

Compact Symbolic Execution^{*}

Jiri Slaby, Jan Strejček, and Marek Trtík

Masaryk University, Brno, Czech Republic
{slaby,strejcek,trtik}@fi.muni.cz

Abstract. We present a generalisation of King’s symbolic execution technique called *compact symbolic execution*. It proceeds in two steps. First, we analyse cyclic paths in the control flow graph of a given program, independently from the rest of the program. Our goal is to compute a so called *template* for each such a cyclic path. A template is a declarative parametric description of all possible program states, which may leave the analysed cyclic path after any number of iterations along it. In the second step, we execute the program symbolically with the templates in hand. The result is a *compact* symbolic execution tree. A compact tree always carry the same information in all its leaves as the corresponding classic symbolic execution tree. Nevertheless, a compact tree is typically substantially smaller than the corresponding classic tree. There are even programs for which compact symbolic execution trees are finite while classic symbolic execution trees are infinite.

1 Introduction

Symbolic execution [16, 13] is a program analysis method originally suggested for enhanced testing. While testing runs a program on selected input values, symbolic execution runs the program on symbols that represent arbitrary input values. As a result, symbolic execution explores all execution paths. On one hand-side, this means that symbolic execution does not miss any error. On the other hand-side, symbolic execution applied to real programs hardly ever finishes as programs typically have a huge (or even infinite) number of execution paths. This weakness of symbolic execution is known as *path explosion problem*. The second weakness of symbolic execution comes from the fact that it calls SMT solvers to decide which program paths are feasible and which are not. The SMT queries are often formulae of theories that are hard to decide or even undecidable. Despite the two weaknesses, there are several successful bug-finding tools based on symbolic execution, for example KLEE [7], EXE [8], PEX [22], or SAGE [11].

This paper introduces the *compact symbolic execution* that partly solves the path explosion problem. We build on the observation that one of the main sources of the problem are program cycles. Indeed, many execution paths differ just in numbers of iterations along program cycles. Hence, before we start symbolic execution, we detect cyclic paths in the control flow graph of a given program and we try to find a *template* for each such a cyclic path. A template is a

^{*} The authors are supported by The Czech Science Foundation, grant P202/12/G061.

declarative parametric description (with a single parameter κ) of all possible program states produced by $\kappa \geq 0$ iterations along the cyclic path followed by any execution step leading outside the cyclic path. The target program locations of such execution steps are called *exits* of the cyclic path.

The compact symbolic execution proceeds just like the classic symbolic execution until we enter a cyclic path for which we have a template. Instead of executing the cyclic path, we can apply the template to jump directly to exits of the cyclic path. At each exit, we obtain a program state with a parameter κ . This parametric program state represents all program states reached by execution paths composed of a particular path to the cycle, κ iterations along the cycle, and the execution step leading to the exit. Symbolic execution then continues from these program states in the classic way again.

Hence, compact symbolic execution reduces the path explosion problem as it explores at once all execution paths that differ only in numbers of iterations along the cyclic paths for which we have templates. As we will see later, a price for this reduction comes in deepening the other weakness of symbolic execution: while SMT queries of standard symbolic execution are always quantifier-free, each application of a template adds one universal quantifier to the SMT queries of compact symbolic execution. Although SMT solvers fail to decide quantified queries significantly more often than queries without quantifiers, our experimental results show that this trade-off is acceptable as compact symbolic execution is able to detect more errors in programs than the classic one. Moreover, future advances in SMT solving can make the disadvantage of compact symbolic execution even smaller.

2 Basic Idea

This section presents basic ideas of compact symbolic execution. To illustrate the ideas, we use a simple program represented by the flowgraph of Figure 1(a). The program implements a standard linear search algorithm. It returns the least index i in the array A such that $A[i]=x$. If x is not in A at all, then the result is -1 . In both cases the result is saved in the variable r . Before we describe the compact symbolic execution, we briefly recall the classic symbolic execution [16].

Classic Symbolic Execution Symbolic execution runs a program over *symbols* representing arbitrary input values. For each input variable v , we denote a symbol passed to it as \underline{v} . A *program state* is a triple (l, θ, φ) consisting of a current program location l in the flowgraph, a *symbolic memory* θ , and a *path condition* φ . θ assigns to each program variable its current symbolic value, i.e. an expression over the symbols. For example, if the first instruction of a program is the assignment $i := 2 * n + x$, then $\theta(i) = 2\underline{n} + \underline{x}$ after its execution. The path condition φ is a quantifier-free first order logic formula representing a necessary and sufficient condition on symbols to drive the execution along the currently executed path. φ is initially *true* and it can be updated at program branchings. For example, in a location with two out-edges labelled by $x > n + 5$ and $x \leq n + 5$, we

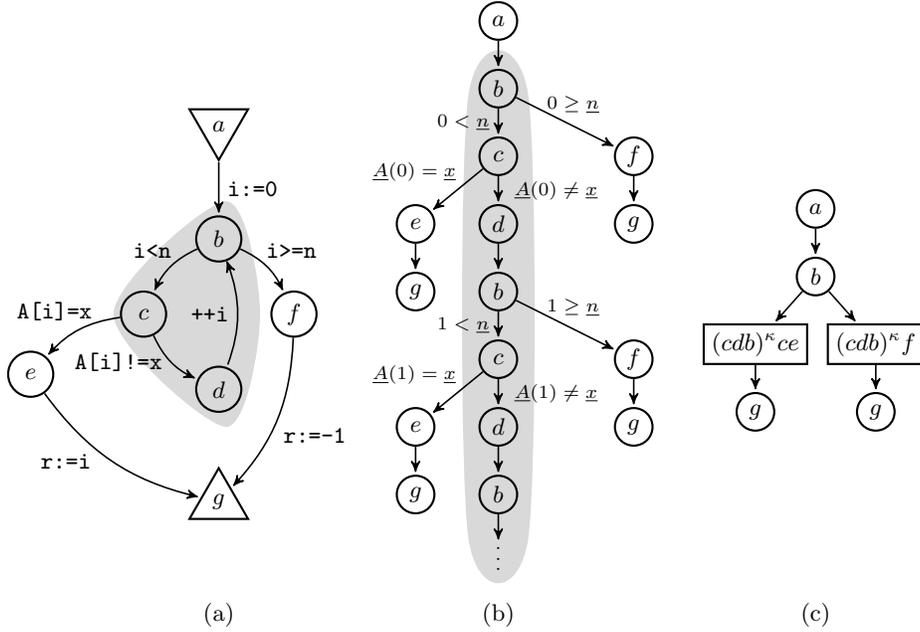


Fig. 1. (a) A flowgraph $\text{linSrch}(A, n, x)$. (b) Classic symbolic execution tree of linSrch . (c) Compact symbolic execution tree of linSrch .

instantiate the conditions with use of the current θ and we check whether the current path condition φ implies their validity. Namely, we ask for validity of implications $\varphi \rightarrow \theta(\mathbf{x}) > \theta(\mathbf{n}) + 5$ and $\varphi \rightarrow \theta(\mathbf{x}) \leq \theta(\mathbf{n}) + 5$. If the first implication is valid, the symbolic execution continues along the first branch. If the second implication is valid, the symbolic execution continues along the second branch. If none of them is valid, it means that we can follow either of the two branches. Hence, the symbolic execution forks in order to execute both branches. In this case, we update the path condition on the first branch to $\varphi \wedge \theta(\mathbf{x}) > \theta(\mathbf{n}) + 5$ and the one on the second branch to $\varphi \wedge \theta(\mathbf{x}) \leq \theta(\mathbf{n}) + 5$. Note that the whole program state is forked into two states in this case.

Due to the forks, symbolic execution is traditionally represented by a tree called *classic symbolic execution tree*. Nodes of the tree are labelled by program states computed during the execution. Edges of the tree correspond to transitions between program states labelling their end nodes. In Figure 1(b), there is a classic symbolic execution tree of the flowgraph from Figure 1(a). For readability of symbolic execution tree figures, nodes are marked only with current program locations instead of full program states. In addition, we label branching edges by instances of the corresponding branching conditions in the flowgraph. These labels allow us to reconstruct the path condition for each node in the tree: it is the conjunction of labels of all edges along the path from the root to the node. Note that contents of symbolic memories are not depicted in the figure.

Overall Effect of Cyclic Paths If we look at the flowgraph of Figure 1(a), we immediately see that locations b, c, d and edges between them form a cyclic path highlighted by a grey region. All executions entering the path (at location b) proceed in the same way: each execution performs κ iterations along the cyclic path (for some $\kappa \geq 0$) and continues either along the edge (b, f) or along the edges (b, c) and (c, e) to leave it. Compact symbolic execution aims to effectively exploit the uniformity of all executions along this cyclic path. To do so, we need to find a unified declarative description of the *effect* of all executions along the cyclic path on a symbolic memory and a path condition. We analyse the cyclic path, together with all the edges allowing to leave it, separately from the rest of the flowgraph. First we introduce symbols for all variables in the isolated part of the program, since they all are now input variables to the part. In our example, we introduce symbols $\underline{n}, \underline{x}, \underline{i}, \underline{A}$ representing the values of the corresponding variables $\mathbf{n}, \mathbf{x}, \mathbf{i}, \mathbf{A}$ at the entry location b , before the first iteration. We emphasise that the introduced symbols do *not* represent inputs to the whole flowgraph, but rather symbolic values of the corresponding variables at the moment of entering the cyclic path at the location b via the edge (a, b) .

Now we study the effect of κ iterations along the cyclic path. One can see that each iteration increases the value of \mathbf{i} by one while values of the other variables keep unchanged. Hence, after κ iterations, the value of \mathbf{i} is $\underline{i} + \kappa$. Formally, the effect of κ iterations of the cycle on values of all variables is described by the following *parametric symbolic memory* $\theta_*[[\kappa]]$ with the parameter κ :

$$\begin{aligned} \theta_*[[\kappa]](\mathbf{n}) &= \underline{n}, & \theta_*[[\kappa]](\mathbf{x}) &= \underline{x}, \\ \theta_*[[\kappa]](\mathbf{i}) &= \underline{i} + \kappa, & \theta_*[[\kappa]](\mathbf{A}) &= \underline{A}. \end{aligned}$$

Further, we formulate a *parametric path condition* $\varphi_*[[\kappa]]$ representing the path condition after κ iterations along the cyclic path. To perform all these κ iterations along the cyclic path, both conditions $\mathbf{i} < \mathbf{n}$ and $\mathbf{A}[\mathbf{i}] \neq \mathbf{x}$ along the path have to be valid in each of κ iterations. Therefore, the path condition after κ iterations has the form

$$\begin{aligned} & \underline{i} < \underline{n} \ \wedge \ \underline{A}(\underline{i}) \neq \underline{x} \ \wedge \\ & \wedge \ \underline{i} + 1 < \underline{n} \ \wedge \ \underline{A}(\underline{i} + 1) \neq \underline{x} \ \wedge \\ & \qquad \qquad \qquad \vdots \\ & \wedge \ \underline{i} + (\kappa - 1) < \underline{n} \ \wedge \ \underline{A}(\underline{i} + (\kappa - 1)) \neq \underline{x}, \end{aligned}$$

where τ -th line, $\tau \in \{0, 1, \dots, \kappa - 1\}$, consists of two predicates which are instances of the conditions $\mathbf{i} < \mathbf{n}$ and $\mathbf{A}[\mathbf{i}] \neq \mathbf{x}$ respectively after τ iterations of the cyclic path, i.e. during the $(\tau + 1)$ -st iteration. Unfortunately, the conjunction above is not a first order formula as its length depends on the parameter κ , whose value can be arbitrary. The conjunction can be equivalently expressed by the following universally quantified formula:

$$\forall \tau (0 \leq \tau < \kappa \rightarrow (\underline{i} + \tau < \underline{n} \ \wedge \ \underline{A}(\underline{i} + \tau) \neq \underline{x})).$$

If we now add to the formula above the obvious fact that we cannot iterate the cyclic path negative number of times (i.e. $\kappa \geq 0$), we get the resulting parametric path condition $\varphi_* \llbracket \kappa \rrbracket$ as

$$\varphi_* \llbracket \kappa \rrbracket = \kappa \geq 0 \wedge \forall \tau (0 \leq \tau < \kappa \rightarrow (\underline{i} + \tau < \underline{n} \wedge \underline{A}(\underline{i} + \tau) \neq \underline{x})).$$

Finally, we use $\theta_* \llbracket \kappa \rrbracket$ and $\varphi_* \llbracket \kappa \rrbracket$ to define symbolic memory $\theta_{bf} \llbracket \kappa \rrbracket$ and path condition $\varphi_{bf} \llbracket \kappa \rrbracket$ describing the effect of κ iterations of the cyclic path followed by leaving it through the edge (b, f) , and similarly $\theta_{ce} \llbracket \kappa \rrbracket, \varphi_{ce} \llbracket \kappa \rrbracket$ with the analogous information for leaving the cyclic path through the edge (c, e) . As the edges $(b, f), (b, c), (c, e)$ do not modify any variable, we immediately get $\theta_{bf} \llbracket \kappa \rrbracket = \theta_{ce} \llbracket \kappa \rrbracket = \theta_* \llbracket \kappa \rrbracket$. Further, $\varphi_{bf} \llbracket \kappa \rrbracket$ and $\varphi_{ce} \llbracket \kappa \rrbracket$ are conjunctions of $\varphi_* \llbracket \kappa \rrbracket$ with the instances of the conditions on the edge (b, f) or on the edges $(b, c), (c, e)$, respectively. Hence, the path conditions $\varphi_{bf} \llbracket \kappa \rrbracket, \varphi_{ce} \llbracket \kappa \rrbracket$ are defined as follows:

$$\begin{aligned} \varphi_{bf} \llbracket \kappa \rrbracket &= \varphi_* \llbracket \kappa \rrbracket \wedge \underline{i} + \kappa \geq \underline{n} \\ \varphi_{ce} \llbracket \kappa \rrbracket &= \varphi_* \llbracket \kappa \rrbracket \wedge \underline{i} + \kappa < \underline{n} \wedge \underline{A}(\underline{i} + \kappa) = \underline{x} \end{aligned}$$

The overall effect of the considered cyclic path with its exit edges is now fully described by a so-called *template* consisting of the entry location b to the cyclic path and two triples $(f, \theta_{bf} \llbracket \kappa \rrbracket, \varphi_{bf} \llbracket \kappa \rrbracket)$ and $(e, \theta_{ce} \llbracket \kappa \rrbracket, \varphi_{ce} \llbracket \kappa \rrbracket)$, one for each exit edge from the cyclic path. Note that the triples have the same structure and meaning as program states in classic symbolic execution. The only difference is that the triples are parametrised by the parameter κ .

Compact Symbolic Execution The template is used during *compact symbolic execution* of the program. The execution starts at the location a of the flowgraph. The *compact symbolic execution tree* initially consists of a single node labelled by the initial state $(a, \theta_I, true)$, where θ_I is the initial symbolic memory assigning to each input variable v the corresponding symbol \underline{v} . Now we execute the instruction $i := 0$ of the flowgraph edge (a, b) using the classic symbolic execution. The tree is extended with a single successor node, say u , labelled with a program state (b, θ', φ') . As we have a template for the location b , we can instantiate it instead of executing the original program. The node u gets one successor for each triple of the template. The triple $(f, \theta_{bf} \llbracket \kappa \rrbracket, \varphi_{bf} \llbracket \kappa \rrbracket)$ generates a successor node labelled by a program state $(f, \theta'_{bf} \llbracket \kappa \rrbracket, \varphi'_{bf} \llbracket \kappa \rrbracket)$. Note that we cannot use $(f, \theta_{bf} \llbracket \kappa \rrbracket, \varphi_{bf} \llbracket \kappa \rrbracket)$ directly as $\theta_{bf} \llbracket \kappa \rrbracket, \varphi_{bf} \llbracket \kappa \rrbracket$ describe executions starting just at the entry location b , while $\theta'_{bf} \llbracket \kappa \rrbracket, \varphi'_{bf} \llbracket \kappa \rrbracket$ have to reflect the effect of the executions starting at a . We create $\theta'_{bf} \llbracket \kappa \rrbracket, \varphi'_{bf} \llbracket \kappa \rrbracket$ by composing $\theta_{bf} \llbracket \kappa \rrbracket, \varphi_{bf} \llbracket \kappa \rrbracket$ with θ', φ' . The composition is precisely described in the following section. In our simple program, θ', φ' reflect only the effect of assignment $i := 0$. Thus, $\theta'_{bf} \llbracket \kappa \rrbracket$ and $\varphi'_{bf} \llbracket \kappa \rrbracket$ equal to $\theta_{bf} \llbracket \kappa \rrbracket$ and $\varphi_{bf} \llbracket \kappa \rrbracket$ respectively, where \underline{i} is replaced by 0. The second triple $(e, \theta_{ce} \llbracket \kappa \rrbracket, \varphi_{ce} \llbracket \kappa \rrbracket)$ of the template generates the successor node labelled with a program state $(e, \theta'_{ce} \llbracket \kappa \rrbracket, \varphi'_{ce} \llbracket \kappa \rrbracket)$ computed analogously using the composition. The symbolic execution then continues from the locations f and e in parallel using the classic symbolic execution. The resulting compact symbolic

execution tree is depicted in Figure 1(c). Observe that the two nodes introduced during template instantiation are drawn with different shape than the others. Moreover, labels of these nodes immediately indicate all paths in the flowgraph whose execution is replaced by the application of the template.

If we compare trees at Figures 1(b) and 1(c), we immediately see that the compact tree is much smaller than the classic one. In particular, the infinite path in the classic tree (highlighted by the grey region) does not appear in the compact one. However, both trees keep the same information in all their leaves. For example, the program state of the left leaf of the compact tree contains the following path condition

$$\varphi[\kappa] = \kappa \geq 0 \wedge \forall \tau (0 \leq \tau < \kappa \rightarrow (\tau < \underline{n} \wedge \underline{A}(\tau) \neq \underline{x})) \wedge \kappa < \underline{n} \wedge \underline{A}(\kappa) = \underline{x}.$$

Let us mark all leaves on the left-hand side of the classic tree as g_0, g_1, g_2, \dots and let $\varphi_0, \varphi_1, \varphi_2, \dots$ be the corresponding path conditions (remember, that each φ_i is the conjunction of labels along the corresponding paths in the tree) and check that φ_i is equivalent to $\varphi[[i]]$ for each $i \geq 0$. For example, for $i = 1$ we have

$$\begin{aligned} \varphi_1 &= 0 < \underline{n} \wedge \underline{A}(0) \neq \underline{x} \wedge 1 < \underline{n} \wedge \underline{A}(1) = \underline{x}, \\ \varphi[[1]] &= 1 \geq 0 \wedge \forall \tau (0 \leq \tau < 1 \rightarrow (\tau < \underline{n} \wedge \underline{A}(\tau) \neq \underline{x})) \wedge 1 < \underline{n} \wedge \underline{A}(1) = \underline{x}, \end{aligned}$$

and hence $\varphi_1 \equiv \varphi[[1]]$. Similarly, each symbolic memory of a node g_i is an instance $\theta[[i]]$ of the parametrized symbolic memory in the left leaf of the compact tree. Analogous relations hold for leaves on the right-hand sides of the compact and the classic symbolic execution trees.

3 Description of the Technique

This section describes the compact symbolic execution in details. For simplicity, we consider only programs represented by a single flowgraph manipulating integer variables and read-only integer arrays. The technique can be extended to handle mutable integer arrays, other data types, and function calls.

3.1 Preliminaries

Besides the terms and notation introduced in the previous section, we use also the following terms and notation.

We write $\theta[\kappa]$ to emphasise that κ is the set of parameters appearing in the symbolic memory θ . Similarly, we write $\varphi[\kappa]$ to emphasise that κ is the set of parameters with free occurrences in the formula φ . We also write $s[\kappa]$ or $(l, \theta, \varphi)[\kappa]$, if $s = (l, \theta[\kappa], \varphi[\kappa])$.

A *valuation* of parameters is a function ν from a finite set of parameters to non-negative integers. By $\theta[\nu]$, $\varphi[\nu]$, and $s[\nu]$ we denote a symbolic memory $\theta[\kappa]$, a formula $\varphi[\kappa]$, and a program state $s[\kappa]$ respectively, where all free occurrences of each $\kappa \in \kappa$ are replaced by $\nu(\kappa)$. If $\kappa = \{\kappa\}$ is a singleton

and $\nu(\kappa) = \nu$, we simply write $\theta[\kappa], \varphi[\kappa], s[\kappa]$ instead of $\theta[\llbracket\kappa\rrbracket], \varphi[\llbracket\kappa\rrbracket], s[\llbracket\kappa\rrbracket]$ and $\theta[\nu], \varphi[\nu], s[\nu]$ instead of $\theta[\llbracket\nu\rrbracket], \varphi[\llbracket\nu\rrbracket], s[\llbracket\nu\rrbracket]$.

If θ is a symbolic memory and φ is a formula or a symbolic expression, then $\theta\langle\varphi\rangle$ denotes φ where all occurrences of all symbols \underline{a} are simultaneously replaced by $\theta(\mathbf{a})$, i.e. by the value of the corresponding variable stored in θ .

When θ_1 and θ_2 are two symbolic memories, then $\theta_1 \diamond \theta_2$ is a *composed symbolic memory* satisfying $(\theta_1 \diamond \theta_2)(\mathbf{a}) = \theta_1\langle\theta_2(\mathbf{a})\rangle$ for each variable \mathbf{a} . Intuitively, the symbolic memory $\theta_1 \diamond \theta_2$ represents an overall effect of a code with effect θ_1 followed by a code with effect θ_2 .

We define *composition of states* $s_1 = (l_1, \theta_1, \varphi_1)$ and $s_2 = (l_2, \theta_2, \varphi_2)$ to be the state $s_1 \diamond s_2 = (l_2, \theta_1 \diamond \theta_2, \varphi_1 \wedge \theta_1\langle\varphi_2\rangle)$. The composed state corresponds to the symbolic state resulting from symbolic execution of the code that produced s_1 immediately followed by the code that produced s_2 .

We often use a dot-notation to denote elements of a program state s : $s.l$ denotes its current location, $s.\theta$ denotes its symbolic memory, and $s.\varphi$ denotes its path condition. Further, if u is a node of a symbolic execution tree, then $u.s$ denotes the program state labelling u and we write $u.l$, $u.\theta$, and $u.\varphi$ instead of $(u.s).l$, $(u.s).\theta$, and $(u.s).\varphi$.

Two program states s_1, s_2 are *equivalent*, written $s_1 \equiv s_2$, if $s_1.l = s_2.l$, the formula $s_1.\theta(\mathbf{a}) = s_2.\theta(\mathbf{a})$ holds for each variable \mathbf{a} , and the formulae $s_1.\varphi$ and $s_2.\varphi$ are equivalent in the logical sense.

Considered integer programs operate in undecidable theories (like Peano arithmetic). We assume that there is a function `satisfiable`(φ) that returns SAT if it can prove satisfiability of φ , UNSAT if it can prove unsatisfiability of φ , and UNKNOWN otherwise.

3.2 Templates and Their Computation

We start with a formal definition of *cycle*, i.e. a cyclic path with a specified entry location and exit edges.

Definition 1 (Cycle) *Let (u, e) be an edge of a flowgraph P , $\pi = e\omega e$ be a cyclic path in P such that ue is not a suffix of π and all nodes in ωe are pairwise distinct, and let $X = \{(u_1, x_1), \dots, (u_n, x_n)\}$ be the set of all edges of P that do not belong to the path π , but their start nodes u_1, \dots, u_n lie on π . Then $C = (\pi, e, X)$ is a cycle in P , the path π is a core of C , e is an entry location of C , all edges in X are exit edges of C , and each location x_i is called an exit location of C .*

We emphasise that the core of a cycle is a cyclic path in a graph sense. Note that a program loop can generate more independent cycles, e.g. if the loop contains internal branching or loop nesting (see [20] for more details).

A template for a cycle (π, e, X) is a pair (e, M) , where M is a set containing one parametric program state for each exit edge of the cycle. A template for a given cycle can be computed by Algorithm 1. The algorithm uses a function `executePath`(P, ρ) which applies classic symbolic execution to instructions on

Algorithm 1: computeTemplate

Input: a program P and a cycle (π, e, X)
Output: a template (e, M) or **null** (if the computation fails)

- 1 $(e, \theta, \varphi) \leftarrow \text{executePath}(P, \pi)$
- 2 **if** $\text{satisfiable}(\varphi) \neq \text{SAT}$ **then return null**
- 3 Set $\theta_* \llbracket \kappa \rrbracket(\mathbf{a}) = \underline{a}$ for each array variable \mathbf{a}
- 4 Set $\theta_* \llbracket \kappa \rrbracket(\mathbf{a}) = \perp$ for each integer variable \mathbf{a}
- 5 **repeat**
- 6 $\text{change} \leftarrow \text{false}$
- 7 **foreach** integer variable \mathbf{a} **do**
- 8 **if** $\theta_* \llbracket \kappa \rrbracket(\mathbf{a}) = \perp$ **then**
- 9 **if** $\theta(\mathbf{a}) = \underline{a} + c$ for some constant c **then**
- 10 $\theta_* \llbracket \kappa \rrbracket(\mathbf{a}) \leftarrow \underline{a} + \kappa \cdot c$
- 11 $\text{change} \leftarrow \text{true}$
- 12 **if** $\theta(\mathbf{a}) = \underline{a} \cdot c$ for some constant c **then**
- 13 $\theta_* \llbracket \kappa \rrbracket(\mathbf{a}) \leftarrow \underline{a} \cdot c^\kappa$
- 14 $\text{change} \leftarrow \text{true}$
- 15 **if** $\theta(\mathbf{a}) = g$ for some symbolic expression g such
 that $\theta_* \llbracket \kappa \rrbracket(\mathbf{b}) \neq \perp$ for each symbol \mathbf{b} in g **then**
- 16 $\theta_* \llbracket \kappa \rrbracket(\mathbf{a}) \leftarrow \text{ite}(\kappa > 0, \theta_* \llbracket \kappa - 1 \rrbracket(g), \underline{a})$
- 17 $\text{change} \leftarrow \text{true}$
- 18 **until** $\text{change} = \text{false}$
- 19 **if** $\theta_* \llbracket \kappa \rrbracket(\mathbf{a}) = \perp$ for some variable \mathbf{a} **then return null**
- 20 $\varphi_* \llbracket \kappa \rrbracket \leftarrow \kappa \geq 0 \wedge \forall \tau (0 \leq \tau < \kappa \implies \theta_* \llbracket \tau \rrbracket(\varphi))$
- 21 $M \leftarrow \emptyset$
- 22 **foreach** $(u, x) \in X$ **do**
- 23 Let ρ be the prefix of π from e to u
- 24 $(x, \theta, \varphi) \leftarrow \text{executePath}(P, \rho x)$
- 25 **if** $\text{satisfiable}(\varphi) = \text{UNKNOWN}$ **then return null**
- 26 **if** $\text{satisfiable}(\varphi) = \text{SAT}$ **then**
- 27 $M \leftarrow M \cup \{(x, \theta_* \llbracket \kappa \rrbracket \diamond \theta, \varphi_* \llbracket \kappa \rrbracket \wedge \theta_* \llbracket \kappa \rrbracket(\varphi))\}$
- 28 **return** (e, M)

the path ρ in the program P and returns the resulting symbolic state (u, θ, φ) , where u is the last location in ρ .

The first part of the algorithm (lines 1–20) tries to derive a parametric symbolic memory $\theta_* \llbracket \kappa \rrbracket$ and a parametric path condition $\varphi_* \llbracket \kappa \rrbracket$, which together describe the symbolic state after κ iterations over the core π of the cycle C , for any $\kappa \geq 0$. Initially, at line 1, we compute the effect of a *single* iteration of the core π and then we check whether the iteration is feasible. If we cannot prove its feasibility, we stop the template computation and return **null**.¹ Otherwise,

¹ It is possible that the iteration is feasible and the chosen SMT solver failed to prove it. However, as parametric path conditions of the resulting template are derived from φ , it is highly probable that any application of the template in compact symbolic execution would lead to failures of the SMT solver. Such a template is useless.

we get a symbolic state (e, θ, φ) , whose elements θ and φ form a basis for the computation of $\theta_*[\kappa]$ and $\varphi_*[\kappa]$.

We compute $\theta_*[\kappa]$ first. As arrays are read-only, we directly set $\theta_*[\kappa](\mathbf{a})$ to \underline{a} for each array variable \mathbf{a} . For integer variables, we initialise $\theta_*[\kappa]$ to an undefined value \perp . Then, in the loop at lines 5–18, we try to define $\theta_*[\kappa]$ for as many variables as possible. For each variable \mathbf{a} , $\theta_*[\kappa](\mathbf{a})$ is defined at most once. Hence, the loop terminates after finite number of iterations. The value of $\theta_*[\kappa](\mathbf{a})$ is defined according to the content of $\theta(\mathbf{a})$ and known values of $\theta_*[\kappa]$. In particular, the conditions at lines 9 and 12 check if the values of \mathbf{a} follow an arithmetic or a geometric progression during the iterations. If they do, we can easily express the exact value of \mathbf{a} after any κ iterations. Note that the case when the value of \mathbf{a} variable is not changed along π at all is a special case of an arithmetic progression ($c = 0$). Obviously, one can add support for other kinds of progression. The condition at line 15 covers the case when each iteration assigns to \mathbf{a} an expression containing only variables with known values of $\theta_*[\kappa]$. The if-then-else expression **ite** $(\kappa > 0, \theta_*[\kappa - 1](g), \underline{a})$ assigned to $\theta_*[\kappa](\mathbf{a})$ says that the value of \mathbf{a} after $\kappa > 0$ iterations is given by the value of expression g where each symbol \underline{b} represents the value of \mathbf{b} at the beginning of the last iteration and thus it must be replaced by $\theta_*[\kappa - 1](\mathbf{b})$. The value of \mathbf{a} after 0 iterations is obviously unchanged, i.e. \underline{a} .

Once we get to line 19, we check whether we succeeded to define $\theta_*[\kappa]$ for all variables. If we failed for at least one variable, then we fail to compute a template for C and we return **null**. Otherwise, at line 20 we compute the formula $\varphi_*[\kappa]$ in accordance with the intuition provided in Section 2.

The second part of the algorithm (lines 21–28) computes the set M of the resulting template. As we already know from Section 2, we try to compute one element of M for each exit edge $(u, x) \in X$. At line 23 we compute a path ρ from the entry location e to u (along π), where we escape from π to the location x . The path ρx is then symbolically executed. If we fail to decide feasibility of the path, we fail to compute a template. If the path is feasible, we can escape π by taking the exit edge (u, x) . Therefore, only in this case we add a new element to M at line 27. The structure of the element follows the intuition given in Section 2.

One can immediately see that the algorithm always terminates. Now we formulate a theorem describing properties of the computed template (e, M) . The theorem is crucial for proving soundness and completeness of compact symbolic execution. Roughly speaking, the theorem says that whenever a node u of the symbolic execution tree of a program P satisfies $u.l = e$, then the subtree rooted in u has the property that each branch to a leaf contains a node w such that $w.s$ corresponds to the composition of $u.s$ and a suitable instance of some program state of the template (L1), and vice versa (L2). A proof of the theorem can be found in the full version of this paper [20].

Theorem 1 (Template Properties) *Let T be a classic symbolic execution tree of P and let $(e, \{(l_1, \theta_1[\kappa], \varphi_1[\kappa]), \dots, (l_n, \theta_n[\kappa], \varphi_n[\kappa])\})$ be a template for a cycle (π, e, X) in P produced by Algorithm 1. Then the following two properties hold:*

- (L1) For each path $\pi = u\omega$ in T leading from a node u satisfying $u.l = e$ to a leaf, there is a node w of ω , an index $i \in \{1, \dots, n\}$, and an integer $\nu \geq 0$ such that $w.s \equiv u.s \diamond (l_i, \theta_i[\nu], \varphi_i[\nu])$.
- (L2) For each node u of T , an index $i \in \{1, \dots, n\}$, and an integer $\nu \geq 0$ such that $u.l = e$ and $(u.\varphi \wedge u.\theta(\varphi_i[\nu]))$ is satisfiable, there is a successor w of u in T such that $w.s \equiv u.s \diamond (l_i, \theta_i[\nu], \varphi_i[\nu])$.

3.3 Compact Symbolic Execution

The compact symbolic execution is formally defined by Algorithm 2. If we ignore the lines marked by \square , then we get the classic symbolic execution. As we focus on compact symbolic execution, we describe the algorithm with \square lines included. The algorithm gets a program P and a finite set p of templates resulting from analyses of some cycles in P . Lines 1–3 create an initial program state, insert it into a queue Q , and create the root of a symbolic execution tree T labelled by the state.

The queue Q keeps all the program states waiting for their processing in the **repeat-until** loop (lines 4–26). The key part of the loop’s body begins at line 9, where we select at most one template of p with entry location matching the actual program location $s.l$. Note that there can be more than one template available at $s.l$ as more cyclic paths can go through the location. We do not put any constraints in the selection strategy. We may for example choose randomly. Also note that we may choose none of the templates (i.e. we select **null**), if there is no template in p for location $s.l$ or even if there are such templates in p . If a template $t = (s.l, M)$ is selected, then we get a fresh parameter (line 12) and replace the original parameter in all tuples of M by the fresh one. This replacement prevents collisions of parameters of already applied templates. The **foreach** loop at lines 14–16 creates a successor state s' for each program state in M . If the template selection at line 9 returns **null**, we proceed to line 18 and compute successor states of the state s by the classic symbolic execution. The successor states with provably satisfiable path conditions are inserted into the queue Q and into the compact symbolic execution tree T in the **foreach** loop at lines 20–22. The successor states with provably unsatisfiable path conditions are ignored as they correspond to infeasible paths. The **foreach** loop at lines 23–25 handles the successor states with path conditions for which we are unable to decide satisfiability; these states are inserted into the resulting tree T as so-called *failed leaves*. A presence of a failed leaf in the resulting tree indicates that applied symbolic execution has failed to explore whole path-space of the executed program. We do not continue computation from these states as there is usually a plethora of other states with provably satisfiable path conditions.

We finish this section by soundness and completeness theorems for compact symbolic execution. We assume that T and T' are classic and compact symbolic execution trees of the program P computed by Algorithm 2 without and with \square -lines respectively. The theorems hold on assumption that our **satisfiable**(φ) function never returns UNKNOWN, i.e. neither T nor T' contains failed leaves. Proofs of both theorems can be found in the full version of this paper [20].

Algorithm 2: executeSymbolically

Input: a program P to be executed
□ and a finite set p of templates computed for cycles in P
Output: a symbolic execution tree T of P (compact tree in □-version)

- 1 $s_0 \leftarrow$ (the starting location of $P, \theta_I, true$)
- 2 Let Q be a queue of states initially containing only s_0
- 3 Insert the root node labelled by s_0 to the empty tree T
- 4 **repeat**
- 5 Extract the first state s from Q
- 6 **if** $s.l$ is either an exit from P or an error location **then**
- 7 **continue**
- 8 $S \leftarrow \emptyset$
- 9 $t \leftarrow \text{chooseTemplate}(s.l, p)$
- 10 **if** $t \neq \text{null}$ **then**
- 11 Let M be the second element of t , i.e. $t = (s.l, M)$
- 12 $\kappa \leftarrow \text{getFreshParam}()$
- 13 Replace all occurrences of the former parameter in M by κ
- 14 **foreach** $(l, \theta[\kappa], \varphi[\kappa]) \in M$ **do**
- 15 $s' \leftarrow s \diamond (l, \theta[\kappa], \varphi[\kappa])$
- 16 Insert s' into S
- 17 **else** /* apply classic symbolic execution step */
- 18 $S \leftarrow \text{computeClassicSuccessors}(P, s)$
- 19 Let u be the leaf of T whose label is s
- 20 **foreach** state $s' \in S$ such that $\text{satisfiable}(s'.\varphi) = \text{SAT}$ **do**
- 21 Insert s' at the end of Q
- 22 Insert a new node v labelled with s' and a new edge (u, v) into T
- 23 **foreach** state $s' \in S$ such that $\text{satisfiable}(s'.\varphi) = \text{UNKNOWN}$ **do**
- 24 Insert a new node v labelled with s' and a new edge (u, v) into T
- 25 Mark the node v in T as a failed leaf
- 26 **until** Q becomes empty
- 27 **return** T

Theorem 2 (Soundness) *For each leaf node $e \in T$ there is a leaf node $e' \in T'$ and a valuation ν of parameters in $e'.s$ such that $e.s \equiv e'.s[\nu]$.*

Theorem 3 (Completeness) *For each leaf node $e' \in T'$ there is a leaf node $e \in T$ and a valuation ν of parameters in $e'.s$ such that $e.s \equiv e'.s[\nu]$.*

Note that in both theorems we discuss only the relation between all *finite* branches of the trees T and T' . Some infinite branches of T (like the one in Figure 1(b)) corresponding to infinite iterations along a cyclic path need not be present in T' . As symbolic execution is typically used to cover as many reachable program locations as possible, missing infinite iterations along cyclic paths can be seen as a feature rather than a drawback.

4 Experimental Results

Implementation We have implemented both classic and compact symbolic execution in an experimental tool called RUDLA. The tool uses our “library of libraries” called BUGST available at SOURCEFORGE [3]. The sources of RUDLA and all benchmarks mentioned below are available in the same repository. The implementation also uses CLANG 2.9 [4], LLVM 3.1 [5], and Z3 4.3.0 [6].

Evaluation Criteria We would like to empirically evaluate and compare the effectiveness of the classic and compact symbolic execution in exploration of program paths. Unfortunately, we cannot directly compare explored program paths or nodes in the constructed trees as a path or a node in a compact symbolic execution tree have a different meaning than a path or a node in a classic symbolic execution tree. To compare the techniques, we fix an exploration method of the trees, namely we choose the breadth-first search as indicated in Algorithm 2, and we measure the time needed by each of the techniques to reach a particular location in an analysed program. Note that for compact symbolic execution we also have to fix a strategy for template selection since there can generally be more than one template related to one program location. We always choose randomly between candidate templates.

Benchmarks and Results We use two collections of benchmarks. The first collection contains 13 programs with a marked target location. As our technique is focused on path explosion caused by loops, all the benchmarks contain typical program loop constructions. There are sequences of loops, nested loops and also loops with internal branching. They are designed to produce a huge number of execution paths. Thus they are challenging for symbolic execution. The target location is chosen to be difficult to reach. The first ten benchmarks have reachable target locations, while the last three do not. For these three benchmarks, all the execution paths must be explored to give an answer.

Experimental results are presented in Table 1. The high numbers of (often infeasible) cycles are due to our translation from LLVM (see [20] for details). We want to highlight the following observations. First, classic symbolic execution was faster only for benchmarks `Hello` and `decode_packets`. Second, the number of states visited by the compact symbolic execution is often several orders of magnitude lower than the number of states visited by the classic one. At the same time we recall that the semantics of a state in classic and compact symbolic execution are different. Finally, presence of quantifiers in path conditions of compact symbolic executions puts high requirements on skills of the SMT solver. This leads to SMT failures, which are not seen in classic symbolic execution.

Algorithm 2 saves SMT failures in the form of failed leaves in the resulting compact symbolic execution tree. Therefore, we may think about subsequent analyses for these leaves. For example, in a failed leaf we may instantiate parameters κ by concrete numbers. The resulting formulae will become quantifier-free and therefore potentially easier for an SMT solver. This way we might be able

Benchmark	Templates			Compact SE			SE	
	Time	Count	Cycles	Time	States	SMTFail	Time	States
hello	12.3	2	126	2.3	187	0	4.5	2262
HW	31.9	4	252	45.4	1048	4	T/O	223823
HWM	48.1	5	336	T/O	5125	24	T/O	162535
matrIR	4.2	4	28	82.9	1234	6	T/O	270737
matrIR_dyn	14.8	5	30	240.5	2472	13	T/O	267636
VM	8.6	6	64	T/O	2274	64	T/O	205577
VMS	4.2	3	32	5.4	466	0	99.8	281263
decode_packets	18.3	5	26	39.9	1276	0	16.3	8992
WinDriver	17.8	5	26	59.2	1370	1	T/O	206903
EQCNT	12.2	3	12	10.6	345	0	T/O	179803
EQCNTex	5.8	4	24	T/O	10581	0	T/O	251061
OneLoop	0.1	1	2	0.1	41	0	T/O	38230
TwoLoops	0.3	2	4	0.1	25	0	T/O	917343
Total time	240			1800			3900	

Table 1. Experimental results of compact and classic symbolic executions. The compact symbolic execution approach is divided into computation of templates and building of compact symbolic execution tree. All the times are in seconds, where 'T/O' identifies exceeding 5 minutes timeout. 'Count' represents the number of computed templates, 'Cycles' shows the number of detected cycles. 'SMTFail' represents the number of failed SMT queries. There was no SMT failure during classic SE of our benchmarks.

	Time	safe	unsafe	timeout	unsupported	points
Compact SE	300+4920	21	25	15	13+5	67
SE	8700	10	27	28	13+1	47

Table 2. Experimental results of compact and classic symbolic executions on 79 SV-COMP 2013 benchmarks in the category 'loops'. Time is in seconds. For compact SE we provide template computation time plus execution time. 'safe' and 'unsafe' report the numbers of programs where the tool decides unreachability and reachability of a marked error location, respectively (all these answers are correct). 'timeout' presents the number of symbolic executions exceeding 5 minutes. 'unsupported' represents the number of compilation failures plus failures during an analysis. 'points' shows the number of points the tools would get according to the SV-COMP 2013 rules.

to explore paths below the failed leaves. But basically, analyses of failed leaves are a topic for our further research. Moreover, as SMT solvers are improving quickly, we may expect that counts of the failures will decrease over time.

The second collection of benchmarks is the whole category 'loops' taken from SV-COMP 2013 (revision 229) [2]. The results are depicted in Table 2.

All the presented experiments were done on a laptop Acer Aspire 5920G (2 × 2GHz, 2GB) running Windows 7 SP1 64-bit.

5 Related Work

The symbolic execution was introduced by King in 1976 [16]. The original concept was generalised in [14] for programs with heap by introducing lazy initialisation of dynamically allocated data structures. The lazy initialisation algorithm was further improved and formally defined in [9]. Another generalisation step was done in [15], where the authors attempt to avoid symbolic execution of library code (called from an analysed program), since such code can be assumed as well defined and properly tested.

In [19, 12], the path explosion problem is tackled by focusing on program loops. The information inferred from a loop allows to talk about multiple program paths through that loop. But the goal is to explore classic symbolic execution tree in some effective manner: more interesting paths sooner. Approaches [10, 1] share the same goal as the previous ones, but they focus on a computation of function summaries rather than on program loops.

Our goal is completely different: instead of guiding exploration of paths in a classic symbolic execution tree, we build a tree that keeps the same information and contains less nodes. In particular, templates of compact symbolic execution have a different objective than summarisation used in [10, 1, 12]. While summarisation basically caches results of some finite part of symbolic execution for later fast reuse, our templates are supposed to *replace* potentially infinite parts of symbolic executions by a single node.

Techniques [17, 18] group paths of classic symbolic execution tree according to their effect on symbolic values of a priori given output variables, and explore only one path per group. We consider all program variables and we explore all program paths (some of them are explored simultaneously using templates).

Finally, in our previous work [21] we compute a non-trivial necessary condition for reaching a given target location in a given program. In other words, the result of the analysis is a first order logic formula. In the current paper, we focus on a fast exploration of as many execution paths as possible. The technique produces a compact symbolic execution tree. Note that, we do not require any target location, since we do not focus on a program location reachability here. Nevertheless, to achieve our goal, we adopted a part of a technical stuff introduced in [21]. Namely, lines 4–18 of Algorithm 1 are similar to the computation of a so-called iterated memory, which is in [21] an over-approximation of the memory content after several iterations in a program loop. In the current technique, the memory content must always be absolutely precise. Moreover, here we analyse flowgraph cycles while [21] summarises program loops.

6 Conclusion

We have introduced a generalisation of classic symbolic execution, called compact symbolic execution. Before building symbolic execution tree, the compact symbolic execution computes *templates* for cycles of an analysed program. A template is a parametric and declarative description of the overall effect of a

related cycle. Our experimental results indicate that the use of templates during the analysis leads to faster exploration of program paths in comparison with the exploration speed of classic symbolic execution. Also a number of symbolic states computed during the program analysis is considerably smaller. On the other hand, compact symbolic execution constructs path conditions with quantifiers, which leads to more failures of SMT queries.

References

1. S. Anand, P. Godefroid, and N. Tillmann. Demand-driven compositional symbolic execution. In *TACAS*, volume 4963 of *LNCS*, pages 367–381. Springer, 2008.
2. D. Beyer. Second competition on software verification. In *TACAS*, volume 7795 of *LNCS*, pages 594–609. Springer, 2013.
3. bugst. <http://sourceforge.net/projects/bugst>.
4. clang. <http://clang.llvm.org>.
5. LLVM. <http://llvm.org>.
6. Z3. <http://z3.codeplex.com>.
7. C. Cadar, D. Dunbar, and D. Engler. KLEE: Unassisted and automatic generation of high-coverage tests for complex systems programs. In *OSDI*, pages 209–224. USENIX Association, 2008.
8. C. Cadar, V. Ganesh, P. M. Pawlowski, D. L. Dill, and D. R. Engler. EXE: Automatically generating inputs of death. In *CCS*, pages 322–335. ACM, 2006.
9. X. Deng, J. Lee, and Robby. Efficient and formal generalized symbolic execution. *Autom. Softw. Eng.*, 19(3):233–301, 2012.
10. P. Godefroid. Compositional dynamic test generation. In *POPL*, pages 47–54. ACM, 2007.
11. P. Godefroid, M. Y. Levin, and D. A. Molnar. Automated whitebox fuzz testing. In *NDSS*, pages 151–166. The Internet Society, 2008.
12. P. Godefroid and D. Luchaup. Automatic partial loop summarization in dynamic test generation. In *ISSTA*, pages 23–33. ACM, 2011.
13. W. E. Howden. Symbolic testing and the DISSECT symbolic evaluation system. *IEEE Trans. Software Eng.*, 3:266–278, 1977.
14. S. Khurshid, C. S. Păsăreanu, and W. Visser. Generalized symbolic execution for model checking and testing. In *TACAS*, volume 2619 of *LNCS*, pages 553–568. Springer, 2003.
15. S. Khurshid and Y. L. Suen. Generalizing symbolic execution to library classes. In *PASTE*, pages 103–110. ACM, 2005.
16. J. C. King. Symbolic execution and program testing. *Commun. ACM*, 19(7):385–394, 1976.
17. D. Qi, H. D. T. Nguyen, and A. Roychoudhury. Path exploration based on symbolic output. In *ESEC/FSE*, pages 278–288. ACM, 2011.
18. R. A. Santelices and M. J. Harrold. Exploiting program dependencies for scalable multiple-path symbolic execution. In *ISSTA*, pages 195–206. ACM, 2010.
19. P. Saxena, P. Poosankam, S. McCamant, and D. Song. Loop-extended symbolic execution on binary programs. In *ISSTA*, pages 225–236. ACM, 2009.
20. J. Slaby, J. Strejček, and M. Trtík. Compact symbolic execution. *CoRR*, abs/1201.4715, 2012. Full version of the ATVA 2013 paper.
21. J. Strejček and M. Trtík. Abstracting path conditions. In *ISSTA*, pages 155–165. ACM, 2012.

22. N. Tillmann and J. de Halleux. Pex – white box test generation for .NET. In *TAP*, volume 4966 of *LNCS*, pages 134–153. Springer, 2008.