Semantic-Directed Clumping of Disjunctive Abstract States *

Huisong Li, François Bérenger, Bor-Yuh Evan Chang, Xavier Rival

To cite this version:


HAL Id: hal-01648679
https://inria.hal.science/hal-01648679
Submitted on 27 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Semantic-Directed Clumping of Disjunctive Abstract States

Huisong Li
INRIA Paris / CNRS / École Normale Supérieure / PSL* Research University
huisong@di.ens.fr

Francois Berenger
INRIA Paris / CNRS / École Normale Supérieure / PSL* Research University
francois.berenger@inria.fr

Bor-Yuh Evan Chang
University of Colorado Boulder
evan.chang@colorado.edu

Xavier Rival
INRIA Paris / CNRS / École Normale Supérieure / PSL* Research University
rival@di.ens.fr

Abstract

To infer complex structural invariants, shape analyses rely on expressive families of logical properties. Many such analyses manipulate abstract memory states that consist of separating conjunctions of basic predicates describing atomic blocks or summaries. Moreover, they use finite disjunctions of abstract memory states in order to account for dissimilar shapes. Disjunctions should be kept small for scalability, though precision often requires keeping additional case splits. In this context, deciding when and how to merge case splits and to replace them with summaries is critical both for precision and efficiency. Existing techniques use sets of syntactic rules, which are tedious to design and prone to failure. In this paper, we design a semantic criterion to clump abstract states based on their silhouette, which applies not only to the conservative union of disjuncts but also to the weakening of separating conjunctions of memory predicates into inductive summaries. Our approach allows us to define union and widening operators that aim at preserving the case splits that are required for the analysis to succeed. We implement this approach in the MemCAD analyzer and evaluate it on real-world C codes from existing libraries dealing with doubly-linked lists, red-black trees, AVL-trees and splay-trees.

Categories and Subject Descriptors D.2.4 [Software Engineering]: Software/Program Verification; F.3.1 [Logics and Meanings of Programs]: Specifying and Verifying Reasoning about Programs; F.3.2 [Logics and Meanings of Programs]: Semantics of Programming Languages-Program analysis

Keywords Static analysis, abstract interpretation, heap abstraction, separation logics, disjunctions, silhouette, clumping of disjuncts

1. Introduction

Over the last two decades, a wide spectrum of analyses have been developed so as to compute invariants for programs manipulating complex, unbounded data structures, and in order to prove memory safety, structural invariance or functional properties. Shape analyses [17, 35] focus on dynamic structures such as linked lists, trees and graphs. Array analyses like [13] deal with contiguous structures indexed by ranges of integers. Dictionary analyses [14, 16] handle structures indexed using sets of keys. All these analyses need to use expressive sets of predicates in order to summarize structures of unbounded size, while keeping information about their shape.

Many of these works [9, 13, 14, 17] implicitly or explicitly use separation [32] to describe sets of concrete memory states, which means they divide memory states into regions and describe each region individually using specific predicates. Typically region predicates either describe a finite region very precisely (namely, cell-by-cell), or summarize an unbounded region using an inductive predicate [4, 9, 17, 38], a property that holds over a possibly empty range of cells [13] or for a set of keys [14]. An abstract state combines region predicates. For instance, [17] partitions memories into list segments. Similarly, [13] partitions arrays into segments. Thus, an abstract state can be viewed as a logical formula described by a grammar of the form:

\[
\text{abstract states: } \exists m(x) ::= p_1 \ast \ldots \ast p_n \\
\text{region predicates: } p(x) ::= p_e (\text{exact description}) \\
\text{summary} p_s
\]

However, such a set of logical predicates is often not sufficient to describe precisely a set of concrete memory states. Indeed, when the program being analyzed may produce states with too different structures, it becomes impossible to abstract them all into a single separating conjunction of base predicates. For instance, Fig. 1(a) shows a pair of concrete memory states that the above predicates cannot describe precisely, yet that may arise in real programs. Both memory states contain a list pointed to by variable \(a\), and two “cursors” \(x, y\) pointing somewhere in that list. The only difference between these two memories is the order of the cursors \(x, y\). A code fragment localizing independently two elements in the list will produce such states. Analyses like [9, 17] instantiate a generic list segment predicate \(Is(a, b)\) to abstract a segment of list starting at \(a\) and finishing with a pointer to \(b\). Using this abstraction, concrete memory state 1 can be abstracted by \(Is(1, x) \ast Is(x, y) \ast Is(y, \text{null})\), yet that predicate cannot abstract the second one in the same time as \(x, y\) appear in the reverse order. Instead, static analyses

* The research leading to these results has received funding from the European Research Council under the European Union’s seventh framework programme (FP7/2007-2013), grant agreement 278673, Project MemCAD, from the ARTEMIS Joint Undertaking no 269335 (see Article II.9 of the JU Grant Agreement), and the U.S. National Science Foundation under grant number CCF-1055066.
Concrete memory state 1:

Concrete memory state 2:

(a) List segments and need for disjunctions

Concrete memory state 1:

Concrete memory state 2:

(b) Tree fragments and need for disjunctions

Figure 1. Structures and disjunctions.

need to handle disjunctions of abstract states:

\[
\text{disjunctive abstract states: } d(\in D) \defeq m \lor \ldots \lor m \\
\text{abstract states: } m(\in M) \defeq p \star \ldots \star p
\]

Then, \(ls(1, x) \star ls(x, y) \star ls(y, 0x0) \lor ls(1, y) \star ls(y, x) \star ls(x, 0x0)\) abstracts the memories of Fig. 1(a). Without a disjunction, the information about either of the cursors would have to be dropped.

The same situation arises for other kinds of structures. For instance, the states of Figure 1(b) can be discriminated by a similar technique using a form of lightweight disjunct clumping, whereas the analysis computations (including the abstract join for clumping) still all take place in the initial, infinite lattice. Similarly, the states of Figure 1(b) can be discriminated by a similar abstraction of paths.

Therefore, we propose to let silhouette abstractions of the abstract states guide the algorithms for clumping and weakening disjuncts. We make the following contributions:

- we set up a path-based silhouette abstraction of abstract states to capture shape similarities and guide weakening in Sect. 3;
- we design algorithms for the clumping (Sect. 4) and the widening (Sect. 5) of disjunctive abstract states; and
- we implement the silhouette-guided algorithms in the MemCAD analyzer [36] and assess their efficiency with the verification of several real-world C libraries manipulating structures such as doubly-linked lists, red-black trees, AVL trees, and splay trees.

2. Overview

In this section, we formalize the disjunct clumping problem and describe the core contribution of the paper at a high level.

Clumping predicates based on their abstraction. As in the introduction, we assume that an abstract state \(m \in M\) is a separating conjunction \(p_0 \star \ldots \star p_n\) of basic region predicates, and we consider an analysis that computes an over-approximation for reachable concrete memories represented as disjunctive abstract states of the form \(d = (m_0 \lor \ldots \lor m_k)\). We let \(\text{join}_d\) denote a computable join operation over the set of abstract states that over-approximates unions of sets of memory states. Thus, it is always sound for the analysis to replace a disjunctive abstract state \((m_0 \lor m_1)\) with a non-disjunctive abstract state \(\text{join}_d(m_0, m_1)\), yet this operation may in general lose some information. Disjunct clumping aims at identifying a partition of a finite set of disjuncts so that each component of the partition can be joined using \(\text{join}_d\) without a significant loss of precision. For instance, as observed in the introduction, joining into a single abstract state abstraction the concrete states of Fig. 1(a) would not allow to represent all segments separately, therefore such disjuncts should not be clumped together.
In this paper, we propose an approach to disjunct clumping that is based on the use of an abstraction of abstract states:

- we shall design a set of logical predicates $\mathcal{S}$ that describe the silhouettes of abstract states in a compact and simple representation, and a computable silhouette equivalence relation $\sim$ that defines which silhouettes are similar;  
- we shall define a computable abstraction function $\theta : \mathcal{M} \rightarrow \mathcal{S}$ that maps an abstract state into its silhouette.

As an example, we defined $\sim$ in the introduction, as a compact abstraction of pointer chains, and used it to build silhouettes.

Using this notion of silhouette, we define the clumping algorithm $\text{clump}$ that inputs a disjunctive abstract state $d$ and returns another disjunctive abstract state $\text{clump}(d)$ that over-approximates $d$ with at most as many (though usually fewer) disjuncts:

1. it inputs a disjunctive abstract state $m_0 \lor \ldots \lor m_n$;
2. it computes $\theta(m_0), \ldots, \theta(m_n)$, and packs together abstract states with the same image by $\theta$ up to equivalence relation $\sim$;
3. it reduces each group into a single abstract state, by repeatedly applying $\text{join}$, and returns a more compact disjunction.

In the rest of this section, we define an instance of clumping of disjunctive predicates, by defining $\mathcal{S}, \theta$ and $\sim$ and we demonstrate how this approach enables the analysis of a complex code fragment, while limiting the size of disjunctions.

### Example: insertion into an AVL tree.

In the following, we study the verification of a function that inserts an element into an AVL tree. AVL trees achieve balancing by enforcing that the heights of two subtrees of a single node differ by at most one, and by storing that difference on each node, so that insertion and removal algorithms can rebalance subtrees incrementally, using “rotation” operations. This property makes insertion and removal involved, thus the preservation of structural invariants and absence of memory errors (such as illegal pointer operations or leakage of subtrees) is difficult to verify. In particular, the insertion and removal functions need to distinguish many cases, which their verification should also discriminate. Thus, verification algorithms will produce large numbers of disjuncts, and could greatly benefit from clumping. Fig. 2 shows an excerpt of an AVL insertion function. While the verification of the full code (from [37]) is presented in Sect. 6, we study here a simplified version for the sake of clarity. We only consider the handling of disjunctions of memory shapes; in particular, we ignore the AVL trees numeric balancing constraints (for our purpose, this restriction has no impact on the analysis).

The fragment in Fig. 2 handles the insertion into a non-empty tree, and carries it out in three phases. First, the node $p$ at which the new element should be inserted is localized, as well as the deepest edge (with source $x$ and target $y$) in the tree where the balancing property is locally broken by this insertion (indeed, this is the only point where a rebalancing will actually be required, since a rotation at this point will prevent any other balancing constraint from being broken). Second, a new node is allocated and inserted at position $p$. Finally, the rebalancing itself is performed. For concision, we consider in detail the first phase only. This phase should compute pointers $x, y, z$. To verify this function, the analysis should compute invariants that characterize precisely the shape of these trees, subtrees and the relative positions of pointers $p, x, y, z$ in all cases, including the corner case where tree $\tau$ consists of a single node, and the insertion is actually done above it.

### Abstract states and analysis.

To express invariants over the code of Fig. 2, the analysis needs summary predicates describing trees of unbounded size as well as cursors inside trees. Thus, we assume a pair of summary predicates $\text{tree}$ and $\text{treeseg}$. More precisely, if $\alpha$ denotes a symbolic address, we let $\text{tree}(\alpha)$ describe a well-formed complete tree stored at address $\alpha$. Similarly, we let $\text{treeseg}(\alpha, \beta)$ describe a tree stored at address $\alpha$ missing a subtree at address $\beta$. Therefore abstract values stand for separating conjunctions of points-to predicates and of summary predicates. We use shape graphs, as shown in Fig. 3 as a more intuitive representation, where nodes stand for symbolic addresses (and the values of variables), and bold edges stand for full summary predicates (edges with no destination node) and for segment predicates (edges with a destination node). The abstract state on the right side of the figure describes states where $\tau$ points to a tree, and $x$ points to a (possibly non-strict) subtree of the tree pointed to by $\tau$. It is made of the separating conjunction of a $\text{treeseg}$ segment summary and a $\text{tree}$ summary. An example concrete memory is shown in the left side of the figure.

Both summary predicates have a natural inductive structure. A region abstracted by $\text{tree}$ is either an empty tree or a tree with a node and two disjoint subtrees. This remark boils down to a pair of fold / unfold rules, that fully characterize $\text{tree}$ (likewise $\text{treeseg}$ is defined by such a pair of rules):

![Figure 3. Abstract states.](image-url)
and at line 14, the reading of field \( d \) of the node pointed to by \( p \) requires unfolding the \( \text{tree} \) summary predicate attached to \( t \). Such unfoldings generate disjunctions of case splits. Conversely, join and widening applied at loop head should fold back case splits, so as to compute a loop invariant.

**Silhouette abstraction.** Fig. 4 describes a few abstract states that are observed at line 25, at the exit of the first loop, and that illustrate the challenges of clumping disjuncts. The figure shows abstract states, the generic form of concrete memories they represent and their silhouettes to be explained below. For concision, and as only the relative positions of cursors matter, we focus on \( t, x, y \) and omit fields \( d, h, a, l, b, p, q \). Labels \( t, x, y \) decorate the nodes that represent their value. Moreover, we show only a sample of 4 among 32 disjuncts. First, abstract state \( m_0 \) abstracts memories where \( x, y \) were not advanced. Second, abstract state \( m_1 \) abstracts memories where \( x \) was advanced to the root of the tree, and the search continued in the left subtree. Third, abstract state \( m_2 \) abstracts memories where the search visits the left subtree of \( t, x \) is advanced into that subtree, and \( y \) is the left child of \( x \). Last, abstract state \( m_3 \) describes a similar condition as \( m_2 \) but when the search visits the right subtree of \( t \) and \( x \) is a left child. Some of these abstract states are very similar to each other and can be joined to reduce the cost without significantly affecting the precision of the analysis. Indeed, \( x \) and \( y \) occupy the same relative positions in \( m_2 \) and \( m_3 \); moreover, in both cases the two cursors point to subtrees of \( t \). Thus, both \( m_2 \) and \( m_3 \) can be approximated by an abstract state with a segment from \( t \) to \( x \) and where \( y \) is the left child of \( x \). Furthermore, \( m_1 \) can also be weakened similarly: the relative positions of \( x, y \) are the same, and the only minor difference is that \( x \) is not a strict subtree of \( t \) since \( t = x \), but this equality can also be described by an (empty) segment. On the other hand, \( m_0 \) abstracts very different memories, where \( x \) is not a subtree of \( t \), so it cannot be described with a segment from \( t \) to \( x \). Any abstract state that over-approximates both \( m_0 \) and \( m_1 \) would discard all information about either \( t \) or \( x \), which would make the proof of structural preservation impossible. Therefore, an ideal clumping would join \( m_1, m_2, m_3 \) together but keep \( m_0 \) separate.

Intuitively, abstract states where the relative positions of cursors \( t, x, y \) into the tree are similar can be clumped together with no severe precision loss. To capture this intuition, the notion of silhouette shown in the last column of Fig. 4 retains only the relative positions of \( t, x, y \) and the access paths between them. As access paths may be of unbounded length, we abstract them with regular expressions describing sequences of fields dereferenced between nodes. For instance, the silhouette \( s_0 \) of \( m_0 \) boils down to a single edge, with a single path labeled by \( 1 \). Even the more complex \( m_3 \) is characterized by only two edges respectively labeled by \( r \cdot (1+r)^* \cdot 1 \) (between \( t \) and \( x \)) and by \( 1 \) (between \( x \) and \( y \)).

**Clumping disjuncts.** Silhouettes make the dissimilarity of \( m_0 \) with the other states in Fig. 4 obvious, due to the incompatible order of cursors. However, we can also see that the silhouettes of \( m_2 \) and \( m_3 \), while not syntactically identical are actually equal up-to a generalization of the path regular expression for the left edge into \( (1+r)^* \). This generalization matches the fact that any structure segment can be weakened into a \( \text{treeseg} \) predicate, whatever the sequence of “left-right” branches it encompasses:

\[
s_{m} = \text{treeseg} (1+r)^* x_1 \cdots y_1
\]

The silhouette of \( m_1 \) also corresponds to a special case of \( s_{m} \). Therefore, we let \( \sim \) denote the similarity of silhouettes up-to generalization of access paths. With this notation, we have:

\[
s_0 \not\sim s_1 \quad s_0 \not\sim s_2 \quad s_0 \not\sim s_3 \quad s_1 \sim s_2 \sim s_3
\]

Therefore, the groups computed here are \( \{m_0\}, \{m_1, m_2, m_3\} \), and the four disjuncts of Fig. 4 are clumped into a disjunctive abstract state composed of only two disjuncts \( m_0 \) and \( m_{1,2,3} \). The
computation of the $m_{1,2,3}$ (that we discuss in the next paragraph) may use any weakening algorithm for abstract states, such as canonical abstraction [35], canonicalization of symbolic heaps [17], or shape graph join [9].

In essence, clumping relies on a weak canonicalization [35] that returns silhouettes, and then selects groups based on an equivalence relation over the set of silhouettes. To ensure the termination of abstract iterates over loops, the finiteness of the clumping relation $\sim$ is required (i.e., it should have a finite set of equivalence classes), though the set of silhouettes may still be infinite. However, clumping may be performed at other points than loop heads (in order to shrink abstract states, without discarding any information), and then it does not require a finite $\sim$. In this point of view, clumping provides a more flexible approach to the handling of disjunctions than canonicalization, since it does not need to project abstract states into a finite set of predicates, and since it may still take advantage of precise binary operators for weakening.

**Clumping region predicates.** The information computed in silhouettes also provides a guideline for weakening abstract states. Indeed, let us consider silhouette $s_i\ldots$, that is weaker than the silhouettes of $m_1, m_2, m_3$. As it contains an edge labeled by $(1 + y)^t$ between $t$ and $x$, it suggests weakening the fragment of these abstract states that are between $t$ and $x$ into a treeseg segment predicate. The soundness of such a weakening can be verified by checking entailment of a fragment of $m_1$ (resp., $m_2, m_3$) and a treeseg predicate. The resulting weaker abstract state $m_{1,2,3}$ is shown in Fig. 5. Here, the role of the silhouette is to provide guidance on how abstract states may be weakened. The advantage of this view to the computation of weakening is to benefit from global semantic information about the structure of abstract states that would be neglected by weakening operators based on syntactic rules only (like the canonical heap abstraction of [17] and the join of [9]).

**3. Silhouette Abstraction and Guided Weakening**

In this section, we formalize silhouettes, and show that they provide a useful abstraction to reason about the weakening of abstract states.

**3.1 Preliminaries: Memories and Abstract States**

Before we formalize silhouettes, we define memories and abstract states. We let $X$ denote a set of pointer variables and $\nu$ denote a set of values (that includes numeric addresses). We let fields (noted as $1, r, \ldots$) denote both field names and offsets. A memory state $\sigma \in \Sigma$ is a partial function from addresses and variables to values. We let $\sigma|_X$ (resp., $\sigma|_{\nu}$) denote the restriction of $\sigma$ to variables (resp., heap addresses). If $x$ (resp., $a$) is a variable (resp., an address), we write $\sigma(x)$ (resp., $\sigma(a)$) for the content of variable $x$ (resp., the cell of address $a$). This basic memory model does not allow computing the address of a variable, but a more general definition based on a classical heap and environment pair could be used without changing the principle of our approach.

An abstract state describes a set of concrete memory states. It is either $\bot$ (denoting $\emptyset$) or a separating conjunction of region predicates that abstract separate memory regions [32] in conjunction with numerical constraints such as equalities and disequalities. The syntax of abstract states is given in Fig. 6. Abstract states utilize a set $\Lambda = \{\alpha, \beta, \ldots\}$ of symbolic addresses to abstract values and heap addresses. A node $n$ is either a variable $x$ or a symbolic address $\alpha$. A region predicate is either $\text{emp}$ describing an empty region, a points-to predicate $n: f \mapsto n'$ (that describes a heap memory cell at base address $n$ with offset $f$ and with content $n'$), or a summary predicate $\text{ind}(n)$ describing a full inductive structure or $\text{indseg}(n, n')$ for a (possibly empty) inductive segment from address $n$ to $n'$. This set of abstract states is parameterized by a user-supplied inductive predicate $\text{ind}$, that characterizes a family of data structures of interest and serves as a template that the analysis will instantiate. It is specified by a disjunction of cases, which boil down to separating conjunctions of points-to and inductive summary predicates (e.g., lists, doubly-linked lists, trees, and trees with parent pointers can be expressed this way). Segment predicate $\text{indseg}$ stands for the segment version of $\text{ind}$ and describes $\text{ind}$-structures with one "hole" (i.e., one missing substructure), and can also be defined by induction. The $\text{ls}$ list segment predicate of [17] is an inductive predicate. The $\text{tree}$ predicate used in Sect. 2 can be defined by induction as well:

$$
\begin{align*}
\text{tree}(n) := & \ \text{emp} \land n = \text{0x0} \\
& \lor n \cdot 1 \mapsto \alpha_1 \land n \cdot r \mapsto \alpha_r \land n \cdot \text{bal} \mapsto \alpha_b \\
& \land n \cdot d \mapsto \alpha_d \land \text{tree}(\alpha_1) \land \text{tree}(\alpha_r) \land n \neq \text{0x0}
\end{align*}
$$

The inductive definition of $\text{treeseg}(n, n')$ has three cases: either the segment is empty and $n = n'$, or it is non-empty and $n'$ in the left subtree of $n$ or it is non-empty and $n'$ in the right subtree of $n$.

Abstract states can also be viewed as shape graphs. For instance, Fig. 3 depicts as a shape graph the abstract state $\text{treeseg}(t, x) \land \text{tree}(x)$. The logical meaning of an abstract state $m$ is defined by its concretization $\gamma_m(\Sigma) \subseteq \Sigma \times (\Lambda \rightarrow \nu)$, as a set of pairs made of a memory state $\sigma$ and a function $\nu$ that maps each symbolic address or variable into its concrete counterpart in $\sigma$. The bottom of Fig. 6 gives the definition of $\gamma_m(n)$ in the main cases. Symbolic addresses are existentially quantified, and thus the concretization of abstract states is unchanged by renaming of symbolic addresses. Similarly, swapping or merging equal nodes preserves the meaning of abstract states.

Abstract transfer functions compute post-conditions for basic statements like assignments, allocation and deallocation operations in a local manner [17]. As an example, we consider abstract state $m = \text{tree}(t) \land t = x \land t \neq \text{0x0}$ and assignment $x := x \cdot \rightarrow 1$. As summary $\text{tree}(t)$ is read, it needs to be unfolded. Due to the side constraint $t \neq \text{0x0}$, the only disjunct created is
\[
t \cdot 1 \mapsto a * t * x \mapsto \beta * \text{tree}(a) * \text{tree}(\beta) \wedge t = x.
\]
Then, the analysis boils down to a local update, which produces the abstract post-condition \( t \cdot 1 \mapsto x * t * x \mapsto \beta * \text{tree}(x) * \text{tree}(\beta) \).

### 3.2 Silhouettes

Abstract states provide a precise description of sets of concrete states. Even summarized regions are characterized by inductive predicates that convey very detailed information about their structure. The purpose of silhouettes as presented in Sect. 2 is to identify similarities among abstract states without looking into the details of the shapes of the structures. Thus, we define silhouettes as graphs, yet with edges that retain less information than region predicates.

As remarked in Sect. 2, information about reachability on pointer paths is relevant to the characterization of groups of abstract states that could be clumped together. Paths can be described using basic regular expressions over fields. We let \(\mathcal{E}\) denote the set of the regular expressions that satisfy the grammar \(e ::= e \mid f \mid (f_0 + \ldots + f_n)^*\) or \(e \cdot e\) which are respectively adequate to describe equalities, points-to, or regions, segment edges, and sequences of edges. This leads us to the following definition:

**Definition 1 (Silhouette).** A silhouette is a graph defined by a set of nodes \(N \subseteq X \cup \emptyset\), and a set of edges \(E\) that are labeled by regular expressions in \(\mathcal{E}\) (we write \((n, e, n')\) for such an edge).

A silhouette collects a conjunction of reachability constraints over paths. Thus, the concretization \(\sigma(e)(s)\) of a silhouette is defined as the set of all pairs \((\sigma, \nu)\) that satisfy all the constraints defined by the edges of \(s\). Deciding whether \((\sigma, \nu)\) satisfies the constraint defined by an edge \((n, e, n')\) boils down to evaluating the values described by \(n, n'\) into values \(e, a'\) and checking whether \(e\) referencing a sequence of fields described by \(n\) starting from \(a\) allows to reach \(a'\). Using the semantics of nodes \([n]I(\).

The following set of rules formalizes this:

\[
\begin{align*}
    (\sigma, \nu) &\models [n][\sigma, \nu], e, [n'](\sigma, \nu) \\
    (\sigma, \nu) &\models (n, e, n') \quad (\sigma, \nu) \models a, (f_0 + \ldots + f_n)^*, a' \\
    (\sigma, \nu) &\models a + f = a' \\
    (\sigma, \nu) &\models a, f, a' \\
    (\sigma, \nu) &\models a \cdot e, a' \\
    (\sigma, \nu) &\models a, e, a' \\
\end{align*}
\]

For example, the concrete memory shown in Fig. 3 can be described by the silhouette shown below with its graphical representation:

\[
s = \{(t, (1 + x)^*, x)\}
\]

### 3.3 Computation of Silhouettes

To allow silhouettes to assist in clumping abstract states, it is crucial to have an efficient way to compute them. First, empty regions, equalities / disequalities to \(0\times0\), and full inductive predicates (of the form \(\text{ind}(\alpha))\) do not contribute to the silhouette. Second, a points-to edge \(n \cdot f \mapsto n'\) simply contributes an edge \((n, f, n')\). Last, a segment predicate \(\text{indseg}(n, n')\) contributes an edge \((n, E(\text{ind}), n')\), where \(E(\text{ind})\) denotes the set of paths that can be induced by a silhouette.

\[
\begin{align*}
    \Pi(n \circ 0x0) &= \emptyset \\
    \Pi(\text{emp}) &= \emptyset \\
    \Pi(n \cdot f \mapsto n') &= \{(n, e, n')\} \\
    \Pi(\text{ind}(n)) &= \Pi(\text{indseg}(n, n')) = \{(n, E(\text{ind}), n')\} \\
    \Pi(p_0 * \ldots * p_k \wedge \ldots \wedge c_i) &= \\
    &\Pi(p_0) \cup \ldots \cup \Pi(p_k) \cup \Pi(c_0) \cup \ldots \cup \Pi(c_n)
\end{align*}
\]

This translation function is sound as it returns a silhouette that describes more memories than the abstract state it is applied to:

**Theorem 1 (Soundness).** The silhouette translation function \(\Pi\) is sound: for all memory state \(m, \gamma_0(m) \subseteq \gamma_0(\Pi(m))\).

The proof proceeds by induction over abstract states. Silhouettes describe conjunctions of constraints, thus can be weakened into coarser approximations of sets of memory states, either by dropping or by weakening some constraints. Given a silhouette \(s = (N, E)\) and a set of nodes \(N' \subseteq N\), we let the restriction of \(s\) to \(N'\) be the silhouette \(s|_{N'}\) defined by the set of nodes \(N'\) and the edges obtained as acyclic concatenations of edges of \(s\) forming paths from \(N'\) to \(N'\). This weakening effectively allows us to ignore some nodes. For instance, if \(X \subseteq X\) is a set of variables that play a special role, then \(\Pi(m)|_{X}\) is a silhouette of \(m\) that will only retain information about the variables in \(X\). We note this silhouette \(\Pi(m, X)\).

**Example 1 (Silhouette computation).** We consider the abstract state \(m_0\) of Fig. 3. Then, \(\Pi(m_0)\) is the silhouette shown at the end of Sect. 3.2. Moreover, \(\Pi(m, (t)) = \emptyset\).

### 3.4 Silhouette-based Weak Entailment Check

The core purpose of the silhouette abstraction is to determine when abstract states are similar enough to be joined together without a significant loss in precision. Since join computes over-approximation of several abstract states, we first study the relationship between entailment check of abstract states and the silhouettes.

**Entailment check for abstract states.** Entailment check of abstract states is usually based on sets of proof rules that establish inclusion locally and express separation, reflexivity, and unfolding of inductive predicates. Rules below describe such a typical system (for clarity, numerical constraints are elided):

\[
\begin{align*}
    m &\models e & m &\models e \\
    m &\models \text{ind}(n) & m &\models \text{indseg}(n, n') & m &\models m' \text{ unfold} \rightarrow m \\
    \text{indseg}(n, n') &\equiv m \models \text{indseg}(n, n') & m_0 &\models m_0' & m_1 &\models m_1' \\
    \text{indseg}(n, n') &\equiv m_0 \models m_0' & m_0 &\models m_1 & m_1 &\models m_1' \\
\end{align*}
\]

An entailment checking algorithm not \(\text{leq}\) implements a proof search based on these rules. It is in general sound and incomplete: when \(\text{leq}_{\gamma_{\sigma}}(m_0, m_1)\) returns \text{true}, then \(m_0 \subseteq m_1\), thus \(\gamma_{\sigma}(m_0) \subseteq \gamma_{\sigma}(m_1)\), yet the reverse implication does not hold in general. Indeed, complete proof search algorithms with backtracking are expensive. Moreover, completeness can in general not be ensured in the presence of more complex sets of numerical constraints [9] than in the restricted set of abstract states used here.

**Silhouette entailment check.** As an alternative to the abstract state entailment check, we propose to first use a weaker and cheaper entailment check on silhouettes, which is based on a classical inclusion of constraints. We let \(\mathcal{L}(e)\) denote the language of \(e\).

**Definition 2 (Silhouette entailment check).** Let \(s_0, s_1 \in \mathcal{S}\), we let \(\text{leq}_{\gamma_{\sigma}}(s_0, s_1)\) return \text{true} if and only if for all edge \((n, e, n')\) of \(s_1\) there exists a (possibly empty) sequence of edges \((n_0, e_1, n_1), \ldots, (n_{k-1}, e_k, n_k)\) in \(s_0\) such that \(n_0 = n, n_k = n'\) and \(\mathcal{L}(e_1 \cdot \ldots \cdot e_k) \subseteq \mathcal{L}(e)\) (or, if the sequence is empty, \(e \in \mathcal{L}(e)\)).

**Theorem 2 (Soundness).** The entailment check \(\text{leq}_{\gamma_{\sigma}}\) is sound: given \(s_0, s_1 \in \mathcal{S}\), if \(\text{leq}_{\gamma_{\sigma}}(s_0, s_1) = \text{true}\), then \(\gamma_{\sigma}(s_0) \subseteq \gamma_{\sigma}(s_1)\).

The most important result on silhouette entailment check is that it is weaker than the rules for abstract states entailment:
Theorem 3 (Weak entailment). Let $m_0, m_1 \in \mathcal{M}$. Then:

$$m_0 \models m_1 \implies \text{leq}_k(\Pi(m_0), \Pi(m_1)) = \text{true}.$$ 

The proof of Th. 2 follows from the structure of silhouette entailment check whereas Th. 3 can be proved by induction on the derivations of $\text{leq}_k$.

From this theorem follows the core principle of our approach: while the most direct way to check if $\gamma_S(m_0) \subseteq \gamma_S(m_1)$ consists in applying the (fairly expensive) $\text{leq}_k(m_0, m_1)$ proof search algorithm, an alternate approach consists in computing $\text{leq}_k(\Pi(m_0), \Pi(m_1))$ first and computing $\text{leq}_k$ only when $\text{leq}_k$ returns true. Indeed, if $\text{leq}_k$ returns false, then $\text{leq}_k$ will also definitely return false, though at a much higher computational cost.

Example 2 (Entailment). We consider the example given in Sect. 2, and the states and silhouettes defined in Fig. 4. Then, $\text{leq}_k(s_0, s_1) = \text{false}$, and indeed $m_0 \not\subseteq m_1$ does not hold. Moreover, $\text{leq}_k(s_1, s_2..) = \text{true}$ and $m_0 \not\subseteq m_{1.2.3}$ holds. On the other hand, among the states in Fig. 4, there is no case where $\text{leq}_k$ returns false and $\not\subseteq$ does not hold.

## 4. Clumping Disjunctions of Abstract States

Silhouettes offer a weak characterization for feasible weakenings. We now put this characterization to work in order to determine when and how disjunctive abstract states may be clumped, without causing a severe loss in precision.

### 4.1 Existing Abstract States Join Operators and Challenges

To clump a disjunctive abstract state, the analysis should identify disjuncts that can be weakened into the same, precise result. This corresponds to a join operator over finite sets of abstract states. Thus, we first study how a join [9, 38] can be computed for a pair of abstract states $m, m'$. First, in order to take advantage of separation [32] and local reasoning, states $m, m'$ should be split into separating conjunctions of regions $m = (m_0 * \ldots * m_k)$ and $m' = (m_0' * \ldots * m_{k'})$ such that $m_i$ and $m_i'$ describe “similar” shapes. Second, for each pair $m_i, m_i'$ a local structural rule should produce a weaker $m_i''$, such that inclusions $m_i \subseteq m_i''$ and $m_i' \subseteq m_i''$ hold. Then, a valid over approximation of both $m$ and $m'$ is defined by $m'' = m_0'' * \ldots * m_{k''}''$. During this second phase, two main kinds of rules are used:

- **Weakening guided by existing region predicates.** The first set of rules searches for cases where $m_i \subseteq m_i', m_i' \subseteq m_i$ holds so that a valid choice for $m_i''$ is $m_i''$ (resp., $m_i$). The most common case is when $m_i = m_i'$. Another important case is when $m_i'' = \text{indseg}(\alpha)$ and $m_i' \subseteq \text{indseg}(\alpha)$; then $m_i$ gets weakened into a summary predicate, that was already present in $m_i'$.

- **Synthesis of summary predicates.** The second set of rules synthesizes new summary predicates, that weaken specific patterns. For instance, the introduction of a segment predicate is a particular case of summary predicate synthesis, that weakens an empty region into a segment:

$$m_i = (\text{emp} \land \alpha = \beta) \implies m_i'' = \text{indseg}(\alpha, \beta)$$

**Example 3 (Abstract states join).** As an example, we let $m = (\text{tree}(x) \land t = x)$ and $m' = (t \cdot 1 \Rightarrow x \land t \cdot r \Rightarrow \alpha \land \text{tree}(x) \land \text{tree}(a))$. Then, $m'$ can be split into two regions: first, $\text{tree}(x)$ is also present into $m$; second, the remainder is included into $\text{treeseg}(t, x)$. Therefore, these two states can be joined into $\text{treeseg}(t, x) \land \text{tree}(x)$.

In practice, the splitting phase and the application of weakening rules are performed concurrently, as the search for a good splitting requires insight about what rules may apply.

### 4.2 Clumping Silhouettes

While the silhouettes of abstract states defined in Sect. 3 provide a weak entailment check, they still adhere too closely to the structure of abstract states to highlight all possibilities for clumping. For instance, Fig. 8 shows excerpts $m_2'$, $m_3'$ of the abstract states $m_2$, $m_3$ of Fig. 4, and the silhouettes of $m_2'$, $m_3'$ are not equal, even though these abstract states could be clumped, as discussed in Sect. 2:

$$\Pi(m_2') = \{(t, \cdot 1, \alpha_0), (\alpha_0, (1 + r)^*, x), (t, r, \alpha_1)\}$$

$$\Pi(m_3') = \{(t, r, \beta_1), (\beta_1, (1 + r)^*, \beta_2), (\beta_2, l, \alpha)\}$$

**Taking advantage of the analysis goal.** As remarked in Sect. 3.3, silhouettes can be made more concise and relevant by collecting only nodes that play an important role in the rest of the analysis. In
As observed in Sect. 4.1, a large part of the join rules, implicitly after generalization, we obtain:

Example 6. Let $\Phi$ be a regular expression (Silhouette generalization) makes such patterns appear more prominently:

\[
\Phi(\Pi(m_1)) \supseteq \Phi(\Pi(m_2))
\]

This result is a direct consequence of Theorem 3. It implies that, whenever a join can be computed by weakening part of $m_1$ into predicates present in $m_2$ and part of $m_2$ into predicates present in $m_1$, then relation $\supseteq$ holds. The contrapositive entails that, when $\supseteq$ does not hold, no precise join can be found using only rules that weaken abstract states based on existing predicates. Therefore, Th. 4 will prevent clumping from attempting to compute some joins that will fail, but not all. Sect. 6 provides evidence that the characterization it provides is actually tight. Moreover, this result does not immediately apply to the cases where join requires the synthesis of summary predicates; Sect. 4.3 discusses such cases.

Example 8 (Clumping relation). We consider the states of Ex. 4. If $X = \{t, u, v\}$ (we observed that joining them would incur a precision loss), we obtain $\Pi(m, X) = \{(t, u, v)\}$ and $\Pi(m', X) = \{(t, v, u)\}$, and $\Phi(\Pi(m, X)) \supseteq \Phi(\Pi(m', X))$ does not hold.

4.3 Silhouette Guided Join of Abstract States

We now show that silhouettes may help synthesizing new summary predicates. For instance, when a segment predicate can be synthesized, it is always done after checking with relation $\sqsubseteq$ that the new predicate actually defines a weakening of the arguments. Since Th. 3 offers a weak entailment check, the silhouette can be used in order to guide introduction of segment predicates.

Definition 5 (Silhouette join). We define the join of silhouette edges $(s_0, e, t_0)$ and $(s_1, e', t_1)$ as the edge $(s_0, e', t_1)$ where $e'$ keeps the same left side as $e$ and $e'$ joins the other part into the smallest and more general regular expression of the form $(e_0 + ... + e_k)^*$. Given two silhouettes $s = (N, E, s') = (N', E')$, we define their silhouette join noted $\text{join}(s, s')$ as the silhouette $s'' = (N'', E'')$ such that $N'' = N \cap N'$ and $E''$ collect the pairwise join of the edges of $s_1|_{N''}$ and of $s_0|_{N''}$.

When two abstract states can be joined into a segment, their silhouette is a refinement of the silhouette of a segment (by Th. 3):

\[
\gamma(s)(\text{join}((\Pi(m, \{n_0, n_1\}), \Pi(m', \{n_0, n_1\}))) \subseteq \gamma(s)((\Pi(m_0, n_1)), \Pi(m_0', n_1)))
\]
This result is beyond the reach of algorithms based on local syntactic
Thus, the above rule applies, and synthesizes a segment predicate:

\[ S_{\text{indseg}} = \text{treeseg}(t, x) \]

This result is beyond the reach of algorithms based on local syntactic
rules such as those of [9, 38].

4.4 Clumping Abstract States

Fig. 9 summarizes the algorithm clump for clumping abstract states. It
takes as input a disjunctive abstract state and the set of live
variables at the current location in the program being analyzed. It
first computes the silhouette of all the abstract states and the
generalized form of these silhouettes. The generalized forms are
used in order to compute \( \triangleright \). The groups of abstract states to be
clumped together are defined by the connected components of this
relation and is denoted \( \sim \) in Sect. 2. Also the silhouette abstraction
denoted \( \theta(m) \) in Sect. 2 is computed as \( \Phi(\Pi(m, X)) \). For each
group, relation \( \triangleright \) suggests a sequence of abstract state joins, that are
expected to preserve silhouettes. These abstract state joins utilize the
silhouettes \( S_0, \ldots, S_n \) in order to enable semantic-guided synthesis
of summary predicates as shown in Sect. 4.3.

This clumping algorithm is sound: it returns a disjunctive abstract
state that over-approximates \( m_0, \ldots, m_n \).

We now propose to discuss, in a general setting, the way the
silhouette abstraction works. We started with an abstraction of heaps
that is used in shape analysis, and identified a cause of precision loss
when computing joins of abstract states. In general, when abstract
states that are too dissimilar and do not join well, the analysis
will lose critical properties about shapes, that cannot be recovered:
indeed, when paths of two abstract states do not match, they cannot
be joined into an abstract state that preserves paths, thus join will
discard important pointer properties. The purpose of the silhouette is
to offer a cheaper characterization of abstract states that the analysis
should not attempt to join by the means of another abstraction,
which focuses on paths (Definition 1). Theorem 3 illustrates this
property: when paths do not match, silhouette inclusion does not
hold, which entails that the inclusion of abstract states cannot hold
either. Theorem 4 extends this property to union.

Remark 1 (Silhouettes in static analysis: a numerical example). We
can extend this approach to other situations. We consider in this
paragaph programs that manipulate integer variables, and perform
basic arithmetic operations. A common goal for static analyses
consists of verifying the absence of division by zero errors. Let us
assume such an analysis relies on interval constraints over program
variables: a division by \( x \) can be proved safe if the analysis computes
for \( x \) either an interval of strictly negative numbers or an interval of
strictly positive numbers. We note that joining together an interval of
strictly negative numbers and an interval of strictly positive numbers
will return an interval that contains zero. Thus, after such a join, a
crucial information will be lost to verify that a division by \( x \) is safe.

The sign abstraction provides an obvious characterization what
interval unions will produce such an imprecise result: the sign
abstract domain is made of four abstract values \( \perp, \ominus, \oplus, \top \) that
denote the empty set, any set of strictly negative integers, any set of
strictly positive integers, and any set of integers. For
instance, if two intervals have sign \( \ominus \), they can be joined without
losing information with respect to zero, whereas the join of an
interval with sign \( \perp \) with an interval with sign \( \ominus \) will cause a
precision loss.

Therefore, a principle similar to silhouettes allows us to enhance
the analysis precision, with respect to the goal of proving variables
are not equal to zero:

- disjunctive states composed of finite disjunctions of intervals
can keep more precise information about sets of states where a
variable \( x \) may be positive or negative;
- the “silhouette” of intervals defined by their sign abstraction
characterizes when joining two intervals will cause a precision
loss, with respect to the property of interest.

5. Widening Disjunctive Abstract States

To infer loop invariants from an abstract pre-condition, the
analysis needs to use a widening operator [10] widen\_D over dis-
junctive abstract states. This operator should over-approximate
concrete union, and ensure that any sequence \((d_{n+1})_n\) of the form
d_{n+1} = widen_d(d_n, d'_n) terminates. In this section, we assume
that widen_d is a widening over abstract states, using similar algo-
rithms as join_p. Such an operator can often be based on join_p [9].

Widening based on silhouette-guided pairing. Intuitively, a
widening of \(d\) with \(d'\) should widen disjuncts of \(d\) with disjuncts of
\(d'\), using some sort of pairing to select which pairs of disjuncts are
to be widened together. Since silhouettes aim at capturing abstract
states that can be joined precisely, we build our widening around a
pairing function that we shall build based on the silhouette.

Definition 6 (Pairing function). Given \(d = m_0 \lor \ldots \lor m_n\) and
d' = \(m_0' \lor \ldots \lor m_n'\), a pairing of \(d, d'\) is a function \(\pi_{d,d'}\) from
\((0, \ldots , n)\) into the powerset of \((0, \ldots , n)\) such that \(i \neq j\) implies
\(\pi(i) \cap \pi(j) = \emptyset\).

If a pairing family \(\pi_{d,d'}\) is defined for all \(d, d' \in \mathbb{D}\), we can
define an operator over disjunctive abstract states that lets the pairing
function define which disjuncts of the right argument are widened
with each disjunct of the left argument, and preserves the disjuncts
of the right argument that do not appear in the pairing.

Definition 7. Let \(d = m_0 \lor \ldots \lor m_n\) and \(d' = m_0' \lor \ldots \lor m'_n\).
We let \(\text{widen}_{\pi_d,d'}(d, d')\) be defined as the abstract state
\(m_0'' \lor \ldots \lor m_n''\), where:

\[
\begin{align*}
&\cdot m_i'' = \text{widen}_{\pi_d}(m_i, \text{join}_p(m_i')) \text{ if } i \leq n; \\
&\cdot n_{i,j}'' = m_j' \text{ otherwise. }
\end{align*}
\]

This operator always returns a sound over-approximation of
its arguments. Yet, depending on the pairing family, it may fail to
guarantee termination. In the following, we define a pairing family
that ensures termination so that as clumping, it should also

A good pairing should map abstract states that produce a precise
widening, in the same way as clumping, for join. It should also
drive the introduction of summary predicates. As observed in
Sect. 4, silhouettes hold information capable of guiding this process.
However, for termination, we need to bound silhouettes. Thus, we
define a pairing that is parameterized by an integer bound \(b\), and that
associates abstract states with silhouettes that are equivalent when
regular expressions are matched upon exceeding length \(b\) (in the
following, we let \(b = 1\)).

Definition 8 (Silhouette-based pairing). We let \(\mathcal{F}\) denote the set
of all field names, and we define the regular expression bounded
abstraction by \(\omega_k(e_1 \ldots e_k) = e_1 \ldots e_k \cdot \mathcal{F}\) if \(k > b\)
and \(\omega_k(e_1 \ldots e_k) = e_1 \ldots e_k \cdot \mathcal{F}^*\) otherwise. We let
function \(\Omega_b\) abstract a silhouette by replacing each enclosing \((n, e, n')\)
by \((n, \omega_b(e), n')\).

Last, if \(s, s'\) are silhouettes, then we write \(s \approx_b s'\) if and only if
\(\text{leq}_{\mathcal{F}}(\Omega_b(s'), \Omega_b(s)) = \text{true}\).

Theorem 7 (Widening). The operator \(\text{widen}_{\pi_d}\) defined using the
pairing of Def. 8 enforces termination of abstract iterates.

The proof relies on the finiteness of the image of \(\Omega_b\), which
tells us that, for any sequence \((d_n)_n\) of widened iterates (such that
d_{n+1} = widen_d(d_n, d'_n)), the disjuncts in \(d_n\) eventually stabilize
to a set corresponding to a fixed, finite set of silhouettes.

Example 11. Fig. 10 displays a few of the disjuncts that arise in
the second and third iteration over the first loop in the analysis
of the code presented in Fig. 2. Disjunct \(m_0\) occurs in the second
iteration whereas disjuncts \(m_1', m_2'\) arise in the third iteration.
For clarity, we only show the nodes on the path between \(t, x\) and
\(y\). In the table below, we show their silhouettes before and after

\[
\begin{align*}
\text{silhouette} & \quad \text{bounded abstraction} \\
m_0 & = \{(t, l, x, (x, l, y)) \} & \{\{(t, l, x, (x, l, y))\}, \{(t, l, x, (x, l, y))\} \}} \\
m_1 & = \{(t, l, x, (x, l, y)) \} & \{(t, l, x, (x, l, y))\} \\
m_2 & = \{(t, l, x, (x, l, y)) \} & \{(t, l, x, (x, l, y))\}
\end{align*}
\]

Then, \(\Pi(m_0) \approx_b \Pi(m_0)\) and \(\Pi(m_0) \approx_b \Pi(m_0)\) hold whereas
\(\Pi(m_0) \approx_b \Pi(m_0)\) does not. Thus, the pairing is defined by \(\pi(0) = \{0, 2\}\). The widening will thus preserve the information which says
whether \(y\) is the left or right child of \(x\).

Static analysis. We now summarize the whole analysis. As a
forward abstract interpretation [10], it starts with an abstract pre-
condition that over-approximates initial memory states with a single
disjunct (either the set of all memory states, with no structure allo-
cated for whole program analysis, or a pre-condition describing the
valid call states for the analysis of a library function) and computes
abstract post-conditions for each statement. Post-conditions of as-
sumption, test, allocation and deallocation statements are computed
locally on each disjunct in parallel, as discussed in Sect. 3.1. When
the analysis computes disjunctive abstract states \(d, d'\) before a con-
trol flow join, it applies clumping (Sect. 4) to \(d \lor d'\). For loops,
standard fixpoint approximation techniques [10] apply, and the
silhouette guided widen operator (Th. 7) enforces convergence of
disjunctive abstract states sequences. The analysis is sound and
returns abstract post-conditions that over-approximate final states.

6. Experimental Evaluation

Research hypotheses. In this section, we empirically evaluate
whether or not semantic-directed clumping is effective in improving
the static analysis of data structures from real-world C libraries. We
implemented clumping inside the MemCAD analyzer [36]. We seek
to provide evidence for or against the following hypotheses:

RH1 (Clumping is effective). Semantic-directed clumping with
guided join is necessary and effective for analyzing data structure
operations from existing, real-world libraries. The underlying
premise of this paper is that inferring disjunctive loop invariants is
necessary to effectively analyze real-world data structures. While
certain code may be reasonably adapted to avoid the need for dis-
junctive invariants, analyzing existing, real-world code typically
involves handling corner cases in separate disjuncts. We assess the
impact of silhouettes on analysis and compare with other techniques.

RH2 (Guided join is necessary). Guided join is necessary to
avoid unacceptable precision loss. Semantic-directed clumping uses
silhouette abstraction to select the disjuncts to join. Guided join then
subsequently uses these same silhouettes to perform the shape join.
We seek to evaluate the need for this latter step.
RH3 (Clumping has low overhead). The overhead of semantic-directed clumping is reasonable. We hope and expect that the additional overhead in computing and comparing silhouettes is outweighed by the benefit of increased precision and improved scalability. However, this must be tested empirically as the number of silhouette comparisons is quadratic in the number of disjuncts.

RH4 (Clumping limits disjunctive explosion). Semantic-directed clumping improves the scalability of disjunctive analysis by limiting the number of disjuncts in the abstract state. Without clumping, the number of disjuncts grows exponentially from control-flow paths and unfolding of inductive predicates. The scalability of disjunctive analysis (and thus most shape analyses) is limited by the exponential growth in the number of disjuncts in the abstract state. Thus the technical challenge is to avoid piling up more and more unnecessary disjuncts while analyzing sequences of operations.

Experimental methodology. To evaluate the effectiveness of semantic-based clumping, we consider 26 benchmarks of varying implementation styles and degrees of difficulty to analyze. These include operations over singly-linked and doubly-linked lists, as well as various binary trees with different kinds of invariants and pointer patterns (search, splay, red black, or AVL and with various sharing patterns). The operations consist of variants of finding an element, inserting, deleting, reversing, and sorting. Notably 21 benchmarks are from external sources. Some of these come with typically simple, user-specified assertions (e.g., \( x \neq \text{NULL} \)). Each benchmark consists of a top-level function implementing a data structure operation with a pre- and a post-condition that specify the preservation of precise shape invariants. Some routines are recursive whereas the others contain nested loops. Table 1 lists these benchmarks with metrics to get a sense of the difficulty to analyze, including presence of recursion, lines of code, numbers of sub-routines, numbers of loops and acyclic paths. The latter gives the number of disjuncts that a naïve path and context-sensitive analysis would have at the exit point. The final metric is the maximum number of simultaneous disjuncts that a naïve path and context-sensitive analysis would have at the exit point. The final metric is the maximum number of simultaneous disjuncts.

The analysis is parameterized by the definition of \texttt{inD} (or infers it for basic list and tree cases). It attempts to prove memory safety, any user-specified assertions, and the (more complex) post-condition given a C program and optionally inductive predicates summarizing memory regions. Clumping is applied at the beginning and end of functions, at loop heads, at loop exits, and at the end of branches. We evaluate and compare the following clumping strategies:

- \texttt{ClumpG} and \texttt{Clump} do silhouette-guided clumping (Sect. 4.4); \texttt{ClumpG} uses guided join (Sect. 4.3), whereas \texttt{Clump} does not;
- \texttt{None} is the baseline technique, that does not compute silhouettes to perform clumping or guided joining;
- \texttt{Canon} and \texttt{CanonG} conservatively model canonicalization operators [17, 35]: they compute silhouettes but only join abstractions when their silhouettes are exactly the same (after folding nodes which are not pointed to by live variables); \texttt{CanonG} uses guided joining whereas \texttt{Canon} does not (we expect these strategies will still compute fewer disjunctions than a purely syntactic canonicalization would).

**RH1: Clumping is effective.** Our analyzer attempts to infer complex disjunctive shape invariants in loops, which is particularly challenging considering the benchmarks shown in Table 1 with not only forward and back pointers but also possibly unbounded sharing patterns. Back pointers (as in doubly-linked lists or trees with parent pointers) require not only forward unfolding of inductive summaries but also backward unfolding [8]. The BSD red-black tree has parent pointers, which is heavily used in rebalancing. The leaves of the GDSL binary search tree all point back to the root node.

We observe a number of implementation idioms that drive a need for disjunctions. First, the maximum number of simultaneous pointers into a data structure instance (Simult. Pointer) is one driving factor as mentioned above, which typically increases as the operations get more complex (e.g., 12 for brbtree-delete). Second, the GDSL doubly-linked list uses sentinel nodes for the list head and tail. Because they are different than normal nodes, they cannot be summarized into the normal list segment, which yields more disjuncts to represent the possible points-to-relationships between

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>LOC</th>
<th>Assert</th>
<th>Fun</th>
<th>Loop</th>
<th>Path</th>
<th>Simult. Pointer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>singly-linked list</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sll-delmin</td>
<td>25</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>sll-delmin$^1$</td>
<td>26</td>
<td>0</td>
<td>2</td>
<td>$^1$</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>sll-delminmax</td>
<td>49</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>248</td>
</tr>
<tr>
<td>sll-delminmax$^1$</td>
<td>52</td>
<td>0</td>
<td>2</td>
<td>$^1$</td>
<td>124</td>
<td>7</td>
</tr>
<tr>
<td><strong>binary search tree</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bstree-find</td>
<td>26</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>bstree-find$^1$</td>
<td>26</td>
<td>0</td>
<td>2</td>
<td>$^1$</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td><strong>Predator singly-linked list</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psll-reverse</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>psll-isort</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>2$^*$</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>psll-bsort</td>
<td>25</td>
<td>0</td>
<td>1</td>
<td>2$^*$</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td><strong>GDSL doubly-linked list (back pointers) with sentinel head and tail</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gdll-findmin</td>
<td>49</td>
<td>14</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>gdll-index</td>
<td>55</td>
<td>14</td>
<td>9</td>
<td>2</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>gdll-findmax</td>
<td>58</td>
<td>14</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>gdll-find</td>
<td>78</td>
<td>26</td>
<td>10</td>
<td>1</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>gdll-delete</td>
<td>107</td>
<td>26</td>
<td>12</td>
<td>1</td>
<td>72</td>
<td>5</td>
</tr>
<tr>
<td><strong>GDSL binary search tree with leaf-to-root and back pointers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gbstree-find</td>
<td>53</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>gbstree-insert</td>
<td>133</td>
<td>15</td>
<td>12</td>
<td>1</td>
<td>7680</td>
<td>5</td>
</tr>
<tr>
<td>gbstree-delete</td>
<td>165</td>
<td>9</td>
<td>15</td>
<td>1</td>
<td>23040</td>
<td>10</td>
</tr>
<tr>
<td><strong>BSD splay tree</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bsplay-find</td>
<td>81</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>56</td>
<td>5</td>
</tr>
<tr>
<td>bsplay-delete</td>
<td>95</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>448</td>
<td>5</td>
</tr>
<tr>
<td>bsplay-insert</td>
<td>101</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>43</td>
<td>5</td>
</tr>
<tr>
<td><strong>BSD red black tree with back pointers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>brbtree-find</td>
<td>29</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>brbtree-insert</td>
<td>177</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>3036</td>
<td>7</td>
</tr>
<tr>
<td>brbtree-delete</td>
<td>329</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>$1.e + 8$</td>
<td>12</td>
</tr>
<tr>
<td><strong>JSW AVL tree</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>javl-find</td>
<td>25</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>26</td>
<td>2</td>
</tr>
<tr>
<td>javl-free</td>
<td>27</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>javl-insert</td>
<td>95</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>$1.e + 8$</td>
<td>6</td>
</tr>
<tr>
<td><strong>summary</strong></td>
<td>1917</td>
<td>126</td>
<td>123</td>
<td>34</td>
<td>2$e + 8$</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Table 1. List of benchmarks, divided into internal, microbenchmarks (top) and benchmarks from external sources (bottom) including the Predator test suite [18], the GNU Data Structure Library (GDSL), the BSD library, and a tutorial implementation of AVL trees (JSW) [37]. Some external libraries include typically simple, user-specified assertions (User Assert). A $^*$ indicates the routine is recursive, while a $^+$ indicates the loops are nested. The subsequent columns provide metrics for the complexity of the code, including the total number of functions (Fun), the number of loops (Loop), the number of acyclic paths (Path), and the maximum number of simultaneous pointers into a data structure instance (Simult. Pointer).
It is clear from the path counts (column Path) that this choice is acyclic. A path may be sufficiently precise in many or even most cases, while maintaining a disjunct, or trace partition, for every node in the tree. For example, delete in a red-black tree requires the materialized node in a tree. For example, delete in a red-black tree requires the materialized node to occur between loops. Disjunctions are needed to represent the cases where the materialized node can occur but becomes irrelevant between second and third loop (after the first and second loop (to be able to track the swap of n). While maintaining a disjunct, or trace partition, for every acyclic path may be sufficiently precise in many or even most cases, it is clear from the path counts (column Path) that this choice is utterly infeasible in practice. The essence of semantic-directed clumping is to identify this sort of relevance or irrelevance of a materialized node by computing silhouettes (i.e., “abstracting the abstraction”). In Fig. 11(a), we compare clumping with guided join with the other strategies showing the number of benchmarks that can be successfully verified memory safe (i.e., free of null or dangling pointer dereference), the user-specified asserts, and the shape preservation post-condition using the different strategies (also see the bottom line of Fig. 11(b)). The bottom line is that clumping with guided join (ClumpG) is significantly more effective (able to verify benchmarks) than the baseline and also more capable than any other strategy. CanonG fails to verify, for example, the particularly complex red-black tree insertion and deletion operations. Clumping with guided join is able to verify these additional benchmarks with comparable analysis times as any other configuration. The analysis time is larger for more complex benchmarks (e.g., for red-black tree delete btree-delete), which as expected raises its average analysis time, but ClumpG is also the only strategy that succeeds on this code. ClumpG is also notably faster than CanonG on javl-insert. Fig. 11(b) shows that ClumpG takes comparable or even less time than other strategies on benchmarks where other strategies succeed (e.g., in the btree benchmarks).

Another measure of effectiveness of ClumpG is that analysis logs show this strategy led to no precision loss in joins.

**RH2: Guided join is necessary.** Fig. 11(a) and 11(b) also show that not only is clumping effective but guided join is also necessary. While Clump does improve on the baseline (11 versus 9), the guided join strategies (ClumpG and CanonG) are significantly better (26 and 21, respectively). Clump and Canon fail to join in the tree benchmarks, as there is a larger search space for the join; guiding appears more critical when the search space for join is larger.

**RH3: Clumping has low overhead.** From the discussion above, we see that ClumpG is comparable in analysis times with any other configuration providing some evidence that clumping has a reasonable overhead. To see this more directly, we show in the graph below, the percentage of the analysis time spent on clumping, join, and other for each benchmark. We find that in all cases the time spent on clumping (which includes the time to compute silhouettes) is a very small percentage of the total analysis time (no more than a few percent), and much smaller than the time spent in join.

**RH4: Clumping limits disjunctive explosion.** Fig. 12(a) drills further down on the verification times from Fig. 11(b) by considering the maximum number of disjuncts produced at a loop head, any program point, and at the exit point of the operation. First, we observe that interestingly, the None baseline configuration only succeeds when the number of disjuncts at loop head is 1. When more than one disjunct is needed for the loop invariant, the silhouette abstraction seems crucial to guiding the inference of a sufficiently precise invariant. Second, in all cases where both ClumpG and CanonG succeed, ClumpG uses fewer disjuncts in the loop invariant, and the difference can be quite significant (e.g., for javl-insert, 3 versus 18). And interestingly, ClumpG is always able to get to a single disjunct at the exit point while proving the post-condition.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>ClumpG</th>
<th>Clump</th>
<th>CanonG</th>
<th>Canon</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>sll-delmin</td>
<td>0.04</td>
<td>⊤</td>
<td>0.05</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>sll-delmin1</td>
<td>0.04</td>
<td>0.05</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>sll-delminmax</td>
<td>0.12</td>
<td>⊤</td>
<td>0.42</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>sll-delminmax1</td>
<td>0.20</td>
<td>0.20</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>bstree-find</td>
<td>0.03</td>
<td>⊤</td>
<td>0.05</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>bstree-find1</td>
<td>0.04</td>
<td>0.04</td>
<td>0.11</td>
<td>0.11</td>
<td>⊤</td>
</tr>
<tr>
<td>psll-reverse</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>⊤</td>
</tr>
<tr>
<td>psll-isort</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>⊤</td>
</tr>
<tr>
<td>psll-isort</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>gdll-findmin</td>
<td>0.61</td>
<td>0.61</td>
<td>0.62</td>
<td>⊤</td>
<td>0.61</td>
</tr>
<tr>
<td>gdll-findmax</td>
<td>0.61</td>
<td>0.61</td>
<td>0.62</td>
<td>⊤</td>
<td>0.61</td>
</tr>
<tr>
<td>gdll-find</td>
<td>0.62</td>
<td>0.62</td>
<td>0.63</td>
<td>0.65</td>
<td>⊤</td>
</tr>
<tr>
<td>gdll-delete</td>
<td>0.62</td>
<td>0.63</td>
<td>0.63</td>
<td>0.64</td>
<td>⊤</td>
</tr>
<tr>
<td>gbstree-find</td>
<td>0.59</td>
<td>⊤</td>
<td>0.59</td>
<td>⊤</td>
<td>0.58</td>
</tr>
<tr>
<td>gbstree-insert</td>
<td>0.65</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>gbtree-delete</td>
<td>1.64</td>
<td>1.71</td>
<td>⊤</td>
<td>1.39</td>
<td>⊤</td>
</tr>
<tr>
<td>btree-insert</td>
<td>0.28</td>
<td>⊤</td>
<td>0.56</td>
<td>0.56</td>
<td>⊤</td>
</tr>
<tr>
<td>btree-delete</td>
<td>0.38</td>
<td>⊤</td>
<td>1.08</td>
<td>1.07</td>
<td>⊤</td>
</tr>
<tr>
<td>brbtree-find</td>
<td>0.36</td>
<td>⊤</td>
<td>0.62</td>
<td>0.62</td>
<td>⊤</td>
</tr>
<tr>
<td>brbtree-insert</td>
<td>1.07</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>brbtree-delete</td>
<td>6.06</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>javl-find</td>
<td>0.19</td>
<td>⊤</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>javl-tree</td>
<td>0.18</td>
<td>0.18</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>javl-insert</td>
<td>1.84</td>
<td>⊤</td>
<td>5.22</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>average (all)</td>
<td>0.66</td>
<td>0.28</td>
<td>0.68</td>
<td>0.39</td>
<td>0.47</td>
</tr>
</tbody>
</table>

(b) Analysis times are only reported if all assertions are successfully verified, including the preservation invariant in post-conditions. ⊤ indicates failure to verify (due to precision loss). Run times (in seconds) measured on one core of a 3.2GHz Intel Xeon with 16GB of RAM. Times are reported as an average of three runs.

**Figure 11.** Successful verification times with different analysis strategies. The clumping with guided join strategy can verify significantly more benchmarks than any other strategy and over 3x more than the None baseline—and all at similar cost in analysis time.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>ClumpG</th>
<th>Clump</th>
<th>CanonG</th>
<th>Canon</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>sll</td>
<td>⊤</td>
<td></td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>dll</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>bst</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>gbst</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>spt</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>rbt</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
<tr>
<td>avl</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
<td>⊤</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>clump</th>
<th>jsv</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>clumpG</td>
<td>⊤</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gbst</td>
<td>⊤</td>
<td>⊤</td>
<td></td>
</tr>
<tr>
<td>avl</td>
<td></td>
<td></td>
<td>⊤</td>
</tr>
</tbody>
</table>
allows one to support disjunctive properties while providing a way

Similarly, [19] introduces the least disjunctive basis of an abstract domain as a compact domain with the same disjunctive properties. Several generic techniques are used to handle disjunctions in static analyses. Disjunctive completion [11] adds support for disjunctions to an abstract domain, but never collapses them, thus is too expensive in practice. Similarly, [19] introduces the least disjunctive basis of an abstract domain as a compact domain with the same disjunctive completion, though this construction does not minimize the size of the representation of the abstract elements. State partitioning [12, 23] attaches abstract states to specific sets of states, which effectively allows one to support disjunctive properties while providing a way to control disjunct numbers via the definition of partitions. Trace partitioning [22, 33] achieves a similar result using information about traces. However, these works provide frameworks, and do not address the problem of finding a criterion to clump disjuncts.

Shape analyses based on three-valued logic [35], and on separation logic [4, 5, 7, 9, 18] are known to require disjunctions. The same goes for array analyses [21]. Several strategies have been designed to limit the number of disjunctions. When summary predicates denote non-empty regions, possibly empty regions can create additional need for disjunctions. Thus, several proposals have been made to let summary predicates denote possibly empty regions [9, 13, 38]. Canonicalization operators [4, 17, 25, 35] collapse abstract states into a smaller, finite lattice, and hereby bound the number of disjunctions. While the analysis may use to represent precision after application of this operation is limited by that of the smaller lattice.

Join operators [9, 38] do not require a smaller lattice and, hence can keep precision higher. To apply join or canonicalization operators efficiently, several strategies have been designed. Arnold [2] groups abstract states that satisfy some inclusion relation. Moreover, [38] performs a partial join that groups disjunctions only when it syntactically verifies the absence of precision loss. By contrast, our work proposes a criterion based on an abstraction of abstract states, and on the fact that this abstraction is cheaper to compute. A more related work is that of Manevich [27], that extends TVLA [25] with a grouping of three-valued abstract states based on a partial graph isomorphism.

Techniques to merge disjunctions have been developed in static analyses for numerical properties as well. For instance, a notion of affinity between polyhedra is used in [31] so as to decide whether they can be joined without too significant a precision loss. This approach is extended in [30] to deal with set properties. Baginna [3] proposed a widening over disjunctions of polyhedra, that tests for implication to better bound the precision loss.

Existing off-line approaches for the parameterization of static analyses and abstract domains (such as the selection of the abstract domain to use, and of the abstract values to keep) include syntactic heuristics based on code patterns [6], machine learning techniques [26, 29], and semantic impact pre-analysis methods [28]. The disjunct clumping problem is tied to the abstract states that arise during the analysis, thus unsurprisingly calls for using an on-line abstraction of these at analysis time.

Finally, we remark that other analyses that abstract structures with summaries [14], heap abstractions [15], rich type systems [24, 34] or quantified logical assertions [20] may benefit from adapted forms of silhouette abstraction when facing the disjunction problem.

8. Conclusion

In this paper, we introduced silhouettes that abstract abstract states. The information enclosed in the silhouettes proves useful not only to clump disjunctions of abstract states, but also to compute better abstract joins. These results were achieved by selecting a definition of silhouettes that provides a weak entailment check over abstract states, and hereby characterizes accurately abstract states that are likely to join well. This characterization is conservative with respect to the standard analysis algorithms: while it will always suggest clumping abstract states that can be joined precisely, it may suggest clumping abstract states that cannot be joined precisely, though we never observed this behavior in the experiments. This situation is quite similar to that of a static analysis that can be proven sound but is conservative in theory, yet computes very precise results in practice. Our experimental evaluation confirms the effectiveness of the silhouette abstraction, which allows our implementation to verify a large collection of challenging benchmarks at a reasonable cost.
Acknowledgments.

We are particularly grateful to Arlen Cox for discussions on the use of silhouettes in analysis and his comments on the paper. We also thank Hugo Illous, Yoonseok Ko, Jiangchao Liu, Kwangkeun Yi, and the anonymous referees for comments helping us improve this paper.

References