CertiStr: A Certified String Solver

Shuanglong Kan
Department of Computer Science
Technische Universität Kaiserslautern
Kaiserslautern, Germany
shuanglong@cs.uni-kl.de

Anthony Widjaja Lin
Department of Computer Science
Technische Universität Kaiserslautern & MPI-SWS
Kaiserslautern, Germany
lin@cs.uni-kl.de

Philipp Rümmer
Department of Information Technology
Uppsala University
Uppsala, Sweden
philipp.ruemmer@it.uu.se

Micha Schrader
Department of Computer Science
Technische Universität Kaiserslautern
Kaiserslautern, Germany
schrader@rhrk.uni-kl.de

Abstract
Theories over strings are among the most heavily researched logical theories in the SMT community in the past decade, owing to the error-prone nature of string manipulations, which often leads to security vulnerabilities (e.g. cross-site scripting and code injection). The majority of the existing decision procedures and solvers for these theories are themselves intricate; they are complicated algorithmically, and also have to deal with a very rich vocabulary of operations. This has led to a plethora of bugs in implementation, which have for instance been discovered through fuzzing.

In this paper, we present CertiStr, a certified implementation of a string constraint solver for the theory of strings with concatenation and regular constraints. CertiStr aims to solve string constraints using a forward-propagation algorithm based on symbolic representations of regular constraints as symbolic automata, which returns three results: sat, unsat, and unknown, and is guaranteed to terminate for the string constraints whose concatenation dependencies are acyclic. The implementation has been developed and proven correct in Isabelle/HOL, through which an effective solver in OCaml was generated. We demonstrate the effectiveness and efficiency of CertiStr against the standard Kaluza benchmark, in which 80.4% tests are in the string constraint fragment of CertiStr. Of these 80.4% tests, CertiStr can solve 83.5% (i.e. CertiStr returns sat or unsat) within 60s.

CCS Concepts: • Theory of computation → Automated reasoning: Formal languages and automata theory.

1 Introduction
Strings are among the most fundamental and commonly used data types in virtually all modern programming languages, especially with the rapidly growing popularity of dynamic languages, including JavaScript and Python. Programs written in such languages often implement security-critical infrastructure, for instance web applications, and they tend to process data and code in string representation by applying built-in string-manipulating functions; for instance, to split, concatenate, encode/decode, match, or replace parts of a string. Functions of this kind are complex to reason about and can easily lead to programming mistakes. In some cases, such mistakes can have serious consequences, e.g., in the case of web applications, cross-site scripting (XSS) attacks can be used by a malicious user to attack both the web server or the browsers of other users.

One promising research direction, which has been intensively pursued in the SMT community in the past ten years, is the development of SMT solvers for theories of strings (dubbed string solvers) including Kaluza [35], CVC4 [28], Z3 [17], Z3-str3 [4], Z3-Trau [9], S3P [39], OSTRICH [13], SLOTH [22], and Norm [1], to name only a few. Such solvers are highly optimized and find application, among others, in bounded model checkers and symbolic execution tools [14, 34], but also in tools tailored to verification of the security properties [3].

It has long been observed that constraint solvers are extremely complicated procedures, and that implementations
are prone to bugs. Defects that affect soundness or completeness are routinely found even in well-maintained state-of-the-art tools, both in real-world applications and through techniques like fuzzing [6, 8, 31]. String solvers are particularly troublesome in this context, since, unlike other SMT theories, the theory of strings requires a multitude of operations including concatenation, regular constraints, string replacement, length constraints, and many others. This inherent intricacy is also reflected in the recently developed SMT-LIB standard for strings (http://smtlib.cs.uiowa.edu/theories-UnicodeStrings.shtml). Many of the string solvers that arose in the past decade relied on inevitably intricate decision procedures, which resulted in subtle bugs in implementations themselves even among the mainstream string solvers (e.g., see [8, 31]). This implies the unfortunate fact that one cannot blindly trust the answer provided by string solvers.

**Contributions.** In this paper, we present CertiStr, the first certified implementation of a string solver for the standard theory of strings with concatenation and regular constraints (a.k.a. word equations with regular constraints [18, 20, 23, 30]). CertiStr aims to solve string constraints by means of the so-called forward-propagation algorithm, which relies on a simple idea of propagating regular constraints in a forward direction in order to derive contradiction, or prove the absence thereof (see Section 2 for an example). Similar ideas are already present in abstract interpretation of string-manipulating programs (e.g., see [32, 42, 43]), but not yet at the level of string solvers, i.e., which operate on string constraints. We have proven in Isabelle/HOL [33] some crucial properties of the forward-propagation algorithm and more specifically with respect to our implementation CertiStr: (i) [TERMINATION]: the algorithm terminates on string constraints without cyclic concatenation dependencies, (ii) [SOUNDNESS]: CertiStr is sound for unsatisfiable results (if forward-propagation detects inconsistencies in a constraint, then the constraint is indeed unsatisfiable), and (iii) [COMPLETEENESS]: CertiStr is complete for constraints satisfying the tree property (i.e., in which on the right-hand side of each equation every variable appears at most once). Here completeness amounts to the fact that, if forward-propagation does not detect inconsistencies for a constraint satisfying the tree property, then the constraint is indeed satisfiable. For the constraints that do not satisfy the tree property and CertiStr cannot decide them as unsatisfiable, CertiStr returns unknown.

In order to facilitate the propagation of regular constraints, we also implement a certified library for Symbolic Non-deterministic Finite Automata (s-NFA), by allowing transition labels to be a set of characters in the (potentially infinite) alphabet, represented by an element of a boolean algebra (e.g. the interval algebra or the BDD algebra), instead of a single character. s-NFAs are especially crucial for an efficient implementation of automata-based string solving algorithms and many string processing algorithms [1, 13, 16, 24, 41], for which reason our certified implementation of an s-NFA library is of independent interests.

Last but not least, we have automatically generated a verified implementation CertiStr in Isabelle/HOL, which we have extensively evaluated against the standard string solving benchmark from Kaluza [35] with around 38000 string constraints. For the first time, we demonstrate that the simple forward-propagation algorithm in fact performs surprisingly well even compared to other highly optimized solvers (which are not verified implementations). In particular, we show that the majority of these constraints (83.5%) are solved by CertiStr; for the rest, the tool either returns unknown, or times out. Moreover, CertiStr terminates within 60 seconds on 98% of the constraints, witnessing its efficiency.

To summarize, our contributions are:

- we developed in Isabelle/HOL the tool CertiStr, the first certified implementation of a string solver.
- we implemented the first certified symbolic automata library, which is crucial for an efficient implementation of many string processing applications [16].
- CertiStr was evaluated over Kaluza benchmark [35], with around 38000 tests. In this benchmark, 83.5% of the tests can be solved with the results sat or unsat. Moreover, the solver terminates within 60 seconds on 98% of the tests, witnessing the surprising competitiveness of the simple forward-propagation algorithm against more complicated algorithms.

## 2 Motivating Example

We start by illustrating the decision procedure implemented by CertiStr. CertiStr uses the so-called forward-propagation algorithm for solving satisfiability of string theory with concatenation and regular membership constraints. To illustrate the algorithm, consider the formula:

\[
\text{domain} \in \text{/[a-zA-Z.]/} + \\
\wedge \text{dir. file} \in \text{/[a-zA-Z0-9.]/} + \\
\wedge \text{path} = \text{dir} + \text{"/" + file} \\
\wedge \text{url} = \text{"http:/% + domain + " + path} \\
\text{[ \\
\wedge \text{url} \in \text{/.*<script>.*/}}
\]

(1) (2) (3) (4) (5)

The formula contains string variables (domain, dir, etc.), and uses both regular membership constraints and word equations with concatenation to model the construction of a URL from individual components. Regular expressions are written using the standard PCRE syntax (https://perldoc.perl.

CPP ’22, January 17–18, 2022, Philadelphia, PA, USA

Shuanglong Kan, Anthony Widjaja Lin, Philipp Rümmer, and Micha Schrader
org/perlre). Each equation should be read as an assignment of the right-hand side term to the left-hand side variable, and is processed in this direction. We initially ignore the constraint (5).

The forward-propagation algorithm propagates regular membership constraints and derives new constraints for the variable on the left-hand side of equations. In this example, the algorithm starts with equation (3) and propagates constraint (2), resulting in the new constraint

\[ path \in /[a-zA-Z0-9.]+\|/[a-zA-Z0-9.]+/ \] (6)

We can propagate this new constraint further, using equation (4) and together with constraint (1), deriving:

\[ url \in /\http://\|/[a-zA-Z-Z.]+\|/[a-zA-Z0-9.]+// \] (7)

At this point, no more propagations are possible. Note that forward-propagation will terminate, in general, whenever no cyclic dependencies exist between the equations, which is the fragment of formulas we consider.

Still ignoring (5), we can observe that forward-propagation has not discovered any inconsistent constraints. In order to conclude the satisfiability of the formula from this, we need a meta-result about forward-propagation: we show that forward propagation is complete for acyclic formulas in which each variable occurs at most once on the right-hand side of equations. Since this criterion holds for the formula (1) \( \land \cdots \land (4) \), we have indeed proven that it is satisfiable.

Consider now also (5), modelling a simple form of injection attack. We can observe that this additional constraint is inconsistent with the derived constraint (7), since the intersection of the asserted regular languages is empty. Forward-propagation detects this conflict as soon as (7) has been computed: each time a new membership constraint is derived, the algorithm will check the consistency with constraints already assumed. Since we also show that forward-propagation is sound w.r.t. inconsistency, the detection of a conflict immediately implies that the formula (1) \( \land \cdots \land (5) \) is unsatisfiable.

In summary, the forward-propagation is defined using two main inference steps: the post-image computation for word equations, and a consistency check for regular constraints. Although the example uses regular expressions, both operations can be defined more easily through a translation to finite-state automata. To obtain a verified string solver, the implementation of both operations has to be shown correct, and in addition meta-results need to be derived about the soundness and completeness of the overall algorithm.

We can also observe that classical finite-state automata, over concrete alphabets, are cumbersome even for toy examples; regular expressions often talk about character ranges that would require a large number of individual transitions. In practice, string constraints are usually formulated over the Unicode alphabet with currently \( 3 \times 2^{16} \) characters, which necessitates symbolic character representation. In our verified implementation, we therefore use a simple form of symbolic automata [15] in which transitions are labelled with character intervals.

### 3 String Constraint Fragment

In this section, we present the string constraint fragment that CertiStr supports and its semantics of satisfiability.

Let \( \Sigma \) be an alphabet and \( \Sigma^* \) denote the set of words over \( \Sigma \). Let \( w_1, w_2 \in \Sigma^* \) be two words, where \( w_2 \) denotes the concatenation of the two words, i.e., \( w_2 \) is appended to the end of \( w_1 \) to get a new string. We consider the following string constraint language over \( \Sigma \):

\[ c ::= x \in A | x = x_1 + x_2 | c_1 \land c_2, \]

where \( x \) denotes a variable ranging over \( \Sigma^* \) and \( A \) denotes an NFA representing a regular language over \( \Sigma \). The language accepted by \( A \) is denoted by \( L(A) \). The semantics of a constraint \( c \) is defined in an obvious way by interpreting \( x \in A \) as a membership of \( x \) in the language \( L(A) \), \( x = x_1 + x_2 \) as a string equation, and \( c_1 \land c_2 \) as a conjunction of \( c_1 \) and \( c_2 \). More precisely, given a function \( \mu \) mapping each variable in the constraint \( c \) to a string in \( \Sigma^* \), we say that \( c \) is satisfied if:

(i) \( \mu(x) \in L(A) \) for each constraint \( x \in A \) in \( c \), and
(ii) \( \mu(x) = \mu(x_1) \mu(x_2) \) for each constraint \( x = x_1 + x_2 \).

The constraint \( c \) is satisfiable if one such solution \( \mu \) for \( c \) exists.

Our string constraint language is as general as word equations with regular constraints [18, 20, 23, 30], which forms the basis of the recently published Unicode string constraint language in SMT-LIB 2.6, and already makes up the bulk of existing string constraint benchmarks. Notice that our restriction to string equation of the form \( x = x_1 + x_2 \) is not a real restriction since general string constraints can be obtained by means of desugaring. For instance, the concatenation of more than two variables, like \( x = x_1 + x_2 + x_3 \) can be desugared as two concatenation constraints: \( (1) x' = x_1 + x_2; x = x' + x_3 \). Moreover, a subset of string constraints with length functions (called monadic length) can also be translated into this fragment with regular constraints. For instance, \( |x| \leq 6 \), where \( |x| \) denotes the length of the word \( x \), can be translated to a regular membership constraint as \( x \in A \), where \( A \) is an NFA that accepts any word whose length is at most 6. Recent experimental evidence shows that a substantial portion of length constraints that appear in practice are essentially monadic [21]. There are also some string constraints using disjunction (\( \lor \)). For instance, \( c_1 \lor c_2 \), where \( c_1 \) and \( c_2 \) are string constraints in our fragment. This also can be solved with CertiStr by first solving \( c_1 \) and \( c_2 \) separately, and then checking whether one of them is satisfiable.

Figure 1 shows the framework of CertiStr. The input of the solver is an SMT file with string constraints. The tool has two parts: (1) the front-end (non-certified) and (2) the back-end (certified). The front-end parses the SMT file, which
We use the following example of string constraints to illustrate our forward-propagation analysis. Forward-propagation for string constraints are not sufficient to decide their satisfiability. The back-end contains two parts: (1) the Automata Library, which contains a collection of automata operations, such as the product and concatenation of two NFAs, and (2) a forward-propagation to check whether the string constraints are satisfiable. The results of the certified string solver can be (1) sat the string constraint is satisfiable, (2) unsat the string constraint is unsatisfiable, and (3) unknown the solver cannot decide whether the string constraint is satisfiable or not. Note that, unknown does not mean that CertiStr is non-terminating. It means the results after executing the forward-propagation for string constraints are not sufficient to decide their satisfiability.

4 Forward-propagation of String Constraints

In this section, we present the algorithm of the forward-propagation and the properties proven for it.

4.1 The Algorithm of Forward-propagation

We use the following example of string constraints to illustrate our forward-propagation analysis.

Example 4.1. \( x_1 \in \mathcal{A}_1 \land x_2 \in \mathcal{A}_2 \land x_3 \in \mathcal{A}_3 \land x_5 = x_3 + x_4 \land x_3 = x_1 + x_2 \).

Figure 2. Dependence graph of variables in Example 4.1

Example 4.1 contains five variables “\( x_1, x_2, x_3, x_4, x_5 \)” and two concatenation constraints “\( x_3 = x_3 + x_4 \land x_3 = x_1 + x_2 \).” We can view these concatenation constraints as a dependence graph shown in Figure 2. For a concatenation constraint \( x_k = x_i + x_j \), it yields two dependence relations: \( x_k \rightarrow x_i \) and \( x_k \rightarrow x_j \) (\( \rightarrow \) denotes \( a \) depends on \( b \)). This means that before propagating the concatenation of the regular constraints of \( x_i \) and \( x_j \), we need to first compute the propagation to \( x_i \) and \( x_j \). Let the initial regular constraint of \( x_i \) be \( \mathcal{A}_k \). Let the regular constraints of \( x_i \) and \( x_j \) be \( \mathcal{A}_i \) and \( \mathcal{A}_j \), respectively. The new regular constraint of \( x_k \) after the propagation can be refined to an automaton, whose language is \( L(\mathcal{A}_k) \cap (L(\mathcal{A}_i) + L(\mathcal{A}_j)) \), where \( L_1 + L_2 \doteq \{ w_1w_2 \mid w_1 \in L_1 \land w_2 \in L_2 \} \). Here, the term “refine” means narrowing the regular constraint of a variable by propagating the concatenation constraints.

The forward-propagation repeats the following computation until all variables are refined:

- detect a set of variables such that for each variable in the set, all its dependence variables have already been refined.
- refine the regular constraints of the variables detected in the last step with respect to their concatenation constraints.

We illustrate the idea of the forward-propagation with Example 4.1. The first iteration of the forward-propagation will detect the set of variables \( \{x_3, x_4, x_5 \} \) since they have no dependence variables and we do not need to refine their regular constraints. The second iteration of the forward-propagation will detect the set of variables: \( \{x_3 \} \), as \( x_3 \) only depends on the variables in \( \{x_1, x_2, x_4 \} \). Its regular constraint will be refined to an automaton, whose language is \( L(\mathcal{A}_3) \cap (L(\mathcal{A}_1) + L(\mathcal{A}_2)) \) after the propagation for the constraint \( x_3 = x_1 + x_2 \). Assume this new regular constraint as \( \mathcal{A}_3' \). The third iteration detects the variable \( x_5 \), whose regular constraint will be refined to an automaton whose language is \( \Sigma^* \cap (L(\mathcal{A}_3') + \Sigma^*) \). Note that we do not specify the initial regular constraints of \( x_4 \) and \( x_5 \) in the string constraint and therefore their initial regular constraints are \( \Sigma^* \) by default.

Algorithm 1 shows the forward-propagation algorithm. It contains three procedures: \texttt{Forward Prop, Ready Set}, and \texttt{Var Lang}. The procedure \texttt{Forward Prop} is the entry point.
of the algorithm and has three parameters: (1) $S$ is the set of variables used in the string constraints, for instance, in Example 4.1, there are five variables: $S = \{x_1, x_2, x_3, x_4, x_5\}$; (2) $\text{Concat} : S \rightarrow 2^{(S \times S)}$ is a partial map which maps a variable $x$ to a set of pairs of variables. For each pair $(x_1, x_2)$ in the set, there is a concatenation constraint: $x = x_1 + x_2$; (3) $\text{Reg} : S \rightarrow N$ is a map from variables to automata. These automata denote the variables’ initial regular constraints. Here $N$ denotes the set of NFAs.

Algorithm 1 The algorithm of forward-propagation

```algorithm
1: procedure Forward Prop($S, \text{Concat}, \text{Reg}$)
2:     $R \leftarrow \emptyset$
3:     while $S \neq \emptyset$ do
4:         $C \leftarrow \text{Ready Set} (S, \text{Concat}, R)$
5:         $\text{Reg} \leftarrow \text{Var Lang} (C, \text{Concat}, \text{Reg})$
6:         $S \leftarrow S - C$; $R \leftarrow R \cup C$
7:     end while
8:     return $\text{Reg}$
9: end procedure

10: procedure Var Lang($C, \text{Concat}, \text{Reg}$)
11:     for each $v \in C$ do
12:         $\mathcal{A} \leftarrow (\text{Reg } v)$
13:         for each $(v_1, v_2) \in \text{Concat } v$ do
14:             $\mathcal{A}' \leftarrow \text{NFA_concat} (\text{Reg } v_1) (\text{Reg } v_2)$
15:             $\mathcal{A} \leftarrow \text{NFA_product } \mathcal{A} \mathcal{A}'$
16:         end for
17:     $\text{Reg} \leftarrow \text{Reg } [v \mapsto \mathcal{A}]$
18:     end for
19:     return $\text{Reg}$
20: end procedure

21: procedure Ready Set($S, \text{Concat}, R$)
22:     $C \leftarrow \emptyset$
23:     for each $v \in S$ do
24:         $D \leftarrow \emptyset$
25:         for each $(v_1, v_2) \in \text{Concat } v$ do
26:             $D \leftarrow D \cup \{v_1, v_2\}$
27:         end for
28:         if $D \subseteq R$ then
29:             $C \leftarrow C \cup \{v\}$
30:         end if
31:     end for
32:     return $C$
33: end procedure
```

In the while loop of the procedure $\text{Forward Prop}$, the computation iteratively refines the regular constraints of variables by propagating the regular constraints of variables on the right hand side of the concatenation constraints. The first step in the loop computes the set $C$ of variables whose dependence variables are all in the set $R$ (computed by the procedure $\text{Ready Set}$). $R$ is initialized empty and it denotes the set of variables that have already been refined by the previous loops. The second step in the loop updates the regular constraints of the variables in $C$ stored in $\text{Reg}$ by propagating the concatenation constraints in $\text{Concat}$ (computed by $\text{Var Lang}$).

At the end of the loop body, the set $S$ is updated by removing the variables in $R$ as they have already been refined and these removed variables are added to the set $R$, i.e., they have already been refined.

The procedure $\text{Ready Set}$ computes the set of variables that are only dependent on the variables in $R$. For each variable $v$, it stores all its dependence variables in $D$ and then checks whether $D$ is a subset of $R$. If $D$ is a subset of $R$ then $v$ is moved to $C$ as all its dependence variables have been refined.

The procedure $\text{Var Lang}$ refines the regular constraints of variables in $C$. In the inner loop of $\text{Var Lang}$, it traverses all concatenation constraints $v = v_1 + v_2$ and refines the regular constraint of $v$ by propagating the concatenation of the regular constraints of $v_1$ and $v_2$. $\text{Reg } [v \mapsto \mathcal{A}]$ means updating the regular constraint of $v$ in $\text{Reg}$ to $\mathcal{A}$. The functions $\text{NFA_concat}$ and $\text{NFA_product}$ denote constructing the concatenation and product of two NFAs, respectively. These two functions will be introduced in Section 5.

Assume a string constraint is stored in $S$, $\text{Concat}$ and $\text{Reg}$, then after calling "$\text{Forward Prop}(S, \text{Concat}, \text{Reg})$", we will get a new $\text{Reg}$ such that for each variable $v$, ($\text{Reg'} v$) stores the refined regular constraint of $v$, which is an NFA. We have two cases to decide whether the string constraint is satisfiable or not:

- If there exists a variable $v$ such that the language of ($\text{Reg'} v$) is empty then the string constraint is unsatisfiable, because the refined regular constraint of $v$ is the empty language.
- If for all variables $v$ such that the language of ($\text{Reg'} v$) is not empty then unfortunately we cannot decide whether the string constraint is satisfiable or not. Consider the string constraint: 

$$
(1) \ y = x + x \wedge (2) \ y \in \{ab\} \wedge (3) \ x \in \{a, b\},
$$

where $ab$, $a$, and $b$ are words, the refined regular constraints of $x$ and $y$ are both not empty. But this string constraint is not satisfiable as constraints (1) and (2) require $x$ to have two different values. We will introduce a property, called tree property later, such that if the string constraint satisfies the tree property then it is satisfiable, otherwise CertiStr returns unknown.

4.2 Theorems for Forward-propagation

Now we introduce the key theorem proven for the algorithm. Firstly we present the definitions of well-formed inputs ($\text{wf}$ in Figure 3) and the acyclic property ($\text{acyclic}$ in Figure 3) for concatenation constraints, which are the premises of the correctness theorem.
definition wf where
wf S Concat Reg =
(dom Concat ≤ S) ∧ (dom Reg = S) ∧
(∀ v v1 v2 v ∈ (dom Concat) ∧ (v1, v2) ∈ (Concat v)
  → v1 ∈ S ∧ v2 ∈ S) ∧ finite S

definition acrylic where
acrylic S Concat l = S = (∪(set l)) ∧ ((l = []))
∨((l = s#l′) → (s ∩ (∪(set l′)) = {})) ∧
(∀ v1 v2 v ∈ s ∧ (v1, v2) ∈ (Concat v)
  → {v1, v2} ⊆ (∪(set l′)) ∧ (acyclic (S − s) Concat l′))

Figure 3. The definitions of wf and acrylic

The wf predicate requires that all variables appearing in
the domain and codomain of Concat should be in the set
S, and the domain of Reg should be the set S, i.e., all variables
should have their initial regular constraints. S should
be finite. Now we consider the definition of acrylic, which
checks whether the dependence graphs of concatenation
constraints are acyclic. Forward Prop cannot solve string
constraints that are not acrylic. For instance, the string con-
straint "x = y1 + y2 ∧ y1 = z1 + x" does not satisfy the acrylic
property, as x =⇒ y1 =⇒ x.

In order to specify the acrylic property, we arrange the
set S of variables to a list in which elements are disjoint
sets of variables in S and the list characterizes the depend-
ence levels among variables. For instance, in Example 4.1,
there are 5 variables and they can be organized in a list:
[{v5}, {v3}, {v1, v2, v4}]. Indeed, the list characterizes the depen-
dence levels among variables. For instance, the variable
v5 only depends on the variables in {v3} and {v1, v2, v4} and
v3 only depends on the variables in {v1, v2, v4} but not {v5}.
The variables in {v1, v2, v4} do not depend on any variables. If
Concat does not satisfy the acrylic property then we cannot
arrange all variables in such a list, as all variables in a cycle
always depend on another variable in the cycle.

The predicate acrylic contains two parts. (1) S = (∪(set l)),
which means the elements in S and l should be exactly the same.
"set l" translates the list l to a set of elements in l and
"∪(set l) = {x | ∃ y ∈ set l ∧ x = y}". (2) Part 2 further has
two cases. The first case is "l = []", i.e., empty list. In this case,
the predicate is obviously true. The second case is "l = s#l'",
i.e., it has at least one element. s is the first element and l' is
the tail. In this case, firstly, it requires the elements in s to be
disjoint with the elements in the sets of the tail l' (checked by
s ∩ (∪(set l')) = ∅). Secondly, let v be a variable in s, for any
two variables v1 and v2 that v depends on, v1 and v2 can only
be in the sets of the tail l'. Thirdly, acrylic (S − s) Concat l'
checks whether the remaining elements in S − s are still acry-
clic w.r.t. Concat and l'. In order to check whether the
variables in S are acrylic w.r.t. Concat, we can specify it as
"∃! acrylic S Concat l'.

Now we are ready to introduce our correctness theorem (Theorem 4.2). The keyword fixes in Isabelle is used to define
universally quantified variables. The keyword assumes is
used to specify assumptions of the theorem and the keyword
shows is used to specify the conclusion of the theo-
rem. The shows part is of the form "M ≤ SPEC (λ Reg'. Φ)",
which means the algorithm M is correct with respect to the
specification Φ. More precisely, the result of applying the pre-
dicate (λ Reg'. Φ) to the output of M should be true.
"NFA accept w A" checks whether w is accepted by A,
which will be introduced in Section 5. The symbol "@" is
the list concatenation operation in Isabelle. We use a list of
elements in Σ to represent a word.

Theorem 4.2 (Correctness of Forward Prop).

fixes S Concat Reg
assumes 1. wf S Concat Reg
  2. (∃! acrylic S Concat l)
shows Forward Prop(S, Concat, Reg) ≤ SPEC (λ Reg'.
  (V v. NFA accept w (Reg' v) ) "=
  (NFA accept w (Reg v) ∧
   (∀ v1 v2 v. (v1, v2) ∈ (Concat v) →
     (∃ w1 w2. w = w1@w2 ∧
       NFA accept w1 (Reg' v1) ∧
       NFA accept w2 (Reg' v2)))))

Theorem 4.2 shows that if S, Concat, Reg are well-formed
and acrylic then after calling Forward Prop(S, Concat, Reg)
we can get Reg', which satisfies: any w is accepted by the
automaton (Reg' v) if and only if
1. w is accepted by the original regular constraint of v,
i.e., (Reg v), and
2. for all variables v1 and v2, (v1, v2) ∈ Concat v implies
   there exist w1, w2 such that w = w1@w2 and w1, for
   i = 1, 2, is accepted by the regular constraint (Reg' vi).

That is, in the new Reg', the concatenation constraints are
used to refine the corresponding variables.

Moreover, in Isabelle, Theorem 4.2 also requires us to
prove the termination of Algorithm 1. The termination is
proven by showing that in the procedure Forward Prop, the
number of variables in S decreases progressively for each
iteration. Since S is finite, the algorithm must terminate.

Even though Theorem 4.2 is the key of the correctness of
the forward-propagation, it does not give us an intuition with
respect to the satisfiability semantics of string constraints
introduced in Section 3. In the following, we will show that
our forward-propagation is sound w.r.t. unsat results.

Firstly we need to formalize the semantics of sat and unsat
for string constraints. Let μ be an assignment, which is a map
S → Σ* from variables in S to words. We define the predicate
sat_str in Figure 4 for the semantics of the satisfiability of
string constraints.
**definition sat_str where**

\[
\text{sat}_{\text{str}}\ S \text{ Concat } \text{Reg } \mu = \\
(\forall \sigma \in S. (\mu \sigma) \in \mathcal{L}(\text{Reg } \nu)) \wedge \\
(\forall \sigma_1 \sigma_2. \sigma \in S \land (\sigma_1, \sigma_2) \in (\text{Concat } \sigma) \rightarrow \\
(\mu \sigma) = (\mu \sigma_1)(\sigma_2))
\]

**definition tree where**

\[
\text{tree } \text{Concat} = \\
(\forall \sigma_1 \sigma_2. (\sigma_1, \sigma_2) \in \text{Concat } \sigma \rightarrow \sigma_1 \neq \sigma_2) \wedge \\
(\forall \sigma_1 \sigma_2 \sigma_3 \sigma_4. (\sigma_1, \sigma_2) \neq (\sigma_3, \sigma_4) \wedge \\
(\sigma_1, \sigma_2) \in \text{Concat } \nu \land (\sigma_3, \sigma_4) \in \text{Concat } \nu \rightarrow \\
\{\sigma_1, \sigma_2\} \cap \{\sigma_3, \sigma_4\} = \emptyset) \wedge \\
(\forall \sigma_1 \sigma_2 \sigma_3 \sigma_4. (\sigma_1, \sigma_2) \neq (\sigma_3, \sigma_4) \land \\
(\sigma_1, \sigma_2) \in \text{Concat } \nu \land (\sigma_3, \sigma_4) \in \text{Concat } \nu \rightarrow \\
\{\sigma_1, \sigma_2\} \cap \{\sigma_3, \sigma_4\} = \emptyset)
\]

**Figure 4.** The definitions of sat_str and tree

The predicate sat_str contains two parts: (1) the assignment \(\mu\) should respect all regular constraints and (2) the assignment \(\mu\) should respect all concatenation constraints.

Theorem 4.3 shows that after executing Forward_PROP, we get a new map \(\text{Reg}'\). If there exists a variable \(\sigma\), such that \(\mathcal{L}(\text{Reg}' \sigma) = \emptyset\), i.e., the language of the refined regular constraint of \(\sigma\) is the empty set, then there is no assignment \(\mu\) that makes the predicate “sat_str S Concat Reg \(\mu\)’ true.

**Theorem 4.3** (Soundness for unsat results).

**fixes** \(S \text{ Concat Reg}\)

**assumes** 1. \(\text{wf } S \text{ Concat Reg}\)
2. \((\exists L. \text{acyclic } S \text{ Concat } l)\)

**shows** Forward_PROP(S, Concat, Reg) \(\leq\) SPEC(\(\lambda\text{Reg}'\)).

\[\left((\forall \sigma. \mathcal{L}(\text{Reg}' \sigma) \neq \emptyset) \rightarrow \\
(\exists \mu. \neg \text{sat_str } S \text{ Concat Reg } \mu)\right)\]

Unfortunately, the theorem holds only for unsat results. As we already analyzed before, the case \(\forall \sigma. \mathcal{L}(\text{Reg}' \sigma) \neq \emptyset\) cannot make us conclude that the string constraint is satisfiable.

In this paper, we characterize a subset of the string constraints, using a property such that the forward-propagation’s satisfiable results are also sound. We call it the tree property.

The intuition behind the tree property is that the dependence graph of a string constraint should constitute a tree or forest.

For example, \(y = x + x\) does not satisfy the tree property, because \(y\) has two edges pointing to a same node \(x\) in the dependence graph. In other words, the node \(x\) has indegree 2. But a tree requires that each node has at most indegree 1. As the map Concat stores the dependence graphs of string constraints, we only need to check Concat for the tree property.

The predicate tree is defined in Figure 4.

This definition has three cases separated by the conjunction operator: (1) the first case requires that for any concatenation constraint \(\sigma = \sigma_1 + \sigma_2\), we have \(\sigma_1 \neq \sigma_2\). (2) The second case requires that for any variable \(\sigma\) with two concatenations \(\sigma = \sigma_1 + \sigma_2\) and \(\sigma = \sigma_3 + \sigma_4\), \(\{\sigma_1, \sigma_2\}\) and \(\{\sigma_3, \sigma_4\}\) are disjoint. (3) The third case requires that for any two different variables \(\sigma\) and \(\sigma'\) with the concatenation constraints \(\sigma = \sigma_1 + \sigma_2\) and \(\sigma' = \sigma_3 + \sigma_4\), \(\{\sigma_1, \sigma_2\}\) and \(\{\sigma_3, \sigma_4\}\) are disjoint.

With this definition we have the following theorem.

**Theorem 4.4** (Completeness for the tree property).

**fixes** \(S \text{ Concat Reg}\)

**assumes** 1. \(\text{wf } S \text{ Concat Reg}\)
2. \((\exists L. \text{acyclic } S \text{ Concat } l)\)
3. \(\text{tree Concat}\)

**shows** Forward_PROP(S, Concat, Reg) \(\leq\) SPEC(\(\lambda\text{Reg}'\)).

\[\left((\forall \sigma. \mathcal{L}(\text{Reg}' \sigma) \neq \emptyset) \rightarrow \\
(\exists \mu. \text{sat_str } S \text{ Concat Reg } \mu)\right)\]

This theorem shows that if Concat satisfies the tree property then we have: \(\text{Reg}'\) satisfies that for all \(\sigma\), the language of \(\text{Reg}' \sigma\) is not empty iff there exists an assignment, which makes the string constraint satisfiable.

In order to check whether a string constraint satisfies the tree property, we propose Algorithm 2. It first stores all the variables, which are on the right hand side of the concatenation constraints, into a list, and then checks whether the list is distinct. If it is not distinct then there must exist a variable with at least indegree 2 and thus Concat cannot satisfy the tree property.

**Algorithm 2** Tree property checking

1: procedure Check_Tree(S, Concat)
2: \(l \leftarrow \[]\)
3: for each \(\sigma \in S\) do
4: \(\quad\text{for each } (\sigma_1, \sigma_2) \in (\text{Concat } \sigma) \text{ do}\)
5: \(\quad\quad l \leftarrow \sigma_1 \# \sigma_2 \# l\)
6: end for
7: end for
8: return distinct \(l\)
9: end procedure

Now we show that if Algorithm 2 returns \(\text{True}\) then the tree property is satisfied.

**Theorem 4.5** (Correctness of Check_Tree).

**fixes** \(S \text{ Concat}\)

**assumes** 1. \(\text{dom Concat } \subseteq S\)
2. \((\forall \sigma \sigma_1 \sigma_2. (\sigma_1, \sigma_2) \in \text{Concat } \sigma \rightarrow \{\sigma_1, \sigma_2\} \subseteq S)\)

**shows** Check_Tree(S, Concat) \(\rightarrow\) tree Concat

With the function Check_Tree, CertiStr can return three results: unsat, sat, and unknown, in which unknown means the string constraint does not satisfy the tree property and the forward-propagation does not refine the regular constraint of any variable to empty.

We now finish introducing the forward-propagation procedure. As it uses many NFA operations, in the following
section, we will introduce our implementation for these NFA operations.

5 Symbolic Automata Formalization

Automata operations are the key to our certified string solver. For instance, we need to check the language emptiness of an NFA as well as construct the concatenation and product of two NFAs. In order to make the string solver practically useful, we should also keep the efficiency in mind.

As argued by D’Antoni et. al [15], the classical definition of NFA, whose transition labels are a character in the alphabet, is not suitable for practical implementation, and they propose s-NFAs, which allow transition labels to carry a set of characters, to address this limitation. In this section, we present our formalization of s-NFAs in Isabelle.

Definition 5.1 (Abstract s-NFAs). An s-NFA is a 5-tuple: 
\( \mathcal{A} = (Q, \Sigma, \Delta, I, F) \), where \( Q \) is a finite set of states, \( \Sigma \) is an alphabet (\( \Sigma \) may be infinite), \( \Delta \subseteq Q \times 2^\Sigma \times Q \) is a finite set of transition relations, \( I \subseteq Q \) is a set of initial states, and \( F \subseteq Q \) is a set of accepting states.

For a transition \( (q, \alpha, q') \) in an s-NFA, the implementation of \( \alpha \) can use various different data structures, such as intervals for string solvers and Binary Decision Diagrams (BDD) for symbolic model checkers. Definition 5.1 of s-NFAs is called abstract s-NFAs because the transition labels are denoted as sets instead of using concrete data structures, like intervals and BDDs.

In order to make our formalization of s-NFAs reusable for different implementation of transition labels, we exploit Isabelle’s refinement framework and divide the formalization of s-NFAs in two levels: (1) the abstract level (introduced in this section) and (2) the implementation level (introduced in the following section). When the correctness of the algorithms at the abstract level is proven, the implementation level does not need to re-prove it, only the refinement relations between these two levels must be shown.

At the abstract level, transition labels are modeled as a set as shown in Definition 5.1. At the implementation level, the sets of characters are refined to concrete data structures, such as intervals or BDDs.

Our formalization of s-NFAs extends an existing classical NFA formalization developed by Tuerk et al. [40]; this formalization is part of the verified CAVA model checker [26]. We extend it to support s-NFAs and further operations, such as the concatenation operation of two NFAs. Figure 5 shows the Isabelle formalization of s-NFAs. An s-NFA is defined by a record with four elements: (1) \( Q \) is the set of states, (2) \( I \) is the set of initial states, (3) \( F \) is the set of accepting states, and (4) \( \Delta \) is the set of transitions with transition labels defined by sets. The definition well-formed is a predicate to check whether an NFA satisfies (1) the states occurring in the sets of initial and accepting states are also in the set \( (Q, \mathcal{A}) \) of states. In Isabelle, the value of a field \( fd \) in a record \( rd \) can be extracted by \( (fd \ rd) \). Therefore, \( (Q, \mathcal{A}) \) denotes the value of the field \( Q \) in \( \mathcal{A} \). (2) The states in the set of transitions should also be in \( (Q, \mathcal{A}) \). (3) \( (Q, \mathcal{A}) \) is a finite set. (4) \( \Delta \) is finite.

For a word \( w = a_1 a_2 \ldots a_n \) and two states \( q \) and \( q' \), the function reachable checks whether there exists a path \( q_0, \alpha_1, q_1, \ldots, q_{n-1}, a_n, q_n \) in the NFA \( \mathcal{A} \) such that \( q = q_0 \land q_1' = q_n \) \( (q_{i-1}, \alpha_i, q_i) \in (\Delta, I) \), \( 1 \leq i \leq n \) and \( \alpha_i \in \alpha_i \), \( 1 \leq j \leq n \). The definition NFA_accept checks whether a word is accepted by an NFA and the definition of \( L \) is the language of an NFA.

Based on this definition of NFAs, a collection of automata operations are defined and their correctness lemmas are proven. Here we only present the operation of the concatenation shown in Figure 6.

NFA_concat Basic constructs the concatenation of two NFAs. The correctness of this definition relies on the fact that the sets of states in the two NFAs are disjoint. In the definition of NFA_concat Basic, the set of transitions of the concatenation is the union of (1) \( (\Delta, I) \), (2) \( (\Delta, F) \), and (3) \( \{ (q, \alpha, q') | 3 q', (q, \alpha, q') \in (\Delta, I) \land q' \in (F, A_1) \times q'' \in (I, A_2) \} \). The transitions in (3) concatenate the language of \( A_1 \) to the language of \( A_2 \).

The set of the initial states in the concatenation is computed by first checking whether there is a state that is in both the sets of initial and accepting states of \( A_1 \). If there exists such a state then the initial states in the concatenation should include the initial states of \( A_2 \), as \( A_1 \) accepts the
definition NFA_concat_basic where
NFA_concat_basic \( A_1, A_2 \equiv \{ \)
\( Q = (Q \ A_1) \cup (Q \ A_2), \)
\( \Delta = (\Delta \ A_1) \cup (\Delta \ A_2) \cup \)
\( \{(q, a, q') \mid \exists q''(q, a, q') \in (\Delta \ A_1) \land q' \in (F \ A_1) \land q'' \in (I \ A_2)\}, \)
\( I = \text{if } ((I \ A_1) \cap (F \ A_1) = \emptyset) \text{ then } (I \ A_1) \)
\( \text{else } ((I \ A_1) \cup (I \ A_2)), \)
\( F = (F \ A_2) \}\)

definition NFA_concat where
NFA_concat \( A_1, A_2, f_1, f_2 = \)
remove_unreachable_states
(NFA_concat_basic
(NFA_rename \( f_1, A_1 \))(NFA_rename \( f_2, A_2 \)))

Figure 6. The definition of concatenation operation

empty word, otherwise the initial states of the concatenation are \( A_1 \)'s initial states.

The definition of NFA_concat firstly renames the states in both NFAs to ensure their sets of states disjoint. In the term "NFA_rename \( f, A \)". \( f \) is a renaming function, for instance, \( f \) can be \( \lambda q.(q, 1) \), which renames any state \( q \) to \( (q, 1) \). "NFA_rename \( f, A \)" renames all states in \( (Q \ A) \) by the function \( f \). Correspondingly, all states in transitions, initial states, and accepting states are also renamed by \( f \). Renaming states makes it easier to ensure the sets of states of two NFAs are disjoint. The function remove_unreachable_states removes all states in an NFA that are not reachable from the initial states of the NFA. This enables us to check the language emptiness of an NFA by only checking the emptiness of its set of accepting states. At the implementation level, the algorithm also only generates reachable states for NFAs. But with remove_unreachable_states used in NFA_concat at the abstract level, the refinement relation between these two levels can be specified as the NFA isomorphism of the two output concatenations of automata at the two levels, instead of language equivalence, which is more challenge to prove.

In order to ensure the correctness of this operation, we need to prove the following lemma in Isabelle.

Lemma 5.2 (Correctness of NFA_concat).

\textbf{fixes} \( A_1, A_2, f_1, f_2 \)
\textbf{assumes} 1. well-formed \( A_1 \)
2. well-formed \( A_2 \)
3. \( (\text{image } f_1 (Q \ A_1)) \cap (\text{image } f_2 (Q \ A_2)) = \emptyset \)
4. \( \text{inj}_\text{on} \ f_1 (Q \ A_1) \land \text{inj}_\text{on} \ f_2 (Q \ A_2) \)
\textbf{shows} \( L(\text{NFA_concat } A_1, A_2, f_1, f_2) = \{w_1 \otimes w_2, w_1 \in L(A_1) \land w_2 \in L(A_2)\} \)

The assumptions well-formed \( A_1 \) and well-formed \( A_2 \) require the input NFAs to be well-formed. In addition, the assumption \( (\text{image } f_1 (Q \ A_1)) \cap (\text{image } f_2 (Q \ A_2)) = \emptyset \) requires that after renaming the states in both \( A_1 \) and \( A_2 \), the new sets of states of the two NFAs are disjoint. The term "image \( f \)" applies the function \( f \) to the elements in the set \( S \) and returns a new set \( \{f \ e, e \in S\} \). The premises \( \text{inj}_\text{on} \ f_1 (Q \ A_1) \) and \( \text{inj}_\text{on} \ f_2 (Q \ A_2) \) require that the functions \( f_1 \) and \( f_2 \) are injective over the sets \( (Q \ A_1) \) and \( (Q \ A_2) \), respectively. The conclusion in "shows" specifies that the language of the concatenation of \( A_1 \) and \( A_2 \) equals the set of the concatenations of the words in \( L(A_1) \) and the words in \( L(A_2) \).

Moreover, some other important functions are defined, for instance:

- The product of two NFAs: \( \text{NFA_product } A_1, A_2 \), which computes the product of \( A_1 \) and \( A_2 \). We proved the theorem: \( L(\text{NFA_product } A_1, A_2) = L(A_1) \cap L(A_2) \).
- Checking the isomorphism of two NFAs:
  \( \text{NFA_isomorphism } A_1, A_2 \equiv (3f. \text{inj}_\text{on} \ f (Q \ A_1) \land \text{NFA_rename } f, A_1 = A_2) \). Some lemmas are proven for it, such as, well-formed \( A_1 \land \text{well-formed } A_2 \land \text{NFA_isomorphism } A_1, A_2 \rightarrow L(A_1) = L(A_2) \).

Section 4 and this section present the abstract level algorithms for the forward-propagation and the NFA operations, respectively. In order to make CertiStr efficient and practically useful, we need to use some efficiently implemented data structures, such as red-black-trees (RBT) and hash maps. This will be introduced in the following section.

6 Implementation-Level Algorithms

In this section, we will first present the relations between abstract concepts, like sets and maps, and the implementation data structures in Isabelle (Subsection 6.1), which is the prerequisite to understand the implementation-level algorithm in Subsection 6.2.

6.1 Isabelle Collections Framework

Isabelle Collections Framework (ICF) [27] provides an efficient, extensible, and machine checked collections framework. The framework features the use of data refinement techniques [25] to refine an abstract specification (using high-level concepts like sets) to a more concrete implementation (using collection data structures, like RBT and hashmaps). The code-generator of Isabelle can be used to generate efficient code.

The concrete data structures implement a collection of interfaces that mimic the operations of abstract concepts. We list the following RBT implementation of set interfaces that will be used to present the implementation-level algorithms:

- \( \text{RBT.empty} \), generates an RBT representation of the empty set.
- \( \text{RBT.mem } e \), checks whether the element \( e \) is in the set represented by RBT \( D \).
In order to support NFA operation implementation, some operations for intervals are implemented. We list the following 3 interval operations here.

- Computing the intersection of two intervals: \( \text{intersection} \left[ I_1, I_2 \right] \left[ I_1', I_2' \right] \triangleq \left[ \text{max}(l_1, l_1'), \text{min}(l_2, l_2') \right] \). The term \( \text{max}(l_1, l_1') \) returns the bigger one of \( l_1 \) and \( l_1' \), and \( \text{min}(l_2, l_2') \) returns the smaller one of \( l_2 \) and \( l_2' \).
- Checking the non-emptiness of the interval \( \left[ I_1, I_2 \right] \): \( \text{mem} \left( I_1, I_2 \right) \triangleq I_1 \land I_2 \).
- Checking the membership of an element \( e \) and the interval \( \left[ I_1, I_2 \right] \): \( \text{mem} \left( e, I_1, I_2 \right) \triangleq e \subseteq I_1 \land e \subseteq I_2 \).

Algorithm 3 The algorithm idea of NFA concatenation

1: \( \text{procedure Construct\_Trans}(S_{A_1}, S_{A_2}, S_{T_2}, S_{F_1}) \)
2: \( S_D \leftarrow RBT\_union S_{A_1}, S_{A_2}; \)
3: \( RBT \_\text{iterator } S_D \)
4: \( (\lambda(q, a, q') S) \)
5: \( \text{if } (RBT\_\text{mem } q' \ S_{F_1}) \text{ then} \)
6: \( RBT\_\text{iterator } S_T \)
7: \( (\lambda q' \ S'). RBT\_\text{insert } (q, a, q') S' \)
8: \( \text{end if}; \)
9: \( \text{return } S_D; \)
10: \( \text{end procedure} \)

11: \( \text{procedure NFA\_Concat\_Impl}(\Delta A_1, \Delta A_2, f_1, f_2) \)
12: \( \text{\hspace{1em} } \Delta A_1, \Delta A_2 \text{ are of type NFA\_rbt} \)
13: \( \Delta A_1' \leftarrow \text{NFA\_Rename\_Impl } f_1 \Delta A_1; \)
14: \( \Delta A_2' \leftarrow \text{NFA\_Rename\_Impl } f_2 \Delta A_2; \)
15: \( \text{if } RBT\_\text{inter } (\Delta A_1') (\Delta A_1') \text{ is empty then} \)
16: \( S_I \leftarrow RBT\_\text{union } (\Delta A_1') (\Delta A_2'); \)
17: \( \text{else} \)
18: \( S_I \leftarrow (\Delta A_1'); \)
19: \( \text{end if}; \)
20: \( \text{S_Q } \leftarrow S_I; S_\Delta \leftarrow RBT\_empty; \text{wl } \leftarrow RBT\_\text{to\_list } S_Q; \)
21: \( S_\Delta' \leftarrow \text{Construct\_Trans}(\Delta A_1', \Delta A_2, f_1, f_2); \)
22: \( \text{while } \text{wl } \neq \text{[]} \text{ do} \)
23: \( \text{q_s } \leftarrow \text{wl} \cdot \text{rm}_\text{first}; \)
24: \( \text{\hspace{1em} } \text{rm}_\text{first}: \text{ removes and returns the first element in } \text{wl} \)
25: \( \text{RBT\_\text{iterator } S_I', S_D} \)
26: \( \text{\hspace{1em} } (\lambda(q, a, q') S) \)
27: \( \text{\hspace{2em} if } q = q_s \land (\text{mem} \ S_Q) \text{ then} \)
28: \( \text{\hspace{3em} if } \neg (\text{RBT\_\text{mem } q' \ S_Q}) \text{ then} \)
29: \( \text{\hspace{4em} RBT\_\text{insert } q' \ S_Q; } \)
30: \( \text{\hspace{4em} wl } \leftarrow \text{wl}\_@\{q'; \}; \)
31: \( \text{\hspace{3em} end if}; \)
32: \( \text{\hspace{3em} RBT\_\text{insert } (q, a, q') S; } \)
33: \( \text{\hspace{3em} end if}; \)
34: \( \text{\hspace{1em} end while} \)
35: \( \text{\hspace{1em} return } (S_Q, S_\Delta, S_I, RBT\_\text{inter } S_Q (\Delta A_2')) \)
36: \( \text{end procedure} \)
Algorithm 3 shows the basic idea of the NFA_concat implementation (the procedure NFA_Concate_Impl). The texts after “⇒” are comments. NFA_Rename_Impl is the implementation of the NFA renaming function NFA_Rename at the abstract level.

The sub-procedure Construct_Tran generates the set of transitions for the concatenation of the two automata, which contains the transitions in both (Δ A1) and (Δ A2), and the transitions that concatenate the two automata (cf. Figure 6 for the abstract algorithm of the concatenation).

In Algorithm 3, Line 16-20 computes the initial states of the concatenation. Line 21 initializes S0, SΔ, and wI, where S0 stores the states of the concatenation NFA, SΔ stores the transitions of the concatenation NFA, wI is a list to mimic a queue that stores the states to be expanded. Line 23 to 37 is the while loop for expanding the reachable states and transitions. Finally Line 38 returns the concatenation of the two NFAs. This algorithm constructs the concatenation of two NFAs with only reachable states, which means that we can check the emptiness of the concatenation NFA by only checking the emptiness of its set of accepting states.

In order to ensure the correctness of the implementation for concatenation, we need to prove the refinement relation between NFA_concat and NFA_Concate_Impl.

In Figure 8, we formalize the refinement relations between the two NFAs at the abstract level and the implementation level. The function NFA_α translates an implementation-level NFA A to an abstract-level NFA of the type NFA. The definition refine_rel defines the refinement relation between an abstract-level NFA A and an implementation NFA A′, which requires the isomorphism between A and (NFA_α A′)

Lemma 6.1 specifies that the concatenation implementation correctly refines the abstract concatenation definition.
Note that the benchmark classification is not in all cases accurate, in the sense that there are some files in unsat/big and unsat/small that are satisfiable, owing to known incorrect results that were produced by the original Kaluza string solver [28, 35].

We ran CertiStr over these 38043 tests on an AMD Opteron 2220 SE machine, running 64-bit Linux and Java 1.8 (for running the Scala front-end of CertiStr). The results are shown in Table 1. The column “total tests” denotes the total number of tests in a group. The columns “sat”, “unknown”, “unsat” denote the numbers of tests for which CertiStr returns sat, unknown, and unsat, respectively. The column “solved%” denotes the percentage of the tests for which the string solver returns sat or unsat. The columns “avg. time(s)” and “timeout” denote the average time for running each test in CertiStr and the number of tests that time out, respectively. The column “CVC4” denotes the average execution time of the state of the art string solver CVC4, which can efficiently solve all the string constraints without timeout. The time limit for solving each test is 60 seconds.

From Table 1, we can conclude that in total, 83.5% of the tests can be solved by CertiStr with the result sat or unsat. The remaining of the tests are unknown or timeout. The groups sat_small and unsat_small have the higher solved percentages, more than 90%, compared with sat_big and unsat_big. This is easy to understand as small tests have a higher probability to satisfy the tree property. The worst group is unsat_big. CertiStr can solve 45% tests in this group. The reason is that in this group, there are a lot of tests with more than 100 concatenation constraints, which significantly increases their probability of violating the tree property. For the groups unsat_big and unsat_small, CertiStr detects 2502 and 7376 tests, respectively, that are indeed satisfiable.

Moreover, in unsat_big, there are 728 tests that time out. After analyzing these tests, we found that a variable with a lot of concatenation constraints (i.e., the variable appears many times on the left-hand sides of the concatenation constraints) can easily yield s-NFA state and transition explosion during the forward-propagation. In order to evaluate such a state explosion problem, we set a test of the form in Example 7.1:

\[ x = x_1 + x_1 \land x = x_2 + x_2 \land x = x_3 + x_3 \land \ldots \]

It is a test case with only the variable \( x \) on the left-hand side. On the right-hand side, there are concatenations of the form \( x_i + x_j \). No regular constraints are in the test.

We ran CertiStr over the test case by increasingly adding more concatenations for the variable \( x \). Table 2 shows the results for solving the test case. Each column contains (1) the number of concatenation constraints (No. Concat), (2) the size of the automaton of the variable \( x \) after executing the forward-propagation (The numbers of states and transitions of the automaton), and (3) the execution time for running the forward-propagation over the test (time).

From the table we can conclude that after adding 16 concatenation constraints for the variable \( x \), the automaton generated for \( x \) contains 65536 states and 43046721 transitions. Solving this test takes 168.28s.

Now consider the 728 tests that time out. These tests have similar constraints, in which some variables have a lot of concatenation constraints. During the forward-propagation, state and transition explosion happens for these tests, therefore they cannot be solved in 60s.

Compared with CVC4, which is not automata-based and incorporates more optimizations over its string theory decision procedure, CertiStr is less efficient. We can also optimize CertiStr further in the future. For instance, we can optimize the language intersection operation \( \Sigma^+ \cap L(\mathcal{A}) \) to \( L(\mathcal{A}) \), and the language concatenation \( \Sigma^+ \Sigma^+ \). We can also implement NFA minimization algorithms to further improve its efficiency. These optimizations can avoid the state explosion problem to some extent. However, integrating such

### Table 1. Experimental results of the Kaluza benchmark

<table>
<thead>
<tr>
<th>group</th>
<th>total tests</th>
<th>sat</th>
<th>unknown</th>
<th>unsat</th>
<th>solved%</th>
<th>avg. time(s)</th>
<th>timeout</th>
<th>CVC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>sat_small</td>
<td>19634</td>
<td>19302</td>
<td>332</td>
<td>0</td>
<td>98.3%</td>
<td>&lt; 0.01</td>
<td>0</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>sat_big</td>
<td>774</td>
<td>521</td>
<td>253</td>
<td>0</td>
<td>67.3%</td>
<td>0.84</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>unsat_small</td>
<td>8775</td>
<td>7376</td>
<td>824</td>
<td>575</td>
<td>91%</td>
<td>0.58</td>
<td>0</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>unsat_big</td>
<td>8860</td>
<td>2502</td>
<td>4880</td>
<td>1478</td>
<td>45%</td>
<td>8.01</td>
<td>728</td>
<td>0.5</td>
</tr>
<tr>
<td>total</td>
<td>38043</td>
<td>29700</td>
<td>6289</td>
<td>2054</td>
<td>83.5%</td>
<td>1.87</td>
<td>728</td>
<td>0.11</td>
</tr>
</tbody>
</table>

### Table 2. Experimental results for state and transition explosion

<table>
<thead>
<tr>
<th>No. Concat</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>states</td>
<td>2048</td>
<td>4096</td>
<td>8192</td>
<td>16384</td>
<td>32768</td>
<td>65536</td>
</tr>
<tr>
<td>transitions</td>
<td>177147</td>
<td>531441</td>
<td>1594323</td>
<td>4782969</td>
<td>14348907</td>
<td>43046721</td>
</tr>
<tr>
<td>time (s)</td>
<td>0.43</td>
<td>1.29</td>
<td>4.10</td>
<td>12.37</td>
<td>50.59</td>
<td>168.28</td>
</tr>
</tbody>
</table>

Example 7.1.
optimizations in a verified solver is challenging, as more cases need to be proven correct.

7.2 The Efforts of Developing CertiStr

In this subsection, we discuss the efforts of developing CertiStr. Table 3 shows the efforts. The rows “Abs Automata Lib” and “Imp Automata Lib” denote abstract level and implementation level automata libraries, respectively. The rows “Abs Forward Prop” and “Imp Forward Prop” denote abstract level and implementation level forward-propagations, respectively. The column “Loc” denotes the lines of Isabelle code for each module. The column “Terms” denotes the number of definitions, functions, locales, and classes. The column “Theorems” denotes the number of theorems and lemmas in a module. The development needs around one person-year efforts.

Table 3. Development effort of certified string solver

<table>
<thead>
<tr>
<th>Module</th>
<th>Loc</th>
<th>Terms</th>
<th>Theorems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abs Automata Lib</td>
<td>4484</td>
<td>100</td>
<td>254</td>
</tr>
<tr>
<td>Imp Automata Lib</td>
<td>8498</td>
<td>270</td>
<td>203</td>
</tr>
<tr>
<td>Abs Forward Prop</td>
<td>6850</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>Imp Forward Prop</td>
<td>2175</td>
<td>41</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>22007</td>
<td>448</td>
<td>514</td>
</tr>
</tbody>
</table>

We now discuss the challenges in developing CertiStr. The challenge in the automata library is the usage of sets as transition labels. Classical NFAs require only the equality comparison for transition labels. But for s-NFAs, we have more operations for labels. For instance, we require the operations seml, intersectionl, nemptyl, meml for intervals. These operations make it harder to prove the correctness of the automata library.

The challenge in the forward-propagation module is the proofs of Theorem 4.2, 4.3, and 4.4. These theorems need around 4000 lines of code to prove.

8 Related Works

String solvers. As introduced in Section 1, there already exist various non-certified string solvers, such as Kaluza [35], CVC4 [28], Z3 [17], Z3-str3 [4, 5], Z3-Trau [9], S3P [39], OSTRICH [11–13], SLOTH [22], and Norn [1], among many others. These solvers are intricate and support more string operations than CertiStr. Some of the solvers, such as Z3-str3 and CVC4, opted to support more string operations and settle with incomplete solvers (e.g., with no guarantee of termination) that could still solve many constraints that arise in practice. Other solvers are designed with stronger theoretical guarantees; for instance, OSTRICH is complete for the straight-line string constraint fragment [29]. CertiStr can solve the string constraints that are out of the scope of the straight-line fragment and guarantees to terminate for the constraints without cyclic concatenation dependencies, but it will return unknown for constraints that do not satisfy the tree property.

Symbolic Automata. CertiStr depends on automata operations for regular expression propagation and consistency checking. An efficient implementation of automata is crucial. In comparison to finite-state automata in the classical sense, symbolic automata [15, 16] have proven to be more appropriate for applications like model checking, natural language processing, and networking. Certified automata libraries [7, 26] can provide trustworthiness for automata-based analysis in applications. However, to the best of our knowledge, existing certified automata libraries are based on the classical definition of automata, which makes them inefficient for practical applications with very large alphabets. Our work in this paper contributes to the development of certified symbolic automata libraries.

Paradigms toward certified constraint solvers. Two main paradigms have emerged for the verification of constraint solvers: (i) the verification of the solver itself, i.e., the development of solvers with machine-checked end-to-end correctness guarantees. For instance, Shi et al. [36] built a certified SMT quantifier-free bit-vector solver. (ii) The generation of certificates, i.e., the actual solver is not verified, but its outputs are accompanied by proof witnesses that can then be independently checked by a verified, trustworthy certificate checker. For instance, Ekici et al. [19] provide an independent and certified checker for SAT and SMT proof witnesses. In our work, we follow the first paradigm.

Security applications based on string analysis. There are various security applications of automata-based analysis and string solvers, such as detecting web security vulnerabilities. Yu et al. [43] investigated the approach to use automata theory for detecting security vulnerabilities, such as XSS, in PHP programs. Their work also depends on a forward analysis. But the forward analysis is over the control flow graphs of PHP programs. CertiStr aims to build a stand-alone string solver and the forward-propagation is only over the string constraints. Many developments of string solvers, string analysis techniques, and automata-theoretic techniques were also directly motivated by security vulnerability detection, e.g., [2, 3, 10, 13, 22, 29, 35, 37–39]. To the best of our knowledge, CertiStr is the first tool that provides certification of a string analysis method that is applicable to security vulnerability detection. We believe there is a need for further development of certified string analyzers: string solving techniques are intricate and hence error-prone, but are applied for security analysis whose correctness is of critical importance.

9 Conclusion and Future Works

In this paper, we present CertiStr, a certified string solver for the theory of concatenation and regular constraints. The backend of CertiStr is verified in Isabelle proof assistant, which provides a rigorous guarantee for the results of the
solver. We ran CertiStr over the benchmark Kaluza to show the efficacy of CertiStr.

As future works, firstly, we plan to support more string operations, such as string replacement and capture groups, which are also widely used in programming languages, such as JavaScript and PHP, increasing the applicability of CertiStr. Secondly, the front end of CertiStr still needs to be verified in Isabelle, especially the correctness of the desugaring from string constraints with monadic length functions and disjunctions to the language of CertiStr.

Acknowledgments
The authors would like to thank anonymous reviewers for their valuable comments. This research was supported in part by the ERC Starting Grant 759969 (AV-SMP), Max-Planck Fellowship, Amazon Research Award, the Swedish Research Council (VR) under grant 2018-04727, and by the Swedish Foundation for Strategic Research (SSF) under the project WebSec (Ref. RT17-0011).

References