Natural Proofs for Structure, Data, and Separation

Xiaokang Qiu Pranav Garg Andrei Ştefănescu P. Madhusudan
University of Illinois at Urbana-Champaign, USA
{qiu2, garg11, stefane1, madhu}@illinois.edu

Abstract

We propose natural proofs for reasoning with programs that manipulate data-structures against specifications that describe the structure of the heap, the data stored within it, and separation and framing of sub-structures. Natural proofs are a subclass of proofs that are amenable to completely automated reasoning, that provide sound but incomplete procedures, and that capture common reasoning tactics in program verification. We develop a dialect of separation logic over heaps, called Divya, with recursive definitions that avoids explicit quantification. We develop ways to reason with heaplets using classical logic over the theory of sets, and develop natural proofs for reasoning using proof tactics involving disciplined unfoldings and formula abstractions. Natural proofs are encoded into decidable theories of first-order logic so as to be discharged using SMT solvers.

We also implement the technique and show that a large class of more than 100 correct programs that manipulate data-structures are amenable to full functional correctness using the proposed natural proof method. These programs are drawn from a variety of sources including standard data-structures, the Schorr-Waite algorithm for garbage collection, a large number of low-level C routines from the Glib library and OpenBSD library, the Linux kernel, and routines from a secure verified OS-browser project. Our work is the first that we know of that can handle such a wide range of full functional verification properties of heaps automatically, given pre/post and loop invariant annotations. We believe that this work paves the way for deductive verification technology to be used by programmers who do not (and need not) understand the internals of the underlying logic solvers, significantly increasing their applicability in building reliable systems.

Categories and Subject Descriptors F.3.1 [Logics and Meanings of Programs]: Specifying and Verifying and Reasoning about Programs: Mechanical verification; D.2.4 [Software Engineering]: Software/Program Verification: Assertion checkers

Keywords heap analysis; data structures; natural proofs; separation logic; SMT solvers

1. Introduction

In recent years, the automated deductive verification paradigm for software verification that combines user written modular contracts and loop invariants with automated theorem proving of the result-
In recent years, separation logic, especially in combination with recursive definitions, has emerged as a much more succinct and natural logic to express properties about structure and separation [29, 36]. However, the validation of verification conditions resulting from separation logic invariants are also complex, and has eluded automatic reasoning and exploitation of SMT solvers (even more so than tools such as Boogie that use classical logic). Again, help from the user in proving the verification conditions are currently necessary—the tools Verifast [20] and Bedrock [13], for instance, admit separation logic specifications but require the user to write low-level lemmas and proof tactics to guide the verification. For example, in verifying an in-place reversal of a linked list, Bedrock would require several lemmas and a hint package be supplied at the level of the code in order for the proof to go through.

The work in this paper is motivated by the opinion that entirely decidable logics are too restrictive, in general, to support the verification of complex specifications of functional correctness for heap manipulating programs, and the other extreme of user-supplied proof tactics and lemmas is too tedious, requiring of the user too much knowledge of the underlying proof systems/decision procedures. Our aim is to build completely automatic, sound, but incomplete proof techniques that can solve a large class of properties involving complex data-structures.

The natural proof methodology: Our proof methodology of natural proofs was first proposed in a paper by Madhusudan et al. on natural proofs for tree data-structures last year at POPL [27]. Natural proofs exploit a fixed set of proof tactics, keeping the expressiveness of powerful logics, retaining the automated nature of proving validity, but giving up on completeness (i.e., giving up decidability, retaining soundness). The idea of natural proofs [27] is to identify a subclass of proofs \( N \) such that (a) a large class of valid verification conditions of real-world programs have a proof in \( N \), and (b) searching for a proof in \( N \) is decidable. In fact, we would even like the search for a proof in \( N \) to be efficiently decidable, possibly utilizing the automatic logic solvers (SMT solvers) that exist today. Natural proofs are hence a single restrictive, however, handling only single trees, with no scope for handling multiple or more complex data-structures and their separation (see section on Related Work for more details).

The aim of this paper is to provide natural proofs for general properties of structure, data, and separation. Our contributions are: (a) we propose Dwał, a dialect of separation logic for heaps, with no explicit (classical) quantification but with recursive definitions, to express second-order properties; (b) show that Dwäł is both powerful in terms of expressiveness, and that the strongest-post of Dwäł specifications with respect to bounded code segments can be formulated in Dwäł; (c) show how Dwäł has been designed so that it can be systematically converted to classical logic using the theory of sets, allowing us to connect the more natural and succinct specifications to more verbose but classical logic, and (d) develop a natural proof mechanism for classical logics with recursion and sets that implement sound but incomplete reductions to decidable theories that can be handled by an SMT solver.

Dwäł: A separation logic with determined heaplets

The primary design principle behind separation logic is the decision to express strict specifications—logical formulas must naturally refer to heaplets (subparts of the heap), and, by default, the smallest heaplets over which the formula needs to refer to. This is in contrast to classical logics (such as FOL) which implicitly refer to the entire heap globally. Strict specifications permit elegant ways to capture how a particular sub-part of the heap changes due to a procedure, implicitly leaving the rest of the heap and its properties unchanged across a call to this procedure. Separation logic is a particular framework for strict specifications, where formulas are implicitly defined on strictly defined heaplets, and where heaplets can be combined using a spatial conjunction operator denoted by \( * \). The frame rule in separation logic captures the main advantage of strict specifications: if the Hoare-triple \( \{ P \} C \{ Q \} \) holds for some program \( C \), then \( \{ P * R \} C \{ Q + R \} \) also holds (with side-conditions that the modified variables in \( C \) are disjoint from the free variables in \( R \)).

Consider, for example, expressing that the location \( x \) is the root of a tree. This is a second-order property and formulations of it in classical logic using set or path quantification are quite complex and not easily amenable to automated verification. We prefer inductive definitions of structural properties without any explicit quantification. The separation logic syntax with recursive definitions and heaplet semantics allows simple quantifier-free formulas to express structural restrictions; for example, tree-ness can be expressed simply as:

\[
\text{tree}(x) \iff (x = \text{nil} \lor \text{emp}) \lor (x \rightarrow (l, r) = \text{tree}(l) \land \text{tree}(r))
\]

We first define a new logic, Dwäł, that permits no explicit quantification, but permits powerful recursive definitions to define integers, sets/multisets of integers, and sets of locations, using least fix-points. The logic Dwäł furthermore has a heaplet semantics and allows the spatial conjunction operator \( * \). However, a key design feature of Dwäł is that the heaplet for recursive formulas is essentially determined by the syntax as opposed to the semantics. In classical separation logic, a formula of the form \( a \land \beta \) says that the heaplet can be partitioned into any two disjoint heaplets, one satisfying \( a \) and the other \( \beta \). In Dwäł, the heaplet for (a complex) formula is determined and hence if there is a way to partition the heaplet, there is precisely one way to do so. We have found that most uses of separation logic to express properties can be written quite succinctly and easily using Dwäł (in fact, it is easier to write such deterministic-heap specifications). The key advantage is that this eliminates implicit existential quantification the separation operator provides. In a verification condition that combines the pre-condition in the negative and the post-condition in the positive, the classical semantics for separation logic invariably introduces universal quantification in the satisfiability query for the negation of the verification condition, which in turn is extremely hard to handle.

In Dwäł, the semantics of a recursive definition \( r(x) \) (such as tree above), requires that the heaplet be determined and defined as the set of all locations reachable from the node \( x \) through a set of pointer-fields \( f_1, \ldots, f_k \) without passing through a set of locations (given by a set of location terms \( t_1, \ldots, t_l \)). While our logical mechanisms can be extended beyond this notion (in deterministic ways), we have found that this captures most common properties required in proving data-structure manipulating programs correct.

Translating Dwäł to classical logic with recursion:

The second key step in our paradigm is a technique to bridge the gap from separation logic to classical logic in order to utilize efficient decision procedures supported by SMT solvers. We show that heaplet semantics and separation logic constructs of Dwäł can be effectively translated to classical logic where heaplets are modeled as sets of locations. We show that Dwäł formulas can be translated into classical logic with free set variables that capture the heaplets corresponding to the strict semantics. This translation does not, of course, yield a decidable theory yet, as recursive definitions are still present (the recursion-free formulas are in a decidable theory). The carefully designed Dwäł logic with determined heaplet semantics ensures that there is no quantification in the resulting formula in classical logic. The heaplets of recursively defined prop-

---

http://plv.csail.mit.edu/bedrock/Tutorial.html
languages, which are defined using the set of all reachable nodes, are translated to recursively defined sets of locations.

**Natural proofs for D**

Finally, we develop a natural proof methodology for Dryad by showing a natural proof mechanism for the equivalent formulas in classical logic. The basic proof tactic that we follow is not just dependent on the formula embodying the verification condition, but also on the precise footprint touched by the program segment being verified. We unfold recursive definitions precisely across footprints, translating them to the frontier of the footprint, and then use a form of formula abstraction that treats recursive formulas on frontier nodes as uninterpreted functions. The resulting formula falls in a logic over sets and integers, which is then decided using the theory of uninterpreted functions and arrays using SMT solvers. The key feature is that heaplets and separation logic constructs, which get translated to recursively defined sets of locations, are unfolded along with other user-defined recursive definitions and formula-abstracted using this uniform natural proof strategy.

While our proof strategy is roughly as above, there are many technical details that are complex. For example, the heaplets defined by pre/post conditions intrinsically specify the modified locations of the heap, which have to be considered when processing procedure calls in order to ensure which recursively defined metrics on locations continue to hold after a procedure call. Also, the final decidable theories that we compile our conditions down to does require a bit of quantification, but it turns out to be in the array property fragment which admits automatic decision procedures.

**Implementation and Evaluation:**

Our proof mechanisms are essentially a class of decidable proof tactics that result in sound but incomplete validation procedures. To show that this class of natural proofs is effective in practice, we provide a prototype implementation of our technique, which handles a low-level programming language with pre-conditions and post-conditions written in Dryad. We show, using a large class of correct programs manipulating lists, trees, cyclic lists, and doubly linked lists as well as multiple data-structures of these kinds, that the natural proof mechanism succeeds in proving the verifications conditions automatically. These programs are drawn from a range of sources, from textbook data-structure routines (binary search trees, red-black trees, etc.) to routines from Glib low-level C-routines used in GTK+/Gnome to implement file-systems, from the Schorr-Waite garbage collection algorithm, to several programs from a recent secure framework developed for mobile applications. Our work is by far the only one that we know of that can handle such a large class of programs, completely automatically. Our experience has been that the user-provided contracts and invariants are easily expressible in Dryad, and the automatic natural proof mechanisms work extremely fast. In fact, contrary to our own expectations, we also found that the tool is useful in debugging: in several cases, when the annotations supplied were incorrect, the model provided by the SMT solver for the natural proof was useful in detecting errors and correcting the invariants/program.

2. Related Work

The natural proof methodology was introduced in [27] (see also [39]), but was exclusively built for tree data-structures. In particular, this work could only handle recursive programs, i.e., no while-loops, and even for tree data-structures, imposed a large number of restrictions on pre/post conditions for methods—the input to a procedure had to be only a single tree, the method can only return a single tree, and even then must havoc the input tree given to it. The lack of handling of multiple structures means that even simple programs like mergesort (that merges two lists), cannot be handled, and simple programs that manipulate two lists or two trees cannot be reasoned with. Also, structures such as doubly-linked lists, trees with parent pointers, etc. are out of scope of this work. Technically, in our present work, we can handle user-defined structures expressible in separation logic, multiple structures and their separation, programs with while-loops, etc., because of our logical treatment of separation logic using classical logic.

There is a rich literature on analysis of heaps in software. We omit discussing literature on general interactive theorem provers (like Isabelle [31]) that require considerable manual guidance. We also omit a lot of work on analyzing shape properties of the heap [6, 13, 18, 28, 41], as they do not handle complex functional properties.

There are several proof systems and assistants for separation logic [32, 39] that incorporate proof heuristics and are incomplete. However, [3] gives a small decidable fragment of separation logic on lists which has been further extended in [11] to include a restricted form of arithmetic. Symbolic execution with separation logic has been used in [4, 5, 8] to prove structural specifications for various lists and tree programs. These tools come hard-wired with a collection of axioms and their symbolic execution engines check the entailment between two formulas modulo these axioms. VeriFast [20], on the other hand, chooses flexibility of writing richer specifications over complete automation, but requires the user to provide inductive lemmas and proof tactics to aid verification. Similarly, Bedrock [15] is a Coq library that aims at mostly automated (but not completely automated) procedures that requires some proof tactics to be given by the user to prove verification conditions. The idea of using regions (sets of locations) for describing heaps in our work also extends to describing frames for function calls, and the use for the latter is similar to implicit dynamic frames [38] in the literature. The crucial difference in our framework is that the implicit dynamic frames are syntactically determined, and amenable to quantifier-free reasoning. A work that comes very close to ours is a paper by Chin et al. [14], where the authors allow user-defined recursive predicates (similar to ours) and build a terminating procedure that reduces the verification condition to standard logical theories. However, their procedure does not search for a proof in a well-defined simple and decidable class, unlike our natural proof mechanism; in fact, the resulting formulas are quantified and incompatible with decidable logics handled by SMT solvers.

In all of the above cited work and other manual and semi-automatic approaches to verification of heap-manipulating programs like [37], inductive definitions of algebraic data-types is extremely common for capturing second-order data-structure properties. Most of these approaches use proof tactics which unroll inductive definitions and do extensive unification to try to match terms to find simple proofs. Our notion of natural proofs is very much inspired by such kinds of semi-automatic and manual heap reasoning that we have seen in the literature.

There is also a variety of verification tools based on classical logics and SMT solvers. Darsny [23] and VCC [16] compile to Boogie [2] and generate VCs that are passed to SMT solvers. This approach requires significant ghost annotations, and annotations that explicitly express and manipulate frames. The Jutao system [43, 44] is one of the first attempts at full functional verification of linked data structures, which integrates a variety of theorem provers, including SMT solvers, and makes the process mostly automated. However, complex specifications combining structure, data and separation usually require more complex provers such as MOSA [21], or even interactive theorem provers such as Isabelle [41] in the worst case. The success of the proof search also relies on users’ manual guidance.
Motivating Example: Heapify

3. Motivating Example

In this section we give intuition into our verification approach through a motivating example. Recall that a max-heap is a binary tree such that for each node \( n \) the key stored at \( n \) is greater than or equal to the keys stored at each of its children. Heaps are often used to implement priority queues. In Figure 1, in the lower right corner, we express the property that a location \( x \) points to a max-heap using recursive definitions \( \text{keys}^A_{pf}(x) \) and \( \text{mheap}^A_{pf}(x) \), with \( \text{pf} \equiv \left(\text{left, right}\right) \). These recursive definitions are written in D\text{rax}, which is formally introduced in Section 4. Intuitively, D\text{rax} extends quantifier free separation logic [22, 25] with recursive predicates and functions. These recursive definitions allow us to express structural and data properties on the heap, like those of max-heap, without explicit quantification.

For a location \( x \), the recursive definition \( \text{keys}^A_{pf}(x) \) returns the set of keys at the nodes of the tree rooted at \( x \); if \( x \) is \text{nil} and the heap is empty, then the empty-set; otherwise, the union of the key stored at \( x \) and the keys stored in the left and right subtrees of \( x \). Similarly, the recursive definition \( \text{mheap}^A_{pf}(x) \) states that \( x \) points to a max-heap if: \( x \) is \text{nil} and the heap is empty; or \( x \) and the heaplets of the left and right subtrees of \( x \) are mutually disjoint (points to a tree) and the key at \( x \) is greater than or equal to the keys of the left and right subtrees of \( x \).

The method heapify in Figure 1 is at the heart of the procedure for deleting the root from a max-heap (removing the node with the maximum priority). If the max-heap property is violated at a node \( x \) while satisfied by its descendants, then heapify restores the max-heap property at \( x \). It does so by recursively descending into the tree, swapping the key of the root with the key at its left or right child, whichever is greater. The precondition \( \varphi_{pre} \) binds the free variable \( K \) to the set of keys of \( x \). The postcondition states that after the procedure call, \( x \) satisfies the max-heap property and the set of keys of \( x \) is unchanged (same as \( K \)).

One of the main aspects of our approach is to reduce reasoning about heaplet semantics and separation logic constructs to reasoning about sets of locations. We use set operations like union, intersection and membership to describe separation constraints on a heaplet satisfying a formula. This translation from D\text{rax} formulas, like those in Figure 1 to formulas in classical logic with recursive predicates and functions is formally presented in Section 5.

Intuitively, we associate a set of locations to each (spatial) atomic formula, which is the domain of the heaplet satisfying that formula. D\text{rax} requires that this heaplet is syntactically determined for each formula. For example, the heaplet associated to the formula \( x \mapsto \ldots \) is the singleton \( \{x\} \); for recursive definitions like \( \text{mheap}^A_{pf}(x) \) and \( \text{keys}^A_{pf}(x) \), the domain of the heaplet is \( \text{reach}_{left,right}(x) \), which intuitively is the set of locations reachable from \( x \) using the pointer fields \( left \) and \( right \), and can be defined recursively.

As shown in Figure 1, \( \varphi_{pre} \) is a conjunction of two formulas. If \( G_{pre} \) is the domain of the heaplet associated to \( \varphi_{pre} \), then the first conjunct requires \( G_{pre} \) to be the disjoint union of the sets \( \{x \} \), reach\text{left,right}(left(x)) and reach\text{left,right}(right(x)). The second conjunct requires \( G_{pre} = \text{reach}_{left,right}(x) \). From these heaplet constraints, we can translate \( \varphi_{pre} \) to the following formula in classical logic over the global heap:

\[
G_{pre} = \{x\} \cup \text{reach}_{left,right}(left(x)) \cup \text{reach}_{left,right}(right(x)) \land \
\neg x \in \text{reach}_{left,right}(left(x)) \land x \notin \text{reach}_{left,right}(right(x)) \land \
\text{reach}_{left,right}(left(x)) \cap \text{reach}_{left,right}(right(x)) = \emptyset \land x \neq \text{nil} \land \
\text{mheap}(left(x)) \land \text{mheap}(right(x)) \land \
\text{keys}(x) = K
\]

Similarly, we translate \( \varphi_{post} \) to

\[
G_{post} = \text{reach}_{left,right}(x) \land \text{mheap}(x) \land \text{keys}(x) = K
\]

Note that the recursive definitions \( \text{mheap} \) and \( \text{keys} \) without the “\( \Delta \)" superscript are in the classical logic (without the heaplet constraint). Hence the recursive predicate \( \text{mheap} \) satisfies

\[
\text{mheap}(x) \leftrightarrow x \neq \text{nil} \land \text{reach}_{left,right}(x) = \emptyset \land \
\text{reach}_{left,right}(left(x)) \cap \text{reach}_{left,right}(left(x)) = \emptyset \land \
\text{reach}_{left,right}(right(x)) \cap \text{reach}_{left,right}(right(x)) = \emptyset \land \
\text{mheap}(left(x)) \land \text{keys}(x) \geq \text{keys}(left(x)) \land \
\text{mheap}(right(x)) \land \text{keys}(x) \geq \text{keys}(right(x))
\]

The right side of Figure 1 presents a basic path from method heapify, corresponding to the case when both children of \( x \) are not \text{nil} and the key of the right child is greater than the keys of the

---

void heapify(loc x) {  
if (x.left = nil)  
s := x.right;  
else if (x.right = nil)  
s := x.left;  
else {  
  lx := x.left;  
  rr := x.right;  
  if (lx.key < rr.key)  
s := x.rightq;  
  assume lx.key < rr.keyq;  
s := x.rightq;  
assumption s x= nilq;  s := x.rightq;  
assumption s x= nilq;  
  if (s.key < x.key)  
t := x.keyq;  
s := x.keyq;  s.x := x.keyq;  
s.x := t;  
heapify(s);  
}  
heapify(s);  
}

Figure 1. Motivating example: Heapify
left child and the root. The subscript of a pointer/data field denotes the timestamp. A key insight is that any basic path touches a finite number of locations and may call some recursive procedures. We refer to the touched locations as the footprint, and to the adjacent locations which are not part of the footprint as the frontier. For this example, the footprint is \{ \text{x, lx, rx} \} (x is known to be equal with \text{rx}) and the frontier is \{ \text{left}_0(\text{lx}), \text{right}_0(\text{lx}), \text{left}_0(\text{rx}), \text{right}_0(\text{rx}) \}. We capture the effect of the path until the call to heapify by

\[
\text{left}_0(x) \neq \text{nil} \land \text{right}_0(x) \neq \text{nil} \land \text{lx} = \text{left}_0(x) \land \text{rx} = \text{right}_0(x) \\
\land \text{key}_0(\text{lx}) < \text{key}_0(\text{rx}) \land x \neq \text{nil} \\
\land \text{key}_0(\text{lx}) > \text{key}_0(\text{rx}) \land t = \text{key}_0(x) \\
\land \text{key}_1 = \text{key}_0(x) \land t = \text{key}_1(x) \\
\land \text{key}_2 = \text{key}_1(x \leftarrow t)
\]

Once we have expressed the verification condition in classical logic with recursive definitions over the global heap, we prove it using the natural proof methodology. We unfold the recursive definitions \(m\text{heap}(x), \text{key}(x)\) and \(\text{reach}\text{left}\text{right}(x)\) for \(x, \text{lx}\) and \(\text{lx}\) (the footprint), thus evaluating them in terms of their values on the frontier. The call to heapify preserves the recursive definitions on locations reachable from \(\text{lx}\), and modifies those on \(\text{rx}\) according to the pre-condition. Finally, we abstract the recursive definitions on the frontier with uninterpreted functions. We decide the resulted formula (which is in a decidable logic) using an SMT solver. Section \ref{app} describes this process in detail.

4. The Logic \textsc{Dryad}

In this section we present our logic \textsc{Dryad} \cite{arunachalam2019dryad}, which redefines the logic \textsc{Heaplet} \cite{arunachalam2018heaplet} on arbitrary data-structures (not just trees), using heaplet semantics and separation logic primitives; the logic hence is a quantifier-free heaplet logic augmented with recursively defined predicates/functions. However, for brevity, we will refer to the new logic we propose as \textsc{Dryad}, and refer to the logic in \cite{arunachalam2018heaplet} as \textsc{Dryad}_tree.

4.1 Syntax

Let us fix a finite set of pointer-fields \(PF\) and a finite set of data-fields \(DF\). A record consists of a set of pointer-fields from \(PF\) and a set of data-fields from \(DF\). Our logic also presumes that locations refer to entire records rather than particular fields, and that address arithmetic is precluded. We will use the term \textit{locations} hence to refer to these records. We assume that every field is defined at every location, i.e., all memory records have the same layout (to simplify the presentation); our logic can easily be extended with record types.

Let \(\textsf{Bool} = \{ \textsf{true}, \textsf{false} \}\) stand for the set of Boolean values, \(\textsf{Int}\) stand for the set of integers and \(\textsf{Loc}\) stand for the universe of locations. For any set \(A\), let \(\textsf{S}(A)\) denote the set of all finite subsets of \(A\), and let \(\textsf{MS}(A)\) denote the set of all finite multisets with elements in \(A\).

The \textsc{Dryad} logic allows expressing quantifier-free first-order properties over heaps/heaplets augmented with recursively defined notions for a location to express second-order properties, denoted as a function \(r: \textsf{Loc} \rightarrow D\). The codomain \(D\) can be \(\textsf{Int}_t, \textsf{S}(\textsf{Loc}), \textsf{S}(\textsf{Int}_t), \textsf{MS}(\textsf{Int}_t)\) or \textsf{Bool}, where \(\textsf{Int}_t\) and \(\textsf{MS}(\textsf{Int}_t)\) extend \textsf{Int} and \(\textsf{MS}(\textsf{Int})\) to lattice domains, respectively, in order to give least fixed-point semantics (explained later in this section). Typical examples of these recursive definitions include the definitions of the height of a tree or the height of black-nodes in the tree rooted at a node (recursively defined integers), the set of nodes reachable from a location following certain pointer fields (recursively defined sets of locations), the set/multiset of keys stored at a particular data-field under nodes reachable from a location (recursively defined set/multiset of integers), and the property that the tree rooted at a node is a binary search tree or a balanced tree or just a tree (recursively defined predicates).

A \textsc{Dryad} formula \(\varphi\) is quantifier-free, but parameterized by a set of recursive definitions \(\text{Def}^R\). The syntax of \textsc{Dryad} logic is given in Figure \ref{fig:logic} where the syntax of formulas is followed by the syntax for recursive definitions. Most symbols in \textsc{Dryad} are common and self-explanatory. Note that the inequality \((< \lor \leq)\) between integer sets/multisets indicates that any integer in the left-hand side is less-than-or-greater-than any integer in the right-hand side. It is also noteworthy that the separating conjunction \((\not\land)\) from separation logic is also allowed, but only if it is not above any negation \((\not\lor)\). We require that every recursive function/predicate used in the formula \(\varphi\) has a unique definition in \(\text{Def}^R\). Each recursive function is parameterized by a set of pointer fields \(\text{pf}\) and a set of program variables \(\text{s}\), denoted as \(\text{pf}^\text{\text{\downarrow}}\). The subscripts are used in defining the semantics of recursive functions in Section \ref{sec:semantics}. We usually simply use \(\text{pf}^\text{\text{\downarrow}}\) when the subscripts are not relevant in the context. Similarly, recursive predicates are denoted as \(p^\text{\text{\downarrow}}\) or simply \(p^\text{\text{\downarrow}}\). The recursive functions are defined using the syntax:

\[
\text{\text{\downarrow}}(x, \text{s}): \{ x_1: \text{s}_1; \ldots ; x_n: \text{s}_n \} \Rightarrow \{ t_1: \text{s}_1; \ldots ; t_m: \text{s}_m \}
\]

where \(\varphi^\text{\text{\downarrow}}(x, \text{s}, t, \text{s})\) is a formula/term in our logic with \(\text{s}\) implicitly existentially quantified. The recursively defined predicates are defined using the syntax: \(\text{\downarrow}(x, \text{s})\), which is a formula in our logic with \(\text{s}\) implicitly existentially quantified. The recursive function syntax above expresses a case-split, with the function evaluating to the first term whose guard evaluates to true. The restrictions on the recursive definitions are:

- Subtraction, set-difference, and negation are disallowed;
- Every variable in \(\text{s}\) should appear in the right hand side of a points-to relation binding it to \(x\) exactly once.

For examples of recursive functions and predicates, see the definitions \(\text{key}^\text{\text{\downarrow}}(x)\) and \(m\text{heap}^\text{\text{\downarrow}}(x)\) in Figure \ref{fig:logic} respectively. The set of program variables \(\text{s}\) parameterizing the definitions is empty in both these definitions and the set of implicitly existentially-quantified variables \(\text{s}\) is \(\{k, l, r\}\).

4.2 Semantics

Our logic is interpreted on models that are \textit{program states}:

\begin{definition}
A program state is a tuple \(C = (R, s, h)\) where
\begin{itemize}
\item \(R \subseteq \textsf{Loc} \setminus \{\textsf{nil}\}\) is a finite set of locations;
\item \(s: \textsf{Vars} \rightarrow \textsf{Int} \cup \textsf{Loc}\) is a store mapping program variables to locations or integers (of appropriate type);
\item \(h: R \times (\textsf{PF} \cup \textsf{DF}) \rightarrow \textsf{Int} \cup \textsf{Loc}\) is a heaplet mapping non-nil locations and each pointer-field/data-field to values of the appropriate type.
\end{itemize}
\end{definition}

Note that the set of locations is, in general, larger than the state \(R\) and hence \(R\) defines a subset of heap locations. The store maps variables to locations (not necessarily in \(R\)), but the heaplet \(h\) gives interpretations for pointer and data-fields only for elements in \(R\).

Given a heaplet \(h\), for every pointer field \(pf\), we denote the projection of \(h\) on \(R \times (\textsf{PF} \setminus \{pf\}) \cup \textsf{DF}\) as \(h \upharpoonright pf\). Similarly, for every data-field \(df\), we denote the projection of \(h\) on \(R \times (\textsf{PF} \cup \textsf{DF}) \setminus \{df\}\) as \(h \upharpoonright df\). Also, for every subset \(S \subseteq R\), we denote the projection of \(h\) on \(S \times (\textsf{PF} \cup \textsf{DF})\) as \(h \upharpoonright S\).

A term/formula with free variables \(F\) is interpreted by interpreting the free variables in \(F\) using the map \(s\) from variables to values. The semantics of \textsc{Dryad} is similar to that of classical Separation Logic (SL). In particular, a term/formula without recursive definitions is interpreted exactly in the same way in \textsc{Dryad} and SL. Hence we first give the semantics of the non-recursive part, followed by the semantics of recursive definitions.
Before defining the semantics of formulas, we define the pure property for terms/formulas. Intuitively, a term/formula is pure if it is independent of the heap. Syntactically, a term/formula is pure if it does not contain \( \text{emp} \), \( \rightarrow \) or any recursive definition. Note that in SL all terms are pure, but in Dravd, a term can be impure if it contains a recursive function \( f^A \).

**Semantics of terms**

Each \( T \)-term evaluates to either a normal value of type \( T \), or to \( \text{undef} \), which is only used in interpreting recursive functions (will be explained later). As a special value, \( \text{undef} \) will be propagated throughout the formula: if a formula \( \varphi \) contains a sub-term that evaluates to \( \text{undef} \), then \( \varphi \) will evaluate to \( \text{false} \) if it appears positively, and will evaluate to \( \text{true} \) otherwise. Intuitively, \( \text{undef} \) cannot help in making the formula true over a model.

The \( \text{Loc} \) terms are evaluated as follows:

\[
\begin{align*}
\llbracket x \rrbracket_{\text{Loc}} &= s(x) \\
\llbracket \text{nil} \rrbracket_{\text{Loc}} &= \text{nil}
\end{align*}
\]

For any binary operator \( op \), \( t \) op \( t' \) is evaluated as follows:

\[
\llbracket op t' t' \rrbracket_{\text{Loc}} = \begin{cases} 
\llbracket t \rrbracket_{\text{Loc}} op \llbracket t' \rrbracket_{\text{Loc}} & \text{if } t \text{ or } t' \text{ is pure} \\
\llbracket t \rrbracket_{\text{Loc}} op \llbracket t' \rrbracket_{\text{Loc}} & \text{if } \llbracket t \rrbracket_{\text{Loc}} \text{ and } \llbracket t' \rrbracket_{\text{Loc}} = \text{undef} \\
\text{undef} & \text{otherwise}
\end{cases}
\]

where \( op \) is interpreted in the natural way.

For singletons, \( [t] \) will evaluate to \( 0 \) if \( t \) evaluates to \( -\infty \) or \( \infty \):

\[
\llbracket [t] \rrbracket_{\text{Loc}} = \begin{cases} 
\text{undef} & \text{if } \llbracket t \rrbracket_{\text{Loc}} = \text{undef} \\
0 & \text{if } \llbracket t \rrbracket_{\text{Loc}} = -\infty \text{ or } \infty \\
[t] & \text{otherwise}
\end{cases}
\]

\( [t]_w \) and \( [t]_l \) evaluate similarly.

**Semantics of formulas**

The formula \( \text{true} \) is always interpreted to be \( \text{true} \):

\[
(R, s, h) \models \text{true}
\]

The formula \( \text{emp} \) asserts that the heap is empty:

\[
(R, s, h) \models \text{emp} \iff R = \emptyset
\]

The formula \( \llbracket t \rrbracket_{\text{Loc}} \rightarrow \llbracket t' \rrbracket_{\text{Loc}} \) asserts that the heap contains exactly one record consisting of fields \( pf_f \) and \( df_f \), at address \( lt \), with values \( lt \) and \( it \), respectively. Formally, the semantics of this formula is given as:

\[
(R, s, h) \models \llbracket t \rrbracket_{\text{Loc}} \rightarrow \llbracket t' \rrbracket_{\text{Loc}} \iff R = \llbracket t \rrbracket_{\text{Loc}} \text{ and } R = \llbracket t' \rrbracket_{\text{Loc}}
\]

Note that, as in separation logic, the above has a strict semantics— the heaplet must be a singleton set and cannot be a larger set.

For binary relations \( t \sim t' \) between integers, sets, and multisets, including equality, the pure property plays an important role. Remember that in SL all terms are pure. To be consistent with SL, if both \( t \) and \( t' \) are pure, it is interpreted in the normal way. Otherwise, \( t \sim t' \) is only defined on the minimum heaplet required by \( t \) and \( t' \), more concretely the union of the heaplet associated with \( t \) and \( t' \).

\[
(R, s, h) \models t \sim t' \iff t \text{ or } t' \text{ is pure and } [t]_C \sim [t']_C
\]

or \( t \) and \( t' \) are impure and there exist \( R_1, R_2 \) s.t. \( R = R_1 \cup R_2 \) and \( [t]_{C[R]} \neq \text{undef} \), \( [t']_{C[R]} \neq \text{undef} \) and \( [t]_{C[R]} \sim [t']_{C[R]} \), where \( \sim \) is interpreted in the natural way.

The semantics of the disjoint conjunction operator \( \ast \) is defined as follows. The formula \( \varphi \ast \varphi' \) asserts that the heap can be split into two disjoint parts in which \( \varphi \) and \( \varphi' \) hold respectively:

\[
(R, s, h) \models \varphi \ast \varphi' \iff \text{there exist } R_0, R_1 \text{ s.t. } R = R_0 \cap R_1 = \emptyset \text{ and } R_0 \cup R_1 = R \text{ and } (R_0, s, h) \models \varphi \text{ and } (R_1, s, h) \models \varphi'
\]

Boolean combinations are defined in the standard way:

\[
\begin{align*}
(R, s, h) \models \varphi \land \varphi' & \iff (R, s, h) \models \varphi \text{ and } (R, s, h) \models \varphi' \\
(R, s, h) \models \varphi' & \iff (R, s, h) \models \varphi \text{ and } (R, s, h) \models \varphi'
\end{align*}
\]

**Semantics of recursive definitions**

The main semantical difference between Dravd and SL is on recursive definitions. We would like to deterministically delineate the heap domain for any recursive definition, so that the heap domain required by any Dravd formula can be syntactically determined.

Given a recursive definition \( rec^A_{pf, df} \), the subscripts \( pf \) and \( df \) play a role in delineating the heap domain. Intuitively, the heap domain for \( rec^A_{pf, df} (l) \) is the set of locations reachable from \( l \) using pointer-fields in \( pf \), but without going through the locations \( l' \). In other words, we want to take the set of locations that lie in between \( l \) and \( l' \). Precisely, this set is determined by a location \( l \) and a program state \((R, s, h)\).
We denote it as \( \text{reachset}_{R,R}^l(l, (R, s, h)) \). Formally it is the smallest set of locations \( L \) satisfying the following two conditions:

1. \( l \) is in the set \( L \) if \( l \) is not in \( R \) and \( l \neq nil \);
2. for each \( c \) in \( L \), with \( c \in R \), and for any pointer \( pf \), if \( h(c, pf) \) is not in \( R \) and is not \( nil \), then \( h(c, pf) \) is also in \( L \).

Note that even though the reach set is defined with respect to the edges in the heaplet, we can determine whether \( R \) includes all nodes reachable from \( l \) without going through \( R \) in the global heap by checking whether \( R = \text{reachset}_{R,R}^l(l, (R, s, h)) \).

For each recursive definition \( \text{rec}_{R}^{\Delta} \), we usually simply denote \( \text{reachset}_{R,R}^l(l, (R, s, h)) \) as \( \text{reachset}^{\Delta} \), as the subscripts are implicitly known.

Now, given a program state \( C = (R, s, h) \) and a recursive function/predicate \( \text{rec}_{R}^{\Delta} \), the semantics on a location \( l \) depends on whether the heap domain \( R \) is exactly the required reach set \( \text{reachset}^{\Delta} \). If this is not true, we simply interpret it as \( \text{undef} \) or \( \text{false} \).

If the heap domain matches the reach set (i.e., \( R = \text{reachset}_{R,R}^l(l, (R, s, h)) \)), the semantics is defined in the natural way (using least fixed-points). The colon operator in the syntax of recursive function \( f_{\epsilon}^{\Delta} \) translates into a nested if-then-else (ITE) operator. Formally, \( f_{\epsilon}^{\Delta} \) is defined as:

\[
\| f_{\epsilon}^{\Delta}(l) \|_C = \begin{cases} 
\text{ITE}(f_{\epsilon}^{\Delta}(l), & \text{ITE}(f_{\epsilon}^{\Delta}(l), \text{ITE}(f_{\epsilon}^{\Delta}(l), \text{ITE}(f_{\epsilon}^{\Delta}(l), \text{ITE}(f_{\epsilon}^{\Delta}(l), \text{ITE}(f_{\epsilon}^{\Delta}(l), \text{ITE}(f_{\epsilon}^{\Delta}(l), \text{ITE}(f_{\epsilon}^{\Delta}(l), 0))))) & \text{if } R = \text{reachset}_{R,R}^l(l, (R, s, h)), \\
\text{undef} & \text{otherwise.}
\end{cases}
\]

where \( R_1 \ldots R_{n+1} \subseteq R \) such that \( f_{\epsilon}^{\Delta}(l) \neq \text{undef} \). In order to give least fixed-point semantics for recursive definitions in the logic, we extend the primitive data types to lattice domains. To the order \( \text{false} \subseteq \text{true} \) forms a complete lattice, and \( S (\text{Loc}) \) and \( S (\text{Int}) \) ordered by inclusion, with join as union and meet as intersection, form complete lattices. Integers and multisets are extended to lattices. Let \((\text{Int}_s, \leq)\) denote the complete lattice, where \( \text{Int}_s = \text{Int} \cup \{ -\infty, \infty \} \), and where the ordering is \( \leq \), then \( \text{max} \), meet is \( \min \). Also, \( S (\text{Int}_s, \leq) \) denote the complete lattice constructed from \( S (\text{Int}) \), where \( S (\text{Int}_s, \leq) = S (\text{Int}) \cup \{ \top \} \), and \( \leq \) extends the inclusion relation with \( S \subseteq \top \) for any \( M \in S (\text{Int}) \). It is easy to see that \((\text{Int}_s, \leq) \) and \((\text{Int}_s, \leq) \) are complete lattices.

Formally, let \( \text{Def} \) consists of definitions of integer functions \( I \), set-of-locations functions \( L \), set-of-integers functions \( S \), and \( \Delta \)-multiset-of-integers functions \( \text{MS} \) and predicates \( P \). Since these definitions could rely on each other, we evaluate them altogether as a function vector \( r^{\Delta} = (f_{\epsilon}^{\Delta}, f_{\epsilon}^{\Delta}, f_{\epsilon}^{\Delta}, f_{\epsilon}^{\Delta}, \text{rec}_{R}^{\Delta}) \).

We take the cartesian product lattice of the individual lattices and take the least fixed-point of \( r^{\Delta} \) to obtain the semantics for each definition. Let select_{\epsilon} \( (\text{ITE}(r^{\Delta}(l), c), c) \) for each recursive definition \( \text{rec}_{R}^{\Delta} \), denote the selection of the coordinate for \( r^{\Delta} \) in \( \text{ITE}(r^{\Delta}) \).

Now we can formally define the semantics of recursive definitions. For any configuration \( C \), the semantics of a recursive function \( f_{\epsilon}^{\Delta} \) is defined as:

\[
\| f_{\epsilon}^{\Delta}(l) \|_C = \begin{cases} 
\text{select}_{\epsilon}(\text{ITE}(r^{\Delta}(l), c), c) & \text{if } R = \text{reachset}_{R,R}^l(l, (R, s, h)), \\
\text{undef} & \text{otherwise.}
\end{cases}
\]

and the semantics of a recursive predicate \( p_{\epsilon}^{\Delta} \) is defined as

\[
\| p_{\epsilon}^{\Delta}(l) \|_C = \begin{cases} 
\text{select}_{\epsilon}(\text{ITE}(r^{\Delta}(l), c), c) & \text{if } R = \text{reachset}_{R,R}^l(l, (R, s, h)), \\
\text{false} & \text{otherwise.}
\end{cases}
\]

**Remark:** Note that we disallow negative operations (subtraction, set-difference and negation) in defining recursive definitions. This syntactical restriction guarantees that each iteration of \( r^{\Delta} \) is monotonic. By Knaster-Tarski theorem, \( r^{\Delta} \) admits a least fixed-point.

**Examples**

The Dryad logic was already used in Section 3 to define max-heaps. Note that the definition of a max-heap is precisely defined on the heaplet that includes the underlying tree nodes of the max-heap only, as the heaplet for a recursive definition is the set of all reachable nodes according to the two pointers.

To clarify the difference between Dryad and SL, consider now this recursive definition:

\[
p_{\text{乾}}^{\Delta}(x) \overset{\text{def}}{=} (x = \text{nil}) \lor \text{emp} \lor \left((x \rightarrow y, z) \ast \left(p_{\text{乾}}^{\Delta}(y) \lor p_{\text{乾}}^{\Delta}(z)\right)\right)
\]

Now consider a global heap that has a tree rooted at \( x \) with pointer fields \( l \) and \( r \). The above recursive formula, in separation logic, will be true on any heap that contains the nodes of a path from \( x \) to nil. However, in Dryad, we require that the heaplet must satisfy the heap constraints of the formula and also be the precise set of locations reachable from \( x \) using the pointer fields \( l \) and \( r \). Consequently, if the tree pointed to by \( x \) plus more than one path, the Dryad formula will be false for any heaplet.

The above example shows the advantage of Dryad; when heaplets are determined, we can avoid quantification. We have not found natural examples where an undetermined heaplet semantics helps in specifying properties of heaps.

Dryad can express structures beyond trees. The main restriction we do impose is that we allow only unary recursive definitions, as this allows us to find simpler natural proofs since there is only one way to unfold the definition across a footprint. However, Dryad can express structures like cyclic lists and doubly-linked lists.

A cyclic-list is captured as \( (v \rightarrow y) \ast \text{iseg}_{\text{乾}}^{\Delta}(y) \). Here, \( v \) is a program variable which denotes the head of the cyclic-list and \( \text{iseg}_{\text{乾}}^{\Delta}(y) \) captures the list segment from \( y \) back to the head \( v \), where the subscripts next and \( v \) indicate that the heaplet of the list segment is the locations that can be reached using the field next, but without going through \( v \).

Another interesting example is a doubly-linked list. We define a doubly-linked list as the following unary predicate:

\[
d_{\text{乾}}^{\Delta}(\text{next}) \overset{\text{def}}{=} (x = \text{nil}) \lor \text{emp} \lor ((y \rightarrow z) \ast d_{\text{乾}}^{\Delta}(z))\]

The first two disjuncts in the definition cover the base case when \( x = \text{nil} \) or the location next to \( x \) is \( \text{nil} \); otherwise, let \( y \) be the location next to \( x \), then the next pointer at \( y \) points to \( x \) and location \( y \) is recursively defined as a doubly-linked list.

### 5. Translation to a Logic over the Global Heap

We now show one of the main contributions of this paper—a translation from Dryad logic to classical logic with recursive predicates and functions, but over the global heap. The formulation of separation logic primitives in the global heap allows us to express complex structural properties, like disjointness of heaplets and trees, using recursive definitions over sets of locations, which are defined locally, and are amenable to unfolding across the footprint and hence amenable to natural proofs.

For example, consider the formula \( mheap^{\Delta}(x) \ast mheap^{\Delta}(y) \), where \( mheap^{\Delta} \) is defined in Section 3. Since the heaplets for \( mheap^{\Delta}(x) \) and \( mheap^{\Delta}(y) \) are precise, it can get translated to an equivalent formula with a free set variable \( G \) that denotes the global heap over which the formula is evaluated:

\[
mheap(x) \land mheap(y) \land (\text{reach}^{mheap}(x) \cap \text{reach}^{mheap}(y) = 0) \land (\text{reach}^{mheap}(x) \cup \text{reach}^{mheap}(y) = G)
\]

where \( mheap \) and \( \text{reach}^{mheap} \) are corresponding recursive definitions in classical logic, which will be defined later in this section. Note that we use italics and remove the \( \Delta \) superscript to show the difference from their counterpart in Dryad.

We assume the Dryad formula to be translated in is disjunctive normal form, i.e., \( \lor \) operators should be above all \( \ast \) and \( \land \) operators.
The domain of the required heaplet for evaluating reached-set expressions is defined recursively in classical logic. The set $\text{rec}^\Delta$ uses the corresponding definitions in classical logic. Translating a recursive definition to classical logic depends on whether the sub-formulas $\text{reach}^\Delta$ are domain-exact but the reverse implication is not true, in general. For example, the formula $(\text{true} \land \text{false}) \lor \text{emp}$ is not domain-exact but the reverse implication is not true, in general. For every program state $C$ with heap domain $\text{Loc}$, and for every interpretation of variables $I$ including a valuation for set-variable $G$, $(C, I) \models T(\varphi, G)$ w.r.t. Def if and only if $(C \mid_G, I \setminus \{G\}) \models \varphi$ w.r.t. Def$^\Delta$.

Figure 3. Domain-exact property and Scope function. Both are defined only for terms and formulas without disjunction and negation. A formula is assumed in its disjunctive normal form.

This is not a real restriction as one can always push the disjunction out. This normal form ensures that for all occurrences of the separation operator in a formula, there exists a unique way of splitting the heap so as to satisfy the * separated sub-formulas. Also, it ensures that this unique heap-split can be determined syntactically from the structure of those sub-formulas.

In our translation, we model the heaplets associated with a formula or a term as a set of locations and all operations on these heaplets are modeled as set operations like set union, set intersections, etc., over set-of-location variables. For example the separating conjunction $P \lor Q$ is translated to the following set constraint: the intersection of the sets associated with the heaplets in the formulas $P$ and $Q$ is empty. Given a formula $\varphi$ in Dvax and its associated heap domain modeled by a set variable $G$, we define an inductive translation $T$ into a classical logic formula $T(\varphi, G)$ in the quantifier-free theory of finite sets, integers and uninterpreted functions. The translated formula is not interpreted on a heaplet, but interpreted on a global heap (i.e., with the heap domain $\text{Loc}$).

The translation uses an auxiliary domain-exact property and an auxiliary scope function. The domain-exact property indicates whether a term evaluates to a well-defined value or a positive formula evaluates to true on a fixed heap domain or not. This is different from the property pure: a pure formula or term is not domain-exact but the reverse implication is not true, in general. For example, the formula $(lt \mapsto it) \ast \text{true}$ is not domain-exact but is also not pure. The scope function maps a term to the minimum heap domain required to evaluate it to a normal value, and maps a positive formula to the minimum heap domain required to evaluate it to true. The domain-exact property and the scope function are defined inductively in Figure 4.

We describe the logic translation in detail in Figure 4. The ITE expression used in the translation is short for "if-then-else". It is just a conditional expression defined as follows: $\text{ITE}(\varphi, t_1, t_2)$ evaluates to $t_1$ if $\varphi$ is true, otherwise evaluates to $t_2$.

In general, our translation reserves an impure term/formula to be evaluated only on the syntactically determined heap domain according to the semantics of Dvax. In particular, when evaluating a recursive formula or predicate $p^\Delta$, we ensure that the heaplet is precisely the reach set $\text{reach}^\Delta(\text{lt})$. For a formula $\varphi \ast \varphi'$, translation to classical logic depends on whether the sub-formulas $\varphi$ and $\varphi'$ are domain-exact or not. If a sub-formula is domain-exact then it is evaluated on its scope. If it is not domain-exact, then it is evaluated on the rest of the heaplet.

Recursive definitions in Dvax are also translated to recursive definitions in classical logic. Translating a recursive definition $\text{rec}^\Delta$ uses the corresponding definitions $\text{rec}$ and $\text{reach}^\Delta$, both of which are defined recursively in classical logic. The set $\text{reach}^\Delta$ represents the domain of the required heaplet for evaluating $\text{rec}^\Delta$, and the $\Delta$-eliminated definition $\text{rec}$ captures the value of $\text{rec}^\Delta$ when the heaplet is restricted to $\text{reach}^\Delta$. Formally, suppose $\text{rec}^\Delta$ is a recursive definition w.r.t. pointer fields $pf$ and stopping locations $s$, then $\text{reach}^\Delta$ is recursively defined as the least fixed-point of

$$
\text{reach}^\Delta(x) \stackrel{\text{def}}{=} \text{ITE}(x = \text{nil} \lor x = \text{nil}, \emptyset, \{x\} \cup \bigcup_{pf} \text{reach}^\Delta(\text{pf}(x)))
$$

For each recursive predicate $p^\Delta$ defined as $p^\Delta(x) \equiv \varphi'(x, x, \delta)$, we define

$$
p(x) \stackrel{\text{def}}{=} T(\varphi'(x, x, \delta), \text{reach}^\Delta(x))
$$

Similarly, for each recursive function $f^\Delta$ defined as

$$
f^\Delta(x) \stackrel{\text{def}}{=} \varphi'_1(x, x, \delta) : \ldots : \varphi'_i(x, x, \delta) : \text{default} : \varphi'_{i+1}(x, \delta)
$$

we define

$$
f(x) \stackrel{\text{def}}{=} \text{ITE}(T(\varphi'_1(x, x, \delta), \text{reach}^\Delta(x)), \ldots, T(\varphi'_{i+1}(x, \delta), \ldots))
$$

where $\varphi'_{i+1}(x, \delta)$ is just the classical logic counterpart of $\varphi'_{i+1}(x, \delta)$, when interpreted in a heap domain within $\text{reach}^\Delta(x)$. Formally it is short for

$$
\text{ITE}(\text{scope}(\varphi'_{i+1}(x, \delta), \text{reach}^\Delta(x)), T(\varphi'_{i+1}(x, \delta), \text{scope}(\varphi'_{i+1}(x, \delta))), \text{undef})
$$

Now for each set of recursive definitions $\text{Def}^\Delta$ in Dvax, we can translate it to a set of recursive definitions $\text{Def}$ in classical logic.

**Theorem 5.1.** Let $\varphi$ be a Dvax formula w.r.t. a set of recursive definitions $\text{Def}^\Delta$. For every program state $C$ with heap domain $\text{Loc}$, and for every interpretation of variables $I$ including a valuation for set-variable $G$, $(C, I) \models T(\varphi, G)$ w.r.t. Def if and only if $(C \mid_G, I \setminus \{G\}) \models \varphi$ w.r.t. Def$^\Delta$. □
We exclude memory errors in order to simplify the presentation. Memory errors can be handled using a similar VC generation for assertions that negate the conditions for memory errors to occur.

Hence we simply use $v$ to denote $s_i(v)$. Moreover, every recursive predicate/function is also indexed by $i$. For example, $p_i$ is the recursive predicate such that $p_i(0)$ is true if $C_i \equiv T^0(p_i(l), \text{reach}^i(l))$. Now for every formula $\varphi$ and every index $i$, we can give the index $i$ to all related pointer fields, data fields, and recursive definitions. We denote the indexed formula as $\varphi[i]$.

We algorithmically derive the verification condition $\psi_{VC}$ corresponding to it in classical logic with recursive definitions on the global heap (the algorithm is quite involved, and is presented in Appendix A in the supplemental material).

**Theorem 6.1.** Given a Hoare-triple $[\psi_{pre}] P [\psi_{post}]$, assume that each procedure call in $P$ satisfies its associated pre- and post-conditions. Then the triple is valid if the formula $\psi_{VC}$ derived above is valid. Moreover, when $P$ contains no procedure calls, the triple is valid if $\psi_{VC}$ is valid.

**Proof.** Presented in Appendix B in the supplemental material. □

### 6.2 Unfolding Across the Footprint

The verification condition obtained above is a quantifier-free formula involving recursive definitions and the reachable sets of the form $\text{reach}^i(x)$, which are also defined recursively. While these recursive definitions can be unfolded ad infinitum, we exploit a proof tactic called unfolding across the footprint. Intuitively, the footprint is the set of locations explored by the program explicitly (not including procedure calls). More precisely, a location is in the footprint if it is dereferenced explicitly in the program. The idea is to unfold the recursive definitions over the footprint of the program, so that recursive definitions on the footprint nodes are, as precisely as possible, to the recursive definitions on frontier nodes. This will enable effective use of the formula abstraction mechanism, as when recursive definitions on frontier nodes are made uninterpreted, the unfolding formulas ensure tight conditions that the frontier nodes have to satisfy.

Furthermore, to enable effective frame reasoning, it is also necessary to strengthen the verification condition with a set of instances of the frame rule. More concretely, we need to capture the fact that a recursive definition (or a field) on a location is unchanged during a segment or procedure call of the program, if the reachable locations (or only the location itself) are not affected by the segment or procedure call.

We incorporate the above facts formally into the verification condition. Let us introduce a macro function $fp$ that identifies the location variables that are in (or aliased to something in) the footprint. The footprint of $P$, $FP$, is the set of dereferenced variables in $P$ (we call a location variable dereferenced if it appears on the left-hand side of a dereferencing operator “\cdot" in $P$). Then $fp(u) \equiv \forall_{aexpr}(v \neq v).

Now we state the unfoldings and framings using a formula $\text{UnfoldAndFrame}$. Assume there are $m$ procedure calls in $P$, then $P$ can be divided into $m + 1$ basic segments (subprograms without procedure calls): $S_0 : q_0 : S_1 : \ldots : q_m : S_m$ where $S_k$ is the $(d + 1)$-th basic segment and $g_j$ is the $d$-th procedure call. Then

$$
\text{UnfoldAndFrame} \equiv \forall \bigwedge_{0 \leq d < m} \forall \bigwedge_{0 \leq j < d} \bigg( (fp(u) \vee u = nil) \Rightarrow (\text{Unfold}^d(u) \land \text{FieldsUnchanged}^d(u)) \bigg) \land \\
( \neg (fp(u) \vee u = nil) \Rightarrow \text{ReUnchanged}^d(u) )
$$

The formula enumerates every recursive definition $rec$ and every index $d$, and for each location $u$ that is either pointed to by a location variable or is $nil$, the formula checks if $u$ is in the footprint, and then unfolds it or frames it accordingly. If $u$ is in the footprint, then we unfold $rec$ for the timestamps before and after $s_d$ (represented by the formula $\text{Unfold}^d(u)$); moreover, all fields of...
6.3 Formula Abstraction

While checking the validity of the strengthened verification condition $\psi_{VC}$ is still undecidable, as we argued before, it is often sufficient to prove it by assuming that the recursive definitions are arbitrary, or uninterpreted. Moreover, the uninterpreted formula falls in the array property fragment [12], whose satisfiability is decidable and is supported by modern SMT solvers such as Z3 [17]. This tactic roughly corresponds to applying unification in proof systems.

To prove $\psi_{VC}$, we first replace each recursive predicate $rec_{i}$ with an uninterpreted predicate $rec_{i}^{\mu}$, and replacing the corresponding reach-set function $reach_{i}^{\mu}$ with an uninterpreted function $reach_{i}^{\mu}$. Let the result formula be $\psi_{VC}^{\mu}$. This conversion, called formula abstraction, is sound: if $\psi_{VC}$ is valid, so is $\psi_{VC}^{\mu}$. When a proof for $\psi_{VC}^{\mu}$ is found, we call it a natural proof for $\psi_{VC}$ (and also for $\psi_{VC}^{\mu}$).

The formula abstraction step is the only step that introduces incompleteness in our framework, but helps us transform the verification condition to a decidable theory. Formula abstraction (combined with unfolding recursive definitions across the footprint) discovers recursive proofs where the recursion is structural recursion on the definitions of the data-structures. The use of these tactics comes from the observation that such programs often have such recursive proofs (see [19] also for use of formula abstractions).

Our goal now is to check the satisfiability of $\neg\psi_{VC}^{\mu}$ in a decidable theory. The resulting formula can be expressed using the theory of maps (to model sets) and corresponding map operations to model set operations. Formulas of the kind $S_{1} \subseteq S_{2}$, where $S_{1}$ and $S_{2}$ are sets/multi-sets of integers, are the only ones that introduce quantification, but they can be translated to formulas in the array property fragment, which is decidable [12]. We hence obtain a formula $\psi^{APF}$ in the array property fragment combined with the theory of uninterpreted functions, maps, and arithmetic (details are in Appendix D in the supplementary material).

Theorem 6.3. Given a Hoare-triple $\{\psi_{pre}\} P \{\psi_{post}\}$, if the derived array formula $\psi^{APF}$ is satisfiable, then the Hoare-triple is valid. □

User-provided axioms:

While natural proofs are often effective in finding recursive proofs that unfold recursive definitions and do unification, they are not geared towards finding relationships between various recursive definitions themselves. We hence require the user to provide certain obvious relationships between the different recursive definitions as axioms. For example, lseg(x,y) + lseg(y,z) = lseg(x,z) is such an axiom saying that a list segment concatenated with a list yields a list. Note that these axioms are not program-dependent, and hence are not program-specific tactics that the user provides. These axioms are necessary typically to relate partial data-structure properties (like list segments) to complete ones (like lists), especially in iterative programs (as opposed to recursive ones), and we can fix them for each class of data-structures. We also allow the use of the separating implication, $\rightarrow$, from separation logic while specifying these axioms. User-defined axioms are instantiated, using the natural proof philosophy, on precisely the footprint nodes uniformly, and get translated to quantifier-free formulas.

7. Experimental Evaluation

We have implemented a prototype of the natural proof methodology for Dryad presented in this paper. The prototype verifier takes as input a set of user-defined recursive definitions, a set of procedure declarations with contracts, and a set of straight-line programs (or basic blocks) annotated with a pre-condition and a post-condition specifying a set of partial correctness properties including structural, data and separation requirements. Both the contracts and pre-/post-conditions are written in Dryad. For each basic block, the verifier automatically generates the abstracted formula $\psi^{APF}$ as described in Section 6 and passes $\psi^{APF}$ to Z3 [17], a state-of-the-art SMT solver, to check the satisfiability in the decidable theory of array property fragment. The front-end of our verifier is based on
ANTLR and our tool is around 4000 lines of C# code. Using the verifier, we successfully proved the partial correctness of 59 routines over a large class of programs involving heap data structures like sorted lists, doubly-linked lists, cyclic lists and trees. Additionally, we pit our natural proofs methodology against real-world programs and successfully verified, in total, 47 routines from different projects including the list and queue implementations in the Glib open source library, the OpenBSD library, the Linux kernel and the memory regions and the page cache implementations from two different operating systems. Experimental details are available at http://www.cs.uiuc.edu/~madhu/dryad/si/.

We conducted the experiments on a machine with a dual-core, 2.4GHz CPU and 6GB RAM. The first part of our experimental results is tabulated in Figure 6. In general, for every routine, we checked the properties formalizing the complete verification of the routines—capturing the precise structure of the resulting heap-structure, the precise change to the data stored in the nodes and the precise heaplet modified by the respective routines.

For every routine, the suffix rec or iter indicates if the routine was implemented recursively or iteratively using while loops. The names for most of the routines are self-descriptive. Routines like find, insert, delete, append, etc. are the natural implementations of the corresponding data structure operations. The routine delete_all for singly-linked lists, sorted lists and doubly-linked lists recursively deletes all occurrences of a particular key in the input list. The max-heap routine heapify accepts an almost max-heap in which the heap property is violated only at the root, both of whose children are max-heaps, and recursively descends the tree to restore the max-heap property. The routine remove_root for binary search trees and treaps is an auxiliary routine which is called to restore the max-heap property. The routine of whose children are max-heaps, and recursively descends the tree.

Figure 7. Results of verifying open-source libraries. (more details at http://www.cs.uiuc.edu/~madhu/dryad/si/)

In an OS kernel, a process address space consists of a set of intervals of linear addresses represented as a memory region. In the ExpressOS implementation, a memory region is implemented as a sorted doubly-linked list where each node of the list with a start and an end address represents an interval included in the address space. We also verified some key components of the Linux implementation of a memory region, present in the file mmap.c. In Linux, a memory region is represented as a red-black tree where each node, again, represents an address interval. We proved methods which find, remove and insert a vma_struct (vma is short for virtual memory address) into a memory region.

It also worth mentioning that in the process of experiments, we did make some unintentional mistakes, in writing both the basic blocks and the annotations. For example, forgetting to free the deleted node, or using ∧ instead of ∨ in the specification between two disjoint heaplets, were common mistakes. In these cases, Z3 provided counter-examples to the verification condition that captured the essence of the bugs, and turned out to be very helpful for us to debug the specification. These debugging hints are usually not available in other incomplete proof systems.

Our experiments show that the natural proof methodology set forth in this paper is successful in efficiently proving full-functional correctness of a large variety of algorithms. Most of the VCs generated for the above examples were discharged by Z3 in a few seconds. To the best of our knowledge, this is the first automatic mechanism that can prove such a wide variety of algorithms correct, handling such complex properties of structure, data and separation.

Acknowledgments

This work is partially funded by NSF CAREER award #0747041, NSF CCF #1018182, and NSA contract H98230-10-C-0294.
References


