FAST: a Transducer-Based Language for Tree Manipulation

Loris D’Antoni
University of Pennsylvania
lorisdan@cis.upenn.edu

Margus Veanes  Benjamin Livshits
David Molnar
Microsoft Research
{margus,livshits,dmolnar}@microsoft.com

Abstract

Tree automata and tree transducers are used in a wide range of applications in software engineering, from XML processing to language type-checking. While these formalisms are of immense practical use, they can only model finite alphabets, and since many real-world applications operate over infinite domains such as integers, this is often a limitation. To overcome this problem we augment tree automata and transducers with symbolic alphabets represented as parametric theories. Admitting infinite alphabets makes these models more general and succinct than their classical counterparts. Despite this, we show how the main operations, such as composition and language equivalence, remain computable given a decision procedure for the alphabet theory.

We introduce a high-level language called FAST that acts as a front-end for the above formalisms. FAST supports symbolic alphabets through tight integration with state-of-the-art satisfiability modulo theory (SMT) solvers. We demonstrate our techniques on practical case studies, covering a wide range of applications.

Categories and Subject Descriptors F.1.1 [Theory of Computation]: Models of Computation, Automata

Keywords Symbolic Tree Transducers, FAST

1. Introduction

This paper introduces FAST, a new language for analyzing and modeling programs that manipulate trees over potentially infinite domains. FAST builds on top of satisfiability modulo theory solvers, tree automata, and tree transducers. Tree automata are used in variety of applications in software engineering, from analysis of XML programs [27] to language type-checking [37]. Tree transducers extend tree automata to model functions over trees, and appear in fields such as natural language processing [31, 33, 34] and XML transformations [32]. While these formalisms are of immense practical use, they suffer from a major drawback: in the most common forms they can only handle finite alphabets.

In order to overcome this limitation, symbolic tree automata (STAs) and symbolic tree transducers (STTs) extend these classical objects by allowing transitions to be labeled with formulas in a specified theory. While the concept is straightforward, traditional algorithms for deciding composition, equivalence, and other properties of finite automata and transducers do not immediately generalize. A notable example appears in [9] where it is shown that while in the classical case allowing finite automata transitions to read subsequent inputs does not add expressiveness, in the symbolic case this extension makes most problems, such as checking equivalence, undecidable. Symbolic tree automata still enjoy the closure and decidability properties of classical tree automata [39] under the assumption that the alphabet theory forms a Boolean algebra (i.e. closed under Boolean operations) and it is decidable. In particular STAs can be minimized and are closed under complement, and intersection, and it is therefore decidable to check whether two STAs are equivalent.

Taking a step further, tree transducers model transformations from trees to trees. A symbolic tree transducer (STT) traverses the input tree in a top-down fashion, processes one node at a time, and produces an output tree. This simple model can capture several scenarios, however in most useful cases it is not closed under sequential composition [22]. In the case of finite alphabets this problem is solved by augmenting the transducer’s rules with regular lookahead [15], that is the capability of checking whether the subtrees of each processed node belong to some regular tree languages. We extend STTs in a similar way, and introduce symbolic tree transducers with regular lookahead (STTRs).

The main theoretical result of this paper is a new composition algorithm for STTRs together with a proof of its correctness. Similarly to the classical case, we show that two STTRs A and B can be composed into a single STTR A◦B if either A is single-valued (for every input produces at most one output), or B is linear (traverses each node in the tree at most once). Remarkably, the algorithm works modulo any decidable alphabet theory that is an effective Boolean algebra.

We introduce the language FAST as a frontend for STAs and STTRs. FAST (Functional Abstraction of Symbolic Transducers) is a functional language that integrates symbolic automata and transducers with Z3 [12], a state-of-the-art solver able to support complex theories that range from data-types to non-linear real arithmetic. We use FAST to model several real world scenarios and analysis problems: we demonstrate applications to HTML sanitization, interference checking of augmented reality applications submitted to an app store, deforestation in functional language compilation, and analysis of functional programs over trees. We also sketch how FAST can capture simple CSS analysis tasks. All such problems require the use of symbolic alphabets. Figure 1 summarizes our applications and the analyses enabling each one. In Section 7 we further contrast FAST with previous DSLs for tree manipulation.
Equivalence

Section 6 discusses why classical tree transducers do not scale causing the output not to be standards compliant.

Contributions summary:
1. a theory of symbolic tree transducers with regular lookahead (STTR), that non-trivially extends the classical theory of tree transducers (§3);
2. a new algorithm for composing STTRs together with a proof of correctness (§4);
3. FAST, a domain-specific language for tree manipulations founded on the theory of STTRs (§3); and
4. five concrete applications of FAST showing how composition of STTR can be beneficial in practical settings (§5).

2. Motivating Example

We use a simple scenario to illustrate the main features of the language FAST, and the analysis enabled by the use of symbolic transducers. We choose to model a basic HTML sanitizer. An HTML sanitizer is a program that traverses an input HTML document and removes or modifies nodes, attributes and values that can cause malicious code to be executed on a server. Every HTML sanitizer works in a different way, but the general structure is usually the following:
1) the input HTML is parsed into a DOM (Document Object Model) tree,
2) the DOM is modified by a sequence of sanitization functions f₁, ⋯, fₙ, and
3) the modified DOM tree is transformed back into an HTML document.

In the following we use FAST to describe some of the functions used during step 2. Each function fᵢ takes as input a DOM tree received from the browser’s parser and transforms it into an updated DOM tree. As an example, the FAST program sani (Figure 2, line 30) traverses the input DOM and outputs a copy of it in which all subtrees in which the root is labeled with the string "script" have been removed, and all the characters "" and "" have been escaped with a "\"".

We informally describe each component of Figure 2. Line 2 defines the data-type HtmlE of our trees. Each node of type HtmlE contains a tag of type string and is built using one of the constructors nil, val, attr, or node.
Each constructor has a number of children associated with it (2 for attr) and all such children are HtmlE nodes. We use the type HtmlE to model DOM trees. Since DOM trees are unranked (each node can have an arbitrary number of children), we will first encode them as ranked trees. We adopt a slight variation of the classical binary encoding of unranked trees (Figure 3). We first informally describe the encoding and then show how it can be formalized in FAST.

Each HTML node n is encoded as an HtmlE element node(x₁, x₂, x₃) with three children x₁, x₂, x₃ where: 1) x₁ encodes the list of attributes of n, 2) x₂ encodes the first child of n in the DOM, 3) x₃ encodes the next sibling of n, and 4) tag contains the node type of n (div, etc.). Each HTML attribute a with value s is encoded as an HtmlE element attr(x₁, x₂) with two children x₁, x₂ where: 1) x₁ encodes the value s (nil if s is the empty string), 2) x₂ encodes the list of attributes following a (nil if a is the last attribute), and 3) tag contains the name of a (id, etc.). Each non-empty string w = s₁⋯sₙ is encoded as an HtmlE element val(x₁) where tag contains the string "s₁", and x₁ encodes the suffix s₂⋯sₙ. Each element nil has tag "", and can be seen as a termination operator for lists, strings, and trees. This encoding can be expressed in FAST (lines 4-13). For example, node Tree (lines 4-7) is the language of correct HTML encodings (nodes): 1) the tree node(x₁, x₂, x₃) is in the language nodeTree if x₁ is in the language attrTree, x₂ is in the language nodeTree, and x₃ is in the language nodeTree; 2) the tree nil is in nodeTree if its tag contains the empty string. The other language definitions are similar.

We now describe the sanitization functions. The transformation remScript (lines 15-19) takes an input tree t of type HtmlE and produces an output tree of type HtmlE: 1) if t = node(x₁, x₂, x₃) and its tag is different from "script",
remScript outputs a copy of $t$ in which $x_2$ and $x_3$ are replaced by the results of invoking remScript on $x_2$ and $x_3$ respectively; 2) if $t = node(x_1, x_2, x_3)$ and its tag is equal to "script", remScript outputs a copy of $x_3$, 3) if $t = nil$, remScript outputs a copy of $t$. The transformation esc (lines 20-27) of type $HtmlE$→$HtmlE$ escapes the characters $'\!'$ and $'\*'$, and it outputs a copy of the input tree in which each node val with tag $"\!\!"$ or $"\*\*\!"$ is pre-pended a node val with tag "$\!\"$.

The transformations remScript and esc are then composed into a single transformation rem_esc (line 29). One might notice that rem_esc also accepts input trees that are not in the language nodeTree, and therefore do not correspond to correct encodings. Therefore, we compute the transformation sani (line 31), which is same as rem_esc, but restricted to only accept inputs in the language nodeTree.

We can now use FAST to analyze the program sani. First, we define the language bad_output (lines 33-36), which accepts all the trees containing at least one node labeled with "script". Next, using transducers composition, we compute the language bad_inputs (line 38) of inputs that produce a bad output. Finally, if bad_inputs is the empty language, sani never produces bad outputs. When running this program in FAST this checking (line 40) fails, and FAST provides the following counterexample:

```
node "[script]" nil nil (node "[script]" nil nil nil)
```

where we omit the attribute for the nil nodes. This is due to a bug in line 18, where the rule does not recursively invoke the transformation remScript on $x_3$. When fixing this bug the assertion becomes valid. In this example we showed how in FAST simple sanitization functions can be first coded independently, and then composed without worrying about efficiency. Finally, the resulting transformation can be analyzed using transducer based techniques.

3. Symbolic Tree Transducers and FAST

The concrete syntax of FAST is shown in Figure 4. FAST is designed for describing trees, tree languages and functions from trees to trees. These are supported using symbolic tree automata (STAs), and symbolic tree transducers with regular lookahead (SSTRs). This section describes these objects and how they describe the semantics of FAST.

3.1 Background

All definitions are parametric with respect to a given background theory, called a label theory, over a fixed background structure with a recursively enumerable universe of elements. Such a theory is allowed to support arbitrary operations (such as addition, etc.), however all the results in the following only require it to be 1) closed under Boolean operations and equality, and 2) decidable (quantifier free formulas with free variables can be checked for satisfiability).

We use $\lambda$-expressions for defining anonymous functions called $\lambda$-terms without having to name them explicitly. In general, we use standard first-order logic and follow the notational conventions that are consistent with [40]. We write $\sigma$ for a type and the universe of elements of type $\sigma$ is denoted by $\sigma$. A $\sigma$-predicate is a $\lambda$-term $Ax.\varphi(x)$ where $x$ has type $\sigma$, and $\varphi$ is a formula whose free variables $FV(\varphi)$ are contained in $\{x\}$. Given a $\sigma$-predicate $\varphi$, $[\varphi]$ denotes the set of all $a \in \sigma$ such that $\varphi(a)$ holds. The set of $\sigma$-predicates is denoted by $[\sigma]$.

This definition illustrates the nondeterministic semantics of FAST: a tree $t$ belongs to bad_output if at least one of the three rules applies.

---

**Example 1.** The FAST program in Figure 2, declares $HtmlE = T_{\Sigma}^{String}$ over $\Sigma = \{nil, val, attr, node\}$, where $\Sigma(nil) = 0, \Sigma(val) = 1, \Sigma(attr) = 2$, and $\Sigma(node) = 3$. For example attrib("a")[$\Sigma(nil)[\sigma], \Sigma(val)\sigma$] is in $T_{\Sigma}^{String}$.

We write $\bar{e}$ for a tuple (sequence) of length $k \geq 0$ and denote the $i$-th element of $\bar{e}$ by $e_i$ for $1 \leq i \leq k$. We also write $(e_1, \ldots, e_n)$ for $\bar{e}$. The empty tuple is () and $(e_1, \ldots, e_n)$ for $\bar{e}$. We use the following operations over $n$-tuples of sets. If $X$ and $Y$ are $k$-tuples of sets then $X \uplus Y \equiv (X_i \uplus Y_i)_{i=1}^k$. If $X$ is a $k$-tuple of sets, $j \in \{1, \ldots, k\}$ and $Y$ is a set then $(X \uplus_j Y)$ is the $k$-tuple $(i=j$ then $X_i \uplus Y$ else $X_i)_{i=1}^k$.

3.2 Alternating Symbolic Tree Automata

We introduce and develop the basic theory of alternating symbolic tree automata, which adds a form of alternation to the basic definition originally presented in [39].

**Definition 1.** An Alternating Symbolic Tree Automaton (Alternating STA) $A$ is a tuple $(Q, T_{\Sigma}^A, \delta)$, where $Q$ is a finite set of states, $T_{\Sigma}^A$ is a tree type, and $\delta \subseteq \bigcup_{\sigma \geq 0} (Q \times \Sigma(k) \times \Psi(\sigma) \times (2^\sigma)^k)$ is a finite set of rules $(q, f, \varphi, \ell)$, where $q$ is the state, $f$ the symbol, $\varphi$ the guard, and $\ell$ the lookahead.

For $q \in Q$, $\delta(q) \equiv \{ r \in \ell \text{ state of } r \text{ is } q \}$. In FAST $\delta(q)$ is

```
lang q : \tau \{ c(\}$) where \varphi(\bar{x}) given \ell(\$) \ldots \}
```

**Example 2.** Consider the following FAST program.

```
type BT[\{ Int(L(0),N(2)) \}]
lang p: BT \{ L(\) where (i > 0) N(c,y, ) given (p x) (p y) \}
lang q: BT \{ L(\) where (odd i) N(c,y, ) given (o x) (o y) \}
lang l: BT \{ N(c,y, ) given (p x) (p y) \}
```

When $\ell(f) = 0$ then $f$ has type $\sigma \rightarrow T_{\Sigma}^A$. 

---
An equivalent STA $A$ over $T_{ST}^\text{int}$ has states $\{o, p, q\}$ and rules

\[
\begin{align*}
\{ (p, o, x > 0), \} & (p, n, x. \text{true}, \{ p \}, \{ p \}), \\
\{ o, L, x. \text{odd}(x) \}, \} & (o, n, x. \text{true}, \{ o \}, \{ o \}), \\
\{ q, n, x. \text{true}, (\theta, \{ p, o \}) \}.
\end{align*}
\]

Since the first subtree in the definition of $q$ is unconstrained, the corresponding component in the last rule is empty. The definition for $q$ has no case for $L$, so there is no rule.

Next, we define the semantics of an STA $A = (Q, T_{ST}^\text{int}, \delta)$.

**Definition 2.** For every state $q \in Q$ the language of $A$ at $q$, is the set

\[
L_A^q \overset{\text{def}}{=} \{f[a](\bar{t}) \in T_{ST}^\text{int} \mid (q, f, \varphi, \bar{t}) \in \delta, a \in \varphi\},
\]

Each subtree lookahead $\bar{t}_i$ above is treated as a conjunction of conditions. If $\bar{t}_i$ is empty then there are no restrictions on the $i$th subtree $t_i$. We extend the definition to all $q \subseteq Q$:

\[
L_A^q = (\bigcap_{p \subseteq Q} L_A^p), \text{ if } q \neq \emptyset; \quad T_{ST}^\text{int}, \text{ otherwise.}
\]

In the following we say STA for alternating STA.\(^5\)

**Definition 3.** $A$ is normalized if for all $(p, f, \varphi, \bar{t}) \in \delta$, and all $i, 1 \leq i \leq |t|$, $\bar{t}_i$ is a singleton set.

For example, the STA in Example 2 is not normalized because of the rule with source $q$. Normalization is a practically useful operation of STAs that is used on several occasions.

**Normalization.** Let $A = (Q, T_{ST}^\text{int}, \delta)$ be an STA. We compute merged rules $(q, f, \varphi, \bar{p})$ over merged states $q \in 2^Q$ where $\bar{p} \in (2^Q)^{|t|}$ for $f \in \Sigma$ by $\delta^f = \bigcup_{p \subseteq Q} \delta^f(p)$ where:

\[
\begin{align*}
\delta^f(\emptyset) &= \{ (\emptyset, f, \emptyset, (\emptyset)^{|t|} ) \}, \\
\delta^f(p \cup q) &= \{ (r \ast s) \mid r \in \delta^f(p), s \in \delta^f(q) \}, \\
\delta^f(p) &= \{ (\langle p, f, (\varphi), \bar{p} \rangle) \mid (p, f, \varphi, \bar{p}) \in \delta \}
\end{align*}
\]

where merge $\ast$ of rules is defined as follows:

\[
(p, f, \varphi, \bar{p}) \ast (q, f, \varphi, \bar{q}) \overset{\text{def}}{=} (p \cup q, f, \varphi \cup \Psi, \bar{p} \cup \bar{q})
\]

We can then define $\text{Normalize}(A)$ as the STA

\[
(2^Q, T_{ST}^\text{int}, \{ (p, f, \bigwedge \varphi, (\{ q \})^{|t|} ) \mid f \in \Sigma, (p, f, \varphi, q) \in \delta^f \})
\]

where the original rules are precisely the ones whose states are singleton sets in $2^Q$.\(^6\) Checking whether $L_A^q \neq \emptyset$ can be done by first normalizing $A$, then removing unsatisfiable guards using the decision procedure of the theory $\Psi(\sigma)$, and finally using emptiness of classical tree automata.

**Proposition 1.** The non-emptiness problem of STAs is decidable if the label theory is decidable.

While normalization is always possible, an STA may be exponentially more succinct than the equivalent normalized

---

\(^5\)When compared to the model in [8], the STAs defined above are “almost” alternating, in the sense that they can only allow disjunctions of conjunctions, rather than arbitrary positive Boolean combinations. Concretely, the lookahead of a rule $r$ corresponds to a conjunction of states, while several rules from the same source state provide a disjunction of cases.

\(^6\)In practice, merged rules are computed lazily starting from the initial state. Merged rules with unsatisfiable guards $\varphi$ are eliminated eagerly. New concrete states are created for all the reachable merged states. Finally, the normalized STA is cleaned by eliminating states that accept no trees, e.g., by using elimination of useless symbols from a context-free grammar [26, p. 88–89].
Let $S$ be an STTR $(Q, q^0, T_S, \Delta)$. The following construction is used to extract an STA from $S$ that accepts all the valid input trees accepted by $S$. Let $t$ be a $k$-rank tree transformer. For $1 \leq t \leq k$ let $S(i, t)$ be the set of all states $x$ such that $q(x)$ occurs in $t$.

**Definition 6.** The domain automaton of $S$, $d(S)$, is the STA $(Q, T_S^d, \{(q, f, \varphi, (t_i \cup S(i, t))_{i=1}^k) \mid q \in T_S^d \land t \in \Delta\})$.

The rules of the domain automaton also take into account the states that occur in the outputs in addition to the lookahead states. For example, the rule in Figure 5 yields the domain automaton rule $(q, g, \lambda x. x < 4, \{(p), x, \{p\})$).

In the following let $T$ be the STTR and let $L_T \overset{\text{def}}{=} L(\tau)$.

**Definition 7.** For all $q \in Q_T$, the transduction of $T$ at $q$ is the function $T^q$ such that, for all $t = f[a](t) \in \mathcal{F}_T$,

\[ T^q(t) \overset{\text{def}}{=} \psi_T(q(t)) \]

where $\psi_T(q(t)) \overset{\text{def}}{=} \bigcup\{\psi_T u(a, t) \mid (q, f, \varphi, \tilde{u}, \tilde{v}) \in [\Delta_T, a \in [\varphi], \bigwedge_{i=1}^k t_i \in L^q_T}\}$

The transduction of $T$ is $T^q = T^q_0$. The definitions are lifted to sets using union. We write $T_T(t, u)$ for $u \in T_T(t)$.

We omit $T$ from $T_T^q$ and $\psi_T$ when $T$ is clear from the context.

**Fast, in a transformation $T^q$ is defined by the statement**

\[ \text{trans } q : \vdash \to \tau \{ f[\tilde{y}] \text{ where } \varphi(x) \text{ given } \ell(\tilde{y}) \to (x, y) \ldots \} \]

where $\ell(\tilde{y})$ denotes the lookahead $(\{ r \mid \tilde{r} y_i \in \ell(\tilde{y})\})_{i=1}^k$.

**Example 3.** Recall the transformation $\text{remScript}$ in Figure 2. These are the corresponding rules. We use $q$ for the state of $\text{remScript}$, and $t$ for a state that outputs the identity transformation. The “safe” case is $\bar{q}(\text{node}[x](y_1, y_2, y_3)) \overset{x=x}{\Rightarrow} \text{node}[x](\bar{q}(y_1), \bar{q}(y_2), \bar{q}(y_3))$.

The “unsafe” case is $\bar{q}(\text{node}[x](y_1, y_2, y_3)) \overset{x=x}{\Rightarrow} \bar{q}(y_3)$, and the “harmless” case is $\bar{q}(\text{nil}[x]() \overset{x=x}{\Rightarrow} \text{true} \overset{x=x}{\Rightarrow} \text{nil}[x]()$.

The following property of STTRs will be used in Section 4.

**Definition 8.** $S$ is single-valued if $\forall t \in \mathcal{F}_T^S, q \in Q_S : |T^q_S(t)| \leq 1$.

**3.4 The Role of Regular Look-ahead**

In this section we briefly describe what motivated our choice of considering STTRs in place of STTs. The main drawback of STTs is that they are not closed under composition, even for very restricted classes. As shown in the next example, when STTs are allowed to delete subtrees, the domain is not preserved by the composition.

**Example 4.** Consider the following Fast program:

**type** $BBT[b : \text{Bool}][L(0), N(2)]$

**trans** $s_1 : BBT \rightarrow BBT$ \{ $L()$ \}

$| N(x, y) \text{ where } b \rightarrow N(b)(s_1 x)(s_1 y) \}

**trans** $s_2 : BBT \rightarrow BBT$ \{ $L() \rightarrow L[true]$ \}

Given an input $t$, $s_1$ outputs the same tree $t$ if all the nodes in $t$ have label $true$. Given an input $t_s$, $s_2$ always outputs $L[true]$. Both transductions are definable using STTs since they do not use lookahead. Now consider the composed transduction $s = s_1 \circ s_2$ that outputs $L[true]$ if all the nodes in $t$ have label $true$. This function cannot be computed by an STT: when reading a node $N[b](x, y)$, if the STT does not produce any output, it can only continue reading one of the two subtrees. This means that the STT cannot check whether the other subtree contains any false labels. However, $s$ can be computed using an STTR that checks that both $x$ and $y$ contain only true labels.

The next example shows how STTRs are simpler than STTs.

**Example 5.** The following STTR describes the function $h$ that negates a node value where the value in its left child is odd, leaves it unchanged otherwise, and is then invoked recursively on the children.

**type** $BT[x : \text{Int}][L(0), N(2)]$

**lang** $\text{oddRoot} : BT \rightarrow N(t_1, t_2)$

**def** $\text{evenRoot} : BT \rightarrow (\text{complement oddRoot})$

**trans** $h : BT \rightarrow BT$

$N(t_1, t_2) \rightarrow \{ \text{oddRoot}(t_1) \rightarrow N[\neg x](h(t_1))(h(t_2)) \}$

$1 \rightarrow N(t_1, t_2) \rightarrow \text{evenRoot}(t_1) \rightarrow N[x](h(t_1))(h(t_2))$

$\text{lL} \rightarrow L[x]$

This function can be expressed using a nondeterministic STT that guesses if the label of the left child is odd or even. Using a deterministic STTR is a more natural solution.

**3.5 Operations on Automata and Transducers**

Fast allows to define new languages and new transformations in terms of previously defined ones. Fast also supports an assertion language for checking simple program properties such as $\text{assert-true} (\text{is-empty} a)$.

**Operations that compute new languages:**

- $\text{intersect}$, $\text{complement}$, etc.: operations over STAs;
- $\text{domain} t$ computes the domain of the STTR $t$ using the operation from Definition 6; and
- $\text{pre-image} t I$ computes an STA accepting all the inputs for which $t$ produces an output belonging to $I$.

**Operations that compute new transformations:**

- $\text{restrict} t I$: constructs a new STTR that behaves like $t$, but is only defined on the inputs that belong to $I$;
- $\text{restrict-out} t I$: constructs a new STTR that behaves like $t$, but is only defined on the inputs for which $t$ produces an output that belongs to $I$; and
- $\text{compose} t_1 t_2$: constructs a new STTR that computes the functional composition $t_1 \circ t_2$ of $t_1$ and $t_2$ (algorithm described in Section 4).

**Assertions:**

- $a \in l_1 l_2$, $\text{is-empty}$: is-empty decision procedures for STAs;
- $\text{type-check} l_1 t l_2$: true iff for every input in $l_1$, $t$ only produces outputs in $l_2$.

Several operations are special applications of composition. For example $\text{restrict-out} q p = \text{compose} q (\text{restrict} I p)$, where $I$ is the identity STTR.
4. Composition of STTRs

Closure under composition is a fundamental property for transducers. Composition is needed as a building block for many operations, such as pre-image computation and output restriction. Unfortunately, as shown in Example 4 and in [22], STTs are not closed under composition. Particularly, when tree rules may delete and/or duplicate input subtrees, the composition of two STT transductions might not be expressible as an STT transduction. This is already known for classical tree transducers and can be avoided either by considering restricted fragments, or by instead adding regular lookahead [2, 14, 16]. In this paper we consider the latter option. Intuitively, regular lookahead acts as an additional child-guard that is carried over in the composition so that even when a subtree is deleted, the child-guard remains in the composed transducer and is not “forgotten”. While deletion can be handled by STTRs, duplication is a much more difficult feature to support. When duplication is combined with nondeterminism, as shown in the next example, it is impossible to express the composition of two STT transductions.

Example 6. Let be the function that, given a tree of type BT (see Example 2) transforms it by nondeterministically replacing some leaves with the value 5. Let be the function that transforms a tree into , where produces the trees and , where the two leaves contain the same value since they are “syntactically” the same run. The function cannot be expressed by an STTR. 

4.1 Composition Algorithm

Algorithms for composing transducers with regular lookahead have been studied extensively [20]. However, as shown in [22], extending classical tree transducers results to the symbolic setting is a far from trivial task. The key property that makes symbolic transducers semantically different and much more challenging than classical tree transducers, apart from the complexity of the label theory itself, is the symbolic setting is a far from trivial task. The key property that makes symbolic transducers semantically different and much more challenging than classical tree transducers, apart from the complexity of the label theory itself, is the symbolic setting.

Example 7. Suppose , the function that, given a tree of type BT (see Example 2) transforms it by nondeterministically replacing some leaves with the value 5. Let be the function that transforms a tree into , where produces the trees and , where the two leaves contain the same value since they are “syntactically” the same run. The function cannot be expressed by an STTR.

4.2 Composition Algorithm

Composition of STT transducers can be handled by STTRs, duplication is a much more difficult feature to support. When duplication is combined with nondeterminism, as shown in the next example, it is impossible to express the composition of two STT transductions.

Example 6. Let be the function that, given a tree of type BT (see Example 2) transforms it by nondeterministically replacing some leaves with the value 5. Let be the function that transforms a tree into , where produces the trees and , where the two leaves contain the same value since they are “syntactically” the same run. The function cannot be expressed by an STTR.

4.3 Composition Algorithm

Algorithms for composing transducers with regular lookahead have been studied extensively [20]. However, as shown in [22], extending classical tree transducers results to the symbolic setting is a far from trivial task. The key property that makes symbolic transducers semantically different and much more challenging than classical tree transducers, apart from the complexity of the label theory itself, is the symbolic setting.

Example 7. Suppose , the function that, given a tree of type BT (see Example 2) transforms it by nondeterministically replacing some leaves with the value 5. Let be the function that transforms a tree into , where produces the trees and , where the two leaves contain the same value since they are “syntactically” the same run. The function cannot be expressed by an STTR.

4.4 Composition Algorithm

Composition of STT transducers can be handled by STTRs, duplication is a much more difficult feature to support. When duplication is combined with nondeterminism, as shown in the next example, it is impossible to express the composition of two STT transductions.
recursive call we have the failing satisfiability check of \( \text{IsSat}(x > 0 \land \text{odd}(x+1) \land \text{odd}(x-2)) \) in line 2(a). So that there exists no choice for which 2(d) is reached in the original call so the set of return values of \( \text{Look}(x > 0, \emptyset, q, t) \) is empty.

In the following we pretend, without loss of generality, that for each rule \( \tau = \langle q, f, \varphi, \ell, t \rangle \) there is a state \( q_r \) that uniquely identifies the rule \( \langle q_r, f, \varphi, \ell, t \rangle \); \( q_r \) is used to refer to the guard and the lookahead of \( \ell \) chosen in line 2(a) in the call to \( \text{Look} \) in 2(b) below, \( q_r \) is not used elsewhere.

**Reduce** \((\gamma_i, L, v) \triangleq \)

1. if \( v = \bar{q}(\bar{g}(y_i)) \) where \( q \in Q_T \) and \( p \in Q_S \) then return \( \langle \gamma_i, L, p, \bar{q}(y_i) \rangle \).
2. if \( v = \bar{q}(\bar{g}(u_0)(\bar{u})) \) where \( q \in Q_T \) and \( g \in \Sigma \) then
   (a) choose \( \tau = \langle q, g, \varphi, u \rangle \) from \( \Delta_T \).
   (b) choose \( \gamma_i, L_i \) from \( \text{Look}(\gamma_i, L_r, q, \bar{g}(u_0)(\bar{u})) \) return \( \gamma_i \).
   (c) choose \( \chi \) from \( \text{Reduce}(\gamma_i, L_i, t(u_0, \bar{u})) \) return \( \gamma_i \).
3. if \( v = g[t_0](\bar{t}) \) where \( g \in \Sigma \) then
   (a) \( \gamma_0 := \gamma_i \), \( L_0 := L \).
   (b) for \( (i = 1; i \leq \mu(\bar{g})); ++i \) 
      choose \( \gamma_i, L_i, u_i \) from \( \text{Reduce}(\gamma_i-1, L_{i-1}, t_i) \).
   (c) return \( \langle \gamma_{i+1}, L_{i+1}, g[t_0](\bar{u}) \rangle \).

There is a close relationship between \( \text{Reduce} \) and Definition 7. We include the case

\[
T^p_T(\bar{g}(t)) \triangleq T^p_T(T^p_S(t)) \quad \text{for} \quad p \in Q_S \text{ and } t \in T^p_S, \quad (1)
\]

that allows states of \( S \) to occur in the input trees to \( T^p_T \) in a non-nested manner. Intuitively this means that rewrite steps of \( T \) are carried out first while rewrite steps of \( S \) are being postponed (called by name). In order to justify the extension (1) we need the following Lemma.

**Lemma 3.** For all \( t \in \Lambda(T^S_S, Q_S, k) \), \( a \in \sigma \), and \( u_1 \in T^S_S \):

1. \( T^p_T(\cup_S(t(a, \bar{u})) \subseteq T^p_T(t(a, \bar{u})) \), and
2. \( T^p_T(\cup_S(t(a, \bar{u}))) = T^p_T(t(a, \bar{u})) \) when \( S \) is single-valued or \( T \) is linear.

**Example 9.** The example shows a case when

\[
T^p_T(\cup_S(t(a, \bar{u}))) \neq T^p_T(t(a, \bar{u})).
\]

Suppose \( p \overset{c}{\rightarrow} \Delta, p \overset{c}{\rightarrow} \Delta, \) and \( q \overset{g}{\rightarrow} \lambda x f[x](\bar{q}(g), \bar{g}(y)). \)

Let \( f = f[\emptyset], c = c[\emptyset], g = g[\emptyset]. \) Then

\[
\bar{q}(\bar{g}(\bar{p}(c))) \overset{\rightarrow}{\bar{g}} \{ \bar{f}(\bar{q}(\bar{p}(c)), \bar{q}(\bar{p}(c))) \}
\]

but

\[
\bar{q}(\bar{g}(\bar{p}(c))) \overset{\rightarrow}{\bar{g}} \{ \bar{f}(\bar{q}(\Delta), \bar{q}(\Delta)), \bar{f}(\bar{q}(\Delta), \bar{q}(\Delta)) \}
\]

where, for example, \( f(\bar{q}(\Delta), \bar{q}(\Delta)) \) is not possible.

The assumptions on \( S \) and \( T \) given in Lemma 3 are the same as in the classical setting, however the proof of Lemma 3 does not directly follow from classical results because either the concrete alphabet \( \Sigma \times \sigma \) is infinite, or else, if \( \sigma \) is encoded as trees, linear rules become non-linear in the classical sense, such as the rule in Figure 5. Theorem 4 uses Lemma 3. It implies that, in general, \( T^p_S \) is an overapproximation of \( T^p_S \) and that \( T^p_S \) captures \( T^p_S \) precisely when either \( S \) behaves as a partial function or when \( T \) does not duplicate its tree arguments.

**Theorem 4.** For all \( p \in Q_S, q \in Q_T \) and \( t \in T^p_S \), \( S \subseteq T^p_S(t) \subseteq T^p_T(t) \), and if \( S \) is single-valued or if \( T \) is linear then \( T^p_S(t) \subseteq T^p_T(T^p_S(t)) \).

**5. Evaluation**

**Fast** can be applied in multiple different applications. We first consider HTML input sanitization for security. Then we show how augmented reality (AR) applications can be checked for conflicts. Next, we show how **Fast** can perform deforestation and verification for functional programs. Finally, we sketch how CSS analysis can be captured in **Fast**.

**5.1 HTML Sanitization**

A central concern for secure web application is untrusted user inputs. These lead to cross-site scripting (XSS) attacks, which, in its simplest form, is echoing untrusted input verbatim back to the browser. Consider bulletin boards that want to allow partial markup such as \( \langle \phi \rangle \) and \( \langle \delta \rangle \) tags or HTML email messages, where the email provider wants rich email content with formatting and images but wants to prevent active content such as JavaScript from propagating through. In these cases, a technique called sanitization is used to allow rich markup, while removing active (executable) content. However, proper sanitization is far from trivial: unfortunately, for both of these scenarios above, there have been high-profile vulnerabilities stemming from careless sanitization of specially crafted HTML input leading to the creation of the infamous Samy worm for MySpace (http://namb.la/po/), and the Yamaner worm for the Yahoo Mail system. In fact, MySpace has repeatedly failed to properly sanitize their HTML inputs, leading to the Month of MySpace Bugs initiative (http://socby.livejournal.com/586.html).

This has lead the emergence of a range of libraries attempting to do HTML sanitization, including PHP Input Filter, HTMLSafe, kses, htmlLawed, Safe HTML Checker, HTML Purifier. Among these, the last one, HTML Purifier (http://htmlpurifier.org) is believed to be most robust, so we choose it as a comparison point for our experiments. Note that HTML Purifier is a tree-based rewriter written in PHP, which uses the HTMLTidy library to parse the input.

We show how **Fast** is expressive enough to model HTML sanitizers, and we argue that writing such programs is easier with **Fast** than with current tools. Our version of an HTML sanitizer written in **Fast** and automatically translated by the **Fast** compiler into C# is partially described in Section 2. Although we can’t argue for the correctness of our implementation (except for the basic analysis shown in Section 2), sanitizers are much simpler to write in **Fast** thanks to composition. In all the libraries mentioned above HTML sanitization is implemented as a monolithic function in order to achieve reasonable performance. In the case of **Fast** each sanitization routine can be written as a single function and all such routines can be then composed preserving the property of traversing the input HTML only once.

**Evaluation** To compare different sanitization strategies in terms of performance, we chose 10 web sites and picked an HTML page from each content, ranging from 20 KB (Bing) to 409 KB in size (Facebook). For speed, the **Fast**-based sanitizer is comparable to HTML Purify. In terms
of maintainability, FAST wins on two counts. First, we can apply analysis to FAST programs that is precise, unlike analyses for PHP. Second, our sanitizer is only 200 lines of FAST code instead of 10000 lines of PHP. While these are different languages, we argue that our approach is more maintainable because FAST captures the high level semantics of HTML sanitization, as well as being fewer lines of code to understand. We manually spot-checked the outputs to determine that both produce reasonable sanitizations.

5.2 Conflicting Augmented Reality Applications

In augmented reality the view of the physical world is enriched with computer-generated information. For example, applications (often called taggers) on the Layar phone AR platform applications provide up-to-date information such as data about crime incidents near the user’s location, information about historical places and landmarks, real estate, and other points of interest.

We call a tagger an AR application that labels elements of a given set with a piece of information based on the properties of such elements. As an example, consider a tagger that assigns to every city a set of tags representing the monuments in such city. A large class of shipping mobile phone AR applications are taggers, including Layar, Nokia City Lens, Nokia Job Lens, and Juniao. We assume that the physical world is represented as a list of elements, and each element is associated with a list of tags (i.e. a tree). Users should be warned if not prevented from installing applications that conflict with others they have already installed. We say that two taggers conflict if they both label the same node of some input tree. In order to detect conflicts we perform the following four-step check for each pair of taggers \((p_1, p_2)\):

- **composition** we compute \(p\), composition of \(p_1\) and \(p_2\);
- **input restriction** we compute \(p'\), a restriction of \(p\) that only accepts trees where each node contains no tags;
- **output restriction** we compute \(p''\), a restriction of \(p'\) that only outputs trees in which some node contains two tags;
- **check** we check if \(p''\) is the empty transducer: if it is not the case, \(p_1\) and \(p_2\) conflict on every input accepted by \(p''\).

**Evaluation:** Figure 6 shows the timing results for conflict analysis. To collect this data, we randomly generated 100 taggers in FAST and checked whether they conflicted with each other. Each tagger we generated conforms to the following properties: 1) it is non-empty; 2) it tags on average 3 nodes; and 3) it tags each node at most once.

The sizes of our taggers varied from 1 to 95 states. The language we used for the input restriction has 3 states, the one for the output 5 states. We analyzed 4,950 possible conflicts and 222 will be actual conflicts. The three plots show the time distribution for the steps of a) composition, b) input restriction, and c) output restriction respectively.

All the compositions are computed in less than 250 ms, and the average time is 15 ms. All the input restrictions are computed in less than 150 ms. The average time is 3.5 ms. All the output restrictions are computed in less than 33,000 ms. The average time is 175 ms. The output restriction takes longer to compute in some cases, due to the following two factors: 1) the input sizes are always bigger: the size of the composed transducers after the input restriction \((p'\) in the list before) vary from 5 to 300 states and 10 to 4,000 rules. This causes the restricted output to have up to 5,000 states and 100,000 rules; and 2) since the conditions in the example are randomly generated, some of them may be complex causing the SMT solver to slow down the computation. The 33,000 ms example contains non-linear (cubic) constraints over reals. The average time of 193 ms per pairwise conflict check is quite acceptable: indeed, adding a new app to a store already containing 10,000 apps will incur an average checking overhead of about 35 minutes.

5.3 Deforestation

Next we explore the idea of deforestation. First introduced by Wadler in 1988 [41], deforestation aims at eliminating intermediate computation trees when evaluating functional programs. For example, to compute the sum of the squares of the integers between 1 and \(n\), the following small program might be used: \(\text{sum (map square (upto 1 n))}\). Intermediate lists created as a result of evaluation are a source of inefficiency. However, it has been observed that transducer composition can be used to eliminate intermediate results. This can be done as long as individual functions are representable as transducers. Unfortunately [41] only considers transformations over finite alphabets. We analysed the performance gain obtained by deforestation in FAST.

**Evaluation:** We considered the function \(\text{map caesar}\) from Figure 8 that replaces each value \(x\) of a integer list with \((x + 5) \mod 26\). We composed the function \(\text{map caesar}\) with itself several times to see how the performance changed when using FAST. Let’s call \(\text{map}^n\) the composition of \(\text{map caesar}\) with itself \(n\) times. We run the experiments on lists containing 4,000 randomly generated elements and we consider up to 512 composed functions. Figure 7 shows the running time with and without deforestation for a list of 4,096 integers used as the input. The running time of the FAST composed version is almost unchanged, even for 512 compositions while
6. A Comparison with Classical Tree Transducers

As we mentioned in the previous section, the HTML sanitization and CSS analysis problems could, in principle, be expressed using existing classical models and do not require symbolic alphabets. In both of these domains the alphabet is finite, and, for example, the sanitizer in Fig. 2 can be represented by classical finite state transducers with regular lookahead. In the next paragraphs we show the benefit of the symbolic representation of the alphabet and argue that the use of classical transducers does not scale in this case.

The HTML sanitization example illustrates some core differences between the symbolic and the classical case. In some respect, the situation is analogous to going from SAT to SMT solving [13], where many of the core propositional techniques remain similar but where a theory specific component adds additional succinctness and expressiveness. Consider our encoding of HTML documents presented in Fig. 3. In our representation each string value is modelled as a list of characters, and this means that each possible character should belong to the input alphabet. The input alphabet \( \Sigma \) therefore needs to include the UTF16 representation of Unicode characters, because UTF16 is used as the standard runtime representation of characters and is the basic building block of strings. Thus, \( \Sigma \) has at least 2\(^{16}\) elements, e.g., as unary function symbols \( f_1 \) for the characters \( c \). If we want to support full Unicode, e.g., including emoticons [38], we need to add additional rules that ensure that consecutive characters \( \ldots (f_1(f_1(\ldots))) \) where \( c \) and \( d \) are surrogates are indeed valid as surrogate pairs. This adds yet another layer of complexity and there are 2\(^{20}\) valid surrogate pairs. In contrast, at the level of strings, that are defined as lists of 16-bit bit-vectors, such checks are straightforward (given a solver that supports lists and bit-vector arithmetic, e.g., Z3 [12]), and involve fairly simple arithmetic operations.

We need to add lookahead automata to all the rules so that the \( \text{tag} \) subtree does not include other symbols besides the character symbols. Such an automaton needs 2\(^{16}\) transitions. The where-condition \( \text{tag} = \text{"script"} \) can be represented by a lookahead automaton, say \( A \), with six transitions. The constraint \( \text{tag} \not= \text{"script"} \) can be represented by the complement \( A' \) of \( A \). Observe that complementation of classical automata over large alphabets is expensive: while \( A \) needs six rules, one per character in the string "script", \( A' \) needs \( 6 * (2^{16} - 1) \) rules. The other string constraints are handled similarly. Besides the additional lookahead tests, transformation rules remain the same, where \( \text{tag} \) is treated as the first subtree. Observe that, a further blowup would occur if we wanted to apply transformations (other than the identity mapping, such as HtmlEncoding) to \( \text{tag} \), in which case we would need explicit rules for all of the 2\(^{16}\) symbols.

The bottom line is that tags are independent of the rest of the tree structure and the two should, if possible, not be mixed. Similar arguments already hold for symbolic finite (word) transducers as a special case of symbolic tree transducers, where a symbolic representation may avoid an exponential blowup compared to an equivalent classical transducer, as demonstrated by the symbolic word transducer implementing UTF8 encoding in [10]. The same argument holds for the domain of CSS analysis.
7. Related Work

Tree transducers. Tree transducers have been long studied, surveys and books are available on the topic [8, 21, 35]. The first models were top-down and bottom-up tree transducers [2, 14], later extended to top-down transducers with regular lookahead in order to achieve closure under composition [15, 16, 20]. Extended top-down tree transducers (XTOP) were introduced in the context of program inversion and allow rules to read more than one node at a time, as long as such nodes are adjacent. When adding lookahead such a model is equivalent to top-down tree transducers with regular lookahead. More complex models, such as macro tree transducers [17], and streaming tree transducers [1] have been introduced to improve the expressiveness at the cost of higher complexity. Due to this reason we don’t consider extending them in this paper.

Symbolic transducers. Symbolic finite transducers (SFTs) over lists, together with a front-end language BEX, were originally introduced in [25] with a focus on security analysis of string sanitizers. The main SFT algorithms, in particular an algorithm for deciding equivalence of SFTs modulo a decidable background theory, are studied in [40]. Variants of SFTs in which multiple input symbols can be read by a single transition are studied in [9] and in [5]. Symbolic tree transducers are originally introduced in [39], where it is wrongly claimed that STTs are closed under composition by referring to a generalization of a proof of the classical case in [21] which is only stated for total deterministic finite tree transducers. In [22] this error is discovered and other properties of STTs are investigated. The main result of [39] is an algorithm for checking equivalence of single-valued linear STTs. For classical transducers, equivalence has been shown to be decidable for deterministic or finite-valued tree transducers [36], streaming tree transducers [1], and MSO tree transformations [18]. We are currently investigating the problem of checking equivalence of single-valued STTRs.

DSL for tree manipulation. Domain specific languages for tree transformation have been studied in several different contexts. VATA [30] is a tree automata library for analyzing tree languages over large alphabets. In VATA transitions are represented symbolically using BDDs, however the library does not support transducers and it is limited to nondeterministic automata over finite (although large) alphabets. TTT [34] and Tiburon [33], are transducers based languages used in natural language processing. TTT allows complex forms of pattern matching, but does not enable any form of analysis. Tiburon supports probabilistic transitions and several transducers algorithms. Both the languages are limited to finite input and output alphabets. ASF+SDF [7] is a term-rewriting language for manipulating parsing trees. ASF+SDF is simple and efficient, but does not support any analysis. In the context of XML processing numerous languages have been proposed for querying (XPath, XQuery [42]), stream processing (STX [3]), and manipulating (XSLT, XDuce [27]) XML trees. While being very expressive, these languages support very limited forms of analysis. Emptiness has been shown decidable for restricted fragments of XPath [4]. XDuce [27] allows to define basic XML transformations, and supports a tree automata based type-checking that is limited to finite alphabets. We plan to extend FAST to better handle XML processing and to identify a fragment of XPath expressible in FAST. However, to the best of our knowledge, FAST is the first language for tree manipulations that supports infinite input and output alphabets while preserving decidable analysis. Table 1 summarizes the relations between FAST and the other domain-specific languages for tree transformations.

Applications. The connection between tree transducers and deforestation was first investigated in [41], and then further investigated in [29]. In this setting deforestation is done via Macro Tree Transducers (MTT) [17]. While being more expressive than Top Down Transducers with regular lookahead, MTTs only support finite alphabets and their composition is very expensive. We are not aware of an actual implementation of the techniques in [29]. Higher-Order Multi-Parameter Tree Transducers (HMTT) [28] are used for type-checking higher-order functional programs. HMTTs enable sound but incomplete analysis of programs which takes multiple trees as input, but only support finite alphabets. Extending our theory to multiple input trees and higher-order functions is an open research direction.

Open problems. Several complexity related questions for STAs and STTRs are open and depend on the complexity of the label theory, but some lower bounds can be established using known results for finite tree automata and transducers. For example, an STA may be exponentially more succinct than the equivalent normalized STA because one can directly express the intersection non-emptiness problem of a set of normalized STAs as the emptiness of a single un-normalized STA. In the classical case, the non-emptiness problem of tree automata is P-co, while the intersection non-emptiness problem is EXPTIME-co [8, Thm 1.7.5]. Recently, new techniques based on antichains have been proposed to check universality and inclusion for nondeterministic tree automata [6]. Whether such techniques translate to our setting is an open research direction. Concrete open problems are decidability of: single-valuedness of STTRs, equivalence of single-valued STTRs, and finite-valuedness of STTRs. Classically these problems are decidable, but some proofs are mathematically quite challenging [36]. Novel algorithms for minimizing and learning symbolic automata over lists have been recently proposed in [11] and [5]. Extending such results to STAs are also unexplored topics.

<table>
<thead>
<tr>
<th>Language σ</th>
<th>Analysis</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAST  ∞</td>
<td>composition, pre-image, language equivalence, determinizing, complement, intersection</td>
<td>Tree-manipulating programs</td>
</tr>
<tr>
<td>VATA  ff</td>
<td>union, intersection, language inclusion</td>
<td>Tree-automata</td>
</tr>
<tr>
<td>Tiburon  ff</td>
<td>composition, type-checking, training weights, language equivalence, determinizing, complement, intersection</td>
<td>NLP</td>
</tr>
<tr>
<td>TTT  ff</td>
<td>-</td>
<td>Parsing</td>
</tr>
<tr>
<td>ASF+SDF  ∞</td>
<td>-</td>
<td>NLP</td>
</tr>
<tr>
<td>XPath  ∞</td>
<td>emptiness for a fragment</td>
<td>XML query (only selection)</td>
</tr>
<tr>
<td>XDuce  ∞</td>
<td>type-checking for navigational part (finite alphabet)</td>
<td>XML transformations</td>
</tr>
</tbody>
</table>

Table 1: Summary of main domain specific languages for tree-manipulating programs and their properties; σ indicates whether the language supports finite (ff) or infinite (∞) alphabets.
8. Conclusions

We introduce Fast, a new domain-specific language for tree manipulation based on symbolic tree automata and symbolic tree transducers. To allow Fast to perform useful program analysis, we design a novel algorithm for composing symbolic tree transducers with regular lookahead and we prove its correctness. Fast strikes a delicate balance between precise analysis and expressiveness, and we show how multiple applications benefit from this analysis. A running version of Fast can be accessed at http://rise4fun.com/Fast/.

Acknowledgements. We thank the anonymous reviewers for their valuable feedback that helped us improving the quality of our paper. Loris D’Antoni did this work as part of an internship at Microsoft Research, and he is supported by NSF Expeditions in Computing award CCF 1138996.

References