R) \diamond \delta(R') \quad (\text{for } \diamond \in \{\&, \mid\})\\ \delta({\sim}R) &= {\sim}\delta(R)\end{aligned}$$

Theorem 4.3 is the correctness theorem of symbolic derivatives. For $L \subseteq \Sigma^*$ and $a \in \Sigma$, recall the classical definition of the *derivative of L wrt a*, $\mathbf{D}_a(L) = \{v \mid av \in L\}$, and for $R \in ERE$ we use *Brzozowski derivatives* $D_a(R) \in ERE$ (modulo $\mathcal{A}$ [26]), and the classical result $\mathbf{L}(D_a(R)) = \mathbf{D}_a(\mathbf{L}(R))$ [9, Theorem 3.1]. Let $\mathbf{D}_a(R) = \mathbf{L}(D_a(R))$.

Theorem 4.3. *$\delta(R)(a) \equiv D_a(R)$.*

Proof. By induction over R. The base cases $\bot$ and ε are trivial.
Base case φ: $\delta(\varphi) = \textbf{IF}(\varphi, \varepsilon, \bot)$. If $a \in [\![\varphi]\!]$
then $\textbf{IF}(\varphi, \varepsilon, \bot)(a) = \varepsilon(a) = \varepsilon = D_a(\varphi)$
else $\textbf{IF}(\varphi, \varepsilon, \bot)(a) = \bot(a) = \bot = D_a(\varphi)$.
Induction case $R \cdot R'$ **and** R **nullable**:

$$\begin{aligned} \mathbf{L}(\delta(R{\cdot}R')(a)) &= \mathbf{L}(\delta(R){\cdot}R' \mid \delta(R')(a))\\ &= \mathbf{L}(\delta(R){\cdot}R'(a) \mid \delta(R')(a))\\ &= \mathbf{L}(\delta(R)(a)){\cdot}\mathbf{L}(R') \cup \mathbf{L}(\delta(R')(a))\\ &\overset{\text{IH}}{=} \mathbf{D}_a(R){\cdot}\mathbf{L}(R') \cup \mathbf{D}_a(R') = \mathbf{D}_a(R{\cdot}R')\end{aligned}$$

Induction case $R \cdot R'$ **and** R **not nullable**:

$$\begin{aligned} \mathbf{L}(\delta(R{\cdot}R')(a)) &= \mathbf{L}(\delta(R){\cdot}R'(a)) = \mathbf{L}(\delta(R)(a)){\cdot}\mathbf{L}(R')\\ &\overset{\text{IH}}{=} \mathbf{D}_a(R){\cdot}\mathbf{L}(R') = \mathbf{D}_a(R{\cdot}R')\end{aligned}$$

Induction case $R*$:

$$\begin{aligned} \mathbf{L}(\delta(R*)(a)) &= \mathbf{L}(\delta(R){\cdot}R*(a)) = \mathbf{L}(\delta(R)(a)){\cdot}\mathbf{L}(R*)\\ &\overset{\text{IH}}{=} \mathbf{D}_a(R){\cdot}\mathbf{L}(R*) = \mathbf{D}_a(R*)\end{aligned}$$

Induction case $R{\diamond}R'$: Let $\diamond \in \{\mid, \&\}$. $\hat{\mid} = \cup$ and $\hat{\&} = \cap$.

$$\begin{aligned} \mathbf{L}(\delta(R{\diamond}R')(a)) &= \mathbf{L}(\delta(R)(a))\ \hat{\diamond}\ \mathbf{L}(\delta(R')(a))\\ &\overset{\text{IH}}{=} \mathbf{D}_a(R)\ \hat{\diamond}\ \mathbf{D}_a(R') = \mathbf{D}_a(R{\diamond}R')\end{aligned}$$

Induction case ${\sim}R$:

$$\mathbf{L}(\delta({\sim}R)(a)) = \Sigma^* \setminus \mathbf{L}(\delta(R)(a)) \overset{\text{IH}}{=} \Sigma^* \setminus \mathbf{D}_a(R) = \mathbf{D}_a({\sim}R)$$

The statement follows by the induction principle. □

A useful property to observe about the proof of Theorem 4.3 is the following corollary.

[3] Function application of (x) binds weakest, so $\tau \diamond \rho(x)$ stands for $(\tau \diamond \rho)(x)$.
[4] Lift rules are discussed in Appendix D.

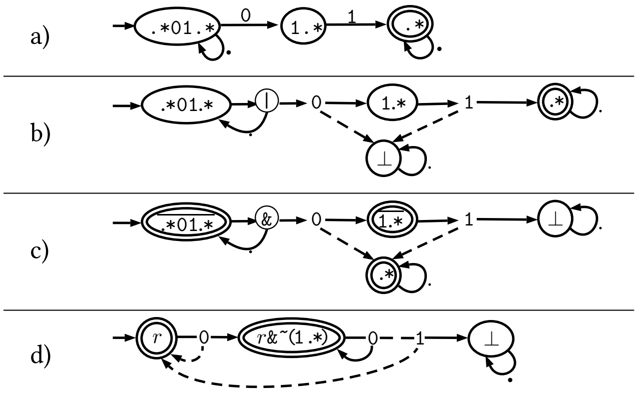


Figure 2. a-c) Symbolic derivations viewed as transitions between regexes; d) DNF form of (c) where $r = \mathtt{\sim(.*01.*)}$.

Corollary 4.4. *If $R \in \mathbb{B}(RE)$ then $\delta(R)(a) \in \mathbb{B}(RE)$.*

Proof. If $R \in \mathbb{B}(RE)$ then lifting is never invoked. Complement and conjunction remain as top level operators only and are never nested within a concatenation or loop. □

Example 4.5. Consider the regex $\mathtt{.*01.*}$ from above. We write individual characters also for the corresponding singleton predicates when this is unambiguous, except that $[\![.]\!] = \Sigma$. We implicitly use the simplification rule that $\textsc{if}(., \tau, \_) \equiv \tau$. Thus, e.g., $\delta(.)$ simplifies to ε (and so $\delta(.)r$ simplifies to r).

$$\begin{aligned} \delta(\mathtt{.*01.*}) &= \delta(\mathtt{.*}){\cdot}\mathtt{01.*}\,|\,\delta(\mathtt{01.*}) \\ &= \delta(.){\cdot}\mathtt{.*01.*}\,|\,\delta(\mathtt{0}){\cdot}\mathtt{1.*} = \mathtt{.*01.*}\,|\,\textsc{if}(\mathtt{0}, \mathtt{1.*}, \bot) \\ \delta(\mathtt{1.*}) &= \textsc{if}(\mathtt{1}, \mathtt{.*}, \bot) \end{aligned}$$

The two transition regexes are shown as classical transitions in Figure 2a where $\bot$ is hidden. The equivalent *complete* view of the transition regexes is shown in Figure 2b where the dashed arrows represent the false branches of conditional regexes. The negation of the complete form is seen in Figure 2c as the dual of Figure 2b, where $\overline{\bot} = \mathtt{.*}$, and $\overline{\mathtt{.*}} = \bot$. A regex q is *final* (has double boundary) when q is nullable. ☒

Algebraic Properties. Transition regexes form a particular kind of an effective Boolean algebra.[5] The regex $\mathtt{.*}$ is treated as the *absorbing* element of | and the *unit* element of &. Conversely, $\bot$ is treated as the unit element of | and the absorbing element of both & and $\cdot$. For example $r \,\&\, \mathtt{.*} = r$ and $\bot \cdot r = \bot$. We also treat |, &, $\cdot$ as associative operators and |, & as commutative idempotent operators. This is important in reducing the number of different but equivalent regexes from arising during search. However, the algebra is *not extensional*, i.e., $\tau \equiv \rho$ does in general not imply $\tau = \rho$.

We exploit this algebra for different algebraic simplifications and normal forms. The most important one is *disjunctive normal form* or *DNF*. Here we consider $\tau = \delta(R)$ for $R \in \mathbb{B}(RE)$ but DNF generalizes to all $R \in ERE$ by using *lift*(τ). For DNF we apply standard laws of distributivity. Perhaps the most relevant case here is $\textsc{if}(\varphi, \tau_1, \tau_2)\&\rho$ that in general expands to $\textsc{if}(\varphi, \tau_1\&\rho, \tau_2\&\rho)$ but is also subject to simplifications discussed next that integrate satisfiability checks of $\mathcal{A}$ into the rules.

1. If $\varphi \wedge \psi \equiv \bot$ then $\textsc{if}(\varphi, \tau, \bot)\&\textsc{if}(\psi, \rho, \bot) \equiv \bot$ else $\textsc{if}(\varphi, \tau, \bot)\&\textsc{if}(\psi, \rho, \bot) \equiv \textsc{if}(\varphi \wedge \psi, \tau \,\&\, \rho, \bot)$.
2. *Cleaning* of unsatisfiable branches of a nested conditional regex. For example if $\tau = \textsc{if}(\varphi, \textsc{if}(\psi, \tau_1, \tau_2), \rho)$ and $\varphi \wedge \psi \equiv \bot$ then τ simplifies to $\textsc{if}(\varphi, \tau_2, \rho)$ or if $\varphi \wedge \neg\psi \equiv \bot$ then τ simplifies to $\textsc{if}(\varphi, \tau_1, \rho)$.
3. It is useful to push complement into $\mathcal{A}$ when possible, e.g., by using the rule $\mathtt{\sim}\textsc{if}(\varphi, \mathtt{.*}, \bot) \equiv \textsc{if}(\neg\varphi, \mathtt{.*}, \bot)$.

Example 4.6. Recall the computation of $\delta(\mathtt{.*01.*})$ from Example 4.5. Let $r = \mathtt{\sim(.*01.*)}$. In Section 2 we showed that $\delta(r)$ can be computed initially as $\mathtt{\sim}\delta(\mathtt{.*01.*})$ and then we take its DNF so that in the end $\delta(r) \equiv \textsc{if}(\mathtt{0}, r \,\&\, \mathtt{\sim(1.*)}, r)$. It is also easy to see that $\delta(\mathtt{\sim(1.*)}) \equiv \textsc{if}(\mathtt{1}, \bot, \mathtt{.*})$. We continue with the regex $r \,\&\, \mathtt{\sim(1.*)}$ and get that

$$\begin{aligned} \delta(r \,\&\, \mathtt{\sim(1.*)}) &= \delta(r) \,\&\, \delta(\mathtt{\sim(1.*)}) \\ &\equiv \textsc{if}(\mathtt{0}, r \,\&\, \mathtt{\sim(1.*)}, r) \,\&\, \textsc{if}(\mathtt{1}, \bot, \mathtt{.*}) \\ &\equiv \textsc{if}(\mathtt{0}, r\&\mathtt{\sim(1.*)} \,\&\, \textsc{if}(\mathtt{1}, \bot, \mathtt{.*}), r\&\textsc{if}(\mathtt{1}, \bot, \mathtt{.*})) \\ &\equiv \textsc{if}(\mathtt{0}, r\&\mathtt{\sim(1.*)}, \textsc{if}(\mathtt{1}, \bot, r)) \end{aligned}$$

where the last equality uses, among other simplifications, the fact that $\mathtt{0} \wedge \mathtt{1} \equiv \bot$ to keep the resulting conditional regex clean. The resulting transitions are shown in Figure 2(d). ☒

When working with the two algebras $\mathcal{A}$ and *TR*, it is important to keep in mind that their Boolean operations have different semantics.[6] For example, the predicate $\neg\varphi$ as a singleton regex denotes the language $\mathbf{L}(\neg\varphi) = \Sigma \setminus [\![\varphi]\!]$, while the regex $\mathtt{\sim}\varphi$ denotes the language $\mathbf{L}(\mathtt{\sim}\varphi) = \Sigma^* \setminus [\![\varphi]\!]$.

We show in Theorem 7.3 that for $R \in \mathbb{B}(RE)$ the number of individual regexes that are formed after computing the fixpoint of all regexes through derivation is *linear* in R.

5 Solving Extended Regular Expression Constraints in SMT

Here we show that derivatives of extended regexes, defined in Section 4, form the basis for a decision procedure that can be integrated in the context of an SMT solver to solve Boolean combinations of *ERE* constraints. The regex solver for *ERE* constraints is part of the sequence theory solver in dZ3. One challenge here is that the problem is not an isolated decision procedure but needs to be integrated into the main satisfiability engine of the solver, and in particular, interact with the solver for the given background theory of characters. We describe our algorithm following our implementation that builds on Z3. A brief overview was given in Section 2.

[5] One can view *TR* as a Boolean algebra over Σ^+ where $f : \Sigma \to 2^{\Sigma^*}$ is represented by $\bigcup_{a\in\Sigma} af(a) \subseteq \Sigma^+$ where $aL = \{av \mid v \in L\}$.

[6] This is also true in the context of SMT where they are distinct primitive operators. Here we avoid ambiguities by not overloading the operators.

We focus here on assertions of the form of $s \in r$, called *membership constraints*, where s is a term whose sort[7] is *sequence over* Σ or Σ^* and r is an *ERE* over Σ^*. Such constraints exist in a broader context of formulas, including possibly other string contraints on s. We assume that regexes are concrete (i.e. there are no variables of type regex or equations between regexes, only membership constraints for concrete regexes). While this restriction is standard, it can be partially relaxed without additional work: for example, *inequivalence* constraints of the form $r \not\equiv r'$ for regexes r, r' (this includes nonemptiness constraints) can also be reduced to membership using the Boolean operators. In particular $r \not\equiv \bot$ iff $\exists x(x \in r)$, and $r_1 \not\equiv r_2$ iff $(r_1$ `&` ~$r_2)$ | $(r_2$ `&` ~$r_1) \not\equiv \bot$.

The regex solver dynamically maintains a *graph* $G = (V, E, F, C)$ with additional derived components *Dead* and *Alive*. The vertices $V \subseteq \mathit{ERE}$ represent the set of all encountered regexes so far, and $E \subseteq V \times V$ is a set of directed edges such that $(v, w) \in E$ implies that $w \in \mathbf{Q}(\delta^{\text{DNF}}(v))$, i.e., w is *derived from* v. In this context $\delta^{\text{DNF}}(v)$ is equivalent to the abstract definition $\delta(v)$ (defined in Section 4) but in a normal form; the required normal form is discussed further below.

We write E^* for the *reflexive and transitive closure* of E and we write $E^*(v)$ for $\{w \mid (v, w) \in E^*\}$, i.e., $E^*(v)$ is the set of all vertices in G that are reachable from v.

- $F \subseteq V$ is a set of *final* vertices (*nullable* regexes).
- $C \subseteq V$ is the set of all *closed* v: $\forall w{\in}\mathbf{Q}(\delta^{\text{DNF}}(v)) : (v, w){\in}E$. In other words, a closed vertex is a vertex all of whose outgoing edges have been added to E.
- $\mathit{Alive} \subseteq V$ is the set of all v s.t. $E^*(v) \cap F \neq \emptyset$.
- $\mathit{Dead} \subseteq V$ is the set of all v s.t. $E^*(v) \subseteq (C \setminus \mathit{Alive})$. In other words, all vertices in *Dead* are dead-end regexes whose status can never change because all of them are closed (have been fully explored).

For modularity, G does not have knowledge of its vertices being regexes, but they are treated as abstract elements. Therefore, for the abstract description here, we consider the sets F and C to be represented explicitly. The event that all immediate (partial) derivatives from v have been added then causes v to be added to the set C. On the other hand, we consider *Alive* and *Dead* to be inferred from (V, E, F, C) rather than being explicitly represented here.

The primary purpose of G is to enable *dead-end detection* and to block search and to infer unsatisfiability of dead-end regexes, as indicated by the BOT-rule in Figure 3a. It is important to note that G is independent of the current logical scope because the property of a vertex in G being dead is independent of other side constraints that may exist on the input sequence s, i.e., this means that any satisfiability checks of branches are performed in a global scope, independent of local assertions. Therefore G can persist across different logical scopes.

$$\frac{\mathit{in\text{-}tr}(s, \mathbf{IF}(\varphi, t, f))}{(\varphi(s_0) \wedge \mathit{in\text{-}tr}(s, t)) \vee (\neg\varphi(s_0) \wedge \mathit{in\text{-}tr}(s, f))}\ (\textsc{ite})$$

$$\frac{\mathit{in\text{-}tr}(s, r)}{\mathit{in}(s_{1..}, r)}\ (\textsc{ere}) \qquad\qquad \frac{\mathit{in\text{-}tr}(s, t_1 \,|\, t_2)}{\mathit{in\text{-}tr}(s, t_1) \vee \mathit{in\text{-}tr}(s, t_2)}\ (\textsc{or})$$

$$\frac{\mathit{in}(s, r) \qquad r \notin G.\mathit{Dead}}{\begin{array}{c}(|s| = 0 \wedge \nu(r)) \vee \\ (|s| > 0 \wedge \mathit{in\text{-}tr}(s, \delta^{\text{DNF}}(r)) \wedge \boldsymbol{\mathit{Upd}}[r{\to}\mathbf{Q}(\delta^{\text{DNF}}(r))])\end{array}}\ (\textsc{der})$$

$$\frac{\mathit{in}(s, r) \qquad r \in G.\mathit{Dead}}{\bot}\ (\textsc{bot})$$

(a) Membership propagation rules for *ERE*s and transition predicates. Here $r \in \mathit{ERE}$. Recall that $\nu(r)$ iff r is *nullable*. All rules are equivalence preserving in their respective contexts. In particular $\mathit{in\text{-}tr}(s, t)$ rules are applied only when $|s| > 0$. An implicit assumption is that $r \in G.V$.

$$\frac{\boldsymbol{\mathit{Upd}}[r{\to}Q] \qquad G = (V, E, F, C)}{G{:=}(V{\cup}Q, E{\cup}\{(r, q) \mid q{\in}Q\}, F{\cup}\{q{\in}Q \mid \nu(q)\}, C{\cup}\{r\})}\ (\textsc{upd})$$

(b) Graph update rule. An implicit assumption is that $r \in G.V$. Observe that the rule has no effect if $r \in G.C$.

Figure 3. Decision procedure propagation rules.

Initially $G = (V_0, \emptyset, \{r \in V_0 \mid r \text{ is nullable}\}, \emptyset)$ where V_0 is some initial set of regexes that occur in initial membership constraints. An unsolved membership constraint $\mathit{in}(s, r)$ trigger a call to the regex solver that performs the steps below.

1. As shown in Figure 3a the DER-rule either allows the solution $s = \varepsilon$ if r is nullable, or it propagates the goal $\mathit{in\text{-}tr}(s, \delta^{\text{DNF}}(r))$ provided that r is not dead and s is nonempty.
2. The propagation rules for $\mathit{in\text{-}tr}(s, \delta^{\text{DNF}}(r))$ create a search space where the leaves of $\delta^{\text{DNF}}(r)$ eventually trigger new membership subgoals for $s_{1..}$ as shown by the ERE-rule.
3. In this process G is incrementally updated, triggered by $\boldsymbol{\mathit{Upd}}[r{\to}Q]$ where Q is the set $\mathbf{Q}(\delta^{\text{DNF}}(r))$ of all the derivative regexes for r and r is consequently closed, as shown by the UPD-rule in Figure 3b.

Transition regex normal form. Ensuring that these rules eventually prove unsatisfiability for regexes r denoting the empty language requires care. Notice that Figure 3a does not contain propagation rules for conjunction (intersection) and negation (complement) of transition regexes. This is because such rules would result in incompleteness. For example, consider the hypothetical rule that we reduce $\mathit{in\text{-}tr}(s, r_1 \& r_2)$ to $\mathit{in\text{-}tr}(s, r_1) \wedge \mathit{in\text{-}tr}(s, r_2)$. Then, if we apply this to the constraint $\mathit{in\text{-}tr}(s,$ `(.*a)&(.*b)`$)$, we obtain two separate constraints which propagate separately, and we never arrive at the required contradiction and conclude the original transition regex is unsatisfiable. More specifically, this would

[7] We say *sort* for *type* as is custom in the context of SMT.

occur after propagating rules DER and then ITE starting from $in(s, \texttt{(a.*a)\&(a.*b)})$, since $\delta^{\text{DNF}}(r) = \textbf{IF}(\texttt{a}, \texttt{(.*a)\&(.*b)}, \bot)$.

To avoid such issues with intersection and complement propagation is why we require that $\delta^{\text{DNF}}(r)$ is a normal form of $\delta(r)$: specifically, we require a DNF form where union and if-then-else are always pushed outwards over complement and intersection, and we enforce this when computing derivatives. In particular this requires using the *lift* rules for $r \in ERE$ (not for $r \in \mathbb{B}(RE)$)[8]. The implication is that when simplifying $in\text{-}tr(s, r)$, after applying ITE and OR as necessary, we can directly apply rule ERE to the conjunctions, which are plain regexes not involving if-then-else.

Using this strategy, we can then prove the following summary theorem about the properties of the membership propagation rules. Here $\vdash$ refers to infererce with respect to the rules in Figure 3a and Figure 3b. Recall that $r \equiv \bot$ means that $\mathbf{L}(r) = \emptyset$. The theorem states that the rules provide a decision procedure for emptiness of *ERE*s modulo any decidable character theory. The proof then uses the property that G represents an accurate reachability graph of the underlying symbolic automaton, where states that end up in $G.Dead$ are eqivalent to $\bot$, and where states may be intersection regexes.

Theorem 5.1. *Let $r \in ERE$ and s be an uninterpreted constant. Then $in(s, r) \vdash \bot$ iff $r \equiv \bot$.*

Complexity. Theorem 5.1 states that the decision procedure is sound and complete for regex emptiness, but does not discuss its complexity. In the worst case, complexity relates to the number of regexes in the space of all derivatives (recursively) of a regex. Studying this is a primary motivation for why we develop a theory of automata corresponding to symbolic extended regexes in Section 7. In particular, we give a complexity bound for the common case in practice of regexes in $\mathbb{B}(RE)$ in Theorem 7.3: for this case, we show that the number of states in an SBFA is linear. As leaves in the DNF $\delta^{\text{DNF}}(r)$ correspond to conjunctions of states in $\mathbb{B}(RE)$, this implies exponential worst-case complexity for the decision procedure here, for $\mathbb{B}(RE)$. For general extended regexes, nonemptiness is known to be non-elementary [44], so we can only hope for concrete complexity bounds in practical subclasses.

Alive and dead state detection. In the implementation the graph G incrementally maintains a DAG of strongly connected componets (SCCs) using the Union-Find datastructure [45] for implementing SCCs, and it implements explicit marking of SCCs corresponding to the *Dead* and *Alive* subsets of V. We let $Find(v)$ denote the SCC that contains v. The event of adding a new batch of edges to E causes an incremental cycle detection algorithm to be executed, that is immediately followed by an algorithm that incrementally updates the DAG of SCCs and propagates the markings of *Dead* and *Alive* vertices.

We implemented a custom variant of incremental cycle detection and SCC maintenance algorithms, that is similar in spirit to the algorithm described in [6]. A unique aspect of our algorithm is that it makes use of an additional *dissimilarity* heuristic asserting that certain states p and q can never belong to the same SCC, denoted by $p \nsim q$, because they can never be both in the same cycle. For example if $p = \texttt{abc}$ then $\delta(p) = \textbf{IF}(\texttt{a}, \texttt{bc}, \bot)$ and, let $q = \texttt{bc}$, trivially $p \nsim q$. This information is used by the DFS search algorithms in our incremental SCC algorithm to prune the search space during cycle detection.

6 Experiments

We have implemented symbolic Boolean derivatives as an extension to Z3, together with the strategies for normalizing derivatives and the sound decision procedure described in section 5. Our solver, dZ3, fully replaces the existing solver in Z3 for regular expression constraints which is based on symbolic automata. We carried out a series of experiments to compare our solver with Z3 and other state-of-the-art string solvers. Our interest is in evaluating the following questions:

Q1 Overall, does dZ3 match the performance of existing regular expression solvers on standard string constraint benchmarks?

Q2 How does dZ3 specifically fare on standard benchmarks which contain *Boolean combinations* of regular expression constraints on the same regex (which are equivalent to Boolean operations on SEREs), compared to the state of the art?

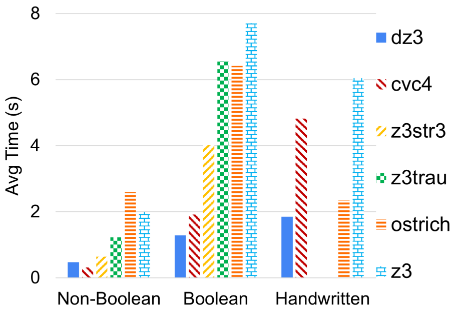
Q3 Finally, how does dZ3 fare on handcrafted difficult examples, designed to showcase the interaction of Boolean operations with other regex operators, compared to the state of the art?

To evaluate Q1, we assembled a collection of standard benchmark suites from the literature: Kaluza, Norn, Slog, and SyGuS-qgen, as collected by SMTlib [41, 42]. We add to this an existing set of benchmarks provided in [8, 40], which we call RegExLib: these ask for the answer to an intersection or containment problem between regular expressions taken from regexlib.com, an online library of regular expressions. From all of these sets, we removed benchmarks that do not contain any regular expression constraints, and some Norn benchmarks which contained existential quantification, as this was not allowed by the stated logic.

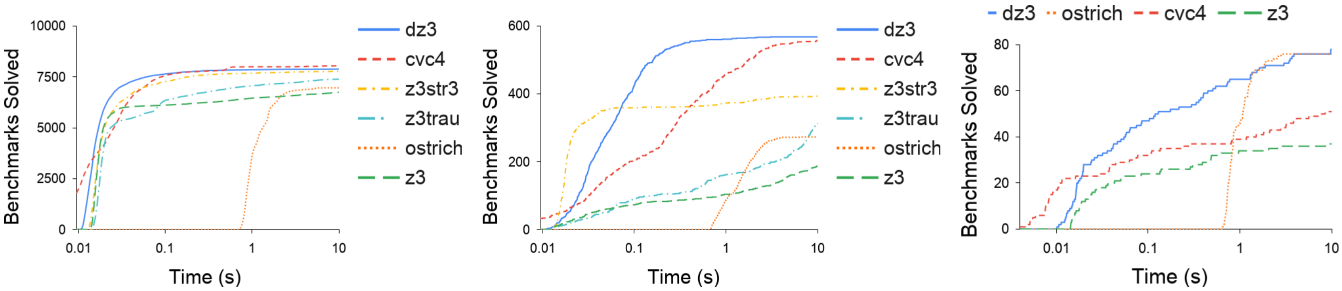
To evaluate Q2, the challenge arises of how to fairly compare with solvers which do not support explicit intersection and complement. To address this issue, we observe that although most standard benchmarks do not explicitly contain intersection and complement, a large number of benchmarks contain multiple regex membership consraints on the same string, which is logically equivalent to (and can be treated as) a Boolean combination. Therefore, we parsed the benchmarks from Q1 to divide them into *simple* benchmarks,

[8] Lift rules are given in Appendix D.

Solver	Solved			Avg (s)			Med (s)		
	NB	B	H	NB	B	H	NB	B	H
dz3	95.6%	**88.1%**	**87.6%**	0.47	**1.28**	**1.85**	**0.016**	0.06	**0.08**
cvc4	**97.6%**	86.4%	57.3%	**0.31**	1.92	4.82	0.019	0.30	3.18
z3str3	94.3%	60.9%	–	0.64	4.02	–	0.018	**0.03**	–
z3trau	89.6%	48.7%	–	1.22	6.56	–	0.020	TO	–
ostrich	84.5%	42.3%	85.4%	2.59	6.41	2.34	1.091	TO	0.92
z3	81.8%	29.0%	41.6%	1.99	7.70	6.05	0.018	TO	TO



(a) Summary of the experimental results on non-Boolean (NB), Boolean (B), and additional handcrafted benchmarks (H): percent of benchmarks solved, average time to solve, and median time to solve. Best solver is in bold. For comparison, errors, wrong answers, and crashes are treated as timeouts (10s). The average time in the table is plotted on the left.



(b) Cumulative number of benchmarks solved on non-Boolean (left), Boolean (middle), and handcrafted (right) benchmarks. The *x*-axis is time on a log-scale, and the *y*-axis shows number of benchmarks solved in that amount of time or less.

Benchmark	Quantity	Benchmark	Quantity	Benchmark	Quantity
Kaluza	5452	Norn	147	Date	20
Slog	1976	SyGuS-qgen	343	Password	34
Norn	813	RegExLib Intersection	55	Boolean + Loops	21
		RegExLib Subset	100	Determinization Blowup	14
Total Non-Boolean	8241	Total Boolean	645	Total Handwritten	89

(c) Benchmarks used for the evaluation. Existing benchmark suites (Kaluza, Slog, Norn, SyGuS, RegExLib) are classified as Boolean if they contain multiple constraints on the same regex.

Figure 4. Results of the experimental evaluation. (A full table of results can be found in the appendix.)

which do not contain multiple regular expression constraints on the same string variable, and *Boolean* benchmarks, which contain at least one instance of multiple regular expression constraints on the same string. Our hypothesis is that our solver is particularly suited to the Boolean case, as it translates such constraints succinctly to SEREs.

To evaluate Q3, we wrote four sets of examples. Unlike in Q2, we incorporate explicit intersection and complement. The first set contains problems involving *date* constraints, where a string is constrained to look like a date, as in Figure 1: the questions ask, e.g. whether one such constraint implies another or whether an intersection of such constraints is satisfiable. Such constraints naturally incorporate Boolean combinations: for example, if the month is February, then the day must not be 30 or 31. The second set contains problems involving *password* constraints, e.g. a password must contain at least one number and a letter, and no more than 20 characters, like the example in Section 2. Third, we have a set of regexes where Boolean operations interact with concatenation and iteration, in particular to create nontrivial unsatisfiable regexes. These also serve to test the dead state elimination described in section 5. Finally, we include classical examples which have small nondeterministic state spaces but blowup when determinized, to test efficiency of derivatives in avoiding determinization: these include variants of `(.*a.{k})&(.*b.{k})` where *k* is constant. Together with the benchmarks for Q1 and Q2, the number of benchmarks from various sources is summarized in Figure 4(c).

For all experiments, we compared dZ3 with a representative list of state-of-the-art and actively maintained solvers: Z3 [20, 49], Z3str3[7, 51], Z3-Trau [1, 50], CVC4 [5, 15], and OSTRICH [14, 35]. We exclude Z3str3 and Z3-Trau from the

Q3 handwritten examples, since explicit intersection and complement are not supported. We ran each solver with a 10s timeout, and compared the answer with the correct label (if provided with the benchmark); otherwise, we compared with the answer provided by a baseline solver that appears to be trained (and sound) for the benchmark set in question: for this purpose we used OSTRICH for the Norn benchmarks and CVC4 for Kaluza, Slog, and SyGuS-qgen (all others were labeled). If the baseline solver did not return a result, we marked the answer as "unchecked" and conservatively considered it correct. An answer of "unknown" is counted as an error. In summary, a correct result can be either sat, unsat, or unchecked, while an incorrect result can be either wrong, a timeout, or an error. We manually inspected solver errors and incorrect answers to ensure that they all appear to be unsupported cases, bugs, or crashes, and never a result of a malformed input (which we correct by replacing the input in question). The experiments were run on a Dell XPS13 with an Intel Core i7 CPU and 16GB of RAM.

Results. The results are summarized in Figure 4. dZ3 shows state-of-the-art performance and is consistently the best or near the best solver —- in terms of average time, median time, or number of benchmarks solved, across all three benchmark sets (Figure 4(a)). dZ3 shows particularly good performance on Boolean and handwritten benchmarks, where only CVC4 (on Boolean) and Ostrich (on handwritten) compare. However, compared to CVC4, dZ3 solves 87% of the handwritten benchmarks rather than 57.3%; and compared to Ostrich, dZ3 solves 88% of the Boolean benchmarks rather than 42.3%. No other solver does consistently well in all three categories. Overall, the plots in Figure 4(b) demonstrate that symbolic Boolean derivatives reach state-of-the-art performance in practice, while on benchmarks with Boolean combinations the solver solves more benchmarks faster than any existing tool.

7 Symbolic Boolean Finite Automata

In order to formally study the efficiency of our *ERE* implementation, and in particular, the state space of the set of derivatives, we explore a connection to automata. In particular, we formally define *symbolic Boolean finite automata* or SBFAs, a variant of alternating automata adapted to the symbolic setting. We show that derivatives of symbolic extended regexes correspond to states in a corresponding SBFA, and in the case of $R \in \mathbb{B}(RE)$, we prove a theorem that the state space size is linear in the size of R. This allows us to analyze the worst-case complexity of our decision procedure. SBFAs will also prove useful in comparing with alternative approaches and existing extensions of automata in Section 8.

SBFA. A *Symbolic Boolean Finite Automaton* or SBFA is a tuple $M = (\mathcal{A}, Q, \iota, F, q_\perp, \Delta)$ where $\mathcal{A}$ is the *alphabet theory*; Q is a finite set of *states*; $\iota \in \mathbb{B}(Q)$ is the *initial state combination*; $F \subseteq Q$ is the set of *final states*; $q_\perp \in Q \setminus F$ is the *bottom state*; $\Delta : Q \to TR_Q$ is the *transition function* such that $\Delta(q_\perp) = q_\perp$, where TR_Q is defined in Section 4.

We lift the *final* condition to $\mathbf{q} \in \mathbb{B}(Q)$ denoted $\nu_F(\mathbf{q})$ as follows: $\nu_F(q)$ iff $q \in F$, $\nu_F(\mathbf{p}|\mathbf{q})$ iff $\nu_F(\mathbf{p})$ or $\nu_F(\mathbf{q})$, $\nu_F(\mathbf{p}\&\mathbf{q})$ iff $\nu_F(\mathbf{p})$ and $\wedge\nu_F(\mathbf{q})$, and $\nu_F({\sim}\mathbf{q})$ iff not $\nu_F(\mathbf{q})$.

The definition of Δ is lifted similarly to $\mathbb{B}(Q) \to TR_Q$.

Semantics. M denotes $\mathbf{M} : \mathbb{B}(Q) \to \Sigma^*$ by the equations

$$\forall \mathbf{q} \in \mathbb{B}(Q) : \ \mathbf{M}(\mathbf{q}) = \{\epsilon \mid \nu_F(\mathbf{q})\} \cup \bigcup_{a \in \Sigma} a \cdot \mathbf{M}(\Delta(\mathbf{q})(a))$$

The *language* accepted by M is $\mathbf{L}(M) = \mathbf{M}(\iota)$.

Construction from Regexes. The construction of an SBFA from a regex $R \in ERE$ starts with the initial state combination $\iota = R$ and computes the rest of the states in Q as the fixpoint of all the states reachable as *terminals* of $\delta(q)$ for $q \in Q$, where, what constitutes as a terminal depends on the state granularity and/or normal form of the intended SBFA. With respect to the granularity that is as assumed below, a terminal of $\textsc{if}(\varphi, \tau, \rho)$ is a terminal of τ or ρ, a terminal of ${\sim}\tau$ is a terminal of τ, and a terminal of $\tau \diamond \rho$ is a terminal of τ or ρ. If $\tau \in RE$ then τ is a terminal. In this case, states (other than potentially ι and ${\sim}\perp = \mathtt{.*}$) are themselves not conjunctions or negations.

The regex $\perp$, that is the bottom state $q_\perp$, and the dual *top* state regex $\mathtt{.*}$ ($= {\sim}\perp$) are called *trivial*. Let $\mathbf{Q}(\tau)$ denote the set of all *nontrivial* terminals of a transition regex τ.

Given a regex R, let $\delta^+(R)$ denote $\mathbf{Q}(\delta(R))$ unioned with all states of derivatives that can be reached from $\mathbf{Q}(\delta(R))$. Formally, $\delta^+(R)$, is the least fixed point of the following equations, where S is a set of regexes,

$$\delta^+(R) = \mathbf{Q}(\delta(R)) \cup \delta^+(\mathbf{Q}(\delta(R))), \qquad \delta^+(S) = \bigcup_{R \in S} \delta^+(R).$$

Observe that $\delta^+(R)$ is the set of regexes reached after *one or more* derivations, which may but need not include R itself, e.g., $\delta^+(\mathtt{b(ab)*}) = \{\mathtt{(ab)*}, \mathtt{b(ab)*}\}$ includes the start regex while $\delta^+(\mathtt{ab}) = \{\mathtt{b}, \varepsilon\}$ does not.

Proposition 7.1. *$\delta^+(R)$ is finite.*

Proof. The are finitely many different states reached in $\delta^+(R)$ because $\mathbf{L}(R)$ is regular and because the various algebraic operations are represented concisely, e.g., & is idempotent, associative, commutative with unit element $\mathtt{.*}$ and absorbing element $\perp$. Similarly for | and $\cdot$. □

SBFA(R). The SBFA of $R \in ERE$ is defined as follows, where $Q = \delta^+(R) \cup \{R, \perp, \mathtt{.*}\}$ and $F = \{q \in Q \mid q \text{ is nullable}\}$.[9]

$$SBFA(R) = (\mathcal{A}, Q, R, F, \perp, \delta{\restriction}Q)$$

The following is the correctness theorem of *SBFA*(*R*).

[9] We write $\delta{\restriction}Q$ to denote δ restricted to the finite set Q — to follow the SBFA definition strictly.

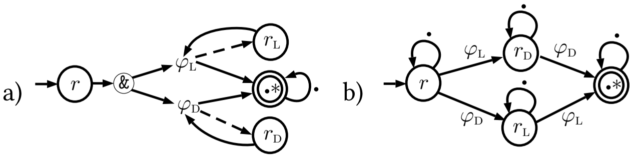


Figure 5. *SBFA*(r); $r = r_L \,\&\, r_D$; r_L = `.*[a-z].*`; r_D = `.*\d.*`.

Theorem 7.2. *Let $R \in ERE$ and $M = SBFA(R)$. Then for all* $\mathbf{q} \in \mathbb{B}(Q_M)$*,* $\mathbf{M}(\mathbf{q}) = \mathbf{L}(\mathbf{q})$*. In particular* $\mathbf{L}(M) = \mathbf{L}(R)$*.*

Proof. The statement follows by proving that $\forall \mathbf{q} \in \mathbb{B}(Q) : v \in \mathbf{M}(\mathbf{q}) \Leftrightarrow v \in \mathbf{L}(\mathbf{q})$ by induction over $|v|$. The base case $v = \epsilon$ follows because $\nu_F(\mathbf{q}) \Leftrightarrow \nu(\mathbf{q})$. The induction case is: $av \in \mathbf{M}(\mathbf{q})$ iff $v \in \mathbf{D}_a(\mathbf{M}(\mathbf{q}))$ iff $v \in \mathbf{M}(\delta(\mathbf{q})(a))$ iff (by the IH) $v \in \mathbf{L}(\delta(\mathbf{q})(a))$ iff (by Theorem 4.3) $v \in \mathbf{L}(D_a(\mathbf{q}))$ iff (by [9, Theorem 3.1]) $av \in \mathbf{L}(\mathbf{q})$. □

Theorem 7.3 is another key result. Here a regex is *normalized* when all concatenations are in right-associative form. A regex is *clean* if it contains no $\bot$ and no unsat predicates. Let $\sharp(R)$ denote the number of predicate nodes in R.

Theorem 7.3. *Let $R \in \mathbb{B}(RE)$. If R is clean and normalized then $|Q_{SBFA(R)}| \leq \sharp(R) + 3$.*[10]

For $R \in ERE$ we do not have a linear bound on $|Q_{SBFA(R)}|$ because the lifting in $(\tau \& \rho){\cdot}R = lift(\tau \& \rho){\cdot}R$ that first transforms $\tau \& \rho$ into DNF, may lead to an exponential blowup.

Example 7.4. Recall r_D = `.*\d.*` from Section 2 and let r_L = `.*[a-z].*` . So r_L matches any string containing at least one lower-case letter. Let φ_L = `[a-z]` and φ_D = `\d`. Then

$$\begin{aligned}
\delta(r_L) &= r_L \,|\, \mathbf{IF}(\varphi_L, .*, \bot) \equiv \mathbf{IF}(\varphi_L, .*, r_L)\\
\delta(r_D) &= r_D \,|\, \mathbf{IF}(\varphi_D, .*, \bot) \equiv \mathbf{IF}(\varphi_D, .*, r_D)\\
\delta(r) &= \delta(r_L) \,\&\, \delta(r_D) = \mathbf{IF}(\varphi_L, .*, r_L) \,\&\, \mathbf{IF}(\varphi_D, .*, r_D)
\end{aligned}$$

SBFA(r) is shown in Figure 5a. The DNF equivalent is shown in Figure 5b where the default operation is disjunction. ⊠

8 Related Work

Here we provide a formal study of the relationship between symbolic derivatives and related formalisms that can be used in the context of decision procedures for *ERE*. In particular, we first compare with classical derivatives of regular expressions and existing extensions. Next, we compare with existing extensions of classical finite automata and symbolic automata. Finally, we discuss work related to string solvers and implementation of the proposed techniques in the context of SMT solvers.

8.1 Relation to Classical Derivatives

The theory of derivatives of regular expressions has evolved in parallel and largely independently of the mainstream automata research. One of the key features of derivatives is that they provide a lazy and a more algebraic perspective on how finite automata and their regular expression counterparts are related; basic theoretical properties between various classical automata and their derivatives are discussed in [2].

The connection between *ERE* (modulo $\mathcal{A}$) and symbolic derivatives was initially studied in-depth in [26], with the main application of language containment in *ERE*. An important side result [26, Section 5] is that classical derivatives do not directly generalize to predicates, and a workaround is to combine *positive* and *negative* derivatives. We have shown here that a remedy is to use *conditionals*.

In the following we discuss the exact relationship to well-established related classical notions, first Brzozowski derivatives [9] and then Antimirov derivatives [3] and its generalization to *ERE* [12]. We show how they relate to $\delta(R)$ for $R \in RE$. Assume Σ is *finite*, let $a \in \Sigma$, and let $R_a = \delta(R)(a)$.

Brzozowski Derivatives. R_a is precisely the *Brzozowski* derivative [9, Theorem 3.1] $D_a(R)$ of R wrt a.[11] If regexes are viewed as DFA states, D_a is the transition function for a.

Antimirov Derivatives. If $R_a = \bot$ then $\partial_a(R) = \emptyset$ else $R_a = |_{i=1}^{n} R_i$ and $\partial_a(R) = \{R_i\}_{i=1}^{n}$ is the *Antimirov* derivative [3, Definition 2.8] of R wrt a as a set of *partial* derivatives R_i. When viewed as states, each R_i corresponds to a separate target state of a transition $(R, \mathtt{a}, R_i)$ of an NFA.

Partial Derivatives of ERE. The Antimirov construction is extended to *ERE* in [12]. The formal construction $\frac{\partial}{\partial_a}(R)$ in [12, Definition 2] inlines negation, inlines concatenation propagation, and inlines conjuction distribution, in the definition of $\frac{\partial}{\partial_a}$ so that the result is essentially an |-set of &-sets. Intuitively $\frac{\partial}{\partial_a}(R) = DNF(R_a)$.

8.2 Relation to Classical Automata

Parallel finite automata by Kozen [28], subsequently renamed to *alternating finite automata* or *AFAs* in [13], and *Boolean finite automata* or *BFAs* by Brzozowski and Leiss [10], were introduced independently (cf [10, p.25]) and use fairly different formalizations and application contexts in doing so. While both work over a finite state space Q and are equivalent classically, their differing notation becomes important symbolically: BFAs use transitons to $\mathbb{B}(Q)$ while AFAs use transitions to 2^{2^Q} encoding $DNF(\mathbb{B}^+(Q))$. We provide a description of SBFAs over finite alphabets as BFAs next.[12]

BFA. Let $M = (\mathcal{A}, Q, \iota, F, q_\bot, \Delta)$ be a SBFA. The equivalent BFA of M is $BFA(M) = (\Sigma, Q, \lambda(q,a).\Delta(q)(a), \iota, F)$.

Proposition 8.1. $\mathbf{L}(M) = L(BFA(M))$ *with L as in [10, p.25].*

8.3 Relation to Symbolic Extensions of Automata

Symbolic alternating finite automata (s-AFAs) [16] and alternating data automata (ADAs) [25] are two orthogonal symbolic extensions of finite automata, in the former case via SFAs and in the latter case via data automata [24].

[10] See Appendix B for a detailed proof.

[11] D_a applies to the whole *ERE* class.

[12] See more discussion on this topic in Appendix C.

Symbolic Alternating Finite Automata. An s-AFA [16] (modulo $\mathcal{A}$) is a generalization of an SFA by allowing transition targets to be elements in $\mathbb{B}^+(Q)$ where Q is a finite set of states. There is an initial state combination $\iota \in \mathbb{B}^+(Q)$, a set of final states $F \subseteq Q$, and a finite set of transitions $\Delta \subseteq Q \times \Psi \times \mathbb{B}^+(Q)$. Let $M_{\text{SAFA}} = (\mathcal{A}, Q, \iota, F, \Delta)$

The equivalent SBFA of M_{SAFA} is defined as follows with a bottom state $q_\perp \notin Q$, and where $OR(\emptyset) = q_\perp$.

$$\begin{aligned}&SBFA(M_{\text{SAFA}}) = (\mathcal{A}, Q \cup \{q_\perp\}, \iota, F, q_\perp,\\ &\quad \{q_\perp \mapsto q_\perp\} \cup \textstyle\bigcup_{q \in Q}\{q \mapsto OR\{\textbf{IF}(\psi, \boldsymbol{p}, q_\perp) \mid (q, \psi, \boldsymbol{p}) \in \Delta\}\})\end{aligned}$$

Proposition 8.2. $\mathbf{L}(SBFA(M_{SAFA})) = \mathscr{L}(M_{SAFA})$

Going from SBFA $M = (\mathcal{A}, Q, \iota, F, q_\perp, \Delta)$ to s-AFA is possible but not easy in general. This is also related to why ~ is not supported in s-AFA [16]. W.l.o.g., assume that Δ does not contain complement. This is achieved by adding negated states $\bar{q}$ to Q and for each negated state $\bar{q}$ letting $\Delta(\bar{q}) = NNF({\sim}\Delta(q))$ where $NNF(\tau)$ computes the *negation normal form* of τ meaning that all negations are pushed down to states. In particular, $NNF({\sim}\textbf{IF}(\varphi, \tau, \rho)) = \textbf{IF}(\varphi, NNF({\sim}\tau), NNF({\sim}\rho))$, and $NNF({\sim}q) = \bar{q}$. The other cases are standard.

The equivalent s-AFA of M is defined as follows where $\tau(\alpha) = \tau(a)$ for some $a \in [\![\alpha]\!]$ — which is well-defined (independednt of choice) due to the local mintermization.

$$\begin{aligned}&SAFA(M) = (\mathcal{A}, Q, NNF(\iota), F,\\ &\quad \{(q, \alpha, \Delta(q)(\alpha)) \mid q \in Q, \alpha \in Minterms(Guards(\Delta(q)))\})\end{aligned}$$

Proposition 8.3. $\mathbf{L}(M) = \mathscr{L}(SAFA(M))$

The problem with this construction is that $|Minterms(\Gamma)|$ can be exponential in $|\Gamma|$ so the construction of $SAFA(M)$ is exponential in the worst case. The same problem arises in s-AFA *normalization* [16] used for complementation.

Alternating Data Automata. This class of automata goes far beyond regular languages because registers are allowed to carry information across state boundaries so that consequtive data elements in traces are functionally related. Data automata, as defined in [24], use registers and have the expressive power of general Turing machines. In an *alternating* data automaton [25], arbitrary Boolean combinations of predicates can be used to relate before and after values of registers. It is stated in [24] that complement of alternating data automata is linear unlike in [16]. We are not aware of work relating *ERE* with ADAs.

Conditional Branching. Conditional transitions (although without Boolean combinations of states) have been used before in a special class of deterministic symbolic transducers called *Branching Symbolic Transducers* or *BSTs* [38]. The main motivation behind BSTs is in the context of data processing pipelines where they preserve condition evaluation order and in this way support more direct and efficient serial code generation. A BST has a finite state space Q, and when the BST acts as a finite state automaton, its rules correspond to a subset of TR_Q without Boolean operations. Conditional transitions are also used in the implementation of MONA [27] where transitions are multi-terminal BDDs whose terminals are states. We apply similar principles in dZ3 to represent transition regexes in a canoniocal way.

8.4 Related Work in SMT

String and regex constraints have been the focus of both SMT and CP solving communities, with several tools being developed over the past decade. A theory of strings with regexes is a standard part of the SMTLIB2 format [46]. String solvers are integrated in the CDCL(T) architecture [34]. From the CP community, the MiniZinc format integrates membership constraints over regular languages presented as either DFAs or NFAs [32]. The solver presented in [29] is closely related to ours in that it relies on Antimirov derivatives to reduce positive regular expression membership constraints. It diverges from our approach as it handles intersection similar to [12], instead of using symbolic derivatives. Consistent with what the empirical evaluation suggests, complementation is not treated in depth and is essentially out of scope of this work. Ostrich is advertised as a symbolic solver for string formulas that come from path constraints [14], and its solver is based on solving for pre-images. Our evaluation suggests that it performs either extremely well, or not at all, depending on categories. While full handling of regexes seems out of scope of z3-Trau, *flat* automata were recently applied [1] for solving symbolic constraints that include string-to-int and int-to-string conversions. Z3Str3 [7] integrates several innovations around string equality solving. Many of the advances previously developed in S3 [47] and now integrated within Z3's default string solver, hence dz3 benefits from these results. ZELKOVA is a tool used internally by Amazon to check AWS policy configurations, it uses a custom NFA engine based extension of Z3 to handle regex constraints [4].

9 Conclusion

In this paper, we generalized the finite-alphabet based work of derivatives to work over a symbolic alphabet and to incorporate Boolean combinations, and showed how to use such symbolic Boolean derivatives to solve regular expression membership constraints in SMT. Our solver, dZ3, achieves state-of-the-art performance on standard benchmark sets, and significant speedup on constraints involving intersection and complement, where no existing solver does consistently well across benchmark sets. While we have experimentally validated the main ideas, many further promising optimizations remain to be explored; in particular taking advatage of algebraic laws of derivatives of *ERE*s and designing heuristics that capture common usage patterns and that can be exploited by CDCL-based solvers.

References

[1] Parosh Aziz Abdulla, Mohamed Faouzi Atig, Yu-Fang Chen, Bui Phi Diep, Julian Dolby, Petr Janku, Hsin-Hung Lin, Lukás Holík, and Wei-Cheng Wu. 2020. Efficient handling of string-number conversion. In *Proceedings of the 41st ACM SIGPLAN International Conference on Programming Language Design and Implementation, PLDI 2020, London, UK, June 15-20, 2020*, Alastair F. Donaldson and Emina Torlak (Eds.). ACM, 943–957. https://doi.org/10.1145/3385412.3386034

[2] Cyril Allauzen and Mehryar Mohri. 2006. A unified construction of the Glushkov, Follow, and Antimirov automata. In *International Symposium on Mathematical Foundations of Computer Science*. Springer, 110–121.

[3] Valentin Antimirov. 1995. Partial Derivatives of Regular Expressions and Finite Automata Constructions. *Theoretical Computer Science* 155 (1995), 291–319.

[4] John Backes, Pauline Bolignano, Byron Cook, Catherine Dodge, Andrew Gacek, Kasper Søe Luckow, Neha Rungta, Oksana Tkachuk, and Carsten Varming. 2018. Semantic-based Automated Reasoning for AWS Access Policies using SMT. In *2018 Formal Methods in Computer Aided Design, FMCAD 2018, Austin, TX, USA, October 30 - November 2, 2018*, Nikolaj Bjørner and Arie Gurfinkel (Eds.). IEEE, 1–9. https://doi.org/10.23919/FMCAD.2018.8602994

[5] Clark Barrett, Christopher L Conway, Morgan Deters, Liana Hadarean, Dejan Jovanović, Tim King, Andrew Reynolds, and Cesare Tinelli. 2011. Cvc4. In *International Conference on Computer Aided Verification*. Springer, 171–177.

[6] Michael A. Bender, Jeremy T. Fineman, Seth Gilbert, and Robert Endre Tarjan. 2011. A New Approach to Incremental Cycle Detection and Related Problems. *CoRR* abs/1112.0784 (2011). http://arxiv.org/abs/1112.0784

[7] Murphy Berzish, Vijay Ganesh, and Yunhui Zheng. 2017. Z3str3: A string solver with theory-aware heuristics. In *2017 Formal Methods in Computer Aided Design (FMCAD)*. IEEE, 55–59.

[8] Nikolaj Bjørner, Vijay Ganesh, Raphael Michel, and Margus Veanes. 2012. An SMT-LIB Format for Sequences and Regular Expressions. In *SMT'12*, P. Fontaine and A. Goel (Eds.). 76–86.

[9] Janusz A. Brzozowski. 1964. Derivatives of regular expressions. *JACM* 11 (1964), 481–494.

[10] J. A. Brzozowski and E. Leiss. 1980. On equations for regular languages, finite automata, and sequential networks. *Theoretical Computer Science* 10 (1980), 19–35.

[11] Tevfik Bultan, Fang Yu, Muath Alkhalaf, and Abdulbaki Aydin. 2017. *String Analysis for Software Verification and Security*. Springer.

[12] Pascal Caron, Jean-Marc Champarnaud, and Ludovic Mignot. 2011. Partial Derivatives of an Extended Regular Expression. In *Language and Automata Theory and Applications, LATA 2011 (LNCS)*, Vol. 6638. Springer, 179–191.

[13] Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *JACM* 28, 1 (1981), 114–133.

[14] Taolue Chen, Matthew Hague, Anthony W Lin, Philipp Rümmer, and Zhilin Wu. 2019. Decision procedures for path feasibility of string-manipulating programs with complex operations. *Proceedings of the ACM on Programming Languages* 3, POPL (2019), 1–30.

[15] CVC4. 2020. (2020). https://github.com/CVC4/CVC4.

[16] Loris D'Antoni, Zachary Kincaid, and Fang Wang. 2018. A Symbolic Decision Procedure for Symbolic Alternating Finite Automata. *Electronic Notes in Theoretical Computer Science* 336 (2018), 79–99.

[17] Loris D'Antoni and Margus Veanes. 2014. Minimization of Symbolic Automata. *ACM SIGPLAN Notices – POPL'14* 49, 1 (2014), 541–553. https://doi.org/10.1145/2535838.2535849

[18] Loris D'Antoni and Margus Veanes. 2020. Automata Modulo Theories. *Commun. ACM* (2020).

[19] James C Davis. 2019. Rethinking Regex engines to address ReDoS. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1256–1258.

[20] Leonardo de Moura and Nikolaj Bjørner. 2008. Z3: An Efficient SMT Solver. In *TACAS'08 (LNCS)*. Springer, 337–340.

[21] Dan Gusfield. 1997. Algorithms on stings, trees, and sequences: Computer science and computational biology. *Acm Sigact News* 28, 4 (1997), 41–60.

[22] J.G. Henriksen, J. Jensen, M. Jørgensen, N. Klarlund, B. Paige, T. Rauhe, and A. Sandholm. 1995. Mona: Monadic Second-order logic in practice. In *TACAS '95 (LNCS)*, Vol. 1019. Springer.

[23] Hossein Hojjat, Philipp Rümmer, and Ali Shamakhi. 2019. On Strings in Software Model Checking. In *APLAS (LNCS)*, A. Lin (Ed.), Vol. 11893. Springer.

[24] R. Iosif, A. Rogalewicz, and T. Vojnar. 2016. Abstraction refinement and antichains for trace inclusion of infinite state systems. In *TACAS'16 (LNCS)*, Vol. 9636. Springer, 71–89.

[25] Radu Iosif and Xiao Xu. 2018. Abstraction Refinement for Emptiness Checking of Alternating Data Automata. In *TACAS'18*, Dirk Beyer and Marieke Huisman (Eds.). Springer, 93–111.

[26] Matthias Keil and Peter Thiemann. 2014. Symbolic Solving of Extended Regular Expression Inequalities. In *FSTTCS'14 (LIPIcs)*. 175–186.

[27] Nils Klarlund, Anders Møller, and Michael I. Schwartzbach. 2002. MONA Implementation Secrets. *International Journal of Foundations of Computer Science* 13, 4 (2002), 571–586.

[28] Dexter Kozen. 1976. On parallelism in Turing machines. In *17th Annual Symposium on Foundations of Computer Science, FOCS'76*. IEEE Xplore, 89–97.

[29] Tianyi Liang, Nestan Tsiskaridze, Andrew Reynolds, Cesare Tinelli, and Clark Barrett. 2015. A Decision Procedure for Regular Membership and Length Constraints over Unbounded Strings?. In *FroCoS 2015: Frontiers of Combining Systems (LNCS)*, Vol. 9322. Springer, 135–150.

[30] Microsoft. 2020. *Azure Resource Manager documentation*. https://docs.microsoft.com/en-us/azure/azure-resource-manager/.

[31] Microsoft. 2020. *.NET regular expressions*. https://docs.microsoft.com/en-us/dotnet/standard/base-types/regular-expressions.

[32] MiniZinc. 2020. https://www.minizinc.org. (2020).

[33] Mehryar Mohri. 1996. On some applications of finite-state automata theory to natural language processing. *Natural Language Engineering* 2, 1 (1996), 61–80.

[34] Robert Nieuwenhuis, Albert Oliveras, and Cesare Tinelli. 2006. Solving SAT and SAT Modulo Theories: From an abstract Davis–Putnam–Logemann–Loveland procedure to DPLL(T). *J. ACM* 53, 6 (2006), 937–977. https://doi.org/10.1145/1217856.1217859

[35] Ostrich. 2020. (2020). https://github.com/uuverifiers/ostrich/.

[36] Scott Owens, John Reppy, and Aaron Turon. 2009. Regular-expression derivatives re-examined. *Journal of Functional Programming* 19, 2 (2009), 173–190.

[37] passwords generator.org. 2020. (2020). https://passwords-generator.org/.

[38] Olli Saarikivi, Margus Veanes, Todd Mytkowicz, and Madan Musuvathi. 2017. Fusing Effectful Comprehensions. In *ACM SIGPLAN Notices – PLDI'17*. ACM.

[39] Reetinder Sidhu and Viktor K Prasanna. 2001. Fast regular expression matching using FPGAs. In *The 9th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM'01)*. IEEE, 227–238.

[40] SMT. 2012. (2012). https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/nbjorner-microsoft.automata.smtbenchmarks.zip.

[41] SMTLib. 2020. (2020). https://clc-gitlab.cs.uiowa.edu:2443/SMT-LIB-benchmarks/QF_S.

[42] SMTLib. 2020. (2020). https://clc-gitlab.cs.uiowa.edu:2443/SMT-LIB-benchmarks/QF_SLIA.

[43] stackoverflow.com. 2020. Regex for password must contain at least eight characters, at least one number and both lower and uppercase letters and special characters. (2020). https://stackoverflow.com/questions/19605150/regex-for-password-must-contain-at-least-eight-characters-at-least-one-number-a.

[44] Larry J Stockmeyer and Albert R Meyer. 1973. Word problems requiring exponential time (preliminary report). In *Proceedings of the fifth annual ACM symposium on Theory of computing*. 1–9.

[45] Robert E. Tarjan. 1975. Efficiency of a good but not linear set union algorithm. *JACM* 22 (1975), 215–225.

[46] Cesare Tinelli, Clark Barrett, and Pascal Fontaine. 2020. (2020). http://smtlib.cs.uiowa.edu/theories-UnicodeStrings.shtml.

[47] Minh-Thai Trinh, Duc-Hiep Chu, and Joxan Jaffar. 2014. S3: A Symbolic String Solver for Vulnerability Detection in Web Applications. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. Association for Computing Machinery, New York, NY, USA, 1232–1243. https://doi.org/10.1145/2660267.2660372

[48] Margus Veanes, Nikolaj Bjørner, and Leonardo de Moura. 2010. Symbolic Automata Constraint Solving. In *Logic for Programming, Artificial Intelligence, and Reasoning. LPAR 2010 (LNCS)*, C.G. Fermüller and A. Voronkov (Eds.), Vol. 6397. Springer, 640–654.

[49] Z3. 2020. (2020). https://github.com/z3prover/z3.

[50] Z3-Trau. 2020. (2020). https://github.com/diepbp/z3-trau.

[51] Z3str3. 2020. (2020). https://sites.google.com/site/z3strsolver/.

A Full experimental results

Figure 6 contains the full experimental results, which were described in Section 6 and summarized in Figure 4.

B Proof of Theorem 7.3

We need the following lemma.

Lemma B.1. *If $X, Z \in RE$ are clean and normalized then $\delta^+(XZ) = \delta^+(X)Z \cup \delta^+(Z)$; if $X = S*$ then $\delta^+(X) = \delta^+(S)X$.*

Proof. We prove by induction over X that

$$\delta^+(XZ) = \delta^+(X)Z \cup \delta^+(Z).$$

It follows from working with normalized regexes that in a concatenation node the first element is not a concatenation and we apply case analysis over the first element, that is not an intersection or complement because here we only consider standard regexes.

Base case $X = \varepsilon$***.*** Follows immediately because $\delta(\varepsilon) = \emptyset$.

Induction case $X = \psi Y$***.***
Then $\delta(XZ) = \mathbf{IF}(\psi, \varepsilon, \bot) \cdot YZ = \mathbf{IF}(\psi, YZ, \bot)$, so $\mathbf{Q}(\delta(XZ)) = \{YZ\}$ and thus

$$\begin{aligned}
\delta^+(XZ) &= \{YZ\} \cup \delta^+(YZ) \\
&\overset{IH}{=} \{YZ\} \cup \delta^+(Y)Z \cup \delta^+(Z) \\
&= (\{Y\} \cup \delta^+(Y))Z \cup \delta^+(Z) \\
&= \delta^+(X)Z \cup \delta^+(Z)
\end{aligned}$$

Induction case $X = (X_1|X_2)Y$***.***
Then $\delta(XZ) = (\delta(X_1YZ) \,|\, \delta(X_2YZ))$, so

$$\begin{aligned}
\delta^+(XZ) &= \delta^+(X_1YZ) \cup \delta^+(X_2YZ) \\
&\overset{2\times IH}{=} \delta^+(X_1Y)Z \cup \delta^+(X_2Y)Z \cup \delta^+(Z) \\
&\overset{2\times IH}{=} (\delta^+(X_1)Y \cup \delta^+(Y))Z \cup \\
&\qquad (\delta^+(X_2)Y \cup \delta^+(Y))Z \cup \delta^+(Z) \\
&= (\delta^+(X_1)Y \cup \delta^+(X_2)Y \cup \delta^+(Y))Z \cup \delta^+(Z) \\
&= (\delta^+(X_1|X_2)Y \cup \delta^+(Y))Z \cup \delta^+(Z) \\
&\overset{(\star)}{=} \delta^+(X)Z \cup \delta^+(Z)
\end{aligned}$$

In $(\star)$, if $Y = \varepsilon$, the equality holds by definition of derivatives of a choice node. If $Y \neq \varepsilon$, then $X_1|X_2$ is smaller than X, and one can apply the IH on $(X_1|X_2)Y$ with $X_1|X_2$ as X and Y as an instance of the universal variable Z in the lemma.

Induction case $X = S*Y$***.***
Then $\delta(X) = \delta(S)X \,|\, \delta(Y)$ because $S*$ is nullable. The proof step uses (1), for any normalized W:

$$\delta^+(S*W) = \delta^+(S)S*W \cup \delta^+(W) \tag{1}$$

Equation (1) is proved first as follows:

$$\begin{aligned}
\delta^+(S*W) &= \delta^+(SS*W) \cup \delta^+(W) \\
&\overset{\text{(IH)}}{=} \delta^+(S)S*W \cup \delta^+(S*W) \cup \delta^+(W) \\
&\overset{\text{(lfp)}}{=} \delta^+(S)S*W \cup \delta^+(W)
\end{aligned}$$

where (lfp) holds because $\delta^+(S*W) \subseteq \delta^+(S)S*W \cup \delta^+(W)$ that can be shown separately. It follows that

$$\begin{aligned}
\delta^+(XZ) &= \delta^+(S*(YZ)) \\
&\overset{(1)}{=} \delta^+(S)S*YZ \cup \delta^+(YZ) \\
&\overset{IH}{=} \delta^+(S)S*YZ \cup \delta^+(Y)Z \cup \delta^+(Z) \\
&= (\delta^+(S)S*Y \cup \delta^+(Y))Z \cup \delta^+(Z) \\
&\overset{(1)}{=} \delta^+(S*Y)Z \cup \delta^+(Z) \\
&= \delta^+(X)Z \cup \delta^+(Z)
\end{aligned}$$

The statement follows by the induction principle. Observe that (1) implies the second part of the lemma by setting $W = \varepsilon$. □

Proof of Theorem 7.3.

Proof. If R is normalized we can use a slighltly condensed definition of $\delta(R)$ that is inlined in the proof. We prove (2)

$$|\delta^+(R)| \leq \sharp(R) \tag{2}$$

by induction over $R = R_1 \cdot Z$ where R_1 is not a concatenation and possibly $Z = \varepsilon$, corresponding to the case that R is not a concatenation or that R is a conjuction or complement.

Base case $R = \varepsilon$***.*** Then $|\delta^+(R)| = 0 = \sharp(R)$.

Induction case $R = \psi Z$***.*** Then $\delta(\psi Z) = \mathbf{IF}(\psi, Z, \bot)$ and so $\delta^+(R) = \{Z\} \cup \delta^+(Z)$. Here $Z \in RE$ counts for one terminal and ψ counts for one predicate node. Thus

$$|\delta^+(R)| = |\delta^+(Z)| + 1 \overset{\text{IH}}{\leq} \sharp(Z) + 1 = \sharp(R).$$

Induction case $R = (X|Y)Z$***.*** Then $\delta(R) = \delta(XZ) \mid \delta(YZ)$ and so $\delta^+(R) = \delta^+(XZ) \cup \delta^+(YZ)$. Then, by Lemma B.1,

$$\delta^+(R) = \delta^+(XZ) \cup \delta^+(YZ) = \delta^+(X)Z \cup \delta^+(Y)Z \cup \delta^+(Z)$$

which implies that (observe that, for a set S, $|S \cdot Z| = |S|$)

$$\begin{aligned}
&|\delta^+(R)| \leq |\delta^+(X)| + |\delta^+(Y)| + |\delta^+(Z)| \overset{\text{IH}}{\leq} \\
&\sharp(X) + \sharp(Y) + \sharp(Z) = \sharp(R).
\end{aligned}$$

Induction case $R = S*Z$***.*** Then $\delta(R) = \delta(S)S*Z \mid \delta(Z)$ and so, by using Lemma B.1, $\delta^+(R) = \delta^+(S)S*Z \cup \delta^+(Z)$. Then

$$|\delta^+(R)| \leq |\delta^+(S)| + |\delta^+(Z)| \overset{\text{IH}}{\leq} \sharp(S) + \sharp(Z) = \sharp(R).$$

Induction case $R = (X\&Y)$***.*** Then $\delta(R) = \delta(X)\&\delta(Y)$ and thus $\delta^+(R) = \delta^+(X) \cup \delta^+(Y)$. It follows that

$$|\delta^+(R)| \leq |\delta^+(X)| + |\delta^+(Y)| \overset{\text{IH}}{\leq} \sharp(X) + \sharp(Y) = \sharp(R).$$

	Solver	Time (s)						Result					
		< .04	< .12	< .37	< 1.1	< 3.3	< 10	sat	unsat	unchk	wrong	tmout	err
Kaluza	dz3	5018	71	48	22	15	10	2608	2576	0	0	268	0
	z3	4325	582	77	30	38	47	2521	2578	0	0	353	0
	z3str3	4439	569	241	22	33	6	2728	2577	5	0	127	15
	z3trau	3998	728	259	104	63	96	2657	2591	0	0	204	0
	cvc4	3744	1323	62	183	6	122	2849	2591	0	0	12	0
	ostrich	0	0	0	1747	2369	65	1665	2516	0	0	0	1271
Slog	dz3	1884	55	23	3	1	0	798	1168	0	0	10	0
	z3	934	101	4	31	30	36	71	1065	0	0	840	0
	z3str3	1143	542	89	36	25	21	784	1072	0	0	120	0
	z3trau	1224	178	186	138	106	61	727	1166	0	1	82	0
	cvc4	1887	61	24	4	0	0	808	1168	0	0	0	0
	ostrich	0	0	0	1363	583	21	800	1167	0	0	1	8
Norn	dz3	366	282	69	13	2	0	594	138	0	0	81	0
	z3	76	103	98	124	51	90	469	73	0	0	274	0
	z3str3	626	2	0	1	0	0	567	62	0	0	187	0
	z3trau	170	78	5	1	1	0	208	47	0	115	0	446
	cvc4	544	132	27	2	30	5	591	149	0	0	73	3
	ostrich	0	0	0	439	377	0	597	219	0	0	0	0
Norn	dz3	82	13	3	1	0	0	67	32	0	0	48	0
	z3	44	30	9	6	3	2	63	31	0	0	53	0
	z3str3	77	0	0	0	0	0	60	17	0	0	70	0
	z3trau	47	50	4	0	0	0	34	67	0	27	0	19
	cvc4	96	25	3	3	1	0	66	62	0	0	19	0
	ostrich	0	0	0	90	57	0	67	80	0	0	0	0
SyGuS-qgen	dz3	126	176	41	0	0	0	331	0	12	0	0	0
	z3	0	0	0	0	14	51	65	0	0	0	278	0
	z3str3	277	4	0	0	0	0	273	0	8	0	41	21
	z3trau	0	0	8	51	24	120	201	0	2	0	105	35
	cvc4	21	17	124	102	62	7	333	0	0	0	10	0
	ostrich	0	0	0	0	0	0	0	0	0	0	0	343
RegExLib Intersection	dz3	6	9	14	4	2	0	26	9	0	0	20	0
	z3	1	2	3	12	9	0	4	23	0	0	28	0
	z3str3	2	1	6	12	6	0	4	23	0	0	28	0
	z3trau	2	0	4	12	8	0	3	23	0	0	29	0
	cvc4	2	9	4	1	1	3	20	0	0	0	35	0
	ostrich	0	0	0	11	34	0	25	20	0	0	0	10
RegExLib Subset	dz3	26	28	27	5	5	0	90	1	0	0	9	0
	z3	0	0	0	2	5	3	7	3	0	0	90	0
	z3str3	0	0	0	2	4	3	6	3	0	0	91	0
	z3trau	0	0	0	0	5	4	6	3	0	0	91	0
	cvc4	17	46	12	2	1	2	80	0	0	0	20	0
	ostrich	0	0	0	12	69	0	72	9	0	0	0	19
Handwr.	dz3	35	14	7	9	7	6	42	36	0	0	10	1
	z3	20	4	4	6	2	1	14	23	0	2	46	4
	cvc4	28	6	3	2	7	5	28	23	0	0	25	13
	ostrich	0	0	0	52	24	0	40	36	0	5	6	2

Figure 6. Full results of the experiments, divided by double lines into non-Boolean benchmarks (regular expression constraints are on separate variables, top), Boolean benchmarks (multiple regular expression constraints on the same variable, middle), and additional handcrafted Boolean examples (bottom).

Induction case $R = {\sim}X$. Here $\delta(R) = {\sim}\delta(X)$. It follows that

$$|\delta^+(R)| = |\delta^+(X)| \overset{\text{IH}}{\le} \sharp(X) = \sharp(R).$$

Equation (2) follows by the induction principle So $Q_{\mathrm{SBFA}(R)} = \delta^+(R) \cup \{R, \bot, .*\}$, where $.* = {\sim}\bot$, and, by (2), $|Q_{\mathrm{SBFA}(R)}| \le |\delta^+(R)| + 3 \le \sharp(R) + 3$. □

C More on Relation to AFAs and BFAs

Algebra $\mathcal{A}$ is assumed to be such that Σ is finite and for each $a \in \Sigma$ there is a predicate $\hat{a}$ such that $[\![\hat{a}]\!] = \{a\}$. In a pure classical setting of finite automata, transition functions are usually parameterized by single characters, so the notion of character predicates seems vacuous. In our translation below we will make use of $\mathcal{A}$, where input predicates such as $\neg\hat{a}$ arise implicitly, because for example, a transition predicate ${\sim}\textbf{IF}(\hat{a}, q, \bot)$ simplifies to $\textbf{IF}(\hat{a}, \overline{q}, \overline{q_\bot})$ that logically corresponds to $\textbf{IF}(\hat{a}, {\sim}q, \bot)\,|\,\textbf{IF}(\neg\hat{a}, q_\top, \bot)$. Perhaps the most common use of predicates is that $\textbf{IF}(\alpha, q, \bot)\,|\,\textbf{IF}(\beta, q, \bot)$ reduces to $\textbf{IF}(\alpha \vee \beta, q, \bot)$, and, analogously, $\textbf{IF}(\alpha, q, \bot)\&\textbf{IF}(\beta, q, \bot)$ reduces to $\textbf{IF}(\alpha \wedge \beta, q, \bot)$.

AFA. Alternating finite automata [13, 28] (AFAs) have a finite *input alphabet* Σ, a finite *set of states* $Q = \{q_i\}_{i=0}^{k-1}$, a *start state* $\iota \in Q$, a set of *final states* $F \subseteq Q$, and a *transition function* $g : Q \to (\Sigma \times \{0,1\}^{(k)}) \to \{0,1\}$. Let $g_p = g(p)$ for $p \in Q$. A sequence $v \in \{0,1\}^{(k)}$ represents the *conjuction*

$$\mathbf{q}_v = \mathit{AND}\{q_i \in Q \mid v_i = 1\}$$

and for $a{\in}\Sigma$, $p{\in}Q$, $\lambda v.g_p(a, v)$ represents the *disjunction*

$$g_{p,a} = \mathit{OR}\{\mathbf{q}_v \mid g_p(a, v) = 1, v \in \{0,1\}^{(k)}\},$$

where $\mathit{OR}(\emptyset) = q_\bot$ is a new state and $\mathit{AND}(\emptyset) = {\sim}q_\bot$. The translation of $M_{\text{AFA}} = (Q, \Sigma, \iota, F, g)$ into an equivalent SBFA is as follows

$$\begin{aligned}\mathit{SBFA}(M_{\text{AFA}}) = (&\mathcal{A}, Q \cup \{q_\bot\}, \iota, F, q_\bot,\\ &\{q_\bot \mapsto q_\bot\} \textstyle\bigcup_{p \in Q}\{p \mapsto \mathit{OR}_{a\in\Sigma}\textbf{IF}(\hat{a}, g_{p,a}, q_\bot)\})\end{aligned}$$

Proposition C.1. $\mathbf{L}(\mathit{SBFA}(M_{\mathit{AFA}})) = L(M_{\mathit{AFA}})$ *with L as in [13].*

BFA. BFAs over Σ have a finite *set of states* Q an *initial function* $\iota \in \mathbb{B}(Q)$, a set of *final states* $F \subseteq Q$, and a *transition function* $\delta : Q \times \Sigma \to \mathbb{B}(Q)$.

We translate $M_{\text{BFA}} = (\Sigma, Q, \delta, \iota, F)$ into an equivalent SBFA as follows. The translation uses the fresh state $q_\bot \notin Q$.

$$\begin{aligned}\mathit{SBFA}(M_{\text{BFA}}) = (&\mathcal{A}, Q \cup \{q_\bot\}, \iota, F, q_\bot,\\ &\{q_\bot \mapsto q_\bot\} \textstyle\bigcup_{p \in Q}\{p \mapsto \mathit{OR}_{a\in\Sigma}\textbf{IF}(\hat{a}, \delta(p, a), q_\bot)\})\end{aligned}$$

where $\hat{a}$ is the predicate in $\mathcal{A}$ such that $[\![\hat{a}]\!] = \{a\}$.

Proposition C.2. $\mathbf{L}(\mathit{SBFA}(M_{\mathit{BFA}})) = L(M_{\mathit{BFA}})$ *with L as in [10].*

Remarks. Observe that the main difference between M_{AFA} and M_{BFA} besides the initial state formula is that g relies essentially on $\mathit{DNF}(\mathbb{B}^+(Q))$ while δ uses the full $\mathbb{B}(Q)$ for state predicates. In that respect, the BFA formulation is closer in spirit to SBFAs. Thus, because of DNF, the size of δ can be exponentially more succinct than g (*if* g is represented explicitly as its type suggests). Negation does not play a big role here because it can be eliminated at a linear cost. Therefore, AFAs and BFAs are to a large extent considered to be equivalent notions. As we know, this is not true in the symbolic case, when comparing SAFAs and SBFAs.

D Lift rules

The lifting rule $\mathit{lift}(\tau)$ propagates the intersection into the leaves and thus lifts the conditionals to the top level. Here we also pass the branch condition ψ that is initally ., that can be maintained to be satisfiable, so that dead branches are eliminated on-the-fly and the resulting transition regex is *clean* — in all conditional regexes all branches are satisfiable. Assume here that τ is in NNF. The NNF rules are specified below.

$$\begin{aligned}\mathit{lift}(\tau) &= \mathit{lift}_{.}(\tau)\\ \mathit{lift}_\psi(\tau) &= \bot \quad \text{if } \psi \equiv \bot\end{aligned}$$

In the remainder ψ is assumed satisfiable ($\psi \not\equiv \bot$).

$$\begin{aligned}\mathit{lift}_\psi(R) &= R \quad \text{if } R \in \mathit{ERE} \text{ and } \psi \equiv .\\ \mathit{lift}_\psi(R) &= \textbf{IF}(\psi, R, \bot) \quad \text{if } R \in \mathit{ERE} \text{ and } \psi \not\equiv .\\ \mathit{lift}_\psi(\textbf{IF}(\varphi, t, f)) &= \textbf{IF}(\varphi, \mathit{lift}_{\psi\wedge\varphi}(t), \mathit{lift}_{\psi\wedge\neg\varphi}(f))\\ \mathit{lift}_\psi(\textbf{IF}(\varphi, t, f)\ \&\ \rho) &= \mathit{lift}_\psi(\textbf{IF}(\varphi, t\ \&\ \rho, f\ \&\ \rho))\\ \mathit{lift}_\psi((\tau_1 \mid \tau_2)\ \&\ \rho) &= \mathit{lift}_\psi(\tau_1\ \&\ \rho) \mid \mathit{lift}_\psi(\tau_2\ \&\ \rho)\end{aligned}$$

NNF. The *negation normal form* of a transition regex τ, $\mathit{NNF}(\tau)$, is defined as follows. The correctness of these rules rests on Lemma 4.2.

$$\begin{aligned}\mathit{NNF}(\textbf{IF}(\varphi, \tau, \rho)) &= \textbf{IF}(\varphi, \mathit{NNF}(\tau), \mathit{NNF}(\rho))\\ \mathit{NNF}({\sim}\textbf{IF}(\varphi, \tau, \rho)) &= \textbf{IF}(\varphi, \mathit{NNF}({\sim}\tau), \mathit{NNF}({\sim}\rho))\\ \mathit{NNF}({\sim}{\sim}\tau) &= \mathit{NNF}(\tau)\\ \mathit{NNF}({\sim}R) &= {\sim}R \quad \text{if } R \in \mathit{ERE}\end{aligned}$$

The remaining cases are standard applications of DeMorgan's rules.