

Exact Heap Summaries from Symbolic Execution

Anonymous

Abstract

One of the fundamental challenges of using symbolic execution to analyze software has been the treatment of dynamically allocated data. State-of-the-art symbolic execution techniques have addressed this challenge by constructing the heap *lazily*, materializing objects on the concrete heap “as needed” and using non-deterministic choice points to explore each feasible concrete heap configuration. Because analysis of the materialized heap locations relies on concrete program semantics, the lazy initialization approach exacerbates the state space explosion problem that limits the scalability of symbolic execution. In this work we present a novel approach for lazy symbolic execution of heap manipulating software which utilizes a fully symbolic heap constructed on-the-fly during symbolic execution. Our approach is 1) *scalable* – it does not create the additional points of non-determinism introduced by existing lazy initialization techniques and it explores each execution path only once for any given set of isomorphic heaps, 2) *precise* – at any given point during symbolic execution, the symbolic heap represents the exact set of feasible concrete heap structures for the program under analysis, and 3) *expressive* – the symbolic heap can represent recursive data structures and heaps resulting from loops and recursive control structures in the code. We report on a case-study of an implementation of our technique in the Symbolic PathFinder tool to illustrate its scalability, precision and expressiveness. We also discuss how test case generation – a common use for symbolic execution results – can benefit from symbolic execution which uses a fully symbolic heap.

Categories and Subject Descriptors CR-number [subcategory]: third-level

General Terms term1, term2

Keywords keyword1, keyword2

1. Introduction

In recent years symbolic execution – a program analysis technique for systematic exploration of program execution paths using symbolic input values – has provided the basis for various software testing and analysis techniques. For each execution path explored during symbolic execution, constraints on the symbolic inputs are collected to create a *path condition*. The set of path conditions

computed by symbolic execution characterize the observed program execution behaviours and can be used as an enabling technology for various applications, e.g., regression analysis [1–6], data structure repair [7], dynamic discovery of invariants [8, 9], and debugging [10].

Initial work on symbolic execution largely focused on checking properties of programs with primitive types, such as integers and booleans [11, 12]. Despite recent advances in constraint solving technologies, improvements in raw computing power, and advances in reduction and abstraction techniques [13, 14] symbolic execution of programs of modest size containing only primitive types, remains challenging because of the large number of execution paths generated during symbolic analysis.

With the advent of object-oriented languages that manipulate dynamically allocated data, e.g., Java and C++, recent work has generalized the core ideas of symbolic execution to enable analysis of programs containing complex data structures with unbounded domains, i.e., data stored on the heap [15–17]. These techniques construct the heap in a lazy manner, deferring materialization of objects on the concrete heap until they are needed for the analysis to proceed. Treatment of heap allocated data then follows concrete program semantics once a heap location is materialized, resulting in a large number of feasible concrete heap configurations, and as a result, a large number of points of non-determinism to be analyzed, further exacerbating the state space explosion problem.

THIS PARA IS NOT QUITE RIGHT BUT THE IDEA IS STARTING TO COME OUT. Although lazy symbolic execution techniques have been instrumental in enabling analysis of heap manipulating programs, they miss an important opportunity to control the state space explosion problem by treating only inputs with primitive types symbolically and materializing a concrete heap. As we show in this work, the use of a fully *symbolic heap* during lazy symbolic execution, can improve the scalability of the analysis while maintaining precision and efficiency. Moreover, the number of path conditions computed by lazy symbolic execution when a symbolic heap is used produces considerably fewer path conditions – a valuable benefit for client analyses that use the results of symbolic execution, e.g., regression analyses.

The key advantages of our approach to lazy symbolic execution using a fully symbolic heap include:

- *Scalability*. Our approach constructs the symbolic heap on-the-fly during symbolic execution and avoids creating the additional points of non-determinism introduced by existing lazy initialization techniques. Moreover, it explores each execution path only once for any given set of isomorphic heaps.
- *Precision*. At any given point during symbolic execution, the symbolic heap represents the exact set of feasible concrete heap structures for the program under analysis
- *Expressiveness*. The symbolic heap can represent recursive data structures and heap structures resulting from loops and recursive control structures in the analyzed code.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CONF 'yy, Month d–d, 20yy, City, ST, Country.
Copyright © 20yy ACM 978-1-nnnn-nnnn-n/yy/mm...\$15.00.
<http://dx.doi.org/10.1145/nnnnnnn.nnnnnnn>

This paper makes the following contributions:

- We present a novel lazy symbolic execution technique for analyzing heap manipulating programs that constructs a fully symbolic representation of the heap on-the-fly during symbolic execution.
- We prove the soundness and completeness of our algorithm...
- We implement our approach in the Symbolic PathFinder tool
- We demonstrate experimentally that our technique improves the scalability of symbolic execution of heap manipulating software over state-of-the-art techniques, while maintaining efficiency and precision.
- We discuss the benefits of using a symbolic heap that can be realized by the client analysis that uses the results of symbolic execution.

2. Background and Motivation

In this section we present the background on state of the art techniques that have been developed to handle data non-determinism arising from complex data structures. We present an overview of lazy initialization and lazier# initialization. We also present a brief description of the two bounding strategies used in symbolic execution in heap manipulating programs. Next we present a motivating examples where current concrete initialization of the heap structures struggle to scale to medium sized program due to non-determinism introduced in the symbolic execution tree. We use this example to motivate the need for a more truly symbolic and compact representation of the heap in a manner similar to that of primitive types.

Generalized symbolic execution technique generates a concrete representation of connected memory structures using only the implicit information from the program itself. In the original lazy initialization algorithm, symbolic execution explores different heap shapes by concretizing the heap at the first memory access (read) to an un-initialized symbolic object. At this point, a non-deterministic choice point of concrete heap locations is created that includes: (a) null, (b) an access to a new instance of the object, and (c) aliases to other type-compatible symbolic objects that have been concretized along the same execution path [?]. The number of choices explored in lazy initialization greatly increases the non-determinism and often makes the exploration of the program state space intractable.

The Lazier# algorithm is an improvement of the lazy initialization and it pushes the non-deterministic choices further into the execution tree. In the case of a memory access to an uninitialized reference location, by default, no choice point is created. Instead, the read returns a unique symbolic reference representing the contents of the location. The reference may assume any one of three states: uninitialized, non-null, or initialized. The reference is returned in an uninitialized state, and only in a subsequent memory access is the reference concretely initialized.

3. Generalized Symbolic Execution with Lazy Initialization

Generalized symbolic execution with lazy initialization is a technique to apply symbolic execution to non-primitive data types (i.e., objects) [17]. The approach initializes each symbolic object to be either null, a new instance of the object, or an alias to an object previously created as part of lazy initialization. Each of these outcomes is a possible choice, and a model checker is able to exhaustively enumerate these choices effectively creating all possible concrete heaps under the choice sets. The approach is first formalized in this section and then extended to describe the precise heap summaries.

```
public class LinkedList {

    /** assume the linked list is valid with no cycles */
    LLNode head;
    Data data0, data1, data2, data3, data4;

    private class Data { Integer val; }

    private class LLNode {
        protected Data elem;
        protected LLNode next; }

    public static boolean contains(LLNode root, Data val) {
        LLNode node = root;
        while (true) {
            if(node.val == val) return true;
            if(node.next == null) return false;
            node = node.next;
        }
    }

    public void run() {
        if(LinkedList.contains(head, data0) &&
           LinkedList.contains(head, data1) &&
           LinkedList.contains(head, data2) &&
           LinkedList.contains(head, data3) &&
           LinkedList.contains(head, data4)) return;
    }
}
```

Figure 1. Linked list

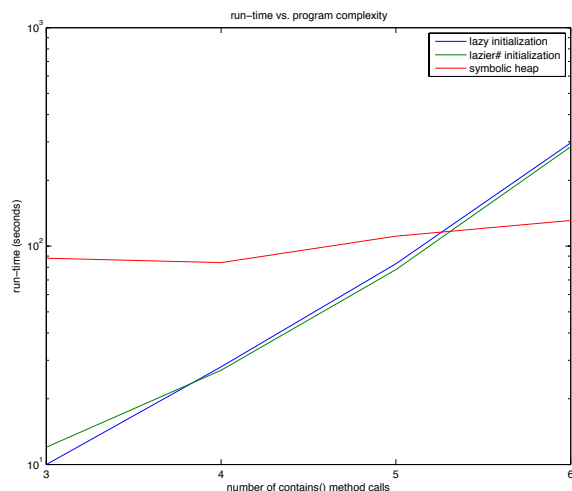


Figure 2. Time versus complexity for the linked list example

Many of these concrete heaps explore redundant control flow paths in the program under test, but since symbolic model checking is stateless, subject to a bound to terminate the search, the model checker needlessly explores many redundant state leading to severe state explosion.

Figure 3 defines the surface syntax for the Javalite language [18]. Figure 4 is the machine syntax. Javalite is syntactic machine defined as rewrites on a string. The semantics use a CEKS model with a (C)ontrol string representing the expression being evaluated, an (E)nvironment for local variables, a (K)ontinuation for what is to be executed next, and a (S)tore for the heap.

The heap is a labeled bipartite graph consisting of references R and locations L in the store. The functions R and L are defined for convenience in manipulating the labeled bipartite graph.

```

P ::= (μ (C m))
μ ::= (CL ...)
T ::= bool | C
CL ::= (class C ([T f] ...) (M ...))
M ::= (T m [T x] e)
e ::= x
    | (new C)
    | (e $f)
    | (x $f := e)
    | (e = e)
    | (if e else e)
    | (var T x := e in e)
    | (e @ m e)
    | (x := e)
    | (begin e ...)
    | v
x ::= this | id
f ::= id
m ::= id
C ::= id
v ::= r | null | true | false
r ::= number
id ::= variable-not-otherwise-mentioned

```

Figure 3. The Javalite surface syntax.

```

φ ::= (φ) | φ ⋈ φ | ¬φ | r = r | r ≠ r
l ::= number
η ::= (mt (η [x → v]))
s ::= (μ L R φg η e k)
k ::= end
    | (* $f → k)
    | (x $f := * → k)
    | (* = e → k)
    | (v = * → k)
    | (if * e else e → k)
    | (var T x := * in e → k)
    | (* @ m e → k)
    | (v @ m * → k)
    | (x := * → k)
    | (begin * (e ...) → k)
    | (pop η k)

```

Figure 4. The machine syntax for Javalite with $\bowtie \in \{\wedge, \vee, \rightarrow\}$.

- $R(l, f)$ maps location-field pairs from the store to a reference in R ; and
- $L(r)$ maps references to a set of location-constraint pairs in the store.

A reference is a node that gathers the possible store locations for an object during symbolic execution. Each store location is guarded by a constraint that determines the aliasing in the heap. Intuitively, the reference is a level of indirection between a variable and the store, and the reference is used to group a set of possible store locations each predicated on the possible aliasing in the associated constraint. For a variable (or field) to access any particular store location associated with its reference, the corresponding constraint must be satisfied.

Locations are boxes in the graphical representation and indicated with the letter l in the math. References are circles in the graphical representation and indicated with the letter r in the math. The special symbol \perp is an uninitialized reference. Edges from lo-

cations are labeled with field names f . Edges from the references are labeled with constraints $\phi \in \Phi$ (it is assumed that Φ is a power set over individual constraints and ϕ is a set of constraints for the edge).

The function $\mathbb{V}\mathbb{S}(L, R, \phi_g, r, f)$ constructs the value-set given a heap, reference, and desired field:

$$\mathbb{V}\mathbb{S}(L, R, \phi_g, r, f) = \{(l' \phi \wedge \phi') \mid \exists l ((l \phi) \in L(r) \wedge \exists r' \in R(l, f) ((l' \phi') \in L(r') \wedge \mathbb{S}(\phi \wedge \phi' \wedge \phi_g)))\}$$

where $\mathbb{S}(\phi)$ returns true if ϕ is satisfiable.

The strengthen function $\mathbb{S}\mathbb{T}(L, r, \phi')$ strengthens every constraint from the reference r with ϕ' and keeps only location-constraint pairs that are satisfiable after this strengthening:

$$\mathbb{S}\mathbb{T}(L, r, \phi') = \{(l \phi \wedge \phi') \mid (l \phi) \in L(r) \wedge \mathbb{S}(\phi \wedge \phi')\}$$

4. Proofs

4.1 Definitions

Definition 1. A *state transition function* \rightarrow_Φ is a mapping $\rightarrow_\Phi: s \mapsto s$, which takes one machine state and transforms it into another machine state. The state transition function with superscript represents composition of the state transition function:

$$s_a \rightarrow_\Phi s_b \rightarrow_\Phi s_c \implies s_a \rightarrow_\Phi^2 s_c$$

Definition 2. A *feasible state sequence* is defined as a sequence of states resulting from repeated application of the state transition relation to some initial state s_0 :

$$\Pi_n = s_0, s_1, \dots, s_n$$

where the relation $s_i \rightarrow_\Phi s_{i+1}$ holds for all $i \in \{i \mid 0 \leq i < n\}$

Definition 3. Given a sequence of states

$$\Pi_n = s_0, s_1, \dots, s_n$$

where

$$s_i = (\mu_i L_i R_i \phi_i \eta_i e_i k_i)$$

the *control flow sequence* of Π_n is the defined as the sequence of tuples

$$\pi_n = \mathbb{C}\mathbb{F}(\Pi_n) = (\eta_0 e_0 k_0), (\eta_1 e_1 k_1), \dots, (\eta_n e_n k_n)$$

Definition 4. Given a state transition function \rightarrow_Φ , an initial state s_0 and a control flow sequence π_n , the *feasible state set* $\zeta = \mathbb{F}\mathbb{S}(\rightarrow_\Phi, s_0, \pi_n)$ is defined as

$$\zeta = \{s \mid \pi_n = \mathbb{C}\mathbb{F}(\Pi_n) \wedge s = \max_s(\Pi) \wedge s_0 \rightarrow_\Phi^{n-1} s\}$$

Definition 5. A *heap homomorphism* $(g h)$ between two states s_x and s_y is defined as a pair of functions $g: r \mapsto r$ and $h: l \mapsto l$ such that for any reference $r \in s_x$, location $l \in s_x$, and field f ,

$$(\phi_x l) \in L_x(r) \Rightarrow (\phi_y h(l)) \in L_y(g(r))$$

and

$$r = R_x(l, f) \Rightarrow g(r) = R_y(h(l), f)$$

If a such a pair of functions exists from one state s_x to another state s_y we say that state s_x is *heap homomorphic* to state s_y , indicated by the notation $(g f): s_x \rightarrow s_y$.

Suppose we take the set of constraints from the image of s_x in s_y under $(g h)$:

$$\chi = \{\phi \mid \exists (r \in R_x, l \in L_x) ((\phi h(l)) \in L_y(g(r)))\}$$

and we conjoin those constraints with the global invariant from s_y :

$$\phi_G = \phi_y \bigwedge_{\phi_i \in \chi} \phi_i$$

VARIABLE LOOKUP $(LR \phi_g \eta x k) \rightarrow$ $(LR \phi_g \eta \eta(x) k)$	NEW $r = \text{fresh}_r() \quad l = \text{fresh}_l(C)$ $R' = R[\forall f \in \text{fields}(C) \ (lf) \mapsto \text{fresh}_r()]$ $\rho = \{r' \mid \exists f \in \text{fields}(C) \ (r' = R'(l, f))\}$ $L' = L[r \mapsto \{(\phi_T l)\}][\forall r' \in \rho \ (r' \mapsto (\phi_T l_{null}))]$ <hr/> $(LR \phi_g \eta (\text{new } C) k) \rightarrow$ $(L' R' \phi_g \eta r k)$	FIELD ACCESS(EVAL) $(LR \phi_g \eta (e \$f) k) \rightarrow$ $(LR \phi_g \eta e (* \$f \rightarrow k))$
FIELD WRITE (EVAL) $(LR \phi_g \eta (x \$f := e) k) \rightarrow$ $(LR \phi_g \eta e (x \$f := * \rightarrow k))$	EQUALS (L-OPERAND EVAL) $(LR \phi_g \eta (e_0 = e) k) \rightarrow$ $(LR \phi_g \eta e_0 (* = e \rightarrow k))$	EQUALS (R-OPERAND EVAL) $(LR \phi_g \eta v (* = e \rightarrow k)) \rightarrow$ $(LR \phi_g \eta e (v = * \rightarrow k))$
EQUALS (BOOL) $v_0 \in \{\text{true}, \text{false}\} \quad v_1 \in \{\text{true}, \text{false}\}$ $v_r = \text{eq?}(v_0, v_1)$ <hr/> $(LR \phi_g \eta v_0 (v_1 = * \rightarrow k)) \rightarrow$ $(LR \phi_g \eta v_r k)$	IF-THEN-ELSE (EVAL) $(LR \phi_g \eta (\text{if } e_0 e_1 \text{ else } e_2) k) \rightarrow$ $(LR \phi_g \eta e_0 (\text{if } * e_1 \text{ else } e_2) \rightarrow k)$	IF-THEN-ELSE (TRUE) $(LR \phi_g \eta \text{true} (\text{if } * e_1 \text{ else } e_2) \rightarrow k) \rightarrow$ $(LR \phi_g \eta e_1 k)$
IF-THEN-ELSE (FALSE) $(LR \phi_g \eta \text{false} (\text{if } * e_1 \text{ else } e_2) \rightarrow k) \rightarrow$ $(LR \phi_g \eta e_2 k)$	VARIABLE DECLARATION (EVAL) $(LR \phi_g \eta (\text{var } T x := e_0 \text{ in } e_1) k) \rightarrow$ $(LR \phi_g \eta e_0 (\text{var } T x := * \text{ in } e_1 \rightarrow k))$	VARIABLE DECLARATION $(LR \phi_g \eta v (\text{var } T x := \text{in } e_1 \rightarrow k)) \rightarrow$ $(LR \phi_g \eta [x \mapsto v] e_1 (\text{pop } \eta k))$
METHOD INVOCATION (OBJECT EVAL) $(LR \phi_g \eta (e_0 @ m e_1) k) \rightarrow$ $(LR \phi_g \eta e_0 (* @ m e_1 \rightarrow k))$	METHOD INVOCATION (ARG EVAL) $(LR \phi_g \eta v_0 (* @ m e_1 \rightarrow k)) \rightarrow$ $(LR \phi_g \eta e_1 (v_0 @ m * \rightarrow k))$	METHOD INVOCATION $(T m [T x] e_m) = \text{lookup}(m)$ $\eta_m = \eta[\text{this} \mapsto v_0][x \mapsto v_1]$ <hr/> $(LR \phi_g \eta v_1 (v_0 @ m * \rightarrow k)) \rightarrow$ $(LR \phi_g \eta_m e_m (\text{pop } \eta k))$
VARIABLE ASSIGNMENT (EVAL) $(LR \phi_g \eta (x := e) k) \rightarrow$ $(LR \phi_g \eta e (x := * \rightarrow k))$	VARIABLE ASSIGNMENT $(LR \phi_g \eta v (x := * \rightarrow k)) \rightarrow$ $(LR \phi_g \eta [x \mapsto v] v k)$	BEGIN (NO ARGS) $(LR \phi_g \eta (\text{begin}) k) \rightarrow$ $(LR \phi_g \eta k)$
BEGIN (ARG0 EVAL) $(LR \phi_g \eta (\text{begin } e_0 e_1 \dots) k) \rightarrow$ $(LR \phi_g \eta e_0 (\text{begin } * (e_1 \dots) \rightarrow k))$	BEGIN (ARGI EVAL) $(LR \phi_g \eta v (\text{begin } * (e_i e_{i+1} \dots) \rightarrow k)) \rightarrow$ $(LR \phi_g \eta e_i (\text{begin } * (e_{i+1} \dots) \rightarrow k))$	BEGIN (ARGN EVAL) $(LR \phi_g \eta v (\text{begin } * (e_n) \rightarrow k)) \rightarrow$ $(LR \phi_g \eta e_n (\text{begin } * () \rightarrow k))$
BEGIN $(LR \phi_g \eta v (\text{begin } * () \rightarrow k)) \rightarrow$ $(LR \phi_g \eta v k)$	NULL $(LR \phi_g \eta \text{null } k) \rightarrow$ $(LR \phi_g \eta r_{null} k)$	POP $(LR \phi_g \eta v (\text{pop } \eta_0 k)) \rightarrow$ $(LR \phi_g \eta_0 v k)$

Figure 5. Javalite rewrite rules that are common to generalized symbolic execution and precise heap summaries.

If the expression ϕ_G is satisfiable, we say that the heap homomorphism $(g \ h) : s_x \rightarrow s_y$ is *valid*.

Definition 6. The *representation relation* is defined as follows: given two states s_ℓ and s_s , $s_\ell \sqsubset s_s$ if and only if $\eta_\ell = \eta_s$, $e_\ell = e_s$, $k_\ell = k_s$, and there exists a valid heap homomorphism $(g \ h) : s_\ell \rightarrow s_s$. The expression $s_\ell \sqsubset s_s$ can be read as "state s_ℓ is represented by state s_s ".

Definition 7. A state s_s is *congruent* to a set of states S if and only if s represents every state in S and represents no other state:

$$s_s \equiv S : s_i \in S \Leftrightarrow s_i \sqsubset s_s$$

Definition 8. A symbolic state s_s is *exact* with respect to an initial state s_0 and control flow sequence π if and only if it is congruent the set of feasible lazy states on the same control flow path:

$$s_s \equiv \mathbb{FS}(\rightarrow_\ell, s_0, \pi)$$

4.2 Theorems

Lemma 1. If symbolic state $s_s = (L_S R_S \phi_g \eta r (* \$f \rightarrow k))$ is exact with respect to some initial state s_0 and control flow path π_n , then the state $s'_s : s_s \rightarrow_s s'_s$ is exact with respect to s_0 and π_{n+1} .

Proof. We will consider two cases for this proof. In the first case, we assume that all of the fields involved in the read are initialized. In the second case we consider the case of uninitialized fields.

Case 1: suppose all of the pertinent fields in s_s are initialized. Take an arbitrary lazy state $s_\ell \sqsubset s_s$. Since s_s is exact, $s_\ell = (L_\ell R_\ell \phi_L \eta r (* \$f \rightarrow k))$, and that $s_\ell \in \mathbb{FS}(\rightarrow_\ell, s_0, \pi_n)$. If we apply the state transition functions to achieve states $s'_\ell : s_\ell \rightarrow_\ell s'_\ell$ and $s'_s : s_s \rightarrow_s s'_s$, we find that:

$$s'_\ell = (L_\ell[r' \mapsto (\phi' l')]) R_\ell \phi_L \eta r' k$$

and

$$s'_s = (L_s[r' \mapsto \mathbb{VS}(L_s, R_s, r, f, \phi_g)]) R_s \phi_g \eta r' k$$

We now show that $s'_\ell \sqsubset s'_s$. Since η , e , and k are identical between s'_s and s'_ℓ , the first condition is met by default. Now we construct functions $g' : g' = g[r' \mapsto r']$ and $h' : h' = h$. Observe that since $s_\ell \rightarrow s_s$, and since R_ℓ and R_s are unchanged from states s_ℓ to s'_ℓ and s_s to s'_s respectively, we are guaranteed that $r = R_\ell(l, f) \Rightarrow g'(r) = R_s(h'(l), f)$. Since $(\phi'_\ell l') = L_\ell(R_\ell(l, f))$, and since $(g \ h)$ is valid, we know that:

$$(\phi_s \wedge \phi'_s l') \in \mathbb{VS}(L_s, R_s, r, f, \phi_g)$$

INITIALIZE (NULL)

$$\begin{array}{l} \Lambda = \{l \mid \exists \phi ((\phi l) \in L(r) \wedge R(l, f) = \perp)\} \quad \Lambda \neq \emptyset \\ r' = \text{fresh}_r() \quad \theta_{\text{null}} = \{(\phi_T l_{\text{null}})\} \quad l_x = \min_l(\Lambda) \\ \phi'_g = (\phi_g \wedge r' = r_{\text{null}}) \\ \hline (LR \phi_g r f) \rightarrow_I (L[r' \mapsto \theta_{\text{null}}] R[(l_x, f) \mapsto r'] \phi'_g r f) \end{array}$$

INITIALIZE (ALIAS)

$$\begin{array}{l} \Lambda = \{l \mid \exists \phi ((\phi l) \in L(r) \wedge R(l, f) = \perp)\} \quad \Lambda \neq \emptyset \\ C = \text{type}(f) \quad r' = \text{fresh}_r() \\ \rho = \{(r_a l_a) \mid \text{isInit}(r_a) \wedge \text{type}(l_a) = C \wedge \exists \phi_a ((\phi_a l_a) \in L(r_a))\} \\ (r_a l_a) \in \rho \quad \theta_{\text{alias}} = \{(\phi_T l_a)\} \quad l_x = \min_l(\Lambda) \\ \phi'_g = (\phi_g \wedge r' \neq r_{\text{null}} \wedge r' = r_a \wedge (\wedge_{(r'_a l_a) \in \rho} (r'_a \neq r_a) \wedge r' \neq r'_a)) \\ \hline (LR \phi_g r f) \rightarrow_I (L[r' \mapsto \theta_{\text{alias}}] R[(l_x, f) \mapsto r'] \phi'_g r f) \end{array}$$

INITIALIZE (NEW)

$$\begin{array}{l} \Lambda = \{l \mid \exists \phi ((\phi l) \in L(r) \wedge R(l, f) = \perp)\} \quad \Lambda \neq \emptyset \\ C = \text{type}(f) \quad r_f = \text{init}_r() \quad l_f = \text{init}_l(C) \\ R' = R[\forall f \in \text{fields}(C) ((l_f f) \mapsto \perp)] \\ \rho = \{(r_a l_a) \mid \text{isInit}(r_a) \wedge \text{type}(l_a) = C \wedge \exists \phi_a ((\phi_a l_a) \in L(r_a))\} \\ \theta_{\text{new}} = \{(\phi_T l_f)\} \quad l_x = \min_l(\Lambda) \\ \phi'_g = (\phi_g \wedge r_f \neq r_{\text{null}} \wedge (\wedge_{(r_a l_a) \in \rho} r_f \neq r_a)) \\ \hline (LR \phi_g r f) \rightarrow_I (L[r_f \mapsto \theta_{\text{new}}] R'[(l_x, f) \mapsto r_f] \phi'_g r f) \end{array}$$

INITIALIZE (END)

$$\begin{array}{l} \Lambda = \{l \mid \exists \phi ((\phi l) \in L(r) \wedge R(l, f) = \perp)\} \quad \Lambda = \emptyset \\ \hline (LR \phi_g r f) \rightarrow_I (LR \phi_g r f) \end{array}$$

Figure 6. The initialization machine, $s ::= (LR \phi_g r f)$, with $s \rightarrow_I^* s'$ indicating stepping the machine until the state does not change.

FIELD ACCESS

$$\begin{array}{l} \{(\phi l)\} = L(r) \quad l \neq l_{\text{null}} \\ (LR \phi_g r f) \rightarrow_I^* (L' R' \phi'_g r f) \\ \{(\phi' l')\} = L'(R'(l, f)) \quad r' = \text{fresh}_r() \\ \hline (LR \phi_g \eta r (* \$f \rightarrow k)) \rightarrow \\ (L'[r' \mapsto (\phi' l')] R' \phi'_g \eta r' k) \end{array}$$

EQUALS (REFERENCE-TRUE)

$$\begin{array}{l} L(r_0) = L(r_1) \quad \phi'_g = (\phi_g \wedge r_0 = r_1) \\ \hline (LR \phi_g \eta r_0 (r_1 = * \rightarrow k)) \rightarrow \\ (LR \phi'_g \eta \text{true } k) \end{array}$$

FIELD WRITE

$$\begin{array}{l} r_x = \eta(x) \quad \{(\phi l)\} = L(r_x) \quad l \neq l_{\text{null}} \\ (LR \phi_g r_x f) \rightarrow_I^* (L' R' \phi'_g r_x f) \\ r' = \text{fresh}_r() \quad \theta = L'(r) \\ \hline (LR \phi_g \eta r (x \$f := * \rightarrow k)) \rightarrow \\ (L'[r' \mapsto \theta] R'[(l, f) \mapsto r'] \phi_g \eta r' k) \end{array}$$

EQUALS (REFERENCE-FALSE)

$$\begin{array}{l} L(r_0) \neq L(r_1) \quad \phi'_g = (\phi_g \wedge r_0 \neq r_1) \\ \hline (LR \phi_g \eta r_0 (r_1 = * \rightarrow k)) \rightarrow \\ (LR \phi'_g \eta \text{false } k) \end{array}$$

Figure 7. Generalized symbolic execution with lazy initialization.

SUMMARIZE

$$\begin{array}{l} \Lambda = \{l \mid \exists \phi ((\phi l) \in L(r) \wedge l \neq l_{\text{null}} \wedge R(l, f) = \perp)\} \quad \Lambda \neq \emptyset \\ C = \text{type}(f) \quad r_f = \text{init}_r() \quad l_f = \text{init}_l(C) \\ R' = R[\forall f \in \text{fields}(C) ((l_f f) \mapsto \perp)] \\ \rho = \{(r_a \phi_a l_a) \mid \text{isInit}(r_a) \wedge (\phi_a l_a) \in L(r_a) \wedge \text{type}(l_a) = C\} \\ \theta_{\text{null}} = \{(\phi l_{\text{null}}) \mid \phi = (r_f = r_{\text{null}})\} \\ \theta_{\text{new}} = \{(\phi l_f) \mid \phi = (r_f \neq r_{\text{null}} \wedge (\wedge_{(r'_a \phi'_a l'_a) \in \rho} r_f \neq r'_a))\} \\ \theta_{\text{alias}} = \{(\phi l_a) \mid \exists r_a (\exists \phi_a ((r_a \phi_a l_a) \in \rho \wedge \phi = (\phi_a \wedge r_f \neq r_{\text{null}} \wedge r_f = r_a \wedge (\wedge_{(r'_a \phi'_a l'_a) \in \rho} (r'_a \neq r_a) \wedge r_f \neq r'_a))))\} \\ \theta = \theta_{\text{alias}} \cup \theta_{\text{new}} \cup \theta_{\text{null}} \quad l_x = \min_l(\Lambda) \\ \hline (LR r f) \rightarrow_S (L[r_f \mapsto \theta] R'[(l_x, f) \mapsto r_f] r f) \end{array}$$

SUMMARIZE (END)

$$\begin{array}{l} \Lambda = \{l \mid \exists \phi ((\phi l) \in L(r) \wedge R(l, f) = \perp)\} \quad \Lambda = \emptyset \\ \hline (LR r f) \rightarrow_I (LR r f) \end{array}$$

Figure 8. The summary machine, $s ::= (LR r f)$, with $s \rightarrow_I^* s'$ indicating stepping the machine until the state does not change.

From this, we may deduce that:

$$(\phi_l l) \in L'_\ell(r') \Rightarrow (\phi_s \wedge \phi'_s h'(l)) \in L'_s(g'(r'))$$

Since r' the only new addition to L'_ℓ and L'_s , we now know that the assertion above holds for all $l \in s'_\ell$. Thus, we have shown that $(g' h')$ is a heap homomorphism from s'_ℓ to s'_s . Furthermore, since $\mathbb{S}(\phi_s \wedge \phi'_s \wedge \phi_g)$ holds true, we know that $(g' h') : s'_\ell \rightarrow s'_s$ is valid. Since there is a valid heap homomorphism, and since

$\eta_\ell = \eta_s$, $e_\ell = e_s$, $k_\ell = k_s$, we by definition 6 know $s'_\ell \sqsubset s'_s$. We have now shown that for any lazy state s_ℓ :

$$s_\ell \in \mathbb{FS}(\rightarrow_\ell, s_0, \pi_n) \Rightarrow s'_\ell \sqsubset s'_s$$

Now, suppose that there exists a state s'_i such that $s'_i \sqsubset s'_s$, but $s'_i \notin \mathbb{FS}(\rightarrow_\ell, s_0, \pi_{n+1})$. Since $s'_i \sqsubset s'_s$, there must exist a valid homomorphism $(g'_i h'_i) : s'_i \rightarrow s'_s$. If $(g'_i h'_i)$ is valid, then there must also exist a valid homomorphism $(g_i h_i) : s_i \rightarrow s_s$, and by

$$\begin{array}{c}
\text{FIELD ACCESS} \\
\frac{\forall(\phi l) \in L(r) \ (l = l_{\text{null}} \rightarrow \neg \mathbb{S}(\phi \wedge \phi_g)) \quad (LR r f) \rightarrow_S^* (L' R' r f) \quad r' = \text{fresh}_r(\cdot)}{(LR \phi_g \eta r (* \$f \rightarrow k)) \rightarrow (L'[r' \mapsto \mathbb{V}\mathbb{S}(L', R', r, f, \phi_g)] R' \phi_g \eta r' k)}
\end{array}$$

$$\begin{array}{c}
\text{FIELD WRITE} \\
\frac{r_x = \eta(x) \quad \forall(\phi l) \in L(r_x) \ (l = l_{\text{null}} \rightarrow \neg \mathbb{S}(\phi \wedge \phi_g)) \quad (LR r_x f) \rightarrow_S^* (L' R' r_x f) \quad \Psi_x = \{(r_{\text{cur}} \phi l) \mid (\phi l) \in L'(r_x) \wedge r_{\text{cur}} \in R(l, f)\} \quad X = \{(r_{\text{cur}} \theta l) \mid \exists \phi ((r_{\text{cur}} \phi l) \in \Psi_x \wedge \theta = \mathbb{S}\mathbb{T}(L', r, \phi) \cup \mathbb{S}\mathbb{T}(L', r_{\text{cur}}, \neg \phi))\} \quad R' = R'[\forall(r_{\text{cur}} \theta l) \in X ((l f) \mapsto \text{fresh}_r(\cdot))] \quad L'' = L'[\forall(r_{\text{cur}} \theta l) \in X (\exists r_{\text{targ}} (r_{\text{targ}} = R'(l, f) \wedge (r_{\text{targ}} \mapsto \theta)))]}{(LR \phi_g \eta r (x \$f := * \rightarrow k)) \rightarrow (L'' R' \phi_g \eta k)}
\end{array}$$

$$\begin{array}{c}
\text{EQUALS (REFERENCES-TRUE)} \\
\frac{\theta_\alpha = \{(\phi_0 \wedge \phi_1) \mid \exists l ((\phi_0 l) \in L(r_0) \wedge (\phi_1 l) \in L(r_1))\} \quad \theta_0 = \{\phi_0 \mid \exists l_0 ((\phi_0 l_0) \in L(r_0) \wedge \forall(\phi_1 l_1) \in L(r_1) \ (l_0 \neq l_1))\} \quad \theta_1 = \{\phi_1 \mid \exists l_1 ((\phi_1 l_1) \in L(r_1) \wedge \forall(\phi_0 l_0) \in L(r_0) \ (l_0 \neq l_1))\} \quad \phi'_g = \phi_g \wedge (\bigvee_{\phi_\alpha \in \theta_\alpha} \phi_\alpha) \wedge (\bigwedge_{\phi_0 \in \theta_0} \neg \phi_0) \wedge (\bigwedge_{\phi_1 \in \theta_1} \neg \phi_1) \quad \mathbb{S}(\phi'_g)}{(LR \phi_g \eta r_0 (r_1 = * \rightarrow k)) \rightarrow (LR \phi'_g \eta \text{true } k)}
\end{array}$$

$$\begin{array}{c}
\text{EQUALS (REFERENCES-FALSE)} \\
\frac{\theta_\alpha = \{(\phi_0 \rightarrow \neg \phi_1) \mid \exists l ((\phi_0 l) \in L(r_0) \wedge (\phi_1 l) \in L(r_1))\} \quad \theta_0 = \{\phi_0 \mid \exists l_0 ((\phi_0 l_0) \in L(r_0) \wedge \forall(\phi_1 l_1) \in L(r_1) \ (l_0 \neq l_1))\} \quad \theta_1 = \{\phi_1 \mid \exists l_1 ((\phi_1 l_1) \in L(r_1) \wedge \forall(\phi_0 l_0) \in L(r_0) \ (l_0 \neq l_1))\} \quad \phi'_g = \phi_g \wedge (\bigwedge_{\phi_\alpha \in \theta_\alpha} \phi_\alpha) \vee ((\bigvee_{\phi_0 \in \theta_0} \phi_0) \vee (\bigvee_{\phi_1 \in \theta_1} \phi_1)) \quad \mathbb{S}(\phi'_g)}{(LR \phi_g \eta r_0 (r_1 = * \rightarrow k)) \rightarrow (LR \phi'_g \eta \text{false } k)}
\end{array}$$

Figure 9. Precise symbolic heap summaries from symbolic execution.

extension, state $s_i \sqsubset s_s$. However, since s_s is exact, $s_i \in \mathbb{FS}(\rightarrow_\ell, s_0, \pi_n)$, and by our previous result, $s'_i \in \mathbb{FS}(\rightarrow_\ell, s_0, \pi_{n+1})$. We have a contradiction. Therefore,

$$s'_i \sqsubset s'_s \Rightarrow s'_i \in \mathbb{FS}(\rightarrow_\ell, s_0, \pi_{n+1})$$

Combining the two previous equations, we find that

$$s'_\ell \in \mathbb{FS}(\rightarrow_\ell, s_0, \pi_n) \Leftrightarrow s'_\ell \sqsubset s'_s$$

By definition 7, $s'_s \equiv \mathbb{FS}(\rightarrow_\ell, s_0, \pi_n)$, and so by definition 8, s'_s is exact. \square

5. Related Work

The related work goes here.

Acknowledgments

Acknowledgments, if needed.

References

- [1] J. Backes, S. Person, N. Rungta, and O. Tkachuk, “Regression verification using impact summaries,” in *Model Checking Software*. Springer, 2013, pp. 99–116.
- [2] P. Godefroid, S. K. Lahiri, and C. Rubio-González, “Statically validating must summaries for incremental compositional dynamic test generation,” in *SAS*, 2011, pp. 112–128.
- [3] S. Person, M. B. Dwyer, S. Elbaum, and C. S. Păsăreanu, “Differential symbolic execution,” in *FSE*, 2008, pp. 226–237.
- [4] S. Person, G. Yang, N. Rungta, and S. Khurshid, “Directed incremental symbolic execution,” in *PLDI*, 2011, pp. 504–515.
- [5] D. A. Ramos and D. R. Engler, “Practical, low-effort equivalence verification of real code,” in *CAV*, 2011, pp. 669–685.
- [6] G. Yang, C. S. Păsăreanu, and S. Khurshid, “Memoized symbolic execution,” in *ISSTA*, 2012, pp. 144–154.
- [7] S. Khurshid, I. García, and Y. L. Suen, “Repairing structurally complex data,” in *SPIN*, 2005, pp. 123–138.
- [8] C. Csallner, N. Tillmann, and Y. Smaragdakis, “Dysy: Dynamic symbolic execution for invariant inference,” in *ICSE*, 2008, pp. 281–290.
- [9] L. Zhang, G. Yang, N. Rungta, S. Person, and S. Khurshid, “Feedback-driven dynamic invariant discovery,” in *ISSTA*, 2014, pp. 362–372.
- [10] K.-K. Ma, K. Y. Phang, J. S. Foster, and M. Hicks, “Directed symbolic execution,” in *SAS*, 2011, pp. 95–111.
- [11] L. A. Clarke, “A system to generate test data and symbolically execute programs,” *IEEE Transactions on Software Engineering*, vol. SE-2, no. 3, pp. 215–222, 1976.
- [12] J. C. King, “Symbolic execution and program testing,” *Communications of the ACM*, vol. 19, no. 7, pp. 385–394, 1976.
- [13] S. Anand, C. S. Pasareanu, and W. Visser, “Symbolic execution with abstraction,” *International Journal on Software Tools for Technology Transfer (STTT)*, vol. 11, pp. 53–67, January 2009.
- [14] P. Godefroid, “Compositional dynamic test generation,” in *POPL*, 2007, pp. 47–54.
- [15] X. Deng, J. Lee, and Robby, “Bogor/Kiasan: A k-bounded symbolic execution for checking strong heap properties of open systems,” in *ASE '06: Proceedings of the 21st IEEE/ACM International Conference on Automated Software Engineering*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 157–166.
- [16] X. Deng, Robby, and J. Hatcliff, “Towards a case-optimal symbolic execution algorithm for analyzing strong properties of object-oriented programs,” in *SEFM '07: Proceedings of the 5th IEEE International*

Conference on Software Engineering and Formal Methods. Washington, DC, USA: IEEE Computer Society, 2007, pp. 273–282.

- [17] S. Khurshid, C. S. Păsăreanu, and W. Visser, “Generalized symbolic execution for model checking and testing,” in *TACAS*, 2003, pp. 553–568.
- [18] S. O. Wesonga, “Javalite - an operational semantics for modeling Java programs,” Master’s thesis, Brigham Young University, Provo UT, 2012.
- [19] M. Felleisen, D. P. Friedman, E. Kohlbecker, and B. Duba, “A syntactic theory of sequential control,” *Theoretical Computer Science*, vol. 52, no. 3, pp. 205 – 237, 1987.