Automatic Data Structure Repair using Separation Logic

Guolong Zheng† Quang Loc Le‡ ThanhVu Nguyen† Quoc-Sang Phan‡
† University of Nebraska-Lincoln {gzheng, tonguyen}@cse.unl.edu
‡ Teesside University q.le@tees.ac.uk
‡ Fujitsu Labs. of America sphan@us.fujitsu.com

ABSTRACT
Software systems are often shipped and deployed with both known and unknown bugs. On-the-fly program repairs, which handle runtime errors and allow programs to continue successfully, can help software reliability, e.g., by dealing with inconsistent or corrupted data without interrupting the running program.

We report on our work-in-progress that repairs data structure using separation logic. Our technique, inspired by existing works on specification-based repair, takes as input a specification written in a separation logic formula and a concrete data structure that fails that specification, and performs on-the-fly repair to make the data conforms with the specification.

The use of separation logic allows us to compactly and precisely represent desired properties of data structures and use existing analyses in separation logic to detect and repair bugs in complex data structures.

We have developed a prototype, called STARFIX, to repair invalid Java data structures violating given specifications in separation logic. Preliminary results show that tool can efficiently detect and repair inconsistent data structures including lists and trees.

1. INTRODUCTION
Software systems are often shipped and deployed with both known and unknown bugs [9]. Many automatic program repair approaches [16,21,23] focus on offline repairs, e.g., they analyze the program code, generate fixes, patch the program, and then recompile and run the patched program. In contrast, on-the-fly repairs [4,22,25] repair a running program, allowing it to recover from errors and continue to run. Although offline repair, e.g., halt a failed program to repair, is ideal in many situations, the ability to handle and recover from runtime errors on-the-fly can improve software reliability considerably, e.g., in situations involving critical systems that cannot be interrupted or dealing with inconsistent/corrupted files [8].

Specification-based repair is a popular on-the-fly repair approach focusing on data structures, such as lists or trees [3,14,24,26]. This approach takes as inputs a specification (e.g., a rep0k predicate [9] or a first-order logic formula) that encodes the required properties of data structures and a program state that fails that specification, and then creates a new program state to satisfy the specification. For example, given a predicate that checks for valid linked lists, this approach can automatically fix bad program states due to corrupted input lists (e.g., by changing some next field, which originally points to an incorrect node, to point to a correct node), thus allowing the program to continue its run correctly.

Our work is inspired by specification-based repair and also focuses on dynamically-allocated data structures (e.g., those created via the new keyword in Java or C and stored in the memory heap). However, instead of using predicates or formulas in first-order logic, we represent desired properties of data structures using heap predicates in separation logic (SL) [11,20]. SL, which extends classical logic, allows for compact and precise representations of heap-based program semantics and reasoning to be localized to small portions of memory. There are several benefits of using SL specifications: (i) SL formulae are specifically designed to describe memory shape properties, which can be difficult to express using predicate or first-order logic formula, (ii) SL heap predicates naturally and succinctly encode the recursive structures of commonly-used and standard data structures (such as lists and trees), and (iii) we can use existing SL analyses such as model checking [4] and predicate unrolling [17,19] to check for bugs and perform on-the-fly repairs.

We present a new technique to repair data structures violating SL specifications. Given an inductive SL predicate and a concrete data structure input, we first check that the data conforms to the given predicate by iteratively unfolding and matching the predicate with the concrete data structures. Checking allows us to both detect bugs (indicated by unmatched results) and localizing faults (identifying unmatched parts of the data). Next, we analyze unmatched results to generate a candidate fix, e.g., computing a new value for an unmatched field, and recheck if the modified data satisfies the SL predicate. If it does not, we backtrack and compute a new fix. The algorithm can generate multiple valid fixes with respect to the given SL specification.

We have developed a repair prototype, called STARFIX, to repair invalid inputs violating given SL predicates. STARFIX is implemented in Java and is designed to check for inconsistent Java data structures. More specifically, STARFIX uses Java StarFinder [2], an extension of the Java Pathfinder platform [1] that supports SL assertions. Currently, STARFIX can detect and repair corrupted data structures including lists and trees.

2. APPROACH AT A GLANCE
Given a specification in SL and a concrete data structure, STARFIX checks whether the data satisfies the predicate. If it is not the case, STARFIX modifies the data to satisfy the predicate. More specifically, STARFIX uses an iterative algorithm that unfolds and matches the heap predicate to the concrete data to find inconsistencies and repairs. STARFIX stops when it reaches a search limit (e.g., defined by the user) or has explored all possible matching attempts.

Heap Predicates. We use SL heap predicates to represent data structures. For example, the following predicates dll and list define circular doubly-linked lists:

\[
\text{pred dll}(head) \equiv (\text{emp} \land \text{head} = \text{null}) \lor (\exists p, n, \text{head} \rightarrow \text{Node}(p, n) \rightarrow \text{dll}(head, p, head, n))
\]
\[
\text{pred list}(h, \text{prev}, \text{cur}, \text{next}) \equiv (\text{emp} \land \text{prev} = \text{cur} = \text{next} = h) \lor (\exists n, \text{next} \rightarrow \text{Node}(\text{cur}, n) \rightarrow \text{list}(h, \text{prev}, \text{next}, n))
\]

Here, \text{emp} indicates that the heap is empty (e.g., the list is null) and head \rightarrow \text{Node}(p, n) indicates that head points to an allocated Node(p) object. The existentially quantified variables p and n represent the previous

\[\text{https://github.com/guolong-zheng/starfix}\]
Figure 1: STARFix repairs the invalid circular doubly-linked list shown in Figure 1a by iteratively generating the repair candidates shown in Figures 1b-1f and finally obtains the correct one in Figure 1d. In the following, we use the conventional terminologies symbolic heap models (or just symbolic heaps) for SL predicates and concrete models for data structure instances.

Example. We illustrate STARFix using the example given in Figure 1 adopted from [9]. Figure 1a shows a data structure that does not satisfy d11, as illustrated by the dashed links: (1) the next field of N2 is N1 but the previous field of N1 is N2 and (2) similarly, the previous field of N3 is N1 but the next field of N1 is N3. Figure 1b shows that to repair the data structure given in Figure 1a, STARFix generates several repairs attempted shown in Figures 1b-1d and finally obtains the correct one in Figure 1d. In the following, we use the conventional terminologies symbolic heap models (or just symbolic heaps) for SL predicates and concrete models for data structure instances.

STARFix uses an iterative, unrolling and matching algorithm to check the concrete model given in Figure 1a for inconsistency with the given symbolic heap represented d11. In the first iteration, STARFix unfolds d11(x) to obtain the symbolic heaps:

\[ \Delta_1 = \exists p_1, n_1, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, p_1, x, n_1) \]

\[ \Delta_2 = \exists p_1, n_1, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, p_1, x, n_1) \]

STARFix then matches these symbolic models with the concrete model \( M_0 \equiv \{ x \rightarrow N0 \} \) (for illustration purpose, we assume that \( x \) points to N0 in Figure 1a). For \( \Delta_1 \), the matching fails because \( \text{list} \) is not null, and we do not continue with this model. For \( \Delta_2 \), the match succeeds because we can find values from \( M_0 \) to concretize \( \Delta_2 \), i.e., we can generate a new concrete model \( M_1 \equiv \{ x \rightarrow N0 ; p_1 \rightarrow N3 ; n_1 \rightarrow N1 \} \). We continue the unfolding and matching process using \( \Delta_2 \) and \( M_1 \).

In the second iteration, we unfold \( \Delta_2 \) and obtain the symbolic heaps:

\[ \Delta_3 = \exists p_1, n_1, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, p_1, x, n_1) \]

\[ \Delta_4 = \exists p_1, n_1, n_2, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, n_2) \]

STARFix prunes \( \Delta_3 \) due to failed match and matches \( \Delta_4 \) to obtain the new concrete model \( M_2 \equiv \{ x \rightarrow N0 ; p_1 \rightarrow N3 ; n_1 \rightarrow N1 ; n_2 \rightarrow N2 \} \). Similarly, in the third iteration, we obtain a matched symbolic heap and create a concrete model (we do not show the unmatched heap):

\[ \Delta_6 = \exists p_1, n_1, n_2, n_3, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, p_1, n_2, n_3) \]

\[ \Delta_7 = \exists p_1, n_1, n_2, n_3, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, n_2, n_3) \]

With this fix, we can continue the unfolding process on \( \Delta_2 \) to obtain:

\[ \Delta_9 = \exists p_1, n_1, n_2, n_3, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, n_2) \]

\[ \Delta_{10} = \exists p_1, n_1, n_2, n_3, x \rightarrow \text{Node}(p_1, n_1) \times \text{list}(x, n_2) \]

The symbolic model \( \Delta_9 \) matches with the generated concrete model shown in Figure 1d. Moreover, we can no longer unfold \( \Delta_9 \) because it contains no inductive predicate. This indicates that the repaired data structure satisfies the specification, i.e., a valid circular doubly-linked list.

Note that STARFix could generate multiple fixes. Other than the fix shown in Figure 1d, STARFix also generates fixes as shown in Figures 1c and 1f both are valid circular doubly linked list. The fix in Figure 1c modifies next field of N0 to N3 and prev field of N3 to N0. The fix in Figure 1f
Predicate defn  $Pred ::= \text{pred } P_i(t_i) \equiv \Theta_i; \ \tau ::= \text{Int } | \ c$
Data structure  $Node ::= \text{data } c_1 \{ t_1 f_1; \cdots ; t_j f_j \}$
Symbolic heap  $\Phi ::= \Delta \ | \ \Phi \land \varnothing \ | \ P \land \bigwedge_{i=1}^{n} \kappa_i$ and $\kappa = \kappa_i \land \kappa_2$
Spatial formula  $\kappa ::= \text{emp } \kappa_1 \ | \ x \rightarrow c(v) \ | \ P(v) \land \kappa_1 \land \kappa_2$
Pure formula  $\pi ::= \text{true } \ | \ \neg \pi \ | \ \exists v. \pi \ | \ \pi_1 \land \pi_2 \ | \ \pi_1 \lor \pi_2$
Arithmetic  $\alpha ::= \ a_i = a_j \ | \ a_i \leq a_j \ | \ k \cdot v \ | \ k \cdot a \ | \ a_1 + a_2 \ | \ -a$

Figure 2: Grammar of separation logic formulas

modifies next field of N1 to N3 and prev field of N3 to N1.

3. BACKGROUND

We briefly describe SL formulas and the Java StarFinder tool, in which our STARFIX is built upon.

3.1 Separation logic

In the last two decades, separation logic formalism [11,20] has been successfully applied to analyzing and verifying heap-manipulating programs. This formalism can support the concise and precise abstraction of shapey data structures via symbolic heaps. The syntax of symbolic heaps in this work is presented in Figure 2. To support type-based heap semantics (like Java), we define concrete heap models via a fixed finite collection Node (using keyword data), a fixed finite collection Fields, a disjoint set Loc of locations (heap addresses), a set of non-address values Val, such that null \in Val and Val \cap Loc = \emptyset (i.e., no pointer arithmetic). We assume a set of integers $Z$ and $Z \subseteq Val$. We use k to denote an integer constant.

In Figure 2, we use $\bar{x}$ to denote a sequence of variables. A formula is a disjunction of symbolic heaps. A symbolic heap is a universally quantified conjunction of a spatial formula and a pure (non-heap) formula. A spatial formula is a separating conjunction of empty predicates emp, points-to predicates $x \rightarrow c(v)$ (where $c \in Node$ and $v$ are variables corresponding to fields of $c$), occurrences of inductive predicates $P(v)$. In STARFIX, a predicate, like $\bar{x}$ (and in contrast to Facebook’s Infer [5]), is defined by the user using symbolic heaps with keyword pred. A pure formula is an arithmetical constraint.

3.2 StarLib

In [17,13], Pham et al. introduced STARLib, the main component for SL functions in JAVA STARFINDER. This library provides JAVA API’s for parsing SL formula, unfolding inductive predicate, dispatching satisfiability, etc. We note that the satisfiability API in STARLib actually invokes the state-of-the-art SL solver presented [15] to discharge the satisfiability problems. In this work, we extend STARLIB with matching function to instantiate existentially quantified variables.

STARFIX uses the popular Unfold–and–Match technique to deal with inductive predicates (which represent unbounded data structures) in SL. This technique has been used for entailment [6], symbolic execution [17,15], satisfiability [15], and frame inference [13]. Intuitively, given a formula with inductive predicates, this technique helps to expose all possible heap structures through unfolding and instantiate the quantified heap-based variables through matching. For model checking problem, this technique first helps instantiate heap structure of a formula and then match this structure against a given concrete heap structure. To our best knowledge, we are the first to apply Unfold–and–Match to the model checking and repair problems.

4. AUTOMATIC DATA STRUCTURE REPAIR

Given a specification $\Delta$ and a concrete data structure $T$, STARFIX checks if $T$ conforms with $\Delta$. If not, STARFIX generates one or multiple modified versions of $T$ that conform with $\Delta$.

4.1 Bug detection

In essence, the specification $\Delta$ describes the set $\Sigma$ of valid symbolic heap configurations. On the other hand, the data structure $T$ is a concrete heap configuration. $T$ contains no bug if there exists $\Sigma$ that can be concretized to $T$. This is a model checking problem. Algorithm 1 describes how STARFIX performs a depth-first search for such a $\sigma$. Similar to SAT solver, it iteratively builds a map $M$, of which each entry is a pair $(o_s, o_c)$ of a symbolic heap object $o_s$ and concrete heap object $o_c$. The concretization is the process of assigning the concrete objects, i.e., values, in $T$ to the symbolic objects in (a derivation of) $\Delta$.

In the beginning, $\Delta$ is pushed on the stack, and $M$ maps the root states of the symbolic and concrete heaps, for instance, $M$ stores (head, $\emptyset$) in our illustrative example. STARFIX then iteratively does the following procedure. First, it takes out the current symbolic heap configuration $\Delta_t$ from the top of the stack. If $\Delta_t$ can be concretized to $T$, the $T$ conforms to $\Delta$, i.e. it contains no bug, and the search ends.

On the other hand, if it determines that $\Delta_t$ cannot be concretized to $T$, it continues the search for different paths (we use the continue instruction as in Java or C/C++). If matching is not decidable at the current state as $\Delta_t$ contains inductive predicates, these predicates need to be unfolded.

Since an inductive predicate is often a disjunction of several symbolic heaps (e.g., one base case and several recursive cases), unfolding $\Delta_t$ will result in a set $S$, and each element of $S$ is pushed on the stack for further searching. Details about the subroutines are explained in the following.

4.1.1 match($\Delta_t, T$)

This function takes two inputs: the symbolic heap $\Delta_t$ and the concrete heap $T$, and it has three possible outputs:

- true: $\Delta_t$ can be concretized to $T$ using $M$. This means $\Delta_t$ does not contain inductive predicates, and all heap variables are contained in $M$. For example, $\Delta_0$ with assignment $M_0^3$ in our illustrative example.
- false: $\Delta_t$ cannot be concretized to $T$. This is caused by a conflict in $M$. For example, the conflict of $n_3$ and $n_1$ in assignment $M_3$ in the illustrative example.
- unknown: neither true nor false. There is no conflict in the partial map $M$, but there are variables of $\Delta_t$ which are not in $M$ as they are defined by inductive predicates. These variables also get concretized, as variable $p_1$ and $n_1$ in $\Delta_2$ in the illustrative example.

In the beginning, $M$ contains only the root states of the symbolic and concrete heaps. match then updates $M$ by comparing the shapes of the symbolic and concrete heaps.
4.1.2 unfold(Δt)

This function is part of the STARLIB library. Its input is a formula Δt, which contains a variable x (or a set of variables) being defined by an inductive predicate. Therefore, it is necessary to unfold the predicate to capture the resources accessed by x, a.k.a. the footprints of x. The unfolding procedure includes the following two sub-procedures: (i) replace one occurrence of inductive predicates by its definition; (ii) rename all existentially quantified variables to avoid clashing. The result of this procedure is a new formula with x typically being defined by a point-to predicate, while the newly renamed variables may still be defined by inductive predicates. As the predicate is often defined by a disjunction of several symbolic heaps, the output of the function is a set of new formulas.

4.2 Automatic repair

When STARFIX determines that T does not conform with Δ, it will generate one or multiple modified versions of T that conforms with Δ. This procedure is defined in Algorithm 2. This algorithm shares many common subroutines with Algorithm 1; the main difference is that match will never return true, and when match returns false, STARFIX will attempt to generate a fix (or fixes), and then continues its search.

Algorithm 2: Repair(Δ, T)

stack ← ∅
Γ ← ∅
stack.push(Δ)
while stack.empty() = false do
  Δt ← stack.pop()
  if match(Δt, T) = false then
    γ ← generateFix(Δt, T)
    Γ ← Γ ∪ {γ}
    S ← unfold(Δt)
    for Δi ∈ S do
      Δt ← stack.push(Δi)
  return Γ

All functions in Alg. 2 have already been explained in the previous section. We now explain how generateFix mutates T to generate fixes that conform with Δ. When match(Δt, T) returns false, there are typically two types of conflict in M.

- \( \exists o_1^1, o_2^1, o_3^1, o_4^1, o_5^1, o_6^1, o_7^1, o_8^1 \in M \land o_1^2 \neq o_2^2 \)
- \( \exists o_9^1, o_10^1, o_11^1, o_12^1, o_13^1 \in M \) but o9 should map to o2 in the concrete heap. For example, the reason explained to make the fix in [12]

The first type of conflict only occurs when \( o_1^1 \) and \( o_2^2 \) are existentially quantified variables. For example, the conflict of n3 and n1 in assignment M3 in the illustrative example. STARFIX can generate two possible fixes by modifying either \( o_1^1 \) or \( o_2^2 \). In order to do so, it needs to backtrack to the state where \( o_1^1 \) (or \( o_2^2 \)) is instantiated and generates a new corresponding concrete heap object \( o_1^2 \) for \( o_1^1 \). To generate \( o_2^2 \), STARFIX considers the following options: (i) null, (ii) initialized objects in the concrete heap, and (iii) a new node.

For the second type of conflict, STARFIX generates only one possible fix by modifying \( o_9 \) in the concrete heap in a way that it can be concretized form \( o_9 \). We perform the similar search on possible values for \( o_9 \) in the similar way.

After fixing, there may be the cases that all symbolic objects are in the map M, while some concrete objects are not in M. In these cases, those concrete objects will be deleted, as they are not bounded by the specification. The fixes are still valid.

5. PRELIMINARY RESULTS

STARFIX is implemented in Java and works with Java bytecode programs. We use Java reflection to collect concrete heaps from running program and STARLIB to parse and unfold SL formulae. Currently, we implement a simple model checker to check the concrete heap with respect to the specification as described in Algorithm 1. In future work, we would extend the solver, e.g., [15], to obtain a more powerful model checker to support more expressive properties involving general inductive definitions, pointer-based (dis)equalities and arithmetic.

To evaluate STARFIX, we manually create corrupted data structures by injecting errors into a correct data structure. For example, we randomly pick nodes in a list and change their next or prev fields to point to null or some random nodes in the list.

In addition to lists (e.g., the doubly-linked list shown in Section 2), STARFIX can repair tree data structures, e.g., a binary tree defined by the predicate:

\[ \text{pred tree(root)} \equiv (\text{emp} \land \text{root} = \text{null}) \lor (\exists l, r. \text{root} \rightarrow \text{Node2}(l, r) \land \text{tree}(l) \land \text{tree}(r)) \]

We can then use the predicate tree to repair tree data structures. For example, Figure 3a shows a corrupted tree that has both the right field of b and the left field of e pointing to g, violating the separating conjunction in tree. When given this tree, STARFIX detects the inconsistency (at iteration 20) and generates two fixes (at iteration 89): changing the right field of b to null [35] and changing the left field of e to null [35]. Both fixes are valid with respect to the tree specification.

6. RELATED WORK

STARFIX is inspired by the line of research on specification-based repair for data structures. In this line of work, Elkarablieh et al. [9] use rep0K predicate functions to check for data structure integrity and mutates invalid data structures to pass these predicates. In addition to rep0K, Zaeem et al. [24-26] support pre/post-conditions using relational first-order logic formulae in the Alloy’s language [12] and use SAT solving to generate repairs. Demsky and Rinard [8] also use specifications written in Alloy’s language and perform repairs by translating constraints into disjunctive normal form and solving using an ad hoc search.

Several works explore strategies to improve efficiency and effectiveness of data structure repairs using rep0K and Alloy specifications, including using execution trace history and unsat core [24] and abstracting and memorizing repairs to apply to similar repairs [26]. Demsky et al. [7] integrates the Daikon dynamic invariant generation [10] to infer desired specifications for data structures (instead of requiring users to manually provide such specifications). We are exploring these ideas to adapt and extend them to STARFIX.

The work in [4] presents a model checking technique for general SL inductive predicates to support runtime verification. Given a concrete model and a specification, the checker relies on a fixed point calculator to compute for the specification a finite set of base pairs each of which is a sub-model of the concrete model. In contrast to [4], STARFIX is based on Unfold-and-Match and focuses on repair, instead of just checking, programs. When the model checker returns false, STARFIX generates fixes to avoid interrupting the running programs.

7. CONCLUSIONS AND FUTURE WORK

We present our work-in-progress on data structure repair using separation logic. By using separation logic, we can compactly and precisely capture desired properties of data structures and use existing techniques in separation logic to detect and repair complex data structures.

Currently, our prototype STARFIX can fix inductive Java data structures
such as trees and lists. We are pursuing four areas to improve the work. First, we would extend an SL satisfiability solver, like [15], to support for a more expressive fragment with general inductive definitions and arithmetic. Secondly, we are evaluating ranking techniques to prioritize fixes as much as possible. Thirdly, we are exploring optimizations, e.g., repair abstraction and history-aware strategy [24, 26], to improve STARFix’s performance. Lastly, we are interested in using dynamic inference to automatically generate required separation logic specifications from good program states, as has been proposed in [7].

Acknowledgements. This work is partially supported by the Google Summer of Code 2018 program.

8. REFERENCES


