Program Analysis (70020)
Overview

Herbert Wiklicky
Department of Computing
Imperial College London

herbert@doc.ic.ac.uk
h.wiklicky@imperial.ac.uk

Autumn 2022

Lectures 11 October and 24 November 2022

Lecture Theatre 145 on Tuesday (2pm-4pm)
and Thursday (9am-11am).

Tutorials typically Tuesdays, second hour.
Coursework Tests on 27 October and 22 November

Material and Notes on
https://www.doc.ic.ac.uk/~herbert/teaching.html
Scientia, Panopto (COMP97146/97128), etc.

Assessment
Coursework Test I: Thu 27 October at 9:00
Coursework Test II: Tue 22 November at 14:00
Examination: Week 11, 12-16 December
Program Analysis

Program analysis is an automated technique for finding out properties of programs without having to execute them.

Static Analysis vs Dynamic Testing

- Compiler Optimisation
- Program Verification
- Security Analysis

Unfortunately, the achieving the aims of (static) program analysis tend to be computationally extremely hard.

Program Properties

In some sense Program Analysis is an impossible task.

Decidable Problems There exists an algorithm or computational process which computes a solution (in finite time) for all instances of the problem.

Halting Problem There is no general computational process or machine which can decide whether or not any given program terminates.

Rice Theorem Any non-trivial program property is undecidable.

The approach is to find terminating algorithms for program analysis while not always finding a “meaningful” solution.
Fermat’s Program – Terminates

1: try ← true;
2: x ← 1;
3: while try do
4:    y ← 1;
5:    while y ≤ x && try do
6:        z ← 1;
7:        while z ≤ y && try do
8:            try ← x^3 + y^3 ≠ z^3
9:            z ← z + 1;
10:       end while
11:    y ← y + 1;
12: end while
13: x ← x + 1;
14: end while

Collatz Problem – Unknown

Take an integer x and compute a sequence of updates:

1: while x ≠ 1 do
2:    if x mod 2 = 0 then
3:        x ← x/2;
4:    else
5:        x ← 3 × x + 1
6:    end if
7: end while

Currently it is unknown whether this terminates for all x.
Techniques

Some techniques used in program analysis include:

- Data Flow Analysis
- Control Flow Analysis
- Types and Effects Systems
- Abstract Interpretation


A First Example

Consider the following fragment in *some* procedural language.

1: \( m \leftarrow 2; \)  
2: \( \textbf{while } n > 1 \textbf{ do} \)  
3: \( m \leftarrow m \times n; \)  
4: \( n \leftarrow n - 1 \)  
5: \( \textbf{end while} \)  
6: \( \textbf{stop} \)

We annotate a program such that it becomes clear about what program point we are talking about.
Claim: This program fragment always returns an even $m$, independently of the initial values of $m$ and $n$.

We can statically determine that in any circumstances the value of $m$ at the last statement will be even for any input $n$.

A program analysis, so-called parity analysis, can determine this by propagating the even/odd or parity information forwards form the start of the program.

Properties

We will assign to each variable one of three properties:

- **even** — the value is known to be even
- **odd** — the value is known to be odd
- **unknown** — the parity of the value is unknown

For both variables $m$ and $n$ we record its parity at each stage of the computation (beginning of each statement).
A First Example

Executing the program with abstract values, parity, for $m$ and $n$.

1: $m \leftarrow 2$; \hspace{1cm} \triangleright \text{unknown}(m) - \text{unknown}(n)$
2: while $n > 1$ do \hspace{1cm} \triangleright \text{even}(m) - \text{unknown}(n)$
3: \hspace{1cm} $m \leftarrow m \times n$; \hspace{1cm} \triangleright \text{even}(m) - \text{unknown}(n)$
4: \hspace{1cm} $n \leftarrow n - 1$ \hspace{1cm} \triangleright \text{even}(m) - \text{unknown}(n)$
5: end while \hspace{1cm} \triangleright \text{even}(m) - \text{unknown}(n)$
6: stop \hspace{1cm} \triangleright \text{even}(m) - \text{unknown}(n)$

Important: We can restart the loop!

A First Example

The first program computes 2 times the factorial for any positive value of $n$. Replacing ‘2’ by ‘1’ in the first statement gives:

1: $m \leftarrow 1$; \hspace{1cm} \triangleright \text{unknown}(m) - \text{unknown}(n)$
2: while $n > 1$ do \hspace{1cm} \triangleright \text{unknown}(m) - \text{unknown}(n)$
3: \hspace{1cm} $m \leftarrow m \times n$; \hspace{1cm} \triangleright \text{unknown}(m) - \text{unknown}(n)$
4: \hspace{1cm} $n \leftarrow n - 1$ \hspace{1cm} \triangleright \text{unknown}(m) - \text{unknown}(n)$
5: end while \hspace{1cm} \triangleright \text{unknown}(m) - \text{unknown}(n)$
6: stop \hspace{1cm} \triangleright \text{unknown}(m) - \text{unknown}(n)$

i.e. the factorial – but then the program analysis is unable to tell us anything about the parity of $m$ at the end of the execution.
Loss of Precision

The analysis of the new program does not give a satisfying result because:

▶  $m$ could be even — if the input $n > 1$, or
▶  $m$ could be odd — if the input $n \leq 1$.

However, even if we fix/require the input to be positive and even — e.g. by some suitable conditional assignment — the program analysis still might not be able to accurately predict that $m$ will be even at statement 5.

Safe Approximations

Such a loss of precision is a common feature of program analysis: many properties that we are interested in are essentially undecidable and therefore we cannot hope to detect (all of) them accurately.

We only aim to ensure that the answers/results we obtain by program analysis are at least safe, i.e.

▶  yes means definitely yes,
▶  no means maybe no.
Data Flow Analysis

The starting point for data flow analysis is a representation of the control flow graph of the program: the nodes of such a graph may represent individual statements – as in a flowchart – or sequences of statements; arcs specify how control may be passed during program execution.

The data flow analysis is usually specified as a set of equations which associate analysis information with program points which correspond to the nodes in the control flow graph. This information may be propagated forwards through the program (e.g. parity analysis) or backwards.

When the control flow graph is not explicitly given, we need a preliminary control flow analysis.

Control Flow Information

\[
\begin{align*}
[z := 1] & \quad 1 \\
[x > 0] & \quad 2 \\
[z := z \cdot y] & \quad 3 \\
[x := x - 1] & \quad 4 \\
\end{align*}
\]

This allows us to determine the predecessors \textit{pred} and successors \textit{succ} of each statement, e.g. \( \text{pred}(2) = \{1, 4\} \).
Reaching Definition

Reaching Definition (RD) analysis determines which set of definitions (i.e. assignments) are current when control reaches a certain program point \( p \).

The analysis can be specified by equations of the form:

\[
RD_{\text{entry}}(p) = \begin{cases} 
RD_{\text{init}} & \text{if } p \text{ is initial} \\
\bigcup_{p' \in \text{pred}(p)} RD_{\text{exit}}(p') & \text{otherwise}
\end{cases}
\]

\[
RD_{\text{exit}}(p) = (RD_{\text{entry}}(p) \setminus \text{kill}_{RD}(p)) \cup \text{gen}_{RD}(p)
\]

Analysis Information

At each program point some definitions get “killed” (those which define the same variable as at the program point) while others are “generated”.

A suitable representation for properties are sets of pairs, where each pair contains a variable \( x \) and a program point \( p \): the meaning of the pair \((x, p)\) is that the assignment to \( x \) at point \( p \) is the current one. The initial value in this case is:

\[
RD_{\text{init}} = \{(x, ?) \mid x \text{ is a variable in the program}\}
\]

Reaching Definitions is a forward analysis and we require the least (most precise) solutions to the set of equations.
For our initial program fragment

\[
\begin{align*}
&m \leftarrow 2^1; \\
&\textbf{while} [n > 1]^2 \textbf{do} \\
&m \leftarrow m \times n^3; \\
&n \leftarrow n - 1^4 \\
&\textbf{end while} \\
&[\text{stop}]^5
\end{align*}
\]

some of the \textit{RD} equations we get are:

\[
\begin{align*}
\text{RD}_{\text{entry}}(1) &= \{(m,?), (n,?)\} \\
\text{RD}_{\text{entry}}(2) &= \text{RD}_{\text{exit}}(1) \cup \text{RD}_{\text{exit}}(4)
\end{align*}
\]
Solving Equations

How can we construct solution to the data flow equations? Answer: Iteratively, by improving approximations/guesses.

INPUT: Control Flow Graph
i.e. initial(p), pred(p).

OUTPUT: Reaching Definitions $RD$.

METHOD: Step 1: Initialisation
Step 2: Iteration

Some Examples

Some examples of data flow analyses — and the possible applications and optimisations they allow for — are:

- Reaching Definitions — Constant Folding
- Available Expressions — Avoid Re-computations
- Very Busy Expressions — Hoisting
- Live Variables — Dead Code Elimination
- Information Flow — Computer Security
- (Probabilistic) Program Slicing
- Shape Analysis — Pointer Analysis — etc.
To illustrate the ideas we shall show how Reaching Definitions can be used to perform Constant Folding.

There are two ingredients to this:

- Replace the use of a variable in some expression by a constant if it is known that the value of that variable will always be a constant.
- Simplify an expression by partially evaluating it: subexpressions that contain no variables can be evaluated.

**Constant Folding I**

\[
RD \vdash [x := a]^{\ell} \triangleright [x := a[y \mapsto n]]^{\ell}
\]

if \[
\begin{cases}
y \in \text{FV}(a) \land (y, ?) \notin RD_{\text{entry}}(\ell) \land \\
\forall (y', \ell') \in RD_{\text{entry}}(\ell) : \\
y' = y \Rightarrow [. . .]^{\ell'} = [y := n]^{\ell'}
\end{cases}
\]

\[
RD \vdash [x := a]^{\ell} \triangleright [x := n]^{\ell}
\]

if \[
\begin{cases}
\text{FV}(a) = \emptyset \land a \text{ is not constant} \land \\
a \text{ evaluates to } n
\end{cases}
\]
Constant Folding II

\[
\begin{align*}
  RD \vdash S_1 \triangleright S'_1 \\
  RD \vdash S_1; S_2 \triangleright S'_1; S_2 \\
  RD \vdash S_2 \triangleright S'_2 \\
  RD \vdash S_1; S_2 \triangleright S_1; S'_2 \\
  RD \vdash S_1 \triangleright S'_1 \\
  RD \vdash \text{if } [b]^\ell \text{ then } S_1 \text{ else } S_2 \triangleright \text{if } [b]^\ell \text{ then } S'_1 \text{ else } S_2 \\
  RD \vdash S_2 \triangleright S'_2 \\
  RD \vdash \text{if } [b]^\ell \text{ then } S_1 \text{ else } S_2 \triangleright \text{if } [b]^\ell \text{ then } S_1 \text{ else } S'_2 \\
  RD \vdash S \triangleright S' \\
  RD \vdash \text{while } [b]^\ell \text{ do } S \triangleright \text{while } [b]^\ell \text{ do } S'
\end{align*}
\]

An Example

To illustrate the use of the transformation consider:

\[
[ x := 10 ]^1; [ y := x + 10 ]^2; [ z := y + 10 ]^3
\]

The (least) solution to the Reaching Definition analysis is:

\[
\begin{align*}
  \text{RD}_{\text{entry}}(1) & = \{(x, ?), (y, ?), (z, ?)\} \\
  \text{RD}_{\text{exit}}(1) & = \{(x, 1), (y, ?), (z, ?)\} \\
  \text{RD}_{\text{entry}}(2) & = \{(x, 1), (y, ?), (z, ?)\} \\
  \text{RD}_{\text{exit}}(2) & = \{(x, 1), (y, 2), (z, ?)\} \\
  \text{RD}_{\text{entry}}(3) & = \{(x, 1), (y, 2), (z, ?)\} \\
  \text{RD}_{\text{exit}}(3) & = \{(x, 1), (y, 2), (z, 3)\}
\end{align*}
\]
Constant Folding

We have for example the following:

\[ RD \vdash [ y := x + 10 ]^2 \triangleright [ y := 10 + 10 ]^2 \]

and therfore the rules for sequential composition allow us to do the following transformation:

\[ RD \vdash [ x := 10 ]^1; [ y := x + 10 ]^2; [ z := y + 10 ]^3 \triangleright [ [ x := 10 ]^1; [ y := 10 + 10 ]^2; [ z := y + 10 ]^3 \]

Transformation

We can continue this kind of transformation and obtain:

\[ RD \vdash [ x := 10 ]^1; [ y := x + 10 ]^2; [ z := y + 10 ]^3 \triangleright [ x := 10 ]^1; [ y := 10 + 10 ]^2; [ z := y + 10 ]^3 \]
\[ \triangleright [ x := 10 ]^1; [ y := 20 ]^2; [ z := y + 10 ]^3 \]
\[ \triangleright [ x := 10 ]^1; [ y := 20 ]^2; [ z := 20 + 10 ]^3 \]
\[ \triangleright [ x := 10 ]^1; [ y := 20 ]^2; [ z := 30 ]^3 \]

after which no more steps are possible.
Additional Issues

The above example shows that optimisation is in general the result of a number of successive transformations.

\[ RD \vdash S_1 \triangleright S_2 \triangleright \ldots \triangleright S_n. \]

This could be costly because one \( S_1 \) has been transformed into \( S_2 \) we might have to *re-compute* the Reaching Definition analysis before the next transformation step can be done.

It could also be the case that different sequences of transformations either lead to different end results or are of very different length.

Correctness

Any Program Analysis should be:

\[ \begin{align*} 
&\triangleright \text{ unambiguously specified,} \\
&\triangleright \text{ efficiently computable,} \\
&\triangleright \text{ most importantly: correct.} 
\end{align*} \]

For example, why not consider in \( RD \) before:

\[ RD_{\text{entry}}(2) = RD_{\text{exit}}(1) \cap RD_{\text{exit}}(4) \]

instead of \( RD_{\text{entry}}(2) = RD_{\text{exit}}(1) \cup RD_{\text{exit}}(4) \).

It requires formal (mathematical) proof whether an analysis (or program transformation) is correct with respect to some model of execution or semantics.
A program is foremost a text but it has intended meaning or semantics describing its execution.

A simple example: Why is \(0.\dot{9} = 0.99999 \ldots = 1\)?

Obviously, these are different strings! However, they have a meaning or semantics as specification of a real number in \(\mathbb{R}\). More concretely, infinite strings refer to the limit of their expansion, so \([0.\dot{9}] = \lim(0.9, 0.99, 0.999, \ldots) = 1 = [1]\).

This course will mostly be concerned with intuitive or light-weight semantics when it comes to the “meaning” of a program and the correctness of a program analysis.

---

Topics Covered – Executive Summary

- Data Flow Analysis
- Monotone Frameworks
- Control Flow Analysis
- Abstract Interpretation
- Probabilistic Programs
- Probabilistic Abstract Interpretation
- Further Topics