July 17, 2011

RT0930

Computer Science; Security 12 pages

Research Report

Path- and Index-sensitive String Analysis based on Monadic Second-order Logic

Takaaki Tateishi, Marco Pistoia, Omer Tripp

IBM Research - Tokyo
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Path- and Index-sensitive String Analysis Based on Monadic Second-order Logic

Takaaki Tateishi
IBM Research - Tokyo
tate@jp.ibm.com

Marco Pistoia
IBM Research - T. J. Watson Research Center
pistoia@us.ibm.com

Omer Tripp
IBM Software Group and Tel Aviv University
omert@il.ibm.com

ABSTRACT

We propose a novel technique for statically verifying the strings generated by a program. The verification is conducted by encoding the program in Monadic Second-Order Logic (M2L). We use M2L to describe constraints among program variables and to abstract built-in string operations. Once we encode a program in M2L, a theorem prover for M2L, such as MONA, can automatically check if a string generated by the program satisfies a given specification, and if not, exhibit a counterexample. With this approach, we can naturally encode relationships among strings, accounting also for cases in which a program manipulates strings using indices. In addition, our string analysis is path sensitive in that it accounts for the effects of string and Boolean comparisons, as well as regular-expression matches.

We have implemented our string-analysis algorithm, and used it to augment an industrial security analysis for Web applications by automatically detecting and verifying *sanitizers*—methods that eliminate malicious patterns from untrusted strings, making those strings safe to use in security-sensitive operations. On the 8 benchmarks we analyzed, our string analyzer discovered 128 previously unknown sanitizers, compared to 71 sanitizers detected by a previously presented string analysis.

Categories and Subject Descriptors

D.2.4 [**Software Engineering**]: Software/Program Verification; D.2.5 [**Software Engineering**]: Testing and Debugging

General Terms

Languages, Security, Verification

Keywords

String Analysis, Static Program Analysis, Web Security

1. INTRODUCTION

String analysis [7, 8, 12, 17, 19, 23, 32] is a particular form of program analysis whose purpose is to infer string values arising at run time. It is often used in the verification of server-side Web applications, where string values used in security-sensitive computations are compared to safe and/or unsafe string patterns to detect potential security vulnerabilities, such as cross-site scripting (XSS), HTTP response splitting (HRS) and Structured Query Language (SQL) injection (SQLi) [2].

1.1 String Analysis for Security

A common way of conducting string analysis is by constructing context-free grammars or regular grammars to approximate strings [8, 23, 32]. With this approach, each built-in string operation is modeled by a grammar transducer. This form of string analysis is suitable for analyzing code that *sanitizes* strings using string-manipulation operations such as Java's `replace` method. Many Web applications fall in this category because they sanitize their inputs by *removing* potentially malicious string patterns or *replacing* them with safe ones. Sanitizers often perform validation against certain patterns, and process the inputs only if validation succeeds. For these validation-based sanitizers, a path-sensitive string analysis is necessary. Conversely, path-insensitive string analyses will conservatively report violations even when proper validation takes place.

In addition, a very large number of Web applications perform sanitization by extracting substrings from input strings, starting at specific indices. Grammar-based string analyses are unable to precisely verify strings that are constructed in this way, and will have to report violations conservatively even when proper index-based sanitization has taken place. Consequently, path-sensitivity and the ability to model index-based string manipulation are essential features when verifying Web applications for security.

1.2 Motivating Example

The Java method `clean` in Figure 1 can be used to prevent an XSS attack. The input to the method can be any possible value, including values potentially under the control of an attacker. In XSS, an attacker typically wraps JavaScript code into a (`<script>`, `</script>`) tag pair, and embeds it into text that becomes part of an online encyclopedia, blog, or social network. Once the text is rendered on other people's browsers, the embedded code is automatically executed on the victims' computers. For this example, we consider the output safe if it does not contain character <.

In order to verify that the program is immune to XSS attacks, we need to prove that no string generated by the program contains <. According to this specification, `clean` is considered a valid XSS sanitizer; when condition `v1.contains(v2)` holds, < is effectively removed from the input string by combining `indexOf` and `substring`, and when that condition does not hold, the string value returned by the method is the same as the input string, which does not contain <. However, a path-insensitive grammar-based string analysis cannot follow this line of reasoning, since it would

```
String clean(String v1){
  String v2 = "<";
  if (v1.contains(v2)) {
    int v3 = v1.indexOf(v2);
    String v4 = v1.substring(0, v3);
    return v4;
  }
  return v1;
}
```

Figure 1: Sanitization against XSS

fail to capture the relationship between `v2` and `v3`, and thus the effect of the ensuing `substring` operation. As a consequence, it produces a resulting grammar that conservatively contains <.

1.3 Our Approach

To abstract string values, we use M2L(Str) (Monadic Second-order Logic on strings) [16]. The effect of branch conditions and dependencies among program variables is abstracted and encoded as M2L formulae. Built-in string operations are also abstracted by M2L formulae, with each formula representing relationships among input and output parameters. In particular, a string operation using an index can be represented naturally by a M2L formula, since M2L(Str) is capable of explicitly mentioning *positions* in a given finite string and can deal with variables ranging over positions (*position variables*) or variables ranging over sets of positions (*position set variables*) on the finite string.

The use of M2L(Str) has the following advantages in addition to enabling index sensitivity combined with path sensitivity:

- *Conservativeness.* M2L(Str) captures not only fixed-size strings but also finite strings (regular languages). This feature is necessary for guaranteeing that our string analysis is conservative, which implies that sanitization code verified by our string analysis can be safely used as a sanitizer, and for conservatively modeling several important built-in string operations such as Java's `replace` method in a manner similar to finite-state transducers that cannot be captured by fixed-size representation.
- *Efficient and effective automaton representation.* We can exploit an automatic theorem prover MONA [20] to implement our string analysis algorithm, where MONA uses the BDD-based automaton representation of M2L formulae. Furthermore, the use of MONA has potential to advance the string analysis implementation in the future, since MONA enables performing separate compilation and generating constraints on input strings (like vulnerability signatures [6, 34]) including counterexamples.

Our analysis consists of the following two automated processes: (i) encoding a string-manipulating method as an M2L formula ϕ_1 that represents possible strings returned by the method, and (ii) encoding a regular expression indicating unsafe strings as an M2L formula ϕ_2, and checking the satisfiability of $\phi_1 \wedge \phi_2$ to verify that the possible strings returned by the method never contain any of unsafe strings, where the method is reported as a sanitizer iff the formula is unsatisfiable. In the first process, the effect of branch conditions and index-based string manipulations are also encoded in M2L, and reflected to the M2L formula ϕ_1. Therefore, our string analysis is both path-sensitive and index-sensitive, and thus we call it PISA (Path- and Index-sensitive String Analysis) in this paper.

1.4 Contributions

This paper makes the following contributions:

(Position variable)	p	$\in$	$\mathsf{Var1}$
(Position-set variable)	P	$\in$	$\mathsf{Var2}$
(Position term)	t	$::=$	$p \mid t+i \mid t-i \mid \$ \mid 0$
(Position-set term)	T	$::=$	$\emptyset \mid \{t,\ldots,t\} \mid \mathsf{all} \mid P \mid T \cup T \mid T \cap T$
		$\mid$	$T \setminus T \mid T^{-1}$
(Formula)	ϕ	$::=$	$`a\text{'}(t) \mid t = t \mid t < t \mid t \leq t$
		$\mid$	$T = T \mid T \subset T \mid T \subseteq T \mid t \in T$
		$\mid$	$\neg\phi \mid \phi \wedge \phi \mid \phi \vee \phi \mid \phi \Rightarrow \phi \mid \phi \Leftrightarrow \phi$
		$\mid$	$\exists p.\phi \mid \forall p.\phi \mid \exists P.\phi \mid \forall P.\phi$

Figure 2: Syntax of M2L(Str)

- *Novel features enabled by M2L.* Our encoding method goes beyond that for regular expressions [20]. Compared to existing string analyses based on bit-vector logic and/or word equation [5, 19, 25], our M2L-based approach can model more string transformations such as replacement and upper-case transformations. Furthermore, to the best of our knowledge, PISA is the first purely static string analysis that simultaneously handles index sensitivity, path sensitivity, and string-replacement operations.
- *Sanitizer detection by path- and index-sensitive string analysis.* String analysis has already been used for sanitizer detection [4]. However, [4] uses an imprecise string analysis, which is neither index nor path sensitive, and compensates for this loss in precision by relying on a complementary dynamic analysis. PISA, on the other hand, is much more precise, which obviates the need for an accompanying dynamic analysis. This also enables scanning applications during the development phase, where they cannot yet be deployed (and thus dynamic analysis cannot be used), which is the optimal stage for detecting security vulnerabilities.
- *Implementation and evaluation.* PISA is fully implemented and is featured in a commercial security product [1]. We evaluated PISA's precision by comparing it with the technique of [23, 12] on 8 open-source benchmarks. We further examined PISA's effectiveness by integrating it into a commercial taint-analysis algorithm. The results show PISA to be far more precise than the previous technique, and also effective in boosting the precision of its client taint analysis.

1.5 Organization

The rest of the paper is organized as follows: In Section 2, we present the overview of our string analysis algorithm. The core string-analysis algorithm is described in Section 3. Then, in Section 4, we extend the core algorithm with index sensitivity and path sensitivity. In Section 5, we extend our analysis to become interprocedural. Section 6 discusses our implementation of the algorithm, as well as experimental results. Section 7 surveys related work, and Section 8 concludes this paper.

2. OVERVIEW

Our string analysis verifies a program by encoding it in M2L(Str) and then checking the satisfiability of an M2L formula. Therefore, we first present the definition of M2L(Str), and then briefly describe how to encode strings and programs in M2L(Str).

2.1 Monadic Second-order Logic on Strings

M2L(Str) is a widely used vehicle for a variety of verification problems [16]. The syntax of M2L(Str) is defined in Figure 2, where $\mathsf{Var1}$ denotes a set of position variables and $\mathsf{Var2}$ a set of position-set variables. The formula 'a'(t) holds if a_i in (finite) string $w = a_0 \cdots a_{n-1}$ is 'a', where i is the interpretation of t. Constants 0 and \$ represent the first and last positions in a string, respectively. The addition $t+i$ of position-term t and natural number i is interpreted as $t + i = j + i \bmod n$, where j is the interpretation of t, and n is the length of string w. $T + i$, where T is a position-set term, results in position set $\{t + i \mid t \in T\}$. $t - i$ and $T - i$ are interpreted similarly.

Its semantics is determined by checking whether an M2L formula ϕ holds on a finite string $w \in \Sigma^*$ and an assignment $\mathcal{I} \in$

$$
\begin{array}{lll}
\mathsf{prog}_{\mathtt{v1}}(V_1) & \equiv & \mathsf{true} \\
\mathsf{prog}_{\mathtt{v2}}(V_2) & \equiv & \text{“<”}(V_2) \\
\mathsf{prog}_{\mathtt{v3}}(v_3, V_1) & \equiv & [\![\mathtt{indexOf}]\!]\,(v_3, V_1, \mathsf{prog}_{\mathtt{v1}}, \mathsf{prog}_{\mathtt{v2}}) \\
\mathsf{prog}_{0}(v_0, V_1) & \equiv & \mathsf{min}(v_0, V_1) \\
\mathsf{prog}_{\mathtt{v1'}}(V_1) & \equiv & \mathsf{prog}_{\mathtt{v1}}(V_1) \wedge [\![\mathtt{contains("<")}]\!]\,(V_1) \\
\mathsf{prog}_{\mathtt{v4}}(V_4) & \equiv & [\![\mathtt{substring}]\!]\,(V_4, \mathsf{prog}_{\mathtt{v1'}}, \mathsf{prog}_{0}, \mathsf{prog}_{\mathtt{v3}}) \\
\mathsf{prog}_{\mathtt{v1''}}(V_1) & \equiv & \mathsf{prog}_{\mathtt{v1}}(V_1) \wedge \neg [\![\mathtt{contains("<")}]\!]\,(V_1),
\end{array}
$$

Figure 3: An example of predicate declarations

($\mathcal{P} \rightarrow 2^{\mathsf{Pos}}$), where $\mathcal{P}$ is the set of free position set variables[1], and 2^{Pos} is the power set of the position set Pos. When the formula holds, we write $w, \mathcal{I} \models \phi$. For example, the M2L formula $\text{‘a’}(0) \wedge \text{‘b’}(1) \wedge \text{‘c’}(2) \wedge \text{‘a’}(3)$ holds on the finite string $w = \text{“abca”}$, which states that character ‘a’ is located at positions 0 and 3, while characters ‘b’ and ‘c’ are located at positions 1 and 2, respectively. Thus, we can write $w, \mathcal{I} \models \text{‘a’}(0) \wedge \text{‘b’}(1) \wedge \text{‘c’}(2) \wedge \text{‘a’}(3)$, where $\mathcal{I} = \{\}$.

2.2 Encoding Programs in M2L(Str)

Our encoding method treats string values using position sets without loss of the order of characters. For example, given the formula $\text{‘a’}(0) \wedge \text{‘b’}(1) \wedge \text{‘c’}(2) \wedge \text{‘a’}(3) \wedge P = \{0\} \wedge Q = \{1, 3\} \wedge R = \{0, 1, 3\}$ holds on $w = \text{“abca”}$ and $\mathcal{I} = \{P \mapsto \{0\}, Q \mapsto \{1, 3\}, R \mapsto \{0, 1, 3\}\}$, the position set variables P,Q, and R can be considered to be the strings “a”, “ba” and “aba”, respectively. In this representation of strings, the concatenation of the two strings represented by P and Q is captured by $P \cup Q$, which is equal to R, without loss of the order of characters, since all of the positions in P are less than any position in Q. Thus, this concatenation relationship can be represented by the predicate:

$$
\begin{array}{l}
\mathsf{concat}(R, P, Q) \equiv \\
(R = P \cup Q) \wedge (\forall p, q \,.\, p \in P \wedge q \in Q \Rightarrow p < q),
\end{array}
$$

where $p < q$ ensures the order of the characters in P and Q. Based on this predicate, we introduce the notation “s”(S) which means that position-set variable S represents string s.

With this encoding of strings, a program is encoded as a set of M2L predicate declarations, where each M2L predicate is declared corresponding to the definition of a program variable. We also use a pre-defined predicate for every string operation to represent each constraint among the return value and the parameters, where any constraint can be abstracted. For example, let us consider the following two-line program.

```
String v1 = "a";
String v2 = v1.concat(v1);
```

Here is the set of predicate declarations when this program is encoded into M2L.

$$
\begin{array}{l}
\mathsf{prog}_{\mathtt{v1}}(V_1) \equiv \text{“a”}(V_1) \\
\mathsf{prog}_{\mathtt{v2}}(V_2) \equiv [\![\mathtt{concat}]\!]\,(V_2, \mathsf{prog}_{\mathtt{v1}}, \mathsf{prog}_{\mathtt{v1}}) \\
\quad \text{where} \quad [\![\mathtt{concat}]\!]\,(R, \mathcal{P}_1, \mathcal{P}_2) \equiv \\
\qquad\qquad \exists P_1, P_2 \,.\, \mathcal{P}_1(P_1) \wedge \mathcal{P}_2(P_2) \wedge \mathsf{concat}(R, P_1, P_2)
\end{array}
$$

Each predicate prog_{v_i} represents the post-condition for the assignment to program variable v_i, where the parameters V_1 and V_2 of the predicates are associated with the program variables v1 and v2, respectively. $[\![\mathtt{concat}]\!]$ is a pre-defined predicate representing an abstraction of the string concatenation operation, where the return value is represented by R, and the parameters are represented by $\mathcal{P}_1$ and $\mathcal{P}_2$ which are instantiated by the predicates associated with the program variables.

With this encoding of programs, the clean method of Figure 1 is encoded as the set of predicate declarations of Figure 3, where we introduce the program variables v1′ and v1″ to distinguish the program variable v1 used in the “then” block of the if-statement from that returned at the end of the method. The variables V_1, V_2, v_3, and V_4 are also M2L variables associated with the program variables v1,v2,v3, and v4, respectively, where upper-case variables V_1,V_2, and V_4 are position set variables each of which represents a string, and the lower-case variable v_3 is a position variable which represents an index. Note that the effect of the branch condition is encoded as constraints on these M2L variables in the declarations of the predicates $\mathsf{prog}_{\mathtt{v1'}}$ and $\mathsf{prog}_{\mathtt{v1''}}$, where $[\![\mathtt{contains("<")}]\!]\,(V_1)$ [2] is the abstraction of the condition contains("<"), which means that the string represented by V_1 contains the string “<”. The predicate prog_0 and $\mathsf{prog}_{\mathtt{v3}}$ means index 0 and an index assigned to the program variable v3, where each of the predicates takes two parameters: a position variable and a position set variables, since we represents an index of a string using the pair of a position and a position set. $[\![\mathtt{indexOf}]\!]$ and $[\![\mathtt{substring}]\!]$ are the abstractions of string operations indexOf and substring as in the case of $[\![\mathtt{concat}]\!]$, respectively. The details of these abstractions are discussed later in Section 3 and 4.

If our only concern is that the string “<” is unsafe, an unsafe specification Unsafe for the program is defined as $\mathsf{Unsafe}(V) \equiv \exists R' \,.\, \mathsf{substr}(R', V) \wedge \text{“<”}(R')$, where $\mathsf{substr}(R', V)$ means that the string R' is a substring of V. Alternatively, we can use the regular expression “.*<.*” which is equivalent to the specification, since we can encode the regular expression into M2L. We then verify that the program never return a string containing the unsafe string by confirming the unsatisfiability of this formula: $\exists V \,.\, (\mathsf{prog}_{\mathtt{v1''}}(V) \vee \mathsf{prog}_{\mathtt{v4}}(V)) \wedge \mathsf{Unsafe}(V)$.

3. CORE ALGORITHM

In this section, we first describe our target language, and then describe the method of encoding strings and regular expressions followed by the method of encoding a program as a set of M2L predicate declarations. We also present a formal discussion of our encoding method, accompanied by a soundness theorem.

3.1 Target Language

We assume that the target program is translated to Static Single Assignment (SSA) form [24, 10]. An SSA program comprises numbered basic blocks, each of which consists of the instructions in Figure 4. In addition, we use the notation op(h) for built-in operator h. For example, x=op(+)(1,2) assigns the result of 1+2 to program variable x. Note that the return instruction and the instruction of calling a user-defined function are introduced only for reflecting actual program languages such as Java, as the core of our string analysis algorithm is intraprocedural. We omit the details of the method of encoding those types of instructions in Section 3. However, this does not restrict us from conducting an interprocedural analysis, since our algorithm can be simply extended so as to conduct the interprocedural analysis by treating assignment relationships between caller's program variables and callee's program variables as described in Section 5.

The following SSA program represents the clean method introduced in Section 1, where 1:, 2:, and 3: represent the basic blocks numbered 1,2, and 3.

```
1:v0 = 0;                2:return v1;
  v2 = "<";              3:v3 = indexOf(v1,v2);
  b1 = v1.contains(v2);    v4 = substring(v1,v0,v3);
  jump b1, 3               return v4;
```

[1] Position variables are handled by using singleton position set variables.

[2] The reason why we do not use $[\![\mathtt{contains}]\!]\,(V_1, V_2)$ is that M2L as well as our encoding method cannot treat equality of strings. Therefore, our encoding method relies on constant propagation analysis to obtain concrete strings such as “<” to avoid this limitation.

(Assignment)	$x = v$	The value v is assigned to program variable x.
(Function call)	$x = f(x_1, \cdots, x_n)$	The result of invoking function f with parameters $x_1, \cdots, x_n$ is assigned to program variable x. f is either a built-in function or a user-defined function.
(ϕ function)	$x = \texttt{phi}(b_1 : x_1, \ldots, b_n : x_n)$	When an immediate predecessor of the current basic block is b_i, program-variable x is assigned the value of x_i. Basic-block numbers are omitted for brevity when possible.
(Conditional jump)	$\texttt{jump}\ x,\ b$	The program jumps to the basic block numbered b if the value of x is true.
(Goto)	$\texttt{goto}\ b$	The program jumps to the basic block numbered b.
(Return)	$\texttt{return}\ x$	The value of the program variable x is returned.

Figure 4: Instructions of the target language

$$
\begin{array}{lcl}
\text{“}a_1 \cdots a_n\text{”}(P) & \equiv & \exists t_1, \cdots, t_n\ . \\
 & & \text{‘}a_1\text{’}(t_1) \wedge \cdots \wedge \text{‘}a_n\text{’}(t_n) \\
 & & \wedge\ t_1 < t_2 \wedge t_2 < t_3 \wedge \cdots \wedge t_{n-1} < t_n \\
 & & \wedge\ P = \{t_1, \cdots t_n\} \\
\mathsf{concat}(R, P, Q) & \equiv & R = P \cup Q \\
 & & \wedge(\forall p, q\ .\ p \in P \wedge q \in Q \Rightarrow p < q) \\
\mathsf{strr}(R, P, p, q) & \equiv & p \leq q \wedge R \subseteq P \wedge (\forall r\ .\ r \in P \\
 & & \qquad \Rightarrow (r \in R \Leftrightarrow p \leq r \wedge r < q)) \\
\mathsf{substrr}(R, P, p, q) & \equiv & \exists p', q'\ .\ p \leq p' \wedge p' \leq q' \wedge q' \leq q \\
 & & \wedge \mathsf{strr}(R, P, p', q') \\
\mathsf{substr}(R, P) & \equiv & \mathsf{substrr}(R, P, \mathsf{min}(P), \mathsf{max}(P) + 1) \\
\mathsf{consecutive}(p, q, R) & \equiv & p < q \wedge p \in R \wedge q \in R \\
 & & \wedge(\forall r\ .\ p < r \wedge r < q \Rightarrow r \notin R)
\end{array}
$$

Figure 5: Utility predicates

$$
\begin{array}{lcl}
\langle\!\langle T \rangle\!\rangle & \rightarrow & \lambda S\ .\ \text{‘}T\text{’}(S) \\
\langle\!\langle T_x \rangle\!\rangle & \rightarrow & \lambda S\ .\ \mathsf{prog}_x(S) \\
\langle\!\langle N_1 N_2 \rangle\!\rangle & \rightarrow & \lambda S\ .\ \exists S_1, S_2\ .\ \langle\!\langle N_1 \rangle\!\rangle(S_1) \wedge \langle\!\langle N_2 \rangle\!\rangle(S_2) \\
 & & \qquad \wedge \mathsf{concat}(S, S_1, S_2) \\
\langle\!\langle N_1 \mid N_2 \rangle\!\rangle & \rightarrow & \lambda S\ .\ \langle\!\langle N_1 \rangle\!\rangle(S) \vee \langle\!\langle N_2 \rangle\!\rangle(S) \\
\langle\!\langle N^\star \rangle\!\rangle & \rightarrow & \lambda S\ .\ \exists P\ .\ \mathsf{min}(S) \in P \wedge \mathsf{max}(S) + 1 \in P \\
 & & \qquad \wedge \forall r, r'\ .\ \mathsf{consecutive}(r, r', P) \\
 & & \qquad\qquad \Rightarrow \exists Q\ .\ \mathsf{strr}(Q, S, r, r') \wedge \langle\!\langle N \rangle\!\rangle(Q)
\end{array}
$$

Figure 6: Encoding regular expressions

$$
\begin{array}{l}
\llbracket \texttt{replace} \rrbracket\,(R, \mathcal{P}_s, \mathcal{P}_x, \mathcal{P}_y) \equiv \\
\bigvee_{v \in V} (\exists S, X, Y\ .\ \mathcal{P}_s(S) \wedge (\forall S'\ .\ \mathsf{substr}(S', S \setminus X) \Rightarrow \neg\text{“}v\text{”}(S')) \\
\qquad \wedge \langle\!\langle v^\star \rangle\!\rangle'(X, S) \wedge \langle\!\langle y^\star \rangle\!\rangle(Y) \wedge \langle\!\langle (vy)^\star \rangle\!\rangle(X \cup Y) \wedge S \cap Y = \emptyset \\
\qquad \wedge R = ((S \setminus X) \cup Y)) \\
\text{where} \\
\langle\!\langle v^\star \rangle\!\rangle'(X, S) \equiv \exists P\ .\ \mathsf{min}(X) \in P \wedge \mathsf{max}(X) + 1 \in P \\
\qquad\qquad \wedge \forall r, r'\ .\ \mathsf{consecutive}(r, r', P) \\
\qquad\qquad \Rightarrow \exists Q\ .\ \mathsf{strr}(Q, X, r, r') \wedge \mathsf{substr}(Q, S) \wedge \langle\!\langle v \rangle\!\rangle(Q)
\end{array}
$$

$$
\begin{array}{l}
\llbracket \texttt{indexOf} \rrbracket\,(p, P, \mathcal{P}_1, \mathcal{P}_2) \equiv \\
\quad \mathcal{P}_1(P) \wedge \bigvee_{v \in V}((\exists P_2\ .\ \text{“}v\text{”}(P_2) \wedge \mathsf{indexOf}(p, P, P_2)) \\
\qquad \wedge(\mathsf{min}(P) \leq p \Rightarrow (\forall P_2, p'\ .\ \text{“}v\text{”}(P_2) \wedge \mathsf{indexOf}(p', P, P_2) \\
\qquad\quad \wedge \mathsf{min}(P) \leq p' \Rightarrow p \leq p'))) \\
\text{where} \\
\mathsf{indexOf}(p, P, Q) \equiv (\mathsf{substr}(Q, P) \Rightarrow ((Q \neq \emptyset \Rightarrow \mathsf{min}(Q) = p) \\
\qquad\qquad\qquad\qquad\qquad\qquad \wedge(Q = \emptyset \Rightarrow \mathsf{min}(P) = p))) \\
\qquad\qquad\qquad\qquad \wedge(\neg\mathsf{substr}(Q, P) \Rightarrow p < \mathsf{min}(P))\,.
\end{array}
$$

$$
\begin{array}{l}
\llbracket \texttt{substring} \rrbracket\,(R, \mathcal{P}_s, \mathcal{P}_n, \mathcal{P}_m) \equiv \\
\exists S, n, m\ .\ \mathcal{P}(S) \wedge \mathcal{P}_n(n, S) \wedge \mathcal{P}_m(m, S) \wedge \mathsf{strr}(R, S, n, m)
\end{array}
$$

$$
\begin{array}{l}
\llbracket \texttt{contains}, 1 \rrbracket\,(R, c, \mathsf{prog}_{v_1}, \mathsf{prog}_{v_2}) \equiv \\
\begin{cases}
\exists P\ .\ \mathsf{prog}_{v_2}(P) \wedge \mathsf{substr}(P, R) & \text{when } c = \mathsf{true} \\
\bigvee_{s \in S} \neg(\exists P\ .\ \text{“}s\text{”}(P) \wedge \mathsf{substr}(P, R)) & \text{when } c = \mathsf{false} \\
\mathsf{true} & \text{otherwise}
\end{cases}
\end{array}
$$

Figure 7: Abstractions of the built-in functions

3.2 Encoding Strings

Our encoding method treats a string value of size m as a position set P of the same size. The positions in P are taken from a “global” position set, $\{0, \ldots, n-1\}$, that represents a word w. Formally, if $w = a_0 \cdots a_{n-1}$ and $P = \{p_0, \cdots, p_{m-1}\}$ is a sorted position set, then P represents the string value s (of size m) iff P satisfies $s = a_{p_0} a_{p_1} \cdots a_{p_{m-1}}$. Given $w =$ “$\texttt{abca}$”, the sets of positions $\{0, 1\}$ and $\{2, 3\}$ represent the strings “$\texttt{ab}$” and “$\texttt{ca}$”, respectively.

Figure 5 lists utility predicates used for the encoding, where “a”(P) and $\mathsf{concat}(R, P, Q)$ are the same as those introduced in Section 2.2. Intuitively, $\mathsf{strr}(R, P, p, q)$ denotes that a string represented by R is the substring represented by P containing all the characters in the range $[p, q)$. $\mathsf{substrr}(R, P, p, q)$ is similar to strr, the difference being that R may be any substring. $\mathsf{min}(P)$ and $\mathsf{max}(P)$ return the minimum and maximum positions in P, respectively. Finally, predicate $\mathsf{consecutive}(p, q, R)$ denotes that positions p and q are consecutive in position set R. This predicate is used to encode the Kleene closure of a regular language (*cf.* $\mathsf{consecutive\_in\_set}$, as described in [20]).

Our algorithm for encoding regular expressions is the same as that of [20], except that we accept program variables as terminal symbols. Figure 6 shows how to encode a set of strings represented by a regular expression r as a predicate denoted by $\langle\!\langle r \rangle\!\rangle$. T represents a terminal symbol (*i.e.*, a character), and T_x represents a terminal symbol associated with program variable x, where prog_x denotes a property of strings possibly assigned to x. N,N_1, and N_2 represent nonterminal symbols. In addition, we use the notation $\lambda S.\phi$, instead of explicitly declaring a new predicate ψ such that $\psi(S) = \phi$.

3.3 Abstracting Built-in Functions

We denote the abstraction of a built-in function f by $\llbracket f \rrbracket$. The parameters of a built-in function are implicitly represented by *higher-order variables*, each of which is instantiated by a predicate representing a property of the relevant actual parameter. Thus, all higher-order variables are instantiated by the end of the encoding process. For example, the higher-order variables $\mathcal{P}_1$ and $\mathcal{P}_2$ used in the encoding of the string-concatenation program in Section 2.2 are instantiated by predicate $\mathsf{prog}_{\mathrm{v1}}$, thus yielding:

$$
\begin{array}{l}
\mathsf{prog}_{\mathrm{v2}}(V_2) \equiv \\
\qquad \exists P_1, P_2\ .\ \mathsf{prog}_{\mathrm{v1}}(P_1) \wedge \mathsf{prog}_{\mathrm{v1}}(P_2) \wedge \mathsf{concat}(V_2, P_1, P_2).
\end{array}
$$

The examples of the abstractions we developed for the built-in functions are listed in Figure 7. Here, we focus only on $\llbracket \texttt{replace} \rrbracket$, which abstracts the Java method `replace`, where `s.replace(x,y)` substitutes all occurrences of `x` in `s` with `y`. Discussion of the other functions is deferred to Section 4, where extensions of the core algorithm needed by the corresponding abstractions are introduced. In our abstraction of `replace`, V is the set of concrete strings possibly assigned to `x`, X represents the set of positions to be removed from position-set S, and Y represents a set of positions to be inserted. Predicate $\langle\!\langle y^\star \rangle\!\rangle$ encodes regular expression $y^\star$, where y is the program variable corresponding to $\mathcal{P}_y$. Predicate $\langle\!\langle (vy)^\star \rangle\!\rangle$ constrains X and Y to guarantee that each pair of removed and inserted positions is consecutive, and predicate $\langle\!\langle v^\star \rangle\!\rangle'$ encodes regular expression $v^\star$, and constrains X to contain only

$$\begin{array}{ll}
\llbracket x := v \rrbracket & \rightarrow \mathsf{prog}_x(R) \equiv \text{“}v\text{”}(R) \\
\llbracket x := f(x_1,\cdots,x_n) \rrbracket & \rightarrow \mathsf{prog}_x(R) \equiv \llbracket f \rrbracket\,(R, \mathsf{prog}_{x_1},\cdots,\mathsf{prog}_{x_n}) \\
\llbracket x := \mathtt{phi}(x_1,\ldots,x_n) \rrbracket & \rightarrow \mathsf{prog}_x(R) \equiv \mathsf{prog}_{x_1}(R) \vee\cdots\vee \mathsf{prog}_{x_n}(R)
\end{array}$$

Figure 8: Encoding Instructions

positions removed from S. The reason why we need to compute the set V of the concrete strings is that our analysis is designed to be conservative. (As will be discussed later, `indexOf` and `contains` pose a similar requirement.) Consider the following Java code fragment:

```
String t = some_condition ? "a" : "b";
String u = "ab".replace(t,"z");
```

where `t`'s value is either “*a*” or “*b*”. However, replacing “a” and “b” in the string “ab” with “z” yields “zz”, while the actual result should be either “az” or “zb”.

In practice, a separate analysis technique (*e.g.*, constant propagation [33]) can be used to obtain the set of concrete values corresponding to the relevant strings, which are then reflected in the abstraction. We emphasize that other string-analysis techniques also require this information. For example, both Minamide's algorithm [23] and JSA [8] approximate the `replace("a","b")` operation with an automaton, but not the function `replace`.

Having reviewed our abstraction of `replace`, we note that in some cases, defining the abstraction of a built-in function using a finite-state transducer, as described in [23], is easier than directly constructing an M2L formula. We use a simple approach to translate a string transducer into an M2L predicate: Since a transducer can be viewed as a finite-state automaton with output characters, we can represent it using the regular-expression notation, while denoting tuples of input characters and output characters as described in [18]. For example, a transducer for `replace("a","b")` can be represented by $((\hat{\mathrm{a}}\mathrm{b})|(\hat{A}A))^\star$, where A matches any character, and the notation $\hat{c}o_1 \ldots o_n$ means that c is an input character followed by the output characters $o_1, \cdots, o_n$. We can then encode this regular expression using the algorithm in Figure 6.

3.4 Encoding Instructions

Encoding an instruction amounts to producing a set of M2L predicate declarations. Similar to the abstraction of a built-in function, we denote the encoding of instruction I by $\llbracket I \rrbracket$.

We say that a left-hand variable is *cyclic* if it is defined depending on itself. Such a cyclic variable can only appear in a program with loops (and recursions). If there are no cyclic variable definitions, then the abstractions in Figure 8 apply. The abstractions of return instructions, calls to user-defined functions, goto instructions, and jump instructions are omitted, since this section assumes an intraprocedural and path-insensitive analysis for simplicity.

When a cyclic variable is affected only by string concatenations, we use the approach of [8]: We first construct a Context-Free Grammar (CFG) for the cyclic variable, where cyclic variables are considered nonterminal symbols, and then approximate the resulting CFG with a regular expression. Finally, we encode the regular expression in M2L using the algorithm shown in Figure 6.

As an example, consider the following loop program and its corresponding SSA program containing cyclic variables `v2` and `v3`:

```
String v0 = "ab";               1:v0 = "ab"
String v1 =                       v1 = toUpperCase(v0)
    v0.toUpperCase();           2:v2 = phi(1:v0,3:v3)
String v2 = v0;                   i0 = length(v2)
while (v2.length()<10) {          b0 = op(<)(i0,10)
  v2 = v2.concat(v1);             b1 = op(neg)(b0)
}                                 jump b1, 4
 ...                            3:v3 = concat(v2,v1)
                                  goto 2
                                4: ...
```

The possible set of strings assigned to `v2` can be represented by CFG $v_2 \rightarrow \mathtt{v0} \mid v_3, v_3 \rightarrow v_2\ \mathtt{v1}$, where `v0` and `v1` are considered terminal symbols (being non-cyclic). This CFG is overapproximated by the regular expression $\mathtt{v0}\ \mathtt{v1}^\star$, which is encoded in M2L as $\langle\!\langle \mathtt{v0}\ \mathtt{v1}^\star \rangle\!\rangle$. Here are the resulting predicate declarations:

$$\begin{array}{l}
\mathsf{prog}_{\mathrm{v0}}(R) \equiv \text{“ab”}(R) \\
\mathsf{prog}_{\mathrm{v1}}(R) \equiv \llbracket \mathtt{toUpperCase} \rrbracket\,(R, \mathsf{prog}_{\mathrm{v0}}) \\
\mathsf{prog}_{\mathrm{v2}}(R) \equiv \langle\!\langle \mathtt{v0}\ \mathtt{v1}^\star \rangle\!\rangle\,(R) \\
\mathsf{prog}_{\mathrm{v3}}(R) \equiv \exists V_1, V_2 . \mathsf{prog}_{\mathrm{v2}}(V_2) \wedge \mathsf{prog}_{\mathrm{v1}}(V_1) \\
\qquad\qquad\qquad \wedge \mathsf{concat}(R, V_2, V_1)
\end{array}$$

When cyclic variable x is affected by other string operations, we use the naïve abstraction $\mathsf{prog}_x(R) \equiv \mathsf{true}$, which holds for any string value.

Alternatively, we could automatically find a loop invariant inv_x, for x, using the character-set abstraction [8], where a position set variable is used as a set of characters by ignoring the order of characters in a string. This predicate can be used instead of prog_x. Note that the *strongest* loop invariant on character sets, in the sense of the *smallest* character set that satisfies cyclic string constraints, is necessary, since our purpose is to check the existence of an unsafe string. Otherwise, the naïve abstraction is allowed to be a loop invariant. In addition, finding the smallest character set requires the subset relation between character sets, whereby the equality of characters is also required. Our experience suggests, however, that computing the smallest character set, which involves checking the equality of characters, is an expensive process whose advantage over the naïve approach is negligible.

3.5 Soundness of the Encoding Method

Here, we describe soundness of our encoding method. Note that only loop-free programs are addressed, since we rely on the grammar-based abstraction by [8, 23] and the trivial widening operation that always yields the naïve abstraction $\mathsf{prog}_x(S) = \mathsf{true}$ to handle loops.

We first introduce the notation $L_{w,\mathcal{I}}(\psi)$ to denote the set of strings represented by R that satisfy $\psi(R)$ given finite string w and assignment $\mathcal{I}$. [3]

DEFINITION 1 (GENERATED LANGUAGE).

$$L_{w,\mathcal{I}}(\psi) \quad \equiv \quad \{s \mid w, \mathcal{I} \models \forall R \,.\, \text{“}s\text{”}(R) \Rightarrow \psi(R)\}$$

The set $\llbracket P \rrbracket$ of predicate declarations, which is obtained by encoding the program P, is sound if for every program variable x, there exists a finite string w and an assignment $\mathcal{I}$, such that $v \in L_{w,\mathcal{I}}(\mathsf{prog}_x)$, where v is a string value assigned to variable x and $\llbracket P \rrbracket = \{\mathsf{prog}_{x_1}, \cdots, \mathsf{prog}_{x_n}\}$.

THEOREM 1 (SOUNDNESS OF THE ENCODING METHOD).

$$\exists w, \mathcal{I} \,.\, \forall x \in \mathit{dom}(\sigma) \,.\, \sigma(x) \in L_{w,\mathcal{I}}(\mathsf{prog}_x)$$

where σ represents a program state (interpreted as a mapping from program variables to values).

The above soundness criterion holds as long as for every string operation f, the abstraction $\llbracket f \rrbracket$ satisfies the following condition:

DEFINITION 2 (SOUND ABSTRACTION OF FUNCTION f).

$$\begin{array}{l}
\forall r, p_1, \cdots, p_n \,.\, r = f(p_1, \cdots, p_n) \\
\Rightarrow \forall w, \mathcal{I}, \psi_1, \cdots, \psi_n \,.\, p_1 \in L_{w,\mathcal{I}}(\psi_1) \wedge \cdots \wedge p_n \in L_{w,\mathcal{I}}(\psi_n) \\
\qquad \Rightarrow \exists w' \,.\, r \in L_{ww',\mathcal{I}}(\lambda R. \llbracket f \rrbracket\,(R, \psi_1, \cdots, \psi_n))
\end{array}$$

where ww' is the concatenation of finite strings w and w', and $L_{w,\mathcal{I}}(\lambda R.\psi(R))$ is short for $\{s \mid w, \mathcal{I} \models \forall R.\text{“}s\text{”}(R) \Rightarrow \psi(R)\}$.

[3] This definition can be viewed as a concretization function in abstract interpretation, where there is no best abstraction as in the case of regular languages [9], which is not a complete partial order.

$$\begin{array}{ll}\llbracket x:=n\rrbracket & \rightarrow\mathsf{prog}_x(p,S)\equiv\mathsf{pos}_n(p,S)\\ \llbracket x:=f(x_1,\cdots,x_n)\rrbracket & \rightarrow\mathsf{prog}_x(p,S)\equiv\llbracket f\rrbracket\,(p,S,\mathsf{prog}_{x_1},\cdots,\mathsf{prog}_{x_n})\\ \llbracket x:=\mathtt{phi}(x_1,\ldots,x_n)\rrbracket & \rightarrow\mathsf{prog}_x(p,S)\equiv\mathsf{prog}_{x_1}(p,S)\vee\cdots\vee\mathsf{prog}_{x_n}(p,S)\end{array}$$

Figure 9: Additional abstraction of instructions for indices

The details of the proof is described in Appendix B, where the proof is done by induction on a transition system that defines the semantics of the SSA program described in Appendix A.

4. INDEX- AND PATH-SENSITIVITY

This section describes how to augment the core algorithm with the index sensitivity and the path sensitivity.

4.1 Handling String Indices

An index is encoded as a position and position-set pair. For example, if string "ace" is encoded as position set $\{0, 2, 4\}$ in M2L, then index 1 into it is encoded as the pair $(2, \{0, 2, 4\})$. More generally, we introduce the following M2L predicates to represent indices: $\mathsf{pos}_0(p, S) \equiv (p = \mathsf{min}(S)), \cdots, \mathsf{pos}_n(p, S) \equiv \mathsf{pos}_{n-1}(p, S \setminus \mathsf{min}(S))$, where $\mathsf{pos}_n(p, S)$ means that position p in position-set S represents index n into a string represented by S.

When encoding instructions, PISA accounts for indices following the rules in Figure 9, where a predicate prog_x takes a position and position-set pair as its arguments. Those rules apply when the left-hand-side variable in an instruction assumes a value representing an index into a string. Note that any numerical expression of the form $n + N$ can be encoded, where n is a variable and N is a constant. However, since M2L cannot directly encode numerical expressions of the form $n + m$, where m is also a variable, for such expressions PISA over-approximate it by $\mathsf{pos}_{\mathsf{any}}(p, S) = p \in S$, which represents an arbitrary index into a string represented by S. This same encoding is also used for a cyclic variable.

With index sensitivity at its disposal, PISA can model string operations such as `indexOf` and `substring`, where `indexOf(s1,s2)` returns the first index in `s1` at which `s2` occurs, whereas `substring(s,n,m)` extract from `s` the substring ranging between indices `n` and `m`. These methods are abstracted as shown in Figure 7. The first and second parameters in the $\llbracket$`indexOf`$\rrbracket$ formula represent a position and a string containing it, respectively, while V is the set of concrete values possibly assigned to `s2` in `indexOf(s1,s2)`. Intuitively, $\mathsf{indexOf}(p, P, Q)$ holds if Q is a substring of P starting at the index represented by (p, P). By requiring $p \leq p'$, we choose the minimal index among the candidates that satisfy $\mathsf{indexOf}(p, P, Q)$. In addition, due to the restriction about the minimal index, we need a set V of concrete string values possibly assigned to `s2` of `indexOf(s1,s2)`. It should be obtained by another analysis as in the case of $\llbracket$`replace`$\rrbracket$ [4]. Otherwise, $\llbracket$`indexOf`$\rrbracket$ involves the same problem as $\llbracket$`replace`$\rrbracket$.

As an example, consider the following SSA program fragment:

```
v0 = 0;  v1 = "a<b";  v2 = "<";
v3 = indexOf(v1,v2);  v4 = substring(v1,v0,v3);
```

We obtain the following set of predicates after expanding the definitions of $\llbracket$`indexOf`$\rrbracket$ and $\llbracket$`substring`$\rrbracket$:

$$\mathsf{prog}_{\mathtt{v0}}(n, S) \equiv \mathsf{pos}_0(n, S) \qquad \mathsf{prog}_{\mathtt{v1}}(R) \equiv \text{“}\mathtt{a<b}\text{”}(R)$$
$$\mathsf{prog}_{\mathtt{v2}}(R) \equiv \text{“}<\text{”}(R)$$
$$\begin{aligned}\mathsf{prog}_{\mathtt{v3}}(p, P) \equiv{}& \mathsf{prog}_{\mathtt{v1}}(P) \wedge (\exists P_2 . \text{“}<\text{”}(P_2) \wedge \mathsf{indexOf}(p, P, P_2))\\ &\wedge(\mathsf{min}(P) \leq p \Rightarrow (\forall P_2, p' . \text{“}<\text{”}(P_2) \wedge \mathsf{indexOf}(p', P, P_2)\\ &\qquad\wedge \mathsf{min}(P) \leq p' \Rightarrow p \leq p'))\end{aligned}$$
$$\begin{aligned}\mathsf{prog}_{\mathtt{v4}}(R) \equiv{}& \exists V_1, v_0, v_3 . \mathsf{prog}_{\mathtt{v1}}(V_1)\\ &\wedge \mathsf{prog}_{\mathtt{v0}}(v_0, V_1) \wedge \mathsf{prog}_{\mathtt{v3}}(v_3, V_1) \wedge \mathsf{substrr}(R, V_1, v_0, v_3)\end{aligned}$$

4.2 Handling Branch Conditions

PISA employs a simple form of path sensitivity, which provides the ability to record the effects of branch conditions on a specific variable by encoding them as M2L predicates. For example, branch-condition `v.equals("a")` constrains variable `v`. Thus, if the test succeeds, we encode the relevant constraint as M2L-predicate $\phi_v(R) = \text{“a”}(R)$. When encoding the true branch of a condition as a set of predicate declarations, we use predicate $\mathsf{prog}'_v(R) = \mathsf{prog}_v(R) \wedge \phi_v(R)$, rather than $\mathsf{prog}_v(R)$, to represent the values possibly assigned to program-variable v.

Figures 10 and 11 present the encoding for path-sensitive analysis, where the notation $\llbracket I\rrbracket^b$ represents encoding instruction I in basic block b. Boolean operators $\wedge$ and $\vee$ are used for notational brevity: For predicates ψ_1 and ψ_2, $\psi_1 \wedge \psi_2$ $[\psi_1 \vee \psi_2]$ represents a predicate ψ such that $\psi(R) = \psi_1(R) \wedge \psi_2(R)$ $[\psi(R) = \psi_1(R) \vee \psi_2(R)]$. In addition, we lift the lambda notation $\lambda R.\psi'$ to represent a predicate ψ such that $\psi(R) = \psi'$, without explicitly declaring predicate ψ.

Path-condition [27, 15] $PC(b_0, b')$ represents a necessary condition for flow from basic block b_0 to b'. We use the notation $PC'(b_0, b, b')$ to distinguish between ϕ-induced assignments, and thus represent a necessary condition for flow from b_0 to b' through b, which is an immediate predecessor of b'. Conditions are formed using Boolean program variables and logical operators.

Figure 11 describes our encoding of the effects of path conditions. Given variable v and basic blocks b and b', such that b is an immediate predecessor of b', we define $C(v, b, b')$. M2L predicate $C(v, b, b')$ represents a necessary condition for v to cause the transition from b to b'. The definition uses function C', which syntactically and recursively transforms the path condition into an M2L predicate. In the figure, $\llbracket f, m\rrbracket\,(t, R, \mathsf{prog}_{v_1}, \ldots, \mathsf{prog}_{v_n})$ represents a predicate restricting R *via* a necessary condition on the m-th parameter of Boolean method f, when f returns t. Such an abstraction should be predefined for each Boolean method, as in the case of built-in functions. In the absence of an abstraction, we default to true. $\mathsf{def}(v')$ represents an instruction, $v' = f(v_1, \cdots, v_n)$, which defines program variable v'.

$$\begin{array}{l}\llbracket x := v\rrbracket^b \rightarrow \mathsf{prog}_x(R)\equiv\text{“}v\text{”}(R)\\ \llbracket x := f(x_1,\cdots,x_n)\rrbracket^b\\ \;\rightarrow \mathsf{prog}_x(R)\equiv\llbracket f\rrbracket\,(R,\mathsf{prog}_{x_1}\wedge C(x_1,b),\cdots,\mathsf{prog}_{x_n}\wedge C(x_n,b))\\ \llbracket x := \mathtt{phi}(b_1 : x_1,\ldots,b_n : x_n)\rrbracket^b\\ \;\rightarrow \mathsf{prog}_x(R)\equiv\bigvee_{i\in\{1,\cdots,n\}}\mathsf{prog}_{x_i}(R)\wedge C(x_i,b_i,b)\end{array}$$

Figure 10: Abstraction for path-sensitive string analysis

$$C(v, b, b') \equiv C'(\mathsf{true}, v, PC'(b_0, b, b'))$$
$$C'(t, v, c) \equiv$$
$$\begin{cases}\lambda R\;.\;\llbracket f, m\rrbracket\,(t, R, \mathsf{prog}_{v_1}, \ldots, \mathsf{prog}_{v_n}) & \\ \qquad\qquad \text{when } c = f(v_1, \ldots, v_n)\text{, and } v = v_m & \\ C'(t, v, \mathsf{def}(v')) \;\; \text{when } c = v'\text{, where } v' \text{ is a program variable} & \\ C'(\mathsf{true}, v, c') \quad \text{when } c = \neg c' \text{ and } t = \mathsf{false} & \\ C'(\mathsf{false}, v, c') \quad \text{when } c = \neg c' \text{ and } t = \mathsf{true} & \\ C'(\mathsf{true}, v, c_1) \vee C'(\mathsf{true}, v, c_2) & \text{when } c = c_1 \vee c_2 \text{ and } t = \mathsf{true}\\ C'(\mathsf{true}, v, c_1) \wedge C'(\mathsf{true}, v, c_2) & \text{when } c = c_1 \wedge c_2 \text{ and } t = \mathsf{true}\\ C'(\mathsf{false}, v, c_1) \wedge C'(\mathsf{false}, v, c_2) & \text{when } c = c_1 \vee c_2 \text{ and } t = \mathsf{false}\\ C'(\mathsf{false}, v, c_1) \vee C'(\mathsf{false}, v, c_2) & \text{when } c = c_1 \wedge c_2 \text{ and } t = \mathsf{false}\\ \lambda R\;.\;\mathsf{true} & \text{otherwise}\end{cases}$$

Figure 11: Constraint on program-variable v when the execution transitions from basic block b to basic block b'

[4] If the minimal index is not required, the abstraction becomes simpler so as not to require the concrete strings, but makes the analysis too conservative to check the existence of unsafe characters.

```
void   f1(String v1) { String s1 = f3(v1); }
void   f2(String v2) { String s2 = f3(v2); }
String f3(String v3) { return v3; }
```

Figure 12: Sample Java Program for Interprocedural Analysis

Figure 10 uses $C(v, b, b')$ to encode ϕ instructions. The variables used by the ϕ statement are each constrained by taking the relevant immediate predecessor of b' into account. For the other instructions, we do not need to consider the immediate predecessor. Thus, we simply use the notation $C(x, b')$, which is equivalent to $\bigvee_{b \in \mathsf{pred}(b')} C(x, b, b')$, where $\mathsf{pred}(b')$ represents a set of immediate predecessors of basic block b'.

With path-sensitivity, we can abstract `contains`, as described in Figure 7. In $[\![\texttt{contains},1]\!]$, S is the set of concrete strings assigned to v_2. Note that $\neg(\exists P\ .\ \mathsf{prog}_{v_2}(P) \wedge \mathsf{substr}(P, R))$ cannot be used when $c = \mathsf{false}$ since the analysis is conservative. As an example, consider the following Java method, along with its corresponding SSA representation, where `op(or)` and `op(neg)` represent logical disjunction and negation, respectively:

```
String clean(String s) {  1: v1 = "<"; v2 = ">";
 if (s.contains("<") ||      b1 = contains(s,v1);
     s.contains(">")) {      b2 = contains(s,v2);
   s = "x";                  z0 = op(or)(b1,b2);
 }                           z1 = op(neg)(z0);
 return s;                   jump z1,3;
}                         2: v4 = "x";
                          3: v5 = phi(1:s,2:v4);
                             return v5;
```

For this method, we obtain the path conditions: $PC'(1,1,2) = \texttt{b1}\vee\texttt{b2}$, $PC'(1,1,3) = \neg(\texttt{b1}\vee\texttt{b2})$, and $PC'(1,2,3) = \texttt{b1}\vee\texttt{b2}$. The resulting predicates are:

$$\begin{aligned}
&C(\texttt{v4},2,3) = C'(\mathsf{true},\texttt{v4},PC'(1,2,3)) = \lambda R\ .\ \mathsf{true}\\
&C(\texttt{s},1,3) = C'(\mathsf{true},\texttt{s},PC'(1,1,3))\\
&\ = \lambda R\ .\ [\![\texttt{contains},1]\!]\,(R,\mathsf{false},\mathsf{prog}_{\texttt{s}},\mathsf{prog}_{\texttt{v1}})\\
&\ \ \wedge\lambda R\ .\ [\![\texttt{contains},1]\!]\,(R,\mathsf{false},\mathsf{prog}_{\texttt{s}},\mathsf{prog}_{\texttt{v2}})\\
&\ = \lambda R\ .\ \neg(\exists P\ .\ \text{“<”}(P) \wedge \mathsf{substr}(P,R))\\
&\ \ \wedge\lambda R\ .\ \neg(\exists P\ .\ \text{“>”}(P) \wedge \mathsf{substr}(P,R)).
\end{aligned}$$

5. INTERPROCEDURAL ANALYSIS

Our interprocedural version of the string analysis relies on a callgraph, where each node of the callgraph contains a set of instructions in the SSA form and these instructions are translated into a set of M2L predicate declarations using our encoding method. The relationships among callgraph nodes are used to obtain possible assignment relationships between caller's program variables and callee's program variables (parameters and return variables). The assignment relationships are encoded in M2L as if those are assignment instructions.

Let us consider a context-insensitive callgraph for three Java's methods shown in Figure 12, where the callgraph has three nodes n_1, n_2, and n_3 for the methods `f1`, `f2`, and `f3`, respectively. We obtain the four assignment relationships `v3=v1`, `v3=v2`, `s1=v3`, `s2=v3`. These assignment relationships are then encoded in M2L using the encoding method described in Section 3.4, where, due to two possible assignment to `v3`, the first two relationships can be encoded in the same way to encode the ϕ-instruction `v3=phi(v1,v2)`.

Note that this approach simply ignores the call stack. Therefore, context-sensitivity of the callgraph affects the precision of the string analysis.

6. IMPLEMENTATION AND EVALUATION

We applied PISA to a production-level taint-analysis engine for the purpose of automatic detection of *user-defined sanitizers* (sanitizers defined in application code). In this section, we describe our sanitizer-detection algorithm, and we then present the implementation of that algorithm. We then discuss two sets of evaluations: In the first set, we investigate how many user-defined sanitizers in the application code of 8 open-source benchmarks PISA was able to detect. In this first set of evaluations, we used two variants of PISA, and we compared PISA to alternative algorithms. In the second set, we examine the overall impact of PISA on a production-level taint-analysis algorithm.

```
void detectSanitizers(
       Set<Method> M, Set<Pattern> P,  // input
       Set<Pair<CGNode,Pattern>> R) {  // output
  CallGraph cg = callgraphOf(M);
  for (Method m : M) {
    Set<CGNode> N = nodes(cg, m);
    for (CGNode n : N) {
      Set<Instruction> I = instructionsOf(n, cg);
      Set<Variable> V = returnVariablesOf(n);
      for (Pattern p : P) {
        boolean r = doStringAnalysis(I, V, p);
        if (r) R.add(new Pair(n, p)); } } } }
```

Figure 13: Outline of the Sanitizer-detection Algorithm

6.1 Sanitizer Detection

Our algorithm for detecting sanitizers consists of two steps. The first step is to find sanitizer candidates based on a syntactic check: An input pattern ranging over method signatures is used to focus the analysis on methods that are likely to act as sanitizers (*e.g.*, methods accepting a single `String` argument and returning a `String` object). Figure 13 shows the algorithm of the second step. This phase consumes the set `M` of candidate methods and the set `P` of unsafe string patterns, and outputs the set `R` of the pairs of the form `(n,p)` such that the method corresponding to callgraph node `n` is a sanitizer for unsafe pattern `p`, where each pattern is used to categorize detected sanitizers. Note that sanitizer categorization is essential for taint analysis, since a method is typically a sanitizer only for certain types of attack. The procedure comprises the following steps:

1. `callgraphOf(M)` builds the callgraph `cg` rooted at the set `M` of sanitizer candidates. Any callgraph-construction algorithm can be chosen to build `cg` (*e.g.*, context insensitive or context sensitive, with various levels of context sensitivity [14, 13]).
2. For each sanitizer candidate, we obtain the set of callgraph nodes representing it.
3. For each node `n`, a set `I` of instructions is then computed using function `instructionsOf(n)`, which returns the set of instructions in n and in all the nodes transitively called by `n`. In addition, for each caller/callee pair, assignment relationships between actual parameters in the caller node and formal parameters in the callee node are included in `I`. This allows us to perform the interprocedural analysis described in Section 5.
4. For each unsafe pattern `p`, return variables of node `n` are extracted by `returnVariablesOf(n)`, and then verified by the string analysis, `doStringAnalysis(I,V,p)`, where the string analysis returns `true` iff the set of potential string values taken by the return variables V of node n never contains the unsafe pattern p. If the string analysis returns `true`, then `n` is reported as a sanitizer for the unsafe string pattern `p`.

6.2 Implementation

We implemented the algorithm described in Section 6.1 as a Java program using the T. J. Watson Libraries for Analysis (WALA) framework [31], which calls MONA [16] as a command-line program to check the satisfiability of M2L formulae. In addition, the implementation embodies the following *sound* optimizations:

Table 1: Statistics on benchmark applications

App. name	Version	Classes	LOC	Candidates
SBM	1.08	143	5,541	15
Blojsom	3.1	255	13,967	51
PersonalBlog	1.2.6	69	5,317	7
Roller	0.9.9	283	41,589	56
SnipSnap	1.0-BETA-1	614	46,962	52
Webgoat	5.1	193	33,906	43
JSPWiki	2.6	503	81,301	91
MVNForum	1.0.2	820	142,954	56
Total				371

- String constants whose length exceeds fixed-size n are over-approximated by a disjunctive regular expression that ranges over a partitioning of the original string, where each partition (except, maybe, the suffix) is of size n. Our experiments used $n = 5$. For example, "`longstringvalue`" is over-approximated by regular expression "`(longs|tring|value)+`".
- The verification of each method has a 30-second time limit, since verifying all of the methods within a reasonable time is more important than verifying a particular method for a long time. If the limit is reached without having concluded the analysis, we conservatively over-approximate the behavior of the method by concluding that it is not a sanitizer.
- We did not define the last-position term , \$, of M2L(Str), when simulating M2L(Str) in MONA, since it was not required by our encoding method.
- The path-condition analysis is run only if the SSA form of the target method consists of less than 50 basic blocks. Otherwise, the analysis is still sound because it conservatively becomes path-insensitive.

6.3 Evaluation

We ran our experiments on top of the Sun Java Runtime Environment (JRE) V1.6.0_06 with 1 GB of maximum heap size using a Lenovo ThinkPad T61p with a Core2 Duo T7800 2.6GHz CPU and 3 GB of RAM. Statistics on the benchmark applications we used are shown in Table 1. The Candidates column reflects the number of methods in application code (*i.e.*, ignoring imported libraries and JUnit test classes) that accept a single parameter of type `String` and return a `String` value. This is the criterion we used for identifying sanitizer candidates. (Note that in some cases, methods that were not intended to be used as sanitizers may be accepted by both the criterion we just described and the ensuing analysis.)

Our study focused on the following four types of attack: XSS, HRS, Log Forging (LOG) and Path Traversal (PATH) [2]. The corresponding regular-expression patterns used in the specification are "`.*[<>].*`" (HTML tags), "`.*[\r\n].*`" (strings of multiple lines), "`.*[\r\n\x08].*`" (strings of multiple lines and backspace) and "`.*\.\./.*`" (strings representing relative paths), respectively.

6.3.1 *Sanitizer Detection*

Our experiment on sanitizer detection comprised four configurations: The first two, PISA/CI and PISA/CS, both embody the PISA algorithm, the difference between them being that the former is based on a context-insensitive (0-CFA) callgraph, whereas the latter relies on a context-sensitive (1-CFA) callgraph [14, 13]. Recall from Section 5 that PISA depends on its underlying callgraph in various ways, and thus the overall accuracy of PISA derives, in part, from the precision of its supporting callgraph. The two remaining candidates are PSA/CI, which is a variant of PISA/CI employing only path sensitivity (and thus lacking the abstract models for `indexOf` and `lastIndexOf`), and the CFG-based string analysis of [12], which is based on [23]; this analysis is neither path- nor index-sensitive, but can handle cyclic variables by computing invariants. This computation is done by iterating the grammar transduction and by using the character-set approximation [8, 23] as a widening operation.

Table 2:Accuracy results of the sanitizer-detection experiment

		XSS			HRS			LOG			PATH			Total		
		TP	FN	Score	TP	FN	Score	TP	FN	Score	TP	FN	Score	TP	FN	Score
SBM	CFG	0	3	0	0	3	0	0	3	0	0	3	0	0	12	0
	PSA/CI	3	0	100	3	0	100	3	0	100	3	0	100	12	0	100
	PISA/CI	3	0	100	3	0	100	3	0	100	3	0	100	12	0	100
	PISA/CS	3	0	100	3	0	100	3	0	100	3	0	100	12	0	100
Blojsom	CFG	0	8	0	1	2	33	0	2	0	1	4	20	2	16	11
	PSA/CI	0	8	0	1	2	33	0	2	0	1	4	20	2	16	11
	PISA/CI	0	8	0	1	2	33	0	2	0	2	3	40	3	15	17
	PISA/CS	0	8	0	1	2	33	0	2	0	2	3	40	3	15	17
PersonalBlog	CFG	0	4	0	0	3	0	0	3	0	0	3	0	0	13	0
	PSA/CI	1	3	25	0	3	0	0	3	0	0	3	0	1	12	8
	PISA/CI	1	3	25	0	3	0	0	3	0	0	3	0	1	12	8
	PISA/CS	1	3	25	0	3	0	0	3	0	0	3	0	1	12	8
Roller	CFG	2	10	17	2	3	40	2	3	40	2	3	40	8	19	30
	PSA/CI	4	8	33	2	3	40	2	3	40	2	3	40	10	17	37
	PISA/CI	4	8	33	2	3	40	2	3	40	2	3	40	10	17	37
	PISA/CS	4	8	33	2	3	40	2	3	40	2	3	40	10	17	37
SnipSnap	CFG	1	3	25	1	3	25	1	3	25	1	7	13	4	16	20
	PSA/CI	4	0	100	4	0	100	4	0	100	4	4	50	16	4	80
	PISA/CI	4	0	100	4	0	100	4	0	100	5	3	63	17	3	85
	PISA/CS	4	0	100	4	0	100	4	0	100	5	3	63	17	3	85
Webgoat	CFG	5	16	24	7	9	44	5	8	38	6	9	40	23	42	35
	PSA/CI	14	7	67	7	9	44	5	8	38	6	9	40	32	33	49
	PISA/CI	14	7	67	7	9	44	5	8	38	8	7	53	34	31	52
	PISA/CS	14	7	67	9	7	56	7	6	54	10	5	67	40	25	62
JSPWiki	CFG	5			5			5			6			21		
	PSA/CI	6			6			6			6			24		
	PISA/CI	6			6			6			6			24		
	PISA/CS	8			8			8			8			32		
MVNForum	CFG	3			4			3			3			13		
	PSA/CI	3			4			3			3			13		
	PISA/CI	3			4			3			3			13		
	PISA/CS	3			4			3			3			13		
Total	CFG	16			20			16			19			71		
	PSA/CI	35			27			23			25			110		
	PISA/CI	35			27			23			29			114		
	PISA/CS	37			31			27			33			128		

Table 2 shows how many sanitizers were detected by PISA on the 8 benchmark applications in comparison with the CFG-based string analyzer. If a method automatically detected by PISA is a true sanitizer, the method is counted as a true positive (TP). Otherwise, it is a false positive (FP). If PISA fails to detect a true sanitizer, that result is counted as a false negative (FN). The score is calculated as $100 \times \mathrm{TP}/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$. Note that the false negatives are counted only for the relatively small 6 benchmark applications, since detecting false negatives requires manually reviewing all of the candidates to find the true sanitizers that are not detected by PISA—a very time-consuming and error-prone operation. In addition, Table 2 shows only true positives and false negatives, since the sanitizer detection reported no false positives. Note that false positives in the sanitizer detection are caused only when the string analysis is not conservative and fails to infer unsafe strings that arise at runtime.

Table 3 shows running time and the statistics reported by MONA. The Abort column shows the total time spent analyzing candidates that were aborted due to either the time limit we set in our analysis or a size limit in MONA. The number of aborted candidates is shown in parentheses. For PISA, the Enc and Ver columns show the time spent on translating the callgraph into a MONA program and the running time of MONA, respectively. For the CFG-based analysis, the Enc column shows the time spent on translating the callgraph into a set of production rules, and the Ver column shows the total time spent on inference and containment checks. The MONA Statistics column shows summaries of the statistics reported by MONA. The Pred and DAG columns show the average/maximum numbers of generated predicates and DAGs (graph representations of formulae), respectively. The State and BDD columns show the average and maximum numbers of elements in the largest sets of states and Binary Decision Diagram (BDD) nodes of minimized

```
static final String PUNCTUATION_CHARS_ALLOWED
                                = " ()&+,-=._$";
static String cleanLink(String link){
  return cleanLink(link, PUNCTUATION_CHARS_ALLOWED);
}
static String cleanLink(String link,
                        String allowedChars){
  if (link == null) return null;
  link = link.trim();
  StringBuffer clean=new StringBuffer(link.length());
  boolean isWord = true; boolean wasSpace = false;
  for (int i = 0; i < link.length(); i++){
    char ch = link.charAt(i);
    if (Character.isWhitespace(ch)) {
      if (wasSpace) continue;
      wasSpace = true;
    } else { wasSpace = false; }
    if (Character.isLetterOrDigit(ch)||
        allowedChars.indexOf(ch) != -1) {
      if (isWord) ch = Character.toUpperCase(ch);
      clean.append(ch); isWord = false;
    } else { isWord = true; }
  }
  return clean.toString();
}
```

```
private static String getFileName(String s) {
 String fileName = new File(s).getName();
 if (fileName.indexOf("/") != -1) {
  fileName =
   fileName.substring(fileName.lastIndexOf("/"),
                      fileName.length());}
 if (fileName.indexOf(".") != -1) {
  fileName =
   fileName.substring(0,fileName.indexOf("."));}
 return fileName;
}

public String getCcnParameter(String name){
  return getRegexParameter(name, "\\d{16}");
}
public String getPhoneParameter(String name){
  return getRegexParameter(str, "[\\d\\s-]+");
}
private String getRegexParameter(String name,
                                 String regexp) {
  String param = getStringParameter(name,regexp);
  if (Pattern.matches(str,regexp)) return str;
  else return "";
}
```

Figure 14: Sanitizers detected by PISA

Table 3: Running Times and MONA Statistics

		Time(sec)				MONA Statistics							
						Average				Maximum			
		Abort	Enc	Ver	Total	Pred	DAG	State	BDD	Pred	DAG	State	BDD
SBM	CFG	0.0(0)	0.1	0.3	0.4								
	PSA/CI	0.0(0)	2.0	6.9	8.9	32	278	187	2,316	134	748	923	11,352
	PISA/CI	0.0(0)	2.0	6.7	8.7	32	278	187	2,316	134	748	923	11,352
	PISA/CS	0.0(0)	2.0	6.8	8.8	32	278	187	2,316	134	748	923	11,352
Blojsom	CFG	0.0(0)	1.8	7.9	9.7								
	PSA/CI	30.3(1)	4.1	17.9	52.3	18	197	49	481	54	615	923	10,679
	PISA/CI	30.2(1)	4.2	18.3	52.7	18	204	49	473	71	841	922	10,647
	PISA/CS	120.3(4)	89.1	19.6	229.0	35	255	55	542	370	1,119	922	10,647
Personal-Blog	CFG	0.0(0)	0.3	0.2	0.5								
	PSA/CI	0.0(0)	0.9	1.9	2.8	23	215	32	235	77	519	129	1,182
	PISA/CI	0.0(0)	0.9	1.9	2.8	23	215	32	235	77	519	129	1,182
	PISA/CS	0.0(0)	0.7	2.0	2.7	23	215	32	235	77	519	129	1,182
Roller	CFG	0.0(0)	4.1	4.5	8.6								
	PSA/CI	0.0(0)	13.2	32.4	45.6	17	222	205	2,912	68	983	8,193	124,926
	PISA/CI	0.0(0)	13.1	33.0	46.1	17	224	205	2,912	68	983	8,193	124,926
	PISA/CS	0.0(0)	18.0	31.4	49.4	17	221	206	2,956	59	594	8,193	124,926
SnipSnap	CFG	0.0(0)	1.8	3.3	5.1								
	PSA/CI	0.0(0)	9.7	20.5	30.2	25	234	129	1,411	126	679	1,908	22,571
	PISA/CI	0.0(0)	9.9	21.0	30.9	25	247	128	1,403	129	679	1,908	22,571
	PISA/CS	0.0(0)	21.2	22.6	43.8	30	262	145	1,601	241	865	1,908	22,571
Webgoat	CFG	0.0(0)	1.4	17.1	18.5								
	PSA/CI	0.0(0)	8.5	33.7	42.2	91	420	75	814	683	1,893	633	8,002
	PISA/CI	0.0(0)	9.2	34.3	43.5	92	434	75	821	684	1,893	633	8,002
	PISA/CS	0.0(0)	8.4	32.7	41.1	72	407	73	803	596	2,401	633	8,002
JSPWiki	CFG	0.0(0)	2.0	23.5	25.5								
	PSA/CI	80.5(10)	396.0	36.7	513.2	41	271	97	1,317	493	1,398	1,539	36,079
	PISA/CI	80.7(10)	394.8	36.7	512.2	41	272	97	1,318	493	1,398	1,539	36,079
	PISA/CS	152.5(9)	388.3	43.8	584.6	43	286	122	1,595	487	1,876	2,064	36,079
MVN-Forum	CFG	30.6(1)	1.3	92.0	123.9								
	PSA/CI	44.7(1)	11.4	59.6	115.7	25	229	89	1,423	124	980	1,406	50,894
	PISA/CI	64.7(2)	11.2	23.9	99.8	25	234	70	698	124	980	552	6,305
	PISA/CS	47.2(1)	18.5	59.5	125.2	29	253	94	1,525	144	1,058	1,406	50,894
Total	CFG	30.6(1)	12.8	148.8	192.2								
	PSA/CI	155.5(12)	445.8	209.6	810.9								
	PISA/CI	175.6(13)	445.3	175.8	796.7								
	PISA/CS	320.0(14)	546.2	218.4	1084.6								

automata, respectively. Note that the number of the predicates is almost equivalent to the number of instructions analyzed by PISA.

PISA/CI versus CFG-based String Analyzer. We summarize the results for the 8 Web applications in the Total row, showing that PISA/CI detected and categorized 114 sanitizers compared to 72 sanitizers detected by the CFG-based string analyzer.

Figure 14 shows several true sanitizers that were successfully detected by PISA, but not detected by the CFG-based string analyzer. Here, we consider method `cleanLink`, defined in class `MarkupParser` of JSPWiki, and method `getFileName`, defined in the class `Course` of Webgoat, mentioned above. The purpose of `cleanLink` is to keep only legal characters (letters, numbers, and characters specified by `PUNCTUATION_CHARS_ALLOWED`) in the link using the branch condition. Method `getFileName` checks for the existence of the illegal characters, and extracts the substring between "`/`" and "`.`" using the methods `substring`, `indexOf`, and `lastIndexOf`.

In terms of efficiency, in our experiments PISA/CI was almost 4.2 times slower than the CFG-based string analyzer. According to our observations, the reason is that PISA/CI has to deal with branch conditions as well as integer and string values. However, when only the verification time was considered, PISA/CI was only 1.4 times slower than the CFG-based string analyzer. Also, the CFG-string analyzer directly uses the automata given as the specification for checking the inferred CFGs, but PISA/CI generates M2L predicate declarations from the regular expressions, and MONA interprets those predicates each time PISA/CI verifies a sanitizer candidate.

PISA/CI versus PSA/CI. In Blojsom, SnipSnap, and Webgoat, PISA/CI detected 4 sanitizers for PATH vulnerabilities that were not detected by PSA/CI, compared to 6 sanitizers for PATH vulnerability that were not detected by the CFG-based string analyzer but detected by PSA/CI. We observed that index-based string operations were used for replacing or removing unsafe substrings for XSS and HRS, but the indices were calculated by loops and/or numerical expressions (*e.g.*, `n + m`, where `n` and `m` are variables) that cannot be encoded in M2L.

PISA/CI versus PISA/CS. For Webgoat and JSPWiki, some true-positive sanitizers detected by PISA/CS were not detected by PISA/CI. Overall, PISA/CS detected 128 sanitizers in the 8 applications. The method `getCcnParameter` of Figure 14 is the simplified version of `ParameterParser.getCcnParameter` of Webgoat, which was detected as a sanitizer by PISA/CS, but which was not be detected by PISA/CI. The context-sensitive call-graph can distinguish among the callers of `getRegexParameter`. Thus, PISA/CS can determine that the return values of `getRegexParameter` called by `getCcnParameter` matched only the regular expression "`\\d16`". This allowed PISA/CS to detect `getCcnParameter` as a sanitizer for HRS. In contrast, PISA/CI could not detect `getCcnParameter` as a sanitizer for HRS since `getRegexParameter` is called by both `getCcnParameter` and `getPhoneParameter` and PISA/CI determined that the return values of `getRegexParameter` matched either "`\\d16`" or "`[\\d\\s]+`". For the same reason, we had other true sanitizers that were not detected by PISA/CI, but were detected by PISA/CS.

```
final String entities[] = {"<", ">"};
final String refs[] = {"&lt;", "&gt;"};
String cleanByLoop(String s) {
  for (int i = 0; i < entities.length; i++)
    s = s.replaceAll(entities[i], refs[i]);
  return s;
}

String removeNonLetter(String str) {
  String ret = "";
  char[] cs = str.toCharArray();
  for (int i = 0; i < cs.length; i++)
    if (Character.isLetter(cs[i]))
      ret = ret + cs[i];
  return ret;
}
```

Figure 15: Sanitizers not detected by PISA

Limitations and False Negatives. The false negatives, which we found manually, were mainly caused by these limitations.

- The method `cleanByLoop` on Figure 15 should be a sanitizer for XSS since it never returns a string value containing < or >. However, neither PISA nor the CFG-based string analyzer can detect it as a sanitizer, even though the CFG-based string analyzer can handle the loop. This is because the resulting CFG inferred by the CFG-based string analyzer contains a string value that comes directly from the value of the parameter `s`. To solve this problem, we might have to unroll the loops while propagating the constant string values. Other examples of the same problem include these methods that were not detected as sanitizers: `ParameterParser.htmlEncode`, `Screen.convertMetachars`, and `HtmlEncoder.encode`.
- PISA's path-sensitivity relies on constraints on individual local variables that are directly checked by built-in Boolean functions. Due to this limitation, method `removeNonLetter` in Figure 15, which is similar to `Macros.removeNonAlphanumeric` in Roller, could not be detected as a sanitizer since PISA cannot determine that `cs[i]` used in the condition is the same as `cs[i]` used in the `then` block.
- We experimented only with 0-CFA and 1-CFA. However, we would need n-CFA (n>1) to detect some sanitizers in the benchmark applications. (E.g.: `Macros.escapeHTML` in Roller).
- Both PISA and the CFG-based string analyzer did not have a complete set of abstractions for built-in string operations (E.g.: `java.text.SimpleDateFormat.format` called by methods in PersonalBlog).

6.3.2 Integration with Taint Analysis

To gain insight on the impact of PISA on the overall precision of the security scanner, we integrated PISA/CS into the taint-analysis algorithm used by the IBM Rational AppScan [1]. Table 4 lists the results we obtained in terms of the number of vulnerable locations (call sites of the sinks) reported by the scanner.

For this study, we weakened our criterion for identifying sanitizer candidates by lifting the requirement that the method's input be a single string argument. (We used this more liberal criterion to guarantee that the candidate filter does not eliminate too many real sanitizers due to their signature.) Consequently, PISA/CS detected 2423 methods as sanitizers. We evaluated four different configurations, each corresponding to a particular combination of whether or not PISA is used and whether or not the set of predefined sanitizers provided as part of the AppScan algorithm is used.

Note that the richer the sanitizer specification is, the more accurate the taint analysis becomes. The four configurations we defined thus allow us to appreciate the effect of using only an automatically generated specification (configuration B) compared to the other three alternatives of (a) using no specification at all (configuration A), (b) using only a manual specification (as authored by a team of security experts), and (c) using the most complete specification we can obtain (configuration D). The numbers in Table 4 confirm that this effect is significant: While configuration A yields 651 reports, of which 136 are ruled out as false findings by configuration D, configuration B ruled out 42 reports overall, which represents 31% of the improvement gained by configuration D in elimination of false reports, and 38% of the improvement gained by configuration C. This means that thanks to PISA, the user can still boost the accuracy of the security tool considerably and in a fully automated fashion, without requiring any time or expertise from the development team. In particular, PISA can account for close to 40% of the benefit from the considerable effort invested by a security team devising a manual specification.

Table 4: Results of the taint-analysis-integration experiment

	w/o pre-defined sanitizers								w/ predefined sanitizers							
	w/o PISA (Configuration A)				w/ PISA (Configuration B)				w/o PISA (Configuration C)				w/ PISA (Configuration D)			
	XSS	HRS	LOG	PATH	XSS	HRS	LOG	PATH	XSS	HRS	LOG	PATH	XSS	HRS	LOG	PATH
SBM	118	4	0	4	115	4	0	4	118	1	0	4	115	1	0	4
Blojsom	1	5	97	14	1	4	94	10	1	4	84	14	1	4	81	10
PersonalBlog	1	0	3	0	1	0	3	0	1	0	3	0	1	0	3	0
Roller	5	0	9	0	4	0	9	0	3	0	9	0	2	0	9	0
SnipSnap	50	6	10	8	50	6	10	7	50	6	10	3	50	6	10	2
Webgoat	8	1	7	4	8	1	7	4	8	1	5	4	8	1	5	4
JSPWiki	27	35	91	19	27	15	88	19	25	14	78	10	25	14	75	7
MVNForum	108	10	2	4	102	10	2	4	70	9	2	2	64	9	2	2
Total	651				609				539				515			

7. RELATED WORK

Many string-analysis algorithms have been presented to date. Java String Analyzer (JSA) was first introduced by Christensen, *et al.* [8]. JSA approximates a string value by a regular language, allowing for statically checking errors in dynamically generated SQL queries. According to the online manual [7], the latest version of JSA has a form of path sensitivity through assertions, which is similar to ours. However, any experimental results of the path sensitivity has not yet been presented. Minamide [23] proposed approximating string values with a CFG and modeling built-in string operations using transducers to check the well-formedness of dynamically generated HTML documents. Wassermann and Su [32] extended Minamide's algorithm to syntactically isolate unsafe substrings from safe substrings in PHP programs. Their string analyzer can also account for regular-expression matches, but their paper does not mention how to deal with branch conditions that consist of string comparisons and Boolean operators as well as regular-expression matches. Yu, *et al.* [35] proposed another string- analysis algorithm for PHP, in which built-in string operations are modeled by the standard operations and a newly introduced *replacement* operation on automata. They also used the MONA's automaton package to implement these operations, and their approach was augmented with a backward analysis [34]. Kieżun, *et al.* [19] presented HAMPI, a string-constraint solver based on quantifier-free bit-vector logic. HAMPI does not have any syntax to explicitly mention positions. It is designed to treat fixed-size CFGs as well as regular grammars as constraints. Hooimeijer and Weimer [17] presented a decision procedure for solving constraints on regular languages, and applied the proposed decision procedure to infer input parameters that create SQL injection vulnerabilities. Fu, *et al.* [11] proposed another constraint solver based on a variation of the word equation, in which string-replacement operations were modeled using finite-state transducers. As a whole, none of these papers addressed the need for index-sensitivity, and only JSA has a form of path sensitivity.

There are a few papers that combine string analysis with symbolic execution. Bjørner, *et al.* [5] proposed to use word equations for checking the feasibility of paths generated by a dynamic symbolic-execution engine Pex [28], while handling the same kind of index sensitivity. In this work, the path constraints are checked by the SMT solver Z3 [36], which is also used by Rex [30]. Saxena, *et al.* [25] recently proposed Kudzu that is a symbolic execution engine combined with a string constraint solver covering bit-vector logic and word equation. However, in the core language models used by the above two techniques, no string-replacement functions, which is essential for making Web applications secure, were not statically modeled. Shannon, *et al.* [26] proposed to treat methods like `indexOf` in their symbolic execution engine by modeling convertion between symbolic strings and symbolic integers. Compared to their approach, our approach could be more precise, since it simultaneously treats string constraints and index constraints without any conversions.

As for sanitizer detection, Balzarotti, *et al.* proposed to use a string analysis for improving the accuracy of their sanitizer detection algorithm, but did not mention how to improve the string analysis algorithm itself except that they used taint propagation.

Various static taint analyses [22, 29, 21] were proposed to date. However, those did not make any use of string analysis, and relied on specifications of sanitizers without any guarantee that the sanitizers configured into the static taint analyzers were correct or that the specifications themselves were complete. Our work can contribute to finding application-specific sanitizers.

Checking satisfiability of an M2L formula could be implemented by exploiting other solvers such as a SAT solver or a SMT solver through constructing a bounded model of M2L(Str) [3] , or symbolic representations of automata [30].

8. CONCLUSION

In this paper, we presented a novel string-analysis technique, which enables unprecedented precision when modeling string operations, thanks to the combination of path and index sensitivity. Our string analysis is conducted by encoding the program in M2L, and relies on the satisfiability checking of an M2L formula. Our technique is motivated by the need for effective security analysis of Web applications, where a robust procedure for detecting and verifying sanitizers is essential. Our evaluation of the proposed approach shows it to compare favorably to the CFG-based string analysis of [12] (discovering 128 *vs.* 71 sanitizers), and have a significant impact on its client taint analysis' precision.

Acknowledgements

We thank Anders Møller and Yinnon Haviv for their invaluable comments and advice on a preliminary version of this paper, Naoshi Tabuchi for his help on our earlier work, the developers of MONA, WALA, and IBM AppScan for providing some of the building blocks for this work, and the ISSTA reviewers for their helpful comments.

9. REFERENCES

[1] IBM Rational AppScan Source Edition. `ibm.com/software/rational/products/appscan/source`.

[2] Open Web Application Security Project (OWASP). `owasp.org/index.php/Category:Attack`.

[3] A. Ayari and D. Basin. Bounded model construction for monadic second-order logics. In *CAV*, 2000.

[4] D. Balzarotti, M. Cova, V. Felmetsger, N. Jovanovic, E. Kirda, C. Kruegel, and G. Vigna. Saner: Composing static and dynamic analysis to validate sanitization in web applications. In *Security and Privacy (Oakland)*, 2008.

[5] N. Bjørner, N. Tillmann, and A. Voronkov. Path feasibility analysis for string-manipulating programs. In *TACAS*, 2009.

[6] D. Brumley, H. Wang, S. Jha, and D. Song. Creating vulnerability signatures using weakest preconditions. In *CSF*, 2007.

[7] A. S. Christensen, A. Feldthaus, and A. Møller. JSA – the Java String Analyzer. `brics.dk/JSA`, 2009.

[8] A. S. Christensen, A. Møller, and M. I. Schwartzbach. Precise analysis of string expressions. In *SAS*, 2003.

[9] P. Cousot and R. Cousot. Formal language, grammar and set-constraint-based program analysis by abstract interpretation. In *FPCA*, 1995.

[10] R. Cytron, J. Ferrante, B. K. Rosen, M. N. Wegman, and F. K. Zadeck. Efficiently computing static single assignment form and the control dependence graph. *TOPLAS*, 1991.

[11] X. Fu and C.-C. Li. A string constraint solver for detecting web application vulnerability. In *SEKE*, 2010.

[12] E. Geay, M. Pistoia, T. Tateishi, B. Ryder, and J. Dolby. Modular string-sensitive permission analysis with demand-driven precision. In *ICSE*, 2009.

[13] D. Grove and C. Chambers. A Framework for Call Graph Construction Algorithms. *TOPLSA*, 2001.

[14] D. Grove, G. DeFouw, J. Dean, and C. Chambers. Call graph construction in object-oriented languages. In *OOPSLA*, 1997.

[15] C. Hammer, R. Schaade, and G. Snelting. Static path conditions for java. In *PLAS*, 2008.

[16] J. G. Henriksen, J. L. Jensen, M. E. Jørgensen, N. Klarlund, R. Paige, T. Rauhe, and A. Sandholm. MONA: Monadic second-order logic in practice. In *TACAS*, 1995.

[17] P. Hooimeijer and W. Weimer. A decision procedure for subset constraints over regular languages. In *PLDI*, 2009.

[18] M. Kay and R. M. Kaplan. Regular models of phonological rule systems. *Computational Linguistics, 20(3)*, 1994.

[19] A. Kieżun, V. Ganesh, P. J. Guo, P. Hooimeijer, and M. D. Ernst. HAMPI: A solver for string constraints. In *ISSTA*, 2009.

[20] N. Klarlund and A. Møller. *MONA Version 1.4 User Manual*. BRICS, 2001. Notes Series NS-01-1. `http://www.brics.dk/mona`.

[21] B. Livshits, A. V. Nori, S. K. Rajamani, and A. Banerjee. Merline: Specification inference for explicit information flow problems. In *PLDI*, 2009.

[22] V. B. Livshits and M. S. Lam. Finding security vulnerabilities in java applications with static analysis. In *USENIX Security*, 2005.

[23] Y. Minamide. Static approximation of dynamically generated web pages. In *WWW*, 2005.

[24] B. K. Rosen, M. N. Wegman, and F. K. Zadeck. Global value numbers and redundant computations. In *POPL*, 1988.

[25] P. Saxena, D. Akhawe, S. Hanna, F. Mao, S. McCamant, and D. Song. A symbolic execution framework for javascript. In *Security and Privacy (Oakland)*, 2010.

[26] D. Shannon, I. Ghosh, S. Rajan, and S. Khurshid. Efficient symbolic execution of strings for validating web applications. In *DEFECTS*, 2009.

[27] G. Snelting. Combining slicing and constraint solving for validation of measurement software. In *SAS*, 1996.

[28] N. Tillmann and J. D. Halleux. Pex: white box test generation for .NET. In *TAP*, 2008.

[29] O. Tripp, M. Pistoia, S. Fink, M. Sridharan, and O. Weisman. TAJ: Effective taint analysis of web applications. In *PLDI*, 2009.

[30] M. Veanes, P. de Halleux, and N. Tillmann. Rex: Symbolic regular expression explorer. Microsoft Research Technical Report MSR-TR-2009-137, 2009.

[31] T. J. Watson Libraries for Analysis, `wala.sf.net/`.

[32] G. Wassermann and Z. Su. Sound and precise analysis of web applications for injection vulnerabilities. In *PLDI*, 2007.

[33] M. N. Wegman and F. K. Zadeck. Constant propagation with conditional branches. *TOPLAS*, 1991.

[34] F. Yu, M. Alkhalaf, and T. Bultan. Generating vulnerability signatures for string manipulating programs using automata-based forward and backward symbolic analyses. In *ASE*, 2009.

[35] F. Yu, T. Bultan, M. Cova, and O. Ibarra. Symbolic string verification: An automata-based approach. In *SPIN Workshop*, 2008.

[36] Z3, `research.microsoft.com/projects/z3`.

x	$\in$	$\mathcal{X}$	
b, p	$\in$	$\mathbf{N}$	
s	$\in$	Σ^*	
bool	$\in$	$\{\mathsf{true}, \mathsf{false}\}$	
v	$::=$	$s \mid p \mid \mathit{bool}$	
I	$::=$	$x = v$	Assignment
	$\mid$	$x = f(\vec{x})$	Function call
	$\mid$	$x = \mathtt{phi}(b : x, \ldots, b : x)$	Φ-instruction
	$\mid$	$x = \mathtt{jump}\ x, b$	Conditional jump
	$\mid$	$x = \mathtt{goto}\ b$	Goto
B	$::=$	$(b, \vec{I})$	Basic block
N	$::=$	$\{B\}$	Set of basic blocks

Figure 16: Syntax of the target language

(CONST)

$$\sigma \vdash (b, b', x = v; \vec{I}) \to \sigma[v/x] \vdash (b, b', \vec{I})$$

(CALL)

$$\sigma \vdash (b, b', x = f(x_1, \cdots, x_n); \vec{I}) \to \sigma[v/x] \vdash (b, b', \vec{I})$$

where $v = f(\sigma(x_1), \cdots, \sigma(x_i))$

(PHI)

$$\sigma \vdash (b, b_i, x = \mathtt{phi}(b_1 : x_1, \cdots, b_n : x_n); \vec{I}) \to \sigma[v_i/x] \vdash (b, b_i, \vec{I})$$

where $v_i = \sigma(x_i)$

(JUMP)

$$\sigma \vdash (b, b', \mathtt{jump}\ x, b'') \to \begin{cases} \sigma \vdash (b'', b, \vec{I}) \\ \quad \text{when } \sigma(x) = \mathsf{true}, \text{ where } (b'', \vec{I}) \in N. \\ \sigma \vdash (b+1, b, \vec{I}) \\ \quad \text{when } \sigma(x) = \mathsf{false}, \text{ where } (b+1, \vec{I}) \in N. \end{cases}$$

(GOTO)

$$\sigma \vdash (b, b', \mathtt{goto}\ x, b'') \to \sigma \vdash (b'', b, \vec{I}) \ \text{ where } (b'', \vec{I}) \in N$$

(BB)

$$\sigma \vdash (b, b', \epsilon) \to \sigma \vdash (b+1, b, \vec{I}) \ \text{ where } (b+1, \vec{I}) \in N$$

Figure 17: Operational semantics

APPENDIX

A. TARGET LANGUAGE

Figure 16 shows the formal syntax of our target language. A program consists of a set of basic blocks N. The meta-variables x and v represent a program variable and a value, respectively, where the value v is a string s, an index p, or a Boolean value *bool*. The guard condition c represents consists of Boolean values, the program variables, and Boolean operators. The meta-variable B is a basic block which is numbered by b and contains the sequence of instructions $\vec{I}$, where we denote an empty sequence by ϵ and a delimiter by ";". We omit the `return` instructions, since we are discussing intraprocedural string analysis.

The semantics of the target language is depicted as transition rules in Figure 17. $\sigma \vdash (b, b', \vec{s})$ denotes a program state, where σ is a mapping from variables to values, b is a current basic block number, b' is the immediate predecessor of the current basic block b, and $\vec{I}$ is a sequence of instructions. In addition, we denotes the transitive closure of $\to$ by $\to^*$.

B. PROOF OF SOUNDNESS

The proof is done by induction on the rules of the operational semantics with the following invariant inv on the program states σ:

$$\begin{array}{l} \mathsf{inv}(\sigma) \equiv \exists w, \mathcal{I}\ .\ \mathsf{inv}'(\sigma, w, \mathcal{I}) \\ \text{where } \mathsf{inv}'(\sigma, w, \mathcal{I}) \equiv \forall x \in \mathit{dom}(\sigma)\ .\ \sigma(x) \in L_{w,\mathcal{I}}(\mathsf{prog}_x) \end{array}.$$

Note that the formula is exactly equivalent to Theorem 1 if we expand the definition of inv'.

Obviously, at the initial state σ_0, $\mathsf{inv}(\sigma_0)$ holds, since $\mathit{dom}(\sigma) = \emptyset$ holds.

Next, for every rule, we prove that, if $\mathsf{inv}(\sigma)$ holds for the program state σ, $\mathsf{inv}(\sigma')$ also holds for a program state σ' that is obtained by a one-step transition from the program state σ.

CONST If $\mathsf{inv}(\sigma)$ holds, there exists a finite string w and an assignment $\mathcal{I}$ that satisfy $\mathsf{inv}'(\sigma, w, \mathcal{I})$. In addition, there exists a finite string w' such that $v \in L_{w',\mathcal{I}}(\text{“}v\text{”})$. Therefore, $v \in L_{ww',\mathcal{I}}(\text{“}v\text{”})$ is also holds, where ww' is the concatenation of the finite strings w and w'. Since $\mathsf{prog}_x(R) \equiv \text{“}v\text{”}(R)$ is the predicate declaration encoded from the CONST rule, $v \in L_{ww',\mathcal{I}}(\mathsf{prog}_x)$ holds. Therefore, taking $\sigma' = \sigma[v/x]$ into account, (since the invariant $\mathsf{inv}'(\sigma', ww', \mathcal{I})$ holds), $\mathsf{inv}(\sigma')$ also holds.

CALL If $\mathsf{inv}(\sigma)$ holds, there exists a finite string w and an assignment $\mathcal{I}$ that satisfy $\mathsf{inv}'(\sigma, w, \mathcal{I})$. Thus, for every parameter $x_i (i = 0, \cdots, n)$ of the function call, $\sigma(x_i) \in L_{w,\mathcal{I}}(\mathsf{prog}_{x_i})$ holds. Here, from Definition 2, there exists a finite string w' that satisfies

$$v \in L_{ww',\mathcal{I}}(\lambda R.\, [\![f]\!]\, (R, \mathsf{prog}_{x_1}, \cdots, \mathsf{prog}_{x_n}))$$

, where v is a return value of the function call. Accordingly, we can obtain $v \in L_{ww',\mathcal{I}}(\mathsf{prog}_x)$ from the fact that a corresponding predicate declaration is

$$\mathsf{prog}_x(R) \equiv [\![f]\!]\, (R, \mathsf{prog}_{x_1}, \cdots, \mathsf{prog}_{x_n}).$$

Therefore, at the program state $\sigma' = \sigma[v/x]$, $\mathsf{inv}(\sigma')$ holds.

PHI $\sigma(x_i) \in L_{w,\mathcal{I}}(prog_{x_i})$ follows from $\mathsf{inv}(\sigma)$, where $x_j (i = 0, \cdots, n)$ is the parameter of the phi-function. Therefore, for the return value v_i, this formula holds:

$$v_i \in L_{w,\mathcal{I}}(\mathsf{prog}_{x_1}) \vee \cdots \vee L_{w,\mathcal{I}}(\mathsf{prog}_{x_n})$$

Since the corresponding predicate declaration is

$$\mathsf{prog}_x(R) \equiv \mathsf{prog}_{x_1}(R) \vee \cdots \vee \mathsf{prog}_{x_n}(R),$$

we obtain the formula $v_i \in L_{w,\mathcal{I}}(\mathsf{prog}_x)$. Thus, at the program state $\sigma' = \sigma[v_i/x]$, $\mathsf{inv}(\sigma')$ holds.

JUMP,GOTO,BB There are no updates on the program state. Therefore, the invariant is preserved.