Static Program Analysis
Lecture 20: Wrap-Up Interprocedural DFA & Pointer Analysis
Winter Semester 2016/17
Thomas Noll
Software Modeling and Verification Group
RWTH Aachen University
https://moves.rwth-aachen.de/teaching/ws-1617/spa/
Recap: Interprocedural Dataflow Analysis – Fixpoint Solution

The Interprocedural Extension

Flow of information:
1. $\hat{\varphi}_l(d \cdot w) = \varphi_l(d) \cdot d \cdot w$
2. $\hat{\varphi}_n(d' \cdot d \cdot w) = \varphi_n(d') \cdot d \cdot w$
3. $\hat{\varphi}_x(d' \cdot d \cdot w) = \varphi_x(d') \cdot d \cdot w$
4. $\hat{\varphi}_r(d' \cdot d \cdot w) = \varphi_r(d', d) \cdot w$

\[\text{[call } \text{P(a,z)}]_{l_r}^c \quad \text{[P(val x, res y)]}_n^l \]

\[\text{(1) } \varphi_l(d) \cdot d \cdot w \quad \text{(2) } \varphi_n(\varphi_l(d)) \cdot d \cdot w \]

\[\text{(3) } \varphi_x(d') \cdot d \cdot w \quad \text{(4) } \varphi_r(\varphi_x(d'), d) \cdot w \]

[\text{[end]}_x^l]
Recap: Interprocedural Dataflow Analysis – Fixpoint Solution

Types of Equations

For an interprocedural dataflow system \( \hat{S} := (\text{Lab}, E, F, (\hat{D}, \hat{\sqcap}), \hat{\iota}, \hat{\varphi}) \), the intraprocedural equation system (cf. Definition 4.9)

\[
\text{AI}_l = \begin{cases} 
\iota & \text{if } l \in E \\
\bigsqcup \{ \varphi_{l'}(\text{AI}_{l'}) | (l', l) \in F \} & \text{otherwise}
\end{cases}
\]

is extended to a system with three kinds of equations (for every \( l \in \text{Lab} \)):

- for actual dataflow information: \( \text{AI}_l \in \hat{D} \)
  - counterpart of intraprocedural AI
- for transfer functions of single nodes: \( f_l : \hat{D} \rightarrow \hat{D} \)
  - extension of intraprocedural transfer functions by special handling of procedure calls
- for transfer functions of complete procedures: \( F_l : \hat{D} \rightarrow \hat{D} \)
  - \( F_l(w) \) yields information at \( l \) if corresponding procedure is called with information \( w \)
  - thus complete procedure represented by \( F_{lx} \) (“procedure summary”)
Recap: Interprocedural Dataflow Analysis – Fixpoint Solution

Formal Definition of Equation System

Dataflow equations:

\[ A_l = \begin{cases} 
l & \text{if } l \in E \\
A_{lc} & \text{if } l = l_r \text{ for some } (l_c, l_n, l_x, l_r) \in \text{iflow} \\
\bigsqcup \{ f_{l'}(A_{l'}) \mid (l', l) \in F \} & \text{otherwise} \end{cases} \]

Node transfer functions (if \( l \) not an exit label):

\[ f_l(w) = \begin{cases} 
\hat{\phi}_{l_r}(\hat{\phi}_{l_x}(F_{l_x}(\hat{\phi}_{l_c}(w)))) & \text{if } l = l_r \text{ for some } (l_c, l_n, l_x, l_r) \in \text{iflow} \\
\hat{\phi}_l(w) & \text{otherwise} \end{cases} \]

Procedure transfer functions (if \( l \) occurs in some procedure):

\[ F_l(w) = \begin{cases} 
w & \text{if } l = l_n \text{ for some } (l_c, l_n, l_x, l_r) \in \text{iflow} \\
\bigsqcup \{ f_{l'}(F_{l'}(w)) \mid (l', l) \in F \} & \text{otherwise} \end{cases} \]

As before: induces monotonic functional on lattice with ACC \( \Rightarrow \) least fixpoint effectively computable
Effectiveness and Correctness

Effectiveness of Fixpoint Iteration

For the fixpoint iteration it is important that the auxiliary functions only operate (at most) on the two topmost elements of the stack:

**Lemma 20.1**

For every $l \in \text{Lab}$, $d \in D$, and $w \in D^*$,

$$f_l(d' \cdot d \cdot w) = f_l(d' \cdot d) \cdot w$$

and

$$F_l(d' \cdot d \cdot w) = F_l(d' \cdot d)w$$

**Proof.**


It therefore suffices to consider stacks with **at most two entries**, and so the fixpoint iteration ranges over “finitary objects”.

7 of 24

Static Program Analysis
Winter Semester 2016/17
Lecture 20: Wrap-Up Interprocedural DFA & Pointer Analysis
Effectiveness and Correctness

Soundness and Completeness

The following results carry over from the intraprocedural case:

**Theorem 20.2**

Let \( \hat{S} := (\text{Lab}, E, F, (\hat{D}, \hat{\subseteq}), \hat{\ast}, \hat{\phi}) \) be an interprocedural dataflow system.

1. (cf. Theorem 6.3)
   \[
   \text{mvp}(\hat{S}) \hat{\subseteq} \text{fix}(\Phi_{\hat{S}})
   \]

2. (cf. Theorem 6.6)
   \[
   \text{mvp}(\hat{S}) = \text{fix}(\Phi_{\hat{S}}) \text{ if all } \hat{\phi}_l \text{ are distributive}
   \]

**Proof.**

Context-Sensitive Interprocedural Dataflow Analysis

Context-Sensitive Interprocedural DFA

- **Observation:** MVP and fixpoint solution maintain *proper relationship between procedure calls and returns*
- **But:** do not distinguish between *different procedure calls*
  - information about calling states *combined for all call sites*
  - procedure body only *analysed once using combined information*
  - resulting information used at *all return points*
  \[ \Rightarrow \text{“context-insensitive”} \]
- **Alternative:** *context-sensitive* analysis
  - separate information for different call sites
  - implementation by “*procedure cloning*” (one copy for each call site)
  - more *precise*
  - more *costly*
Pointer Analysis

- **So far:** only static data structures (variables)
- **Now:** pointer (variables) and dynamic memory allocation using heaps

**Problem:**
- Programs with pointers and dynamically allocated data structures are error prone
- Identify subtle bugs at compile time
- Automatically prove correctness

**Interesting properties of heap-manipulating programs:**
- No null pointer dereference
- No memory leaks
- Preservation of data structures
- Partial/total correctness
### Pointer Analysis

#### The Shape Analysis Approach

- **Goal:** determine the possible shapes of a dynamically allocated data structure at given program point
- **Interesting information:**
  - data types (to avoid type errors, such as dereferencing `nil`)
  - aliasing (different pointer variables having same value)
  - sharing (different heap pointers referencing same location)
  - reachability of nodes (garbage collection)
  - disjointness of heap regions (parallelisability)
  - shapes (lists, trees, absence of cycles, ...)
- **Concrete questions:**
  - Does `x.next` point to a shared element?
  - Does a variable `p` point to an allocated element every time `p` is dereferenced?
  - Does a variable point to an acyclic list?
  - Does a variable point to a doubly-linked list?
  - Can a loop or procedure cause a memory leak?
- **Here:** basic outline; details in [Nielson/Nielson/Hankin 2005, Sct. 2.6]
Introducing Pointers

Extending the Syntax

Syntactic categories:

<table>
<thead>
<tr>
<th>Category</th>
<th>Domain</th>
<th>Meta variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic expressions</td>
<td>AExp</td>
<td>a</td>
</tr>
<tr>
<td>Boolean expressions</td>
<td>BExp</td>
<td>b</td>
</tr>
<tr>
<td>Selector names</td>
<td>Sel</td>
<td>sel</td>
</tr>
<tr>
<td>Pointer expressions</td>
<td>PExp</td>
<td>p</td>
</tr>
<tr>
<td>Commands (statements)</td>
<td>Cmd</td>
<td>c</td>
</tr>
</tbody>
</table>

Context-free grammar:

\[
\begin{align*}
a & ::= z \mid x \mid a_1+a_2 \mid \ldots \mid p \mid \text{nil} \in AExp \\
b & ::= t \mid a_1=a_2 \mid b_1 \land b_2 \mid \ldots \mid \text{is-nil}(p) \in BExp \\
p & ::= x \mid x.sel \\
c & ::= [\text{skip}]' \mid [p := a]' \mid c_1;c_2 \mid \text{if } [b]' \text{ then } c_1 \text{ else } c_2 \text{ end} \mid \\
    & \quad \text{while } [b]' \text{ do } c \text{ end} \mid [\text{malloc } p]' \in Cmd
\end{align*}
\]
Example 20.3 (List reversal)

Program that reverses list pointed to by \( x \) and leaves result in \( y \):

\[
x \rightarrow n_1 \xrightarrow{\text{next}} n_2 \xrightarrow{\text{next}} n_3 \xrightarrow{\text{next}} \diamond
\]

\[
y \rightarrow \diamond
\]

\[
z
\]

\[
x \rightarrow n_1 \xrightarrow{\text{next}} n_2 \xrightarrow{\text{next}} n_3 \xrightarrow{\text{next}} \diamond
\]

\[
y \rightarrow \diamond
\]

\[
z \rightarrow \diamond
\]
Introducing Pointers

Heap Configurations

Definition 20.4 (Heap configuration)

A (concrete) heap configuration is given by

\[ H = (Nod, Sel, Var, \sigma, \rightarrow) \]

where

- **Nod** is a finite set of (concrete) nodes
- **Sel** is a finite set of selector names
- **Var** is a finite set of program variables
- \( \sigma : Var \rightarrow \mathbb{Z} \cup \text{Nod}^\diamond \) is a variable valuation (with \( \text{Nod}^\diamond := \text{Nod} \cup \{\diamond\} \))
- \( \rightarrow : \text{Nod} \times \text{Sel} \rightarrow \text{Nod}^\diamond \) is a (concrete) heap
  - notation: \( n_1 \xrightarrow{sel} n_2 \) for \( ((n_1, sel), n_2) \in \rightarrow \)
Shape Graphs

Shape Graphs I

**Approach:** representation of (infinitely many) concrete heap states by (finitely many) abstract shape graphs

- abstract nodes \( X \) = sets of variables
- interpretation: \( x \in X \) iff \( x \) points to concrete node represented by \( X \)
- \( \emptyset \) represents all concrete nodes that are not directly addressed by pointer variables
- \( x, y \in X \) (with \( x \neq y \)) indicate aliasing (as \( x \) and \( y \) point to the same concrete node)
- if \( x.\text{sel} \) and \( y \) refer to the same heap address and if \( X, Y \) are abstract nodes with \( x \in X \) and \( y \in Y \), this yields abstract edge \( X \xrightarrow{\text{sel}} Y \) (similarly for \( X = \emptyset \) or \( Y = \emptyset \))
- transfer functions transform (sets of) shape graphs
Example 20.5 (List reversal; cf. Example 20.3)

Concrete heap

```
x -> n₁ next n₂ next n₃ next ♦
ych
z
x -> n₁ next n₂ next n₃ next ♦
y ♦
z
x -> n₁ next n₂ next n₃ next ♦
y
```

Shape graph

```
{x} next ∅
{x} next ∅
{x} next ∅
```
Shape Graphs

Definition 20.6 (Shape graph)

A shape graph

\[ G = (\text{Abs}, \Rightarrow) \]

consists of

- a set \( \text{Abs} \subseteq 2^{\text{Var}} \) of abstract locations and
- an abstract heap \( \Rightarrow \subseteq \text{Abs} \times \text{Sel} \times \text{Abs} \)
  - notation: \( X \xrightarrow{\text{sel}} Y \) for \( (X, \text{sel}, Y) \in \Rightarrow \)

with the following properties:

Disjointness: \( X, Y \in \text{Abs} \Rightarrow X = Y \) or \( X \cap Y = \emptyset \)
(a variable can refer to at most one heap location)

Determinacy: \( X \neq \emptyset \) and \( X \xrightarrow{\text{sel}} Y \) and \( X \xrightarrow{\text{sel}} Z \Rightarrow Y = Z \)
(target location is unique if source node is unique)

\( \text{SG} \) denotes the set of all shape graphs.
Shape Graphs

From Heap Configurations to Shape Graphs I

**Definition 20.7**

Given a heap configuration \( H = (Nod, Sel, Var, \sigma, \rightarrow) \), the corresponding shape graph \( G = (Abs, \Longrightarrow) \) is defined by

- \( Abs := \{ \sigma^{-1}(n) \mid n \in Nod \} \)
  \[= \{ \{ x \in Var \mid \sigma(x) = n \} \mid n \in Nod \} \]
- For all \( X, Y \in Abs \) and \( sel \in Sel \):
  \[ X \xrightarrow{sel} Y \Longleftrightarrow \exists n_x, n_y \in Nod : \sigma^{-1}(n_x) = X, \sigma^{-1}(n_y) = Y, n_x \xrightarrow{sel} n_y \]

**Remark:** yields Galois connection between sets of heap configurations and sets of shape graphs, both ordered by \( \subseteq \)
Remark: the following example shows that determinacy can only be postulated if $X \neq \emptyset$:

- Concrete:

  $\begin{align*}
  y & \rightarrow \bullet \xleftarrow{\text{sel}} \bullet \\
  z & \rightarrow \bullet \xleftarrow{\text{sel}} \bullet
  \end{align*}$

- Abstract:

  $\begin{align*}
  Y = \{y\} & \xleftrightarrow{\text{sel}} X = \emptyset & \xrightarrow{\text{sel}} Z = \{z\}
  \end{align*}$
Example 20.8

Let $G = (\text{Abs}, \longrightarrow)$ be a shape graph. Then the following concrete heap properties can be expressed as conditions on $G$:

- $x \neq \text{nil}$
  $\iff$ $\exists X \in \text{Abs} : x \in X$

- $x = y \neq \text{nil}$ (aliasing)
  $\iff$ $\exists Z \in \text{Abs} : x, y \in Z$

- $x.\text{sel1} = y.\text{sel2} \neq \text{nil}$ (sharing)
  $\rightarrow$ $\exists X, Y, Z \in \text{Abs} : x \in X, y \in Y, X \seleq Z \seldeq Y$
  (“$\seleq$” only valid if $Z \neq \emptyset$)