A Local Shape Analysis based on Separation Logic

Dino Distefano[1], Peter W. O'Hearn[1], and Hongseok Yang[2]

[1] Queen Mary, University of London
[2] Seoul National University

Abstract. We describe a program analysis for linked list programs where the abstract domain uses formulae from separation logic.

1 Introduction

A shape analysis attempts to discover the shapes of data structures in the heap at program points encountered during a program's execution. It is a form of pointer analysis which goes beyond the tabulation of shallow aliasing information (e.g., can these two variables be aliases?) to deeper properties of the heap (e.g., is this an acyclic linked list?).

The leading current shape analysis is that of Sagiv, Reps and Wilhelm, which uses very generic and powerful abstractions based on three-valued logic [16]. Although powerful, a problem with this shape analysis is that it behaves in a global way. For example, when one updates a single abstract heap cell this may require also the updating of properties associated with all other cells. Furthermore, each update of another cell might itself depend on the whole heap. This global nature stems from the use of certain instrumentation predicates, such as ones for reachability, to track properties of nodes in the heap: an update to a single cell might alter the value of a host of instrumentation predicates.

In contrast, separation logic provides an approach to reasoning about the heap that has a strong form of locality built in [13]. Typically, one reasons about a collection of cells in isolation, and their update does not necessitate checking or updating cells that are held in a different component of a separating conjunction. It thus seems reasonable to try to use ideas from separation logic in program analysis, with an eye towards the central problem of modularity in the analysis.

Our technical starting point is recent work of Berdine, Calcagno and O'Hearn [5], who defined a method of symbolic execution of certain separation logic formulae called symbolic heaps. Their method is not, by itself, suitable as an abstract semantics because there are infinitely many symbolic heaps and there is no immediate way to guarantee convergence of fixed-point calculations. Here, we obtain a suitable abstract domain by working with (a variation on) their method of symbolic execution, and adding to it an abstraction or widening operator which converts any symbolic heap to one in a certain "canonical form". This abstraction method is an adaptation of work in [7, 8] to the symbolic heaps of Berdine et. al. In contrast to unrestricted symbolic heaps we show that there are only finitely many canonical forms, resulting in termination of the fixed-point calculation used in the abstract semantics of while loops.

Our abstract domain uses linked lists only. Other abstractions based on separation logic might be considered as well.

After defining the analysis we turn to locality. We describe a sense in which the abstract semantics obeys the Frame Rule of separation logic, and we identify a notion of footprint as an input-output relation that mentions only those symbolic heap cells accessed by a program. The footprint provides a sound over-approximation of a program's entire (abstract) meaning, in many cases an exact representation. The results on locality give a way to infer sound results on large states from those explicitly obtained on small ones, suggesting further possible developments on interprocedural and concurrency analyses.

1.1 Related Work

In work on heap analysis (see [14] for discussion) much use has been made of a "storeless semantics" where the model is built from equivalence classes of paths rather than locations. The storeless semantics has the pleasant property that it is garbage collecting by its very nature, but it is also extremely complex. This makes it highly nontrivial to see that a particular analysis based on it is sound. In contrast, here we work directly with a store model, and soundness is almost immediate. The abstraction we use is defined by rewrite rules which are all sound implications in separation logic, and the symbolic execution rules are derived from true Hoare triples.

Recent work on shape analysis [14, 15] might be regarded as taking some steps towards separation logic. Early on in separation logic there was an emphasis on what was referred to as "local reasoning": reasoning concentrates on the cells accessed during computation [11]. In [14, 15] a version of the local reasoning idea is employed in an interprocedural analysis, where a procedure summary is constructed which involves only the (abstract) cells reachable from input parameters or variables free in a procedure. The method of applying a procedure does not, however, explicitly utilize a separating conjunction operator $*$; one might say that (the resourceful form of) local reasoning is adopted (or altered), but the formal apparatus of separation logic is not.

In this paper we reciprocate by taking some steps towards shape analysis. Our intention initially was full reciprocation: to build an interprocedural analysis. But, after labouring for the better part of a year, we decided to aim lower: to define an abstract domain and abstract post operator, together with an account of its locality, for a language without procedures. In doing this we have been influenced by shape analysis, but have not adopted the formal apparatus of shape graphs or 3-valued logic. We hope that this paper can serve as a springboard for further developments in local interprocedural and modular concurrency analysis.

We want to make clear that we do not claim that our analysis is superior, in a practical sense, to existing shape analyses. Although it works on small examples, we have not attempted detailed timing comparisons. Also, from a methodological point of view, in the framework of [16] different abstractions are obtained in a uniform way, where a notion of "canonical abstraction" results once instrumentation predicates are nailed down. In contrast, here we have just one particular set of rewrite rules that have been hand-built; how this might be turned into a more general scheme is not obvious.

Nonetheless, we believe that research on how separation logic, or more particularly, the local reasoning idea, might be used in program analysis is of interest because it suggests a genuinely different approach which has promise for the central problem of obtaining modular analyses. A very good example of this is the recent work of Amtoft et. al. [2, 1] which uses local reasoning in information flow analysis (this is a more shallow form of analysis than shape analysis, but they are successful in formulating a very modular analysis).

Finally, in work carried out independently of (and virtually in parallel to) that here, Magill et. al. have defined a method of inferring invariants for linked list programs in separation logic [9]. They also utilize a symbolic execution mechanism related to [5], and give rewrite rules to attempt to find fixed points. There are many detailed differences: (i) they use a different basic list predicate than we do and, as they point out, have difficulty dealing with acyclic lists, where that is a strong point of our analysis; (ii) they do a predicate abstraction of arithmetic operations, where we do not; (iii) and they use an embedding into Presburger arithmetic to help decide implications and Hoare triples, where we do not provide a method for deciding implications (or Hoare triples); (iv) their algorithm does not always terminate, where ours does. But, there is remarkable similarity.

2 Semantic Setting

We first describe the general semantic setting for this work. Following the framework of abstract interpretation [6], we will work with complete lattices D: The semantics of a command c will be given by a continuous function $[\![c]\!]\colon D \to D$.

If we are given a programming language with certain primitive operations p, together with conditionals, sequencing and while loops, then to define the semantics we must specify the meaning $[\![p]\!]$ of each primitive operation as well as a continuous function

$$\mathsf{filter}(b)\colon D \to D$$

for each boolean. Typically, D is built from subsets of a set of states, and the filter function removes those elements that are not consistent with b's truth. The semantics extends to the rest of the language in the usual way.

$$\begin{aligned} [\![c\,;c']\!] &= [\![c]\!]\,;[\![c']\!] \\ [\![\mathbf{if}\ b\ \mathbf{then}\ c\ \mathbf{else}\ c']\!] &= (\mathsf{filter}(b)\,;[\![c]\!]) \sqcup (\mathsf{filter}(\neg b)\,;[\![c']\!]) \\ [\![\mathbf{while}\, b\, \mathbf{do}\, c]\!] &= \lambda d.\ \mathsf{filter}(\neg b)\big(\mathit{fix}\ \lambda d'.\ d \sqcup ([\![c]\!] \circ \mathsf{filter}(b))(d')\big) \end{aligned}$$

One way to understand the semantics of **while** is to view d' as a loop invariant. The d in the lhs of $\sqcup$ means that the loop invariant d' should be implied by the precondition, and the rhs of $\sqcup$ means that d' is preserved by the body. (Here, the fixed-point operator has been moved inward from its usual position in semantics, so that it applies to predicates instead of two command denotations.)

Our domains D will be constructed using a powerset operation. If S is a set we denote by $\mathcal{P}(S)$ the "topped" powerset of S, that is, the set of subsets of $S \cup \{\top\}$. Here, $\top \notin S$ is a special element that corresponds to memory fault (accessing a dangling pointer). If we were to take logical implications between elements of $\mathcal{P}(S)$ into account then we would make $\{\top\}$ the top element and equate all sets containing $\top$. For simplicity in this paper we just use the subset order.

Given a relation $p \Longrightarrow : S \leftrightarrow S \cup \{\top\}$, with membership notated $\sigma, p \Longrightarrow \sigma'$, we can lift it to a function $p^\dagger : \mathcal{P}(S) \to \mathcal{P}(S)$ by

$$p^\dagger X = \{\sigma' \mid \exists \sigma \in X.\, (\sigma, p \Longrightarrow \sigma') \text{ or } (\sigma = \sigma' = \top)\}.$$

The semantics of primitive commands will be given by first specifying an execution semantics $\Longrightarrow$ and then lifting it to $\mathcal{P}(S)$.

Every semantics we work with will have two additional properties: that $\{\top\}$ is mapped to $\{\top\}$ and that it preserves unions. Because of this we could in fact work with a corresponding map $[\![c]\!]_\dagger : S \to \mathcal{P}(S)$ instead of $[\![c]\!] : \mathcal{P}(S) \to \mathcal{P}(S)$.

3 Concrete and Symbolic Heaps

Throughout this paper we assume a fixed *finite* set Vars of program variables (ranged over by $x, y, \ldots$), and an infinite set Vars' of primed variables (ranged over by $x', y', \ldots$). The primed variables will not be used within programs, only within logical formulae (where they will be implicitly existentially quantified).

Definition 1. *A symbolic heap $\Pi \mid \Sigma$ consists of a set Π of equalities and a set Σ of heap predicates. The equalities $E{=}F$ involve variables $x, y, \ldots$, primed variables $x', y', \ldots$, and* nil. *The elements of Σ are of the form*

$$E \mapsto F \qquad \mathsf{ls}(E, F) \qquad \mathsf{junk}.$$

We use $\mathcal{SH}$ to denote the set of consistent *symbolic heaps. (For the definition of consistency, see below.)*

The first two heap predicates are "precise" in the sense of [12]; each cuts out a unique piece of (concrete) heap. The points-to assertion $E \mapsto F$ can hold only in a singleton heap, where E is the only active cell. Similarly, when a list segment holds of a given heap, the path it traces out is unique, and goes through all the cells in the heap. This precise nature of the predicates is helpful when accounting for deallocation.

The junk predicate is used in the canonicalization phase of our analysis to swallow up garbage. It is crucial for termination of our analysis, and it has the useful property to reveal memory leaks.

Besides the heap formulae, symbolic heaps also keep track of equalities involving pointer variables and nil.

Table 1 Semantics of Symbolic Heaps

$$\mathcal{C}[\![x]\!] = s(x) \qquad \mathcal{C}[\![x']\!] = s(x') \qquad \mathcal{C}[\![\mathsf{nil}]\!]s = \mathsf{nil}$$

$s, h \vDash \{E \mapsto F\}$	iff $h = [\mathcal{C}[\![E]\!]s \mapsto \mathcal{C}[\![F]\!]s]$
$s, h \vDash \{\mathsf{ls}(E, F)\}$	iff there is a nonempty acyclic path from $\mathcal{C}[\![E]\!]s$ to $\mathcal{C}[\![F]\!]s$ in h and this path contains all heap cells in h
$s, h \vDash \{\mathsf{junk}\}$	iff $h \neq \emptyset$
$s, h \vDash \Sigma_0 * \Sigma_1$	iff $\exists h_0, h_1.\ h = h_0 * h_1$ and $s, h_0 \vDash \Sigma_0$ and $s, h_1 \vDash \Sigma_1$
$s \vDash \{E{=}F\}$	iff $\mathcal{C}[\![E]\!]s = \mathcal{C}[\![F]\!]s$
$s \vDash \Pi_0 \cup \Pi_1$	iff $s \vDash \Pi_0$ and $s \vDash \Pi_1$
$s, h \vDash \Pi \mid \Sigma$	iff $\exists \boldsymbol{v'}.\ (s(\boldsymbol{x'} \mapsto \boldsymbol{v'}) \vDash \Pi)$ and $(s(\boldsymbol{x'} \mapsto \boldsymbol{v'}), h \vDash \Sigma)$ where $\boldsymbol{x'}$ is the collection of primed variables in $\Pi \mid \Sigma$

We often use the notation $\Sigma * P$ for the (disjoint) union of a formula P onto the spatial part of a symbolic heap, and we similarly use $\Pi \wedge P$ in the pure part.

The meaning of a symbolic heap corresponds to a formula

$$\exists x'_1 x'_2 \ldots x'_n.\ \Big(\bigwedge\nolimits_{P \in \Pi} P\Big) \wedge \Big(\bigstar_{Q \in \Sigma} Q\Big),$$

in separation logic, where $\{x'_1, \ldots, x'_n\}$ is the set of all the primed variables in Σ and Π. More formally, the meaning of a symbolic heap is given by a forcing relation $s, h \vDash \Pi \mid \Sigma$, where s is a stack and h a (concrete) heap.

$$\mathsf{Values} = \mathsf{Locations} \cup \{\mathsf{nil}\} \qquad \mathsf{Heaps} = \mathsf{Locations} \rightharpoonup_f \mathsf{Values}$$
$$\mathsf{Stacks} = (\mathsf{Vars} \cup \mathsf{Vars'}) \to \mathsf{Values} \qquad \mathsf{States} = \mathsf{Stacks} \times \mathsf{Heaps}$$

The semantics is given in Table 1. The operation $h_0 * h_1$ there is the union of heaps with disjoint domains. We give the semantics for the singleton sets in the pure and spatial parts, and then for unions. There, the clause for list segments is given informally, but corresponds to the least predicate satisfying

$$\mathsf{ls}(E, F) \Longleftrightarrow E \neq F \wedge \big(E \mapsto F \vee (\exists x'. E \mapsto x' * \mathsf{ls}(x', F))\big).$$

Our analysis will require us to be able to answer some questions about symbolic heaps algorithmically: whether two expressions are equal, whether they are unequal, whether the heap is inconsistent, and whether a cell is allocated.

$$\Pi \vdash E{=}F \qquad \Pi \mid \Sigma \vdash E{\neq}F \ \ (\text{when } \mathsf{Vars'}(E, F) = \emptyset)$$
$$\Pi \mid \Sigma \vdash \mathsf{false} \qquad \Pi \mid \Sigma \vdash \mathtt{Allocated}(E) \ \ (\text{when } \mathsf{Vars'}(E) = \emptyset)$$

$\Pi \vdash E{=}F$ is easy to check. It just considers whether E and F are in the same equivalence class induced by the equalities in Π.

The other operators use subroutine $\mathsf{allocated}$, which takes Σ and an expression E, and decides whether Σ implies that E points to an allocated cell, by a "nontrivial reason": $\mathsf{allocated}$ ignores the case where Σ is not satisfiable and implies all formulae.

$$\mathsf{allocated}(\Sigma, E) = \exists E'.\ (E \mapsto E' \in \Sigma) \text{ or } (\mathsf{ls}(E, E') \in \Sigma).$$

We then define the other querying operators as follows:

$$\begin{aligned}
\Pi \mathrel{¦} \Sigma \vdash \mathsf{false} \iff & (\exists E.\ \ \Pi \vdash E{=}\mathsf{nil} \text{ and } \mathsf{allocated}(\Sigma, E)), \text{ or} \\
& (\exists E, F.\ \ \Pi \vdash E{=}F \text{ and } \mathsf{ls}(E, F) \in \Sigma), \text{ or} \\
& \left(\begin{array}{l}\exists E, F.\ \ \Pi \vdash E{=}F \text{ and } \Sigma \text{ contains two distinct} \\ \quad \text{predicates whose lhs's are, respectively, } E \text{ and } F\end{array}\right) \\
\Pi \mathrel{¦} \Sigma \vdash E{\neq}F \iff & (E{=}F \wedge \Pi \mathrel{¦} \Sigma) \vdash \mathsf{false} \\
\Pi \mathrel{¦} \Sigma \vdash \mathtt{Allocated}(E) \iff & \Pi \mathrel{¦} \Sigma \vdash \mathsf{false}, \text{ or} \\
& \exists E'.\ \Pi \vdash E{=}E' \text{ and } \mathsf{allocated}(\Sigma, E')
\end{aligned}$$

These definitions agree with what one would obtain from a definition in terms of the forcing relation $\models$, but they are simple syntactic checks that do not require calling a theorem prover.

The rules that define our analysis will preserve consistency of symbolic heaps (that $\Pi \mathrel{¦} \Sigma \nvdash \mathsf{false}$). In particular, inconsistent heaps introduced in branches of if statements or as a result of tests in a while loop will be filtered out.

4 Concrete and Symbolic Execution Semantics

The grammar of commands for the programming language used in this paper is given by

$$\begin{array}{ll}
b ::= E{=}E \mid E{\neq}E & \\
p ::= x{:=}E \mid x{:=}[E] \mid [E]{:=}F \mid \mathbf{new}(x) \mid \mathbf{dispose}(E) & \text{Primitive Commands} \\
c ::= p \mid c\,;c \mid \mathbf{while}\ b\ \mathbf{do}\ c \mid \mathbf{if}\ b\ \mathbf{then}\ c\ \mathbf{else}\ c & \text{Commands}
\end{array}$$

We do not consider commands that contain any primed variables amongst their expressions. We include only a single heap dereferencing operator $[\cdot]$ which refers to the "next" field. In the usual way, our experimental implementation ignores commands that access fields other than "next" (say, a data field), and treats any boolean conditions other than those given as nondeterministic.

4.1 Concrete Semantics

The execution rules for the primitive commands are as follows, where in the faulting rule (the last rule) we use notation for atomic commands that access heap cell E:

$$A(E) ::= [E]{:=}F \mid x{:=}[E] \mid \mathbf{dispose}(E)$$

Concrete Execution Rules

$$\frac{\mathcal{C}[\![E]\!]s = n}{s, h,\ x{:=}\,E \implies (s \mid x \mapsto n), h} \qquad \frac{\mathcal{C}[\![E]\!]s = \ell \quad h(\ell) = n}{s, h,\ x{:=}\,[E] \implies (s \mid x \mapsto n), h}$$

$$\frac{\mathcal{C}[\![E]\!]s = \ell \quad \mathcal{C}[\![F]\!]s = n \quad \ell \in \mathbf{dom}(h)}{s, h,\ [E]{:=}\,F \implies s, (h \mid \ell \mapsto n)} \qquad \frac{\ell \notin \mathbf{dom}(h)}{s, h,\ \mathbf{new}(x) \implies s, (h \mid \ell \mapsto n)}$$

$$\frac{\mathcal{C}[\![E]\!]s = \ell}{s, h * [\ell \mapsto n],\ \mathbf{dispose}(E) \implies s, h} \qquad \frac{\mathcal{C}[\![E]\!]s \notin \mathbf{dom}(h)}{s, h,\ A(E) \implies \top}$$

Notice the tremendous amount of nondeterminism in **new**: it picks out *any* location not in the domain of the heap, and *any* value n for its contents.

The concrete semantics is given in the topped powerset $\mathcal{P}(\mathsf{States})$, where the filter map is

$$\mathsf{filter}(b)X \;=\; \{(s,h) \in X \mid \mathcal{C}[\![b]\!]s \;=\; \text{true}\} \cup \{\top \mid \top \in X\}$$

where $\mathcal{C}[\![b]\!]s \in \{\text{true,false}\}$ just checks equalities by looking up in the stack s.

With these definitions we may then set $\mathcal{C}[\![p]\!] \;=\; p^\dagger$ and by the recipe of Section 2 we obtain the concrete semantics

$$\mathcal{C}[\![c]\!] : \mathcal{P}(\mathsf{States}) \to \mathcal{P}(\mathsf{States})$$

of every command c.

4.2 Symbolic Semantics

The symbolic execution semantics $\sigma, A \implies \sigma'$ takes a symbolic heap σ and an atomic command, and transforms it into an output symbolic heap or $\top$. In these rules we require that the primed variables x', y' be fresh.

Symbolic Execution Rules

$$\begin{array}{lll}
\Pi \mid \Sigma, & x{:=}\,E & \implies x{=}E[x'/x] \wedge (\Pi \mid \Sigma)[x'/x] \\
\Pi \mid \Sigma * E{\mapsto}F, & x{:=}\,[E] & \implies x{=}F[x'/x] \wedge (\Pi \mid \Sigma * E{\mapsto}F)[x'/x] \\
\Pi \mid \Sigma * E{\mapsto}F, & [E]{:=}\,G & \implies \Pi \mid \Sigma * E{\mapsto}G \\
\Pi \mid \Sigma, & \mathbf{new}(x) & \implies (\Pi \mid \Sigma)[x'/x] * x{\mapsto}y' \\
\Pi \mid \Sigma * E{\mapsto}F, & \mathbf{dispose}(E) & \implies \Pi \mid \Sigma
\end{array}$$

$$\frac{\Pi \mid \Sigma \nvdash \mathtt{Allocated}(E)}{\Pi \mid \Sigma,\ A(E) \implies \top}$$

Rearrangement Rules

$$P(E,F) \;::=\; E\mapsto F \mid \mathsf{ls}(E,F)$$

$$\frac{\Pi_0 \mathrel{¦} \Sigma_0 * P(E,G),\ A(E) \implies \Pi_1 \mathrel{¦} \Sigma_1}{\Pi_0 \mathrel{¦} \Sigma_0 * P(F,G),\ A(E) \implies \Pi_1 \mathrel{¦} \Sigma_1}\ \Pi_0 \vdash E{=}F$$

$$\frac{\Pi_0 \mathrel{¦} \Sigma_0 * E\mapsto x' * \mathsf{ls}(x',G),\ A(E) \implies \Pi_1 \mathrel{¦} \Sigma_1}{\Pi_0 \mathrel{¦} \Sigma_0 * \mathsf{ls}(E,G),\ A(E) \implies \Pi_1 \mathrel{¦} \Sigma_1}$$

$$\frac{\Pi \mathrel{¦} \Sigma * E\mapsto F,\ A(E) \implies \Pi' \mathrel{¦} \Sigma'}{\Pi \mathrel{¦} \Sigma * \mathsf{ls}(E,F),\ A(E) \implies \Pi' \mathrel{¦} \Sigma'}$$

The execution rules that access heap cell E are stated in a way that requires their pre-states to explicitly have $E\mapsto F$. Sometimes the knowledge that E is allocated is less explicit, such as in $E{=}x \mathrel{¦} x\mapsto y$ or $\mathsf{ls}(E,F)$, and we use rearrangement rules to put the pre-state in the proper form. The first rearrangement rule simply makes use of equalities to recognize that a dereferencing step is possible, and the other two correspond to unrolling a list segment.

In contrast to the concrete semantics, the treatment of allocation is completely deterministic[3]. However, a different kind of nondeterminism results in rearrangement rules that unroll list segments.

All that is left to define the symbolic (intermediate) semantics $\mathcal{I}[\![c]\!]\colon \mathcal{P}(\mathcal{SH}) \to \mathcal{P}(\mathcal{SH})$ by the recipe before is to define the filter map. It tosses in the equality for the $E{=}F$ case, but does not do so for the $E{\neq}F$ case because we do not have inequalities in our symbolic domain.

$$\mathsf{filter}(E{=}F)X = \{\top \mid \top{\in}X\} \cup \{(E{=}F \wedge \Pi \mathrel{¦} \Sigma) \mid \Pi \mathrel{¦} \Sigma \in X \text{ and } \Pi \mathrel{¦} \Sigma \nvdash E{\neq}F\}$$
$$\mathsf{filter}(E{\neq}F)X = \{\top \mid \top{\in}X\} \cup \{(\Pi \mathrel{¦} \Sigma) \in X \mid \Pi \nvdash E{=}F \text{ and } \Pi \mathrel{¦} \Sigma \nvdash \mathsf{false}\}$$

To state the sense in which the symbolic semantics is sound we define the "meaning function" $\gamma\colon \mathcal{P}(\mathcal{SH}) \to \mathcal{P}(\mathsf{States})$:

$$\gamma(X) = \begin{cases} \mathsf{States} \cup \{\top\} & \text{if } \top \in X \\ \{(s,h) \mid \exists \Pi \mathrel{¦} \Sigma \in X.\ (s,h) \models \Pi \mathrel{¦} \Sigma\} & \text{otherwise} \end{cases}$$

Theorem 2. *The symbolic semantics is a sound overapproximation of the concrete semantics.*

$$\forall X \in \mathcal{P}(\mathcal{SH}).\ \mathcal{C}[\![c]\!](\gamma(X)) \subseteq \gamma(\mathcal{I}[\![c]\!]X).$$

5 The Analysis

The domain $\mathcal{SH}$ of symbolic heaps is infinite. Even though there are finitely many program variables, primed variables can be introduced during symbolic

[3] Provided we fix a deterministic way to choose fresh primed variables.

Table 2 Abstraction Rules

$$\frac{}{E{=}x' \wedge \Pi \mid \Sigma \rightsquigarrow (\Pi \mid \Sigma)[E/x']}\ (\mathsf{St1}) \qquad \frac{}{x'{=}E \wedge \Pi \mid \Sigma \rightsquigarrow (\Pi \mid \Sigma)[E/x']}\ (\mathsf{St2})$$

$$\frac{x' \notin \mathsf{Vars}'(\Pi, \Sigma)}{\Pi \mid \Sigma * P(x', E) \rightsquigarrow \Pi \mid \Sigma \cup \mathsf{junk}}\ (\mathsf{Garbage1})$$

$$\frac{x', y' \notin \mathsf{Vars}'(\Pi, \Sigma)}{\Pi \mid \Sigma * P_1(x', y') * P_2(y', x') \rightsquigarrow \Pi \mid \Sigma \cup \mathsf{junk}}\ (\mathsf{Garbage2})$$

$$\frac{x' \notin \mathsf{Vars}'(\Pi, \Sigma, E, F) \qquad \Pi \vdash F{=}\mathsf{nil}}{\Pi \mid \Sigma * P_1(E, x') * P_2(x', F) \rightsquigarrow \Pi \mid \Sigma * \mathsf{ls}(E, \mathsf{nil})}\ (\mathsf{Abs1})$$

$$\frac{x' \notin \mathsf{Vars}'(\Pi, \Sigma, E, F, G, H) \quad \Pi \vdash F{=}G}{\Pi \mid \Sigma * P_1(E, x') * P_2(x', F) * P_3(G, H) \rightsquigarrow \Pi \mid \Sigma * \mathsf{ls}(E, F) * P_3(G, H)}\ (\mathsf{Abs2})$$

execution. For example, in a loop that includes allocation we can generate formulae $x{\mapsto}x' * x'{\mapsto}x'' \cdots$ of arbitrary length.

In order to ensure fixed-point convergence we perform abstraction. The abstraction we consider is specified by a collection of rewrite rules which perform abstraction by gobbling up primed variables. This is done by merging lists, swallowing single cells into lists, and abstracting two cells by a list. We also remove primed variables from the pure parts of formulae, and we collect all garbage into the predicate junk.

5.1 Canonicalization Rules

The canonicalization rules are reported in Table 2. We again use the notation $P(E, F)$ to stand for an atomic formula either of the form $E{\mapsto}F$ or $\mathsf{ls}(E, F)$.

The most important rules are the last two. The sense of abstraction that these rules implement is that we ignore any facts that depend on a midpoint in a list segment, unless it is named by a program variable. There is a subtlety in interpreting this statement, however. One might perhaps have expected the last rule to leave out the $P_3(G, H)$ $*$-conjunct, but this would result in unsoundness; as Berdine and Calcagno pointed out [4, 5], we must know that the end of a second list segment does not point back into the first if we are to concatenate them. We are forced, by considerations of soundness, to keep some primed midpoints, such as in the formula $\mathsf{ls}(x, x') * \mathsf{ls}(x', y)$, to which no rewrite rule applies.

Notice the use of a $\cup$ rather than a $*$ on the rhs of the (Garbage1) and (Garbage2) rules. This has the effect that when more than one unreachable node named by a primed variable is present, all of them get put into the unique junk node. Having the junk node helps in obtaining a finite abstract semantics.

5.2 The Algorithm

We say that $\Pi \mid \Sigma$ is a canonical symbolic heap if it is consistent (i.e., $\Pi \mid \Sigma \nvdash \mathsf{false}$) and no canonicalization rule applies to it, and we denote by $\mathcal{CSH}$ the set of all such. We can immediately observe:

Lemma 3 (Strong Normalization). $\rightsquigarrow$ *has no infinite reduction sequences.*

This, together with the results in the next section, would be enough to define a terminating analysis. But, there are many distinct reduction sequences and to try all of them in an analysis would lead to a massive increase in nondeterminism. We have not proven a result to the effect that choosing different reduction sequences matters in the final result (after applying the meaning function γ, but neither have we found examples where the difference can be detected. So, in our implementation we have chosen a specific strategy which applies the equality rules, followed by (Garbage1), followed by abstraction rules, followed by (Garbage2). In the theory, we just presume that we have a function (rather than relation)

$$\mathsf{can}\colon \mathcal{SH} \to \mathcal{CSH}$$

which takes a symbolic heap $\Pi \mid \Sigma$ and returns a canonical symbolic heap $\Pi' \mid \Sigma'$ where $\Pi \mid \Sigma \rightsquigarrow^* \Pi' \mid \Sigma'$.

[We remark that $\mathsf{can}(\Pi \mid \Sigma)$ is not the best (logically strongest) canonical heap implied by $\Pi \mid \Sigma$. A counterexample is $\{\} \mid \{x \mapsto x', x' \mapsto y, y \mapsto \mathsf{nil}\}$. This symbolic heap is reduced to $\{\} \mid \{\mathsf{ls}(x, y), y \mapsto \mathsf{nil}\}$ by the canonicalization, but implies another symbolic heap $\{\} \mid \{x \mapsto x', x' \mapsto z', y \mapsto \mathsf{nil}\}$, which is not (logically) weaker than $\{\} \mid \{\mathsf{ls}(x, y), y \mapsto \mathsf{nil}\}$. We believe that this "problem" is fixable; we conjecture that there is a preorder $\sqsubseteq$ on $\mathcal{SH}$ such that (i) $\sqsubseteq$ is a sub preorder of the logical implication and (ii) $\mathsf{can}(\Pi \mid \Sigma)$ is the smallest canonical heap greater than or equal to $\Pi \mid \Sigma$ with respect to $\sqsubseteq$. As of this writing we have not succeeded in proving this conjecture. If true, it would perhaps open the way to a study pinpointing where precision is and is not lost (as in, e.g., [3]) using Galois connections. Although valuable, such questions are secondary to our more basic aim of existence (soundness and termination) of the analysis.]

Let $\mathtt{in}\colon \mathcal{P}(\mathcal{CSH}) \to \mathcal{P}(\mathcal{SH})$ denote the inclusion function. We define the abstract semantics for each primitive command p by the equation

$$\mathcal{A}[\![p]\!] \;=\; \mathtt{in}\,;\mathcal{I}[\![p]\!]\,;(\mathsf{can}^{\dagger}).$$

The filtering map in the abstract semantics is just the restriction of the symbolic one to $\mathcal{CSH}$. Then, by the recipe from Section 2 we obtain a semantics

$$\mathcal{A}[\![c]\!] : \mathcal{P}(\mathcal{CSH}) \to \mathcal{P}(\mathcal{CSH})$$

for every command.

The soundness of the abstract semantics relies on the soundness of the rewriting rules.

Lemma 4 (Soundness of $\rightsquigarrow$). *If $\Sigma \mid \Pi \rightsquigarrow \Sigma' \mid \Pi'$ then $\Sigma \mid \Pi \vdash \Sigma' \mid \Pi'$.*

The statement of soundness of the abstract semantics is then the same as for the symbolic semantics, except that we quantify over $\mathcal{P}(\mathcal{CSH})$ instead of $\mathcal{P}(\mathcal{SH})$.

Theorem 5. *The abstract semantics is a sound overapproximation of the concrete semantics.*

$$\forall X \in \mathcal{P}(\mathcal{CSH}).\ \mathcal{C}[\![c]\!](\gamma(X)) \subseteq \gamma(\mathcal{A}[\![c]\!]X).$$

Here are some examples of running the analysis on particular pre-states, taken from an implementation of it in OCaml.

Example 1. This is the usual program to reverse a list. Here 0 is used to denote nil and $x \rightarrow tl$ is used instead of $[x]$.

```
Program: while (c!=0) {n=c->tl; c->tl=p; p=c; c=n; };
Precondition: {true}|{ls(c,0)}
Invariant: {p=0}|{ls(c,0)}  OR  {c=n AND n=0}|{p|->0}  OR
           {c=n}|{p|->0 * ls(n,0)}  OR  {c=n AND n=0}|{ls(p,0)}  OR
           {c=n}|{ls(p,0) * ls(n,0)}
Postcondition: {c=0 AND c=n AND n=0}|{ls(p,0)}  OR
               {c=0 AND c=n AND n=0}|{p|->0}
```

Given a linked list as a precondition, the analysis calculates that the postcondition might be a linked list or a single points-to fact. The postcondition has some redundancy, in that we could remove the second disjunct without affecting the meaning; this is because we have used the subset ordering on sets of states, rather than one based on implication. The analysis also calculates the pictured loop invariant along the way.

When we apply the analysis to the same program but give a circular linked list as input, we get the following (we omit the calculated invariant, which has 11 disjuncts).

```
Precondition: {true}|{ls(c,c_) * ls(c_,c)}
Postcondition: {c=0 AND c=n AND n=0}|{p|->p_ * ls(p_,p)}  OR
               {c=0 AND c=n AND n=0}|{p|->p_ * p_|->p}
```

Example 2. This is the program to dispose a list.

```
Program: while (c!=0) {t=c; c=c->tl; dispose(t); };
Precondition: {true}|{ls(c,0)}
Invariant: {c=0}|{emp}  OR  {true}|{ls(c,0)}
Postcondition: {c=0}|{emp}
```

We properly get emp on termination. If we leave out the dispose instruction, it returns postcondition $c = \mathsf{nil} \mid t \mapsto \mathsf{nil} * \mathsf{junk}$ (showing memory leak). When we run the analysis on this program on a circular list or $\mathsf{ls}(c, d)$ it reports a memory fault.

In addition to these examples we have run the analysis on a range of other small programs, such as list append, list copy, and programs to insert and delete from the middle of a list. For what it's worth, all of them completed in milliseconds running on a PowerBook G4.

6 Termination

Although the abstract semantics exists, we have not yet established that the "algorithm" it determines always terminates. We do that by showing that the abstract domain $\mathcal{CSH}$, consisting of the normal forms of the rewriting rules, is finite.

To gain some insight into the nature of the canonical symbolic heaps here are some examples, where the pure part Π is empty (and left out).

Irreducible	**Reducible**
$\mathsf{ls}(x,x') * \mathsf{ls}(y,x') * \mathsf{ls}(x',\mathsf{nil})$	$\mathsf{ls}(x,x') * \mathsf{ls}(x',y') * \mathsf{ls}(y',\mathsf{nil})$
$\mathsf{ls}(x,x') * \mathsf{ls}(x',x)$	$\mathsf{ls}(x,y') * \mathsf{ls}(y',x') * \mathsf{ls}(x',x)$
$\mathsf{ls}(x,x')$	$\mathsf{ls}(x',x)$
$\mathsf{ls}(x,x') * \mathsf{ls}(x',y)$	$\mathsf{ls}(x,x') * \mathsf{ls}(x',y) * \mathsf{ls}(y,z)$

In the first element of the first row, variable x' is shared (pointed to by x and y), and this blocks the application of rule (Abs1) because of its variable condition. On the other hand, the second element can be reduced, in fact twice, to end up with $\mathsf{ls}(x,\mathsf{nil})$. The second row contains two cycles, one of (syntactic) length two and the other of length three. The first of these cannot be reduced. We would need to know that $x{=}\mathsf{nil}$ to apply (Abs1) and we cannot, because x in $\mathsf{ls}(x,x')$ cannot be nil or else we would have an inconsistent formula. The second in this row can, however, be reduced, to the first. In the third row x' is reachable variable that possibly denotes a dangling pointer and there is no way to eliminate it. In the second it is not reachable, and can be removed using the (Garbage1) rule. In the final row, first x' points to a possibly dangling variable y. We cannot gobble x' up because to do so would result in unsoundness, and the rule (Abs2) is arranged to prevent this. If we tack on another heap formula to ensure that y is not dangling then (Abs2) can apply.

Based on these ideas we can characterize the normal forms of $\rightsquigarrow^*$ using "graphical" ideas of path and reachability, as well as conditions about sharing, cycles, and dangling pointers.

Definition 6. *1. A* path *in $\Pi \mid \Sigma$ is a sequence of expressions $E_0, E_1, \ldots, E_n$ such that*

$$\forall i \in \{1,\ldots,n\}.\ \exists E, E'.\ \ \Pi \vdash E_{i-1}{=}E \ \text{ and } \ \Pi \vdash E_i{=}E' \ \text{ and } \ P(E,E') \in \Sigma.$$

Reachability between expressions is defined in the usual way: E is reachable from *E'* in *$\Pi \mid \Sigma$ if and only if there is a path in $\Pi \mid \Sigma$ that starts from E and ends in E'.*

2. *An expression E in $\Pi \mid \Sigma$ is* shared *if and only if Σ contains two distinct elements $P_0(E_0, E_0')$ and $P_1(E_1, E_1')$ such that $\Pi \vdash E{=}E_0'$ and $\Pi \vdash E{=}E_1'$.*
3. *A primed variable x' in a cycle (a path from E to itself) is an* internal node *if and only if it is not shared.*

4. E *is called* possibly dangling *in* $\Pi \mid \Sigma$ *if and only if*
   (a) $\Pi \not\vdash E{=}\mathsf{nil}$,
   (b) *there exists some* E' *such that* $\Pi \vdash E{=}E'$ *and* E' *has right occurrences in* Σ*, and*
   (c) *there are no expressions* F' *such that* $\Pi \vdash E{=}F'$ *and* F' *has left occurrences in* Σ.
5. E points to a possibly dangling expression *if and only if there are* E', F *such that* $\Pi \vdash E{=}E'$, $P(E', F) \in \Sigma$*, and* F *possibly dangles.*

Definition 7 (Reduced Symbolic Heap). *A symbolic heap* $\Pi \mid \Sigma$ *is* reduced *if and only if*

1. Π *does not contain primed variables;*
2. *every primed variable* x' *in* Σ *is reachable from some unprimed variable; and*
3. *for every reachable variable* x'*, either*
   (a) x' *is shared, or*
   (b) x' *is the internal node of a cycle of length precisely two, or*
   (c) x' *points to a possibly dangling variable, or*
   (d) x' *is possibly dangling.*

In (b) of this definition the length refers to the syntactic length of a path, not the length of a denoted cycle. For example, $\mathsf{ls}(x, x') * \mathsf{ls}(x', x)$ has syntactic length two, even though it denotes cycles of length two or greater.

This definition of reduced heaps is not particularly pretty; its main point is to give us a way to prove termination of our analysis.

Proposition 8 (Canonical Characterization). *When a symbolic heap* $\Pi \mid \Sigma$ *is consistent,* $\Pi \mid \Sigma$ *is reduced if and only if* $\Pi \mid \Sigma \not\rightsquigarrow$.

We consider the formulae in $\mathcal{CSH}$ as being equivalent up to α-renaming, that is, renaming of primed variables. With this convention, we can show $\mathcal{CSH}$ finite.

Proposition 9. $\mathcal{CSH}$ *is finite.*

The proof of this proposition proceeds by first showing a lemma that bounds the number of primed variables in any reduced form (this bound is very coarse). In essence, the no sharing part of the definition of "reduced" stops there being infinitely many possible primed variables (starting from a fixed finite set of program variables). We have obtained a coarse bound of $7n$. This bound then limits the number of atomic predicates that can appear in such formulae, giving us finiteness. The overall bound one obtains is exponential. Again, our argument (in the appendix) produces only a coarse[4], extremely large, bound of $2^{(129n^2+18n+2)}$.

This the leads us to

Theorem 10. *The algorithm specified by* $\mathcal{A}[\![\cdot]\!]$ *always terminates.*

[4] E.g., in this bound we do not exclude inconsistent states which would be a very large part of them.

7 Locality

We now describe locality properties of the semantics, beginning with an example. Suppose that we have a queue, represented in memory as a list segment from c to d. An operation for getting an element is

```
x = c; c = c->tl; /* get from left of queue, put in x */
```

The list segment might not be the whole storage, of course. In particular, we might have an additional element pointed to by d which is (perhaps) used to place an element into the queue. When we run our tool on an input reflecting this state of affairs we obtain

```
Precondition: {true}|{ls(c,d) * d|->d_}
Postcondition: {c=d}|{x|->d * d|->d_}  OR
               {true}|{x|->c * ls(c,d) * d|->d_}
```

However, it is clear that the $d \mapsto d_{\_}$ information is irrelevant, that a run of the tool on the smaller input gives us all the information we need.

```
Precondition: {true}|{ls(c,d)}
Postcondition: {c=d}|{x|->d}  OR  {true}|{x|->c * ls(c,d)}
```

In fact, the behaviour of the tool in the first case follows from that in the second, using the Frame Rule of separation logic.

This example is motivated by the treatment of a concurrent queue in [10]. The fact that we do not have to consider the cell d when inserting is crucial for a verification which shows that the two ends of a nonempty queue can be manipulated concurrently. To produce such results from an analysis, rather than a by-hand proof, we would similarly like to avoid the need to analyze the entire state including the cell d.

We can give a theoretical account of the locality of our analysis using the following notions. First, we define a notion of $*$ on entire symbolic heaps.

$$(\Pi_1 \mathrel{\vdots} \Sigma_1) * (\Pi_2 \mathrel{\vdots} \Sigma_2) \;=\; (\Pi_1 \cup \Pi_2 \mathrel{\vdots} \Sigma_1 * \Sigma_2).$$

This is a partial operation, which is undefined when $\Sigma_1 * \Sigma_2$ is undefined, or when $(\Pi_1 \cup \Pi_2 \mathrel{\vdots} \Sigma_1 * \Sigma_2)$ is inconsistent, or when some primed variable appears both in $\Pi_1 \mathrel{\vdots} \Sigma_1$ and in $\Pi_2 \mathrel{\vdots} \Sigma_2$. We extend this to $\mathcal{SH} \cup \{\top\}$ by stipulating $(\Pi \mathrel{\vdots} \Sigma) * \top \;=\; \top \;=\; \top * (\Pi \mathrel{\vdots} \Sigma)$. It then lifts to a total binary operation on $\mathcal{P}(\mathcal{SH})$ by

$$X * Y \;=\; \{\sigma_1 * \sigma_2 \mid \sigma_1 \in X, \sigma_2 \in Y\}.$$

To formulate the locality property we suppose a fixed set *Mod* of modified variables, that appear to the left of ≔ or in $\mathbf{new}(x)$ in a given command c.

Theorem 11 (Frame Rule). *$\forall X, Y \in \mathcal{P}(\mathcal{CSH})$, if $\mathsf{Vars}(Y) \cap Mod = \emptyset$ then*

$$\gamma(\mathcal{A}[\![c]\!](X * Y)) \subseteq \gamma((\mathcal{A}[\![c]\!]X) * Y).$$

There are two reasons why we get an overapproximation $\subseteq$ rather than exact match here. First, and trivially, there might be states in X where c faults, returns $\top$, while it never does for states in $X * Y$. The second reason is best understood by example. When the program

$$\mathbf{new}(x)\ ;\ (\mathbf{if}\ x{=}y\ \mathbf{then}\ z{:=}\,a\ \mathbf{else}\ z{:=}\,b)\ ;\ \mathbf{dispose}(x)$$

is run in the empty heap, it returns a pair of post-states, one where $z{=}a$ and the other where $z{=}b$. But when run in $y{\mapsto}y'$ the **if** branch is ruled out and we only get $z{=}b \mid y{\mapsto}y'$ as a possible conclusion. However, we get the state $z{=}a \mid y{\mapsto}y'$ as an additional possibility starting from $y{\mapsto}y'$, when we put the small output together with $y{\mapsto}y'$ using $*$. Although precision can be lost when passing to smaller states, it is often an acceptable loss.

For a given command c and symbolic heap σ we define

1. $\mathsf{safe}(c,\sigma)$ iff $\top \notin \mathcal{A}[\![c]\!]\{\sigma\}$
2. $\sigma_1 \preceq \sigma_3$ iff $\exists \sigma_2.\, \sigma_3 \;=\; \sigma_1 * \sigma_2$
3. $\sigma \prec \sigma'$ iff $\sigma \preceq \sigma'$ and $\sigma \neq \sigma'$
4. $\mathsf{onlyaccesses}(c,\sigma)$ iff $\mathsf{safe}(c,\sigma)$ and $\neg\exists \sigma' \prec \sigma.\, \mathsf{safe}(c,\sigma')$.

The notion of accesses is coarse. For example, $\mathsf{onlyaccesses}([x]{:=}\,y, \mathsf{ls}(x,\mathsf{nil}))$ holds, even though a single cell can be picked out of the list segment. A stronger notion of accesses, and hence footprint, might be formulated taking implications between symbolic heaps into account as well as $\preceq$.

The footprint is partial function $\mathsf{foot}(c){:}\,\mathcal{CSH} \rightharpoonup \mathcal{P}(\mathcal{CSH})$,

$$\mathsf{foot}(c)\sigma \;=\; \text{if } (\mathsf{onlyaccesses}(c,\sigma)) \text{ then } (\mathcal{A}[\![c]\!]\{\sigma\}) \text{ else (undefined)}.$$

The point of the footprint is that, as a set of pairs, it can be compact compared to the entire meaning. For the disposelist program the footprint has three entries, with preconditions $\{\} \mid \{\mathsf{ls}(c,\mathsf{nil})\}$, $\{\} \mid \{c{\mapsto}\mathsf{nil}\}$ and $\{c{=}\mathsf{nil}\} \mid \{\}$. The entire meaning has 16 entries, corresponding to the number of canonical symbolic heaps over a single input variable c.

To express the sense in which the footprint is a sound representation of the semantics of c we show how any potential footprint can be "fleshed out" by applying the idea behind the Frame Rule. Again, let Mod be the set of modified variables in a given command c, and for each symbolic heap $\Pi \mid \Sigma$, let $\mathsf{unaffectedEqs}(\Pi \mid \Sigma)$ be the set of equalities $E{=}F$ in Π such that $\mathsf{Vars}(E{=}F) \cap Mod \,=\, \emptyset$. If $f{:}\,\mathcal{CSH} \rightharpoonup \mathcal{P}(\mathcal{CSH})$, then $\mathsf{flesh}(f){:}\,\mathcal{CSH} \to \mathcal{P}(\mathcal{CSH})$ is defined as follows:

$$\begin{array}{rcl}
\mathsf{validSplit}(\sigma_0,\sigma_1,\sigma) & \iff & \sigma_0 * \sigma_1 = \sigma \text{ and } \mathsf{Vars}(\sigma_1) \cap Mod = \emptyset \text{ and} \\
& & \sigma_0 \in \mathbf{dom}(f) \text{ and } \mathsf{unaffectedEqs}(\sigma_1) = \mathsf{unaffectedEqs}(\sigma) \\
\mathsf{flesh}(f)\sigma & = & \text{if } (\neg\exists \sigma_0,\sigma_1.\, \mathsf{validSplit}(\sigma_0,\sigma_1,\sigma)) \text{ then } \{\top\} \\
& & \text{else let } \sigma_0',\sigma_1' \text{ be symbolic heaps s.t. } \mathsf{validSplit}(\sigma_0',\sigma_1',\sigma) \\
& & \quad\text{in } \mathcal{P}(\mathsf{can})\big(f(\sigma_0') * \{\sigma_1'\}\big)
\end{array}$$

The fleshing out picks one access point, and adds as many $*$-separated invariants as possible to the access point.

Theorem 12. *The footprint is a sound overapproximation of the abstract semantics:*

$$\forall X \in \mathcal{P}(\mathcal{CSH}).\ \gamma(\mathcal{A}[\![c]\!]X) \subseteq \gamma(\mathsf{foot}(c)^{\dagger}X).$$

Although theoretically incomplete, the footprint very often does give an accurate representation of a program's abstract meaning.

The calculation of whole footprints is, of course, not realistic. A more practical way to employ the footprint idea would be, given an input state σ, to look at substates on which a procedure or command does not produce a fault. In interprocedural analysis, we might record the input-output behaviour on as small states as possible when tabulating a procedure summary. This would be similar to [15], but would not involve entire reachable substates. In concurrency, we would look for disjoint substates of an input state on which to run parallel commands: if these input states were safe for the commands in question, then we could soundly avoid (many) interleavings during symbolic execution. We hope to report on these matters at a later time.

References

[1] T. Amtoft, S. Bandhakavi, and A. Banerjee. A logic for information flow analysis of pointer programs. 33rd POPL, to appear, 2006.

[2] T. Amtoft and A. Banerjee. Information flow analysis in logical form. 11th Static Analysis Symposium, LNCS3184, pp100-115, 2004.

[3] Thomas Ball, Andreas Podelski, and Sriram K. Rajamani. Boolean and Cartesian abstraction for model checking C programs. *7th TACAS, LNCS*, 2031:268–283, 2001.

[4] J. Berdine, C. Calcagno, and P. O'Hearn. A decidable fragment of separation logic. Proceedings of FSTTCS, LNCS 3328, Chennai, December, 2004.

[5] Josh Berdine, Cristiano Calcagno, and Peter W. O'Hearn. Symbolic execution with separation logic. In K. Yi, editor, *APLAS 2005*, volume 3780 of *LNCS*, 2005.

[6] P. Cousot and R. Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. 4th ACM Symposium on Principles of Programming Languages. pages 238–252, 1977.

[7] D. Distefano. *On model checking the dynamics of object-based software: a foundational approach*. PhD thesis, University of Twente, 2003.

[8] D. Distefano, A. Rensink, and J.-P. Katoen. Who is pointing when to whom: on model-checking pointer structures. CTIT Technical Report TR-CTIT-03-12, Faculty of Informatics, University of Twente, March 2003.

[9] S. Magill, A. Nanevski, E. Clarke, and P. Lee. Inferring invariants in Separation Logic for imperative list-processing programs. Draft, July 2005, 2005.

[10] P. O'Hearn. Resources, concurrency and local reasoning. *Theoretical Computer Science*, 2006. to appear. Preliminary version appeared in CONCUR'04, LNCS 3170, 49–67.

[11] P. O'Hearn, J. Reynolds, and H. Yang. Local reasoning about programs that alter data structures. In *Proceedings of 15th Annual Conference of the European Association for Computer Science Logic*, LNCS, pages 1–19. Springer-Verlag, 2001.

[12] P. W. O'Hearn, H. Yang, and J. C. Reynolds. Separation and information hiding. In *31st POPL*, pages 268–280, 2004.
[13] J. C. Reynolds. Separation logic: A logic for shared mutable data structures. In *17th LICS*, pp 55-74, 2002.
[14] N. Rinetzky, J. Bauer, T. Reps, S. Sagiv, and R. Wilhelm. A semantics for procedure local heaps and its abstractions. *32nd POPL*, pp296–309, 2005.
[15] N. Rinetzky, M. Sagiv, and E. Yahav. Interprocedural shape analysis for cutpoint-free programs. In *12th International Static Analysis Symposium (SAS)*, 2005.
[16] M. Sagiv, T. Reps, and R. Wilhelm. Parametric shape analysis via 3valued logic. *ACM Trans. Program. Lang. Syst.*, 24(3):217–298, 2002.

Appendix

This appendix gives the proofs of the key elements that ensure termination.

Proof of Proposition 8

Proposition 8
When a symbolic heap $\Pi \mid \Sigma$ is consistent, $\Pi \mid \Sigma$ is reduced if and only if $\Pi \mid \Sigma \not\rightsquigarrow$.

Proof
[Only if Part]. Suppose that $\Pi \mid \Sigma$ is reduced; we need to show that no rule applies. By condition 1 of Definition 7 neither (St1) nor (St2) can apply, and none of (Garbage1) and (Garbage2) can apply by condition 2.

For (Abs1) to apply, note that x' in the rule is not shared because of the assumption $x' \notin \mathsf{Vars}'(\Sigma, \Pi, E, F)$ in the rule. It neither dangles nor points to a dangler because $\mathsf{ls}(x', F)$ and $\Pi \vDash F{=}\mathsf{nil}$. It is not in a cycle because $\mathsf{ls}(x', F)$ and $\Pi \vDash F{=}\mathsf{nil}$ together preclude any cycle in consistent symbolic heaps. Thus, if $\Pi \mid \Sigma$ is reduced (Abs1) cannot apply.

For (Abs2) to apply, we similarly must have that x' is not shared, doesn't dangle, and doesn't point to a dangler. It can't be in a cycle of length two either, because the only way to get a cycle through x' in the rule is to go from E through x' to $F{=}G$ and on to H. We know that G and E cannot be equal because they are lhs's that are $*$-separated and $\Pi \mid \Sigma$ is consistent. That rules out one way of getting a cycle of length two. The only other is to start from x', go through $F{=}G$, and come back to x' if $H{=}x'$. But $H{=}x'$ is ruled out by the variable condition in the rule. Thus, (Abs2) cannot apply.

[If Part]. If $\Pi \mid \Sigma$ is not reduced, we want to show that a rule applies. If condition 1 of Definition 7 fails then we can apply one of (St1) or (St2). If condition 2 does not hold, we have a primed variable x' in Σ that is not reachable from unprimed variables. Suppose that the (Garbage1) rule cannot fire on this $\Pi \mid \Sigma$. Since the (Garbage1) rule cannot fire, x' should be in a cycle. We make three observations about this cycle. First, all the variables in the cycle as well as all the variables from which we can reach x' should be primed, because x' is not reachable from unprimed variables. Second, the length of the cycle of x' is at least 2. For, if the length is one, then there would be some primed variable y' such that x' is reachable from y' and y' occurs in only one predicate $P(y', E)$ of Σ; so, (Garbage1) can be applied to y', contradicting to our assumption that (Garbage1) is not applicable. Third, x' is reachable only from some cells in the cycle. If y' is not in the cycle but it can reach to x', then (Garbage1) can be applied to some primed variable z', which contradicts to our assumption. Because of these observations, it is sufficient to consider two subcases: when the length of the cycle is greater than 2, and when the length of the cycle is 2. In the first subcase, we can apply (Abs2), because the third observation ensures that every primed variables in the cycle occurs twice, one in the lhs and the other in the rhs. The same property about the primed variable in the cycle also ensures that in the second subcase, we can apply (Garbage2).

Because of what we have just shown, we suppose that condition 1 and 2 do hold, but condition 3 does not. We need to argue that either (Abs1) or (Abs2) applies. In more detail,

> *if* x' is not shared, not the internal node of a cycle of length two, not a dangler, and not pointing to a dangler,
> *then* we can apply one of the two rules.

Since x' is reachable from unprimed variables and it is not in Π, we must have $P_1(E, x') \in \Sigma$ for some $E \not\equiv x'$. Since x' is not a dangler, we must have $P_2(x', F) \in \Sigma$. Now, $P_1(E, x')$ and $P_2(x', F)$ must be distinct elements of Σ, because E and x' are syntactically different. So we have at least the $*$-combination $P_1(E, x') * P_2(x', F)$ as part of Σ.

Since x' is not shared, it cannot appear as an rhs other than in $P_1(E, x')$. And it cannot appear as any lhs other than in $P_2(x', F)$ because that would imply that Σ is inconsistent. So, x' is not in Σ, other than in $P_1(E, x')$ and $P_2(x', F)$. Next, there are two subcases.

The first subcase is that $\Pi \vdash F{=}\mathsf{nil}$. Then, since we have shown that x' is not in Σ, other than in $P_1(E, x')$ and $P_2(x', F)$, and x' is syntactically different from E, F, the variable condition in rule (Abs1) holds and it can fire.

The second subcase is that $\Pi \not\vdash F{=}\mathsf{nil}$. This, combined with the assumption that x' does not point to a dangler, implies that there is $P_3(G, H) \in \Sigma$ with $\Pi \vdash F{=}G$. Predicate $P_3(G, H)$ is distinct from $P_2(x', F)$. If $P_3(G, H)$ and $P_2(x', F)$ were identical, then both G and F would be x'. But, x' appears as an rhs only in $P_1(E, x')$, so it is different from F and G. The requirement that x' is not the internal node of a cycle of length two implies that $P_3(G, H)$ cannot be the same as $P_1(E, x')$, because otherwise we would have a cycle from x' to E and back to x'. So, $P_1(E, x')$, $P_2(x', F)$ and $P_3(G, H)$ are all distinct elements of Σ, and we have a $*$-combination $P_1(E, x') * P_2(x', F) * P_3(G, H)$. Here x' is different from H, because x' appears as an rhs only in $P_1(E, x')$. Then, since we have shown earlier that x' is not in Σ, other than in $P_1(E, x')$ and $P_2(x', F)$, and that x' is different from E, F, G, H, the variable condition in rule (Abs2) holds and it can fire. □

Proposition 9 $\mathcal{CSH}$ is finite.

Proof The proof uses a lemma which bounds the number of primed variables in any reduced form (this bound is very coarse).

> Lemma: If $\Pi \mid \Sigma \in \mathcal{CSH}$ then the number of primed variables in Σ is bounded by $7n$, where n is the number of program variables.

Let X_s, X_c, X_p, X_d be sets of primed variables in $\Pi \mid \Sigma$, such that

- $x' \in X_s$ iff x' is shared,
- $x' \in X_c$ iff x' is the internal node of a cycle of length 2,
- $x' \in X_p$ iff x' points to a possibly dangling variable,
- $x' \in X_d$ iff x' is possibly dangling.

Then, $\{X_s, X_c{-}X_s, X_p{-}X_s, X_d{-}X_s\}$ is the partitioning of the set of primed variables in $\Pi \mid \Sigma$. We denote the cardinality of each partition by $m_s{=}|X_s|$, $m_c{=}|X_c{-}X_s|$, $m_p{=}|X_p{-}X_s|$, and $m_d{=}|X_p{-}X_s|$. Since $\Pi \mid \Sigma$ is consistent, the lhs's of the predicates in Σ are all different primed or unprimed variables, so the number of those predicates is bounded above by $n + m_s + m_c + m_p + m_d$. On the other hand, since every primed variable x' is reachable, it should occur on the lhs of some predicate in Σ, and if x' is shared, it should occur at least twice on the lhs's. Therefore, the number of the predicates in Σ is bounded below by $2m_s + m_c + m_p + m_d$. By combining the obtained lower and upper bounds of $|\Sigma|$, we get $2m_s + m_c + m_p + m_d \le n + m_s + m_c + m_p + m_d$, and this inequality gives the upper bound n on m_s:

$$m_s \le n.$$

We will now compute the upper bound for the other primed variables. For the upper bound of $m_c + m_p$, we use the fact that if x' is in $X_c{-}X_s$ or $X_p{-}X_s$, then Σ should contain $P(E, x')$ where E is a unprimed variable or a shared primed variable. Thus, the number of such x' is bounded by $m_s + n$, because by consistency of $\Pi \mid \Sigma$, the lhs's of the predicates in Σ are never the same. The upper bound of $m_c + m_p$ is:

$$m_c + m_p \le m_s + n \le 2n.$$

Finally, we compute the upper bound of m_d. For this, we use the fact that if x' is in $X_d{-}X_s$, then Σ should have $P(E, x')$ where E is a shared primed variable, or unprimed variable, or a unshared primed variable that possibly points to a dangling variable. Thus, the number of such x' is bounded by $m_s + n + m_p$. So, the upper bound of m_d is:

$$m_d \le m_s + n + m_p \le n + n + 2n = 4n.$$

Combining the obtained upper bounds together, we get $7n$ as the upper bound on the number of primed variables. This completes the proof of the lemma.

For the proof of the proposition first note that if $\Pi \mid \Sigma \in \mathcal{CSH}$ then Π can contain at most as $(n+1) \times (n+1)$ formulas, where n is the number of program variables, because there are $(n+1)$-many expressions. There are therefore at most $2^{(n^2+2n+1)}$ different Π's.

By the lemma, the number of variables (primed and unprimed) appearing in Σ is then bounded by $8n$. Thus, for each predicate P, there are $8n \times (8n+1)$ different atomic formulae of the form $P(E, E')$. Since each atomic formula is junk, or $\mathsf{ls}(E, E')$, or $E \mapsto E'$, there are $1 + 2 \times 8n \times (8n+1)$ atomic formulae in total. There are then at most $2^{(128n^2+16n+1)}$ Σ's.

Putting these observations together, we have $2^{(n^2+2n+1)} + 2^{(128n^2+16n+1)}$ as a bound on the number of symbolic heaps in $\mathcal{CSH}$. (We have argued a coarse bound for simplicity.) $\square$