## 



## [ HOW TO ANALYZE LANGUAGES AUTOMATICALLY ]



## ト Abstract

## - "Data-Flow Analysis":

In this $3 * 3$ hour mini course we will look at data-flow analysis. Rather than just look at the classical "monotone framework" analyses (which are usually synonymous with teaching data-flow analysis: reaching definitions, live variables, available expressions, and very busy expressions), we will instead take one step backwards and look at the general theory and practice behind these analyses. The idea is that you will then learn how to design your own customized data-flow analyses for automatically analyzing whatever aspects of programming languages you want to. (From this perspective, the monotone framework analyses are just special cases.)

## Keywords:

- undecidability, approximation, control-flow graphs, partial-orders, lattices, transfer functions, monotonicity, [how to solve] fixed-point equations - and how all of these things combine to enable you to design data-flow analyses.



## - Agenda

## - Introduction:

- Undecidability, Reduction, and Approximation
- Data-flow Analysis:
- Quick tour of everything \& running example
- Control-Flow Graphs:
- Control-flow, data-flow, and confluence
- "Science-Fiction Math":
- Lattice theory, monotonicity, and fixed-points
$\square$
- Putting it all together...: (next wednesday)
- Example revisited


## - Notes on Static Analysis



## - Quiz: Optimization?

- If you want a fast C-program, should you use:
- LOOP 1:

```
for (i = 0; i < N; i++) {
a[i] = a[i] * 2000;
    a[i] = a[i] / 10000;
}
```

- LOOP 2 (optimized by programmer):

```
b = a;
for (i = 0; i < N; i++) {
    *b = *b * 2000;
    *b = *b / 10000;
    b++;
}
```

■ i.e., "array-version" or "optimized pointer-version" ?

## - Answer:

- Results (of running the programs):

| LOOP | opt. level | SPARC | MIPS | Alpha |
| :--- | :---: | :---: | :---: | :---: |
| \#1 (array) | no opt | 20.5 | 21.6 | 7.85 |


| \#2 (ptr) | no opt | 19.5 | 17.6 | 7.55 |
| :--- | :---: | :---: | :---: | :---: |

- Compilers use highly sophisticated static analyses for optimization! (you'll learn how to do this!!!)
- Recommendation: focus on writing clear code for people (and compilers) to understand!


## - Data-Flow Analysis

- Purpose (of Data-Flow Analysis):
- Gather information (on running behavior of program)
- " $\forall$ program points"
- Usage (of static analysis):
- Basis for subsequent...:


## Error Detection

Static Analysis


Optimization

## $\vdash$ Analyses for Error Detection

- Example Analyses:

■ "Symbol Checking":

- Catch (dynamic) symbol errors
- "Type Checking":
- Catch (dynamic) type errors

■ "Initialized Variable Analysis":

- Catch unintialized variables

■ ...
■...
■...
■ ...

## - Analyses for Optimization

- Example Analyses:

■ "Constant Propagation Analysis":

- Precompute constants (e.g., replace '5* $x+z$ ' by '42')
- "Live Variables Analysis":
- Dead-code elimination (e.g., get rid of unused variable 'z')
- "Available Expressions Analysis":
- Avoid recomputing already computed exprs (cache results)

■...
■...
■...
■...

## Conceptual Motivation

- Undecidability
- Reduction principle
- Approximation


## - Rice's Theorem (1953)

"Any interesting problem about the runtime behavior of a program* is undecidable"

- Examples:


## - does program ' $P$ ' always halt when run?

- is the value of integer variable ' $x$ ' always positive?
- does variable ' $x$ ' always have the same value?
- which variables can pointer 'p' point to?
- does expression ' $E$ ' always evaluate to true?
- what are the possible outputs of program 'P'?

■...

## - Undecidability (self-referentiality)

. Consider "The Book-of-all-Books":

- This book contains the titles of all books that do not have a self-reference (i.e. don't contain their title inside)


## "The Bible"

- Finitely many books; i.e.:
- We can sit down \& figure out whether to include or not...
- Q: What about "The Book-of-all-Books";
- Should it be included or not?


## - "Self-referential paradox" (many guises):

- e.g. "This sentence is false" $\leftarrow$


## $\vdash$ Termination Undecidable!

- Assume termination is decidable (in Java);

■i.e. $\exists$ some program, halts: Program $\rightarrow$ bool

```
bool halts(Program p) { ... }
```

```
-- P}\mp@subsup{P}{0}{}.java --
Program p
if (halts(po)) loop();
else halt();
```

- Q: Does $\mathrm{P}_{0}$ loop or terminate...? :)
- Hence: halts cannot exist!

■ i.e., "Termination is undecidable" *) for turing-complete languages

## - Rice's Theorem (1953)

"Any interesting problem about the runtime behavior program* is undecidable"

- Examples:
*) written in a turing-complete language


## - does program 'P' always halt?

- is the value of integer variable ' $x$ ' always positive?
- does variable 'x' always have the same value?
- which variables can pointer 'p' point to?
- does expression 'E' always evaluate to true?
- what are the possible outputs of program 'P'?


## Reduction:

solve always-pos $\Rightarrow$ solve halts

1) Assume ' $x$-is-always-pos( $(P)$ ' is decidable
2) Given $P$ (here's how we could solve 'halts $\left.(P)^{\prime}\right)$ :
3) Construct (veeeeery clever) reduction program $R$ :
```
|
```

4) Run "supposedly decidable" analysis: res $=x$-is-always-positive $(R)$
5) Deduce from result:
if (res) then $\boldsymbol{P}$ loops!; else $\boldsymbol{P}$ halts $\quad:-)$
6) THUS: ' $x$-is-always-pos(P)' must be undecidable!

## - Reduction Principle

- Reduction principle (in short):
$\phi(P)$ undecidable $\wedge[$ solve $\psi(P) \Rightarrow$ solve $\phi(P)]$ $\psi(P)$ undecidable
- Example:
reduction
'halts $(P)$ ' undecidable $\wedge$ [solve 'x-is-always-pos $(P)$ ' $\Rightarrow$ solve 'halts $(P)$ '] ' $x$-is-always-pos(P)' undecidable
- Exercise:
- Carry out reduction + whole explanation for:

■ "which variables can pointer 'q' point to?"

## Answer

1) Assume 'which-var-q-points-to(P)' is decidable:
2) Given $P$ (here's how to (cleverly) decide halts( $P$ )):
3) Construct (veeeeery clever) reduction program $R$ :

| -- R.java -- |
| :---: |
| ```ptr q = 0xffff; P /* insert program P (assume w/o 'q') */ q = null;``` |

4) Run 'which-var-q-points-to(R)' = res
5) If (null $\in$ res) $P$ halts! else; $P$ loops! :-)
6) THUS:
'which-var-q-points-to( $P$ )' must be undecidable!

## $\vdash$ Undecidability

- Undecidability means that...:


■ ...no-one can decide this line (for all programs)!

- However(!)...


# - "Side-Stepping Undecidability" 

However, just because it's undecidable, doesn't mean there aren't (good) approximations! Indeed, the whole area of static analysis works on "side-stepping undecidability":

- Compilers use safe approximations (computed via "static analyses") such that:



## ト"Side-Stepping Undecidability"

However, just because it's undecidable, doesn't mean there aren't (good) approximations! Indeed, the whole area of static analysis works on "side-stepping undecidability":

- Unsafe approximation:

- For testing it may be okay to "abandon" safety and use unsafe approximations:


Here are some programs for you to (manually) consider !

## F "Slack"

- Undecidability means: "there'll always be a slack":

- However, still useful:
(possible interpretations of "Dunno?"):
- Treat as error (i.e., reject program):
- "Sorry, program not accepted!"
- Treat as warning (i.e., warn programmer):
. "Here are some potential problems: ..."


## - Soundness \& Completeness

## - Soundness:



- Analysis reports no errors $\Rightarrow$ Really are no errors
- Completeness:

- Analysis reports an error $\Rightarrow$ Really is an error
...or alternative (equivalent) formulation, via "contra-position":

$$
\mathbf{P} \Rightarrow \mathbf{Q} \equiv \neg \mathbf{Q} \Rightarrow \neg \mathbf{P}
$$

- Really are error(s) $\Rightarrow$ Analysis reports error(s)
- Really no error(s) $\Rightarrow$ Analysis reports no error(s)


## - Example: Type Checking

- Will this program have type error (when run)?
- void f() \{
var b;
if $\frac{(\langle E X P\rangle)}{b=42 ;}$
\} else \{ b = true;
\}
if (b) ...; // error is b is '42' \}

■ Undecidable (because of reduction):

- Type error $\Leftrightarrow$ <EXP> evaluates to true


## F Example: Type Checking

- Hence, languages use static requirements:

- All variables must be declared
- And have only one type (throughout the program)
- This is (very) easy to check (i.e., "type-checking")


## - Agenda

## - Introduction:

- Undecidability, Reduction, and Approximation
- Data-flow Analysis:
- Quick tour \& running example
- Control-Flow Graphs:
- Control-flow, data-flow, and confluence
- "Science-Fiction Math":
- Lattice theory, monotonicity, and fixed-points
- Putting it all together...:
- Example revisited



## 5' Crash Course on Data-Flow Analysis



Claus Brabrand ((( brabrand@itu.dk )))<br>Associate Professor, Ph.D.<br>((( Programming, Logic, and Semantics )))<br>로를 University of Copenhagen

## - Data-Flow Analysis

- IDEA: "Simulate runtime execution at compile-time using abstract values"
- We (only) need 3 things:
- A control-flow graph
- A lattice
- Transfer functions
- Example: "(integer) constant propagation"


## $\vdash$ Control-flow graph

We (only) need 3 things:

- A control-flow graph
- A lattice
- Transfer functions

Given program:



## - A Lattice

We (only) need 3 things:<br>- A control-flow graph<br>- A lattice<br>- Transfer functions

- Lattice L of abstract values of interest and their relationships (i.e. ordering " $\leq$ "):

- Induces least-upper-bound operator: ப
- for combining information



## - Agenda

## - Introduction:

- Undecidability, Reduction, and Approximation
- Data-flow Analysis:
- Quick tour \& running example
- Control-Flow Graphs:
- Control-flow, data-flow, and confluence
- "Science-Fiction Math":
- Lattice theory, monotonicity, and fixed-points
- Putting it all together...:
- Example revisited


## - Control Structures

- Control Structures:
- Statements (or Expr's) that affect "flow of control":
- if-else:
[syntax] if ( Exp ) Stm St $_{1}$ else Stm $_{2}$
[semantics] The expression must be of type boolean; if it evaluates to true, Statement-1 is executed,

- if:
[syntax] if ( Exp ) Stm
[semantics] The expression must be of type boolean; if it evaluates to true, the given statement is executed, otherwise not.


## $\vdash$ Control Structures (cont'd)

## - while:

[syntax] while (Exp ) Stm
[semantics] The expression must be of type boolean; if it evaluates to false, the given statement is skipped, otherwise it is executed and afterwards the expression is evaluated again. If it is still true, the statement is executed again. This is continued until the expression evaluates to false.

- for:




## - Control-flow graph

Given program:

```
int x = 1;
if (a>b) {
} else {
}
print(x,y);
```



## Exercise: Draw a Control-Flow Graph for:

```
public static void main ( String[] args ) {
    int mi, ma;
    if (args.length == 0)
    System.out.println("No numbers");
    else {
        mi = ma = Integer.parseInt(args[0]);
        for (int i=1; i < args.length; i++) {
        int obs = Integer.parseInt(args[i]);
        if (obs > ma)
            ma = obs;
        else
                        if (mi < obs) mi = obs;
        }
        System.out.println("min=" + mi + "," +
        "max=" + ma);
}}
```



## - Control Structures (cont'd²)

- do-while: exercise
- do Stm while (Exp );

■ "?:"; "conditional expression":

- Exp ${ }_{1}$ ? Exp : $_{2} \operatorname{Exp}_{3}$

■ "||"; "lazy disjunction" (aka., "short-cut ${ }^{\prime \prime}$ ):
$-\operatorname{Exp}_{1}| | \operatorname{Exp}_{2}$
■ "\&\&"; "lazy conjunction" (aka., "short-cut ^"):

- $\operatorname{Exp}_{1}$ \&\& $\operatorname{Exp}_{2}$
- switch:
switch (Exp ) \{ Swb* \}


## - Control Structures (cont'd ${ }^{3}$ )

- try-catch-finally (exceptions):
- try $\operatorname{stm}_{1}$ catch ( Exp ) Stm ${ }_{2}$ finally Stm $_{3}$
- return / break / continue:
"ret return ; Exp
- e.g.; $f(x)$

■ "recursive method invocation":

- e.g.; $f(x)$

■ "virtual dispatching":

- e.g.; $f(x)$


## $\vdash$ Control Structures (cont'd ${ }^{4}$ )

- "function pointers":
- e.g.; (*f)(x)
- "higher-order functions":
- e.g.; $\lambda f . \lambda x .(f x)$
- "dynamic evaluation":
- e.g.; eval(some-string-which-has-been-dynamically-computed)
- Some constructions (and thus languages) require a separate control-flow analysis for determining control-flow in order to do data-flow analysis


## - Agenda

- Introduction:
- Undecidability, Reduction, and Approximation
- Data-flow Analysis:
- Quick tour \& running example
- Control-Flow Graphs:
- Control-flow, data-flow, and confluence
- "Science-Fiction Math":
- Lattice theory, monotonicity, and fixed-points
- Putting it all together...:
- Example revisited


## MATH

## - Agenda

- Relations:
- Crossproducts, powersets, and relations
- Lattices:
- Partial-Orders, least-upper-bound, and lattices
- Monotone Functions:
- Monotone Functions and Transfer Functions
- Fixed Points:
- Fixed Points and Solving Recursive Equations
- Putting it all together...:
- Example revisited


## - Crossproduct: 'x'

- Crossproduct (binary operator on sets):
- Given sets:
- $\boldsymbol{A}=\{0,1\}$
- $B=\{$ true, false $\}$
- $\boldsymbol{A} \times \boldsymbol{B}=\{(0$, true $),(0$, false $),(1$, true $),(1$, false $)\}$
- i.e., creates sets of pairs
- Exercise:

$$
\begin{aligned}
& ■ \boldsymbol{A} \times \boldsymbol{A}=\{(0,0),(0,1),(1,0),(1,1)\} \\
& ■ \boldsymbol{Z} \times \boldsymbol{Z}=\{(0,0),(0,1),(0,1), \ldots,(1,0),(1,1), \ldots,(42,87), \ldots\} \\
& ■(\boldsymbol{A} \times \boldsymbol{A}) \times \boldsymbol{B}=\{((0,0), \text { true }),((0,1), \text { true }), \ldots,((1,1), \text { false })\}
\end{aligned}
$$

## ト Powersets : ' $\mathcal{P}(\mathbf{S})$ '

- Powerset (unary operator on sets):
- Given set "S = \{ A, B \}";
$-\mathscr{P}(\boldsymbol{S})=\{\boldsymbol{\emptyset},\{\mathrm{A}\},\{\mathrm{B}\},\{\mathrm{A}, \mathrm{B}\}=\mathrm{S}\}$
- i.e., creates the set of all subsets (of the set)
- Note: $\mathrm{X} \subseteq \boldsymbol{S} \quad \Leftrightarrow \quad \mathrm{X} \in \mathscr{P}(\boldsymbol{S})$
- Exercise:
$\square \mathscr{P}(\boldsymbol{Z})=\{\varnothing,\{0\},\{1\},\{2\}, \ldots,\{0,1\}, \ldots\{13,42,87\}, \ldots \boldsymbol{Z}\}$
$\bullet \mathscr{P}(\boldsymbol{Z} \times \boldsymbol{Z})=\{\boldsymbol{\varnothing},\{(0,0)\},\{(1,1)\}, \ldots,\{(0,0),(3,2),(4,9)\}, \ldots \boldsymbol{Z} \times \boldsymbol{Z}\}$
- If a set $\boldsymbol{S}$ has $|\boldsymbol{S}|$ elements;
- How many elements does $\mathscr{P}(\mathbf{S})$ have? Answer: $\mathbf{2}^{|\mathbf{S}|}$ ' $\mathscr{P}(S)$ ' is (therefore) often written ' $2 S$,


## - Relations

- Example: "equals" relation:

■ Signature: ${ }^{\prime}=’ \subseteq \mathbf{Z} \times \mathbf{Z} \quad$...same as saying: $\quad=' \in \mathscr{P}(\mathbf{Z} \times \mathbf{Z})$

- Relation is: equals $=\{(0,0),(1,1),(2,2),(3,3),(4,4), \ldots\}$
- Written as: $2=2$ as a short-hand for: $(2,2) \in{ }^{\prime}=\prime$
- ... and as: $2 \neq 3$ as a short-hand for: $(2,3) \notin{ }^{\prime}=$
- Example: "less-than" relation:

■ Signature: $<^{\prime} \subseteq \mathbf{Z} \times \mathbf{Z} \quad$...same as saying: $\quad{ }^{\prime} \in \mathcal{P}(\mathbf{Z} \times \mathbf{Z})$
■ Relation is: less-than $=\{(0,1),(0,2),(0,3), \ldots,(1,2),(1,3), \ldots\}$

- Written as: $7<8$ as a short-hand for: $\quad(7,8) \in$ ' $\quad$ '

■ $\ldots$ and as: $8 \nless 7$ as a short-hand for: $(8,7) \notin$ '<’

## $\vdash$ Exercises

- Example: "less-than-or-equal-to" relation:

■ Signature: $\quad \leq \leq \subseteq \mathbf{Z} \times \mathbf{Z} \quad$...same as saying: $\quad \leq \prime \in \mathscr{P}(\mathbf{Z} \times \mathbf{Z})$
■ Relation is: ‘$\leq=\{(0,0),(0,1),(0,2), \ldots,(1,1),(1,1), \ldots,(2,3), \ldots\}$

- Written as: $2 \leq 3$ as a short-hand for: $(2,3) \in ‘ \leq$

■ $\ldots$ and as: $3 \nless 2$ as a short-hand for: $(3,2) \notin \leq^{\prime}$

- Example: "is-congruent-modulo-3" relation:

■ Signature: ${ }^{\prime} \equiv_{3}{ }^{\prime} \subseteq \mathbf{Z} \times \mathbf{Z} \quad$...same as saying: ${ }^{\prime} \equiv_{3}{ }^{\prime} \in \mathscr{P}(\mathbf{Z} \times \mathbf{Z})$

- Relation is: ${ }^{\prime} \equiv_{3}^{\prime}=\{(0,0),(0,3),(0,6), \ldots,(1,1),(1,4), \ldots,(6,9), \ldots\}$
- Written as: $6 \equiv_{3} 9$ as a short-hand for: $(6,9) \in{ }^{\prime} \equiv_{3}^{\prime}$
- ... and as: $7 \#_{3} 8$ as a short-hand for: $(7,8) \notin{ }^{\prime} \equiv_{3}^{\prime}$


## - Equivalence Relation

- Let ' $\sim$ ' be a binary relation over set $A$ :
- '~' $\subseteq A \times A$
- ~ is an equivalence relation ifs:
- Reflexive:
- $\forall x \in A: \quad x \sim x$
- Symmetric:

$$
\forall x, y \in A: \quad x \sim y \Leftrightarrow y \sim x
$$

- Transitive:

$$
\forall x, y, z \in A: \quad x \sim y \wedge y \sim z \Rightarrow x \sim z
$$

## - Agenda

## - Relations:

- Crossproducts, powersets, and relations
- Lattices:
- Partial-Orders, least-upper-bound, and lattices
- Monotone Functions:
- Monotone Functions and Transfer Functions
- Fixed Points:
- Fixed Points and Solving Recursive Equations
- Putting it all together...:
- Example revisited


## - Partial-Order

- A Partial-Order is a structure $(S, \sqsubseteq)$ :
- $S$ is a set
- '드' is a binary relation on $S$ (i.e., ' $\underline{\underline{-}} \subseteq \mathbf{S} \times \mathbf{S}$ ) satisfying:
- Reflexivity:
- $\forall x \in \mathbf{S}: \quad x \sqsubseteq x$
- Transitivity:
$\forall x, y, z \in \boldsymbol{S}: \quad x \sqsubseteq y \wedge y \sqsubseteq z \Rightarrow x \sqsubseteq z$
- Anti-Symmetry:
- $\forall x, y \in \boldsymbol{S}: \quad x \sqsubseteq y \wedge y \sqsubseteq x \Rightarrow x=y$


## - Visualization: Hasse Diagram

Partial-Order (S, ㄷ): $\Leftrightarrow$ Hasse Diagram:

- Reflexive:

$$
\forall x \in \mathbf{S}: \quad x \sqsubseteq x
$$

- Transitive:
$\forall x, y, z \in \boldsymbol{S}: \quad x \sqsubseteq y \wedge y \sqsubseteq z \Rightarrow x \sqsubseteq z$
- Anti-Symmetric:

$$
\forall x, y \in S: \quad x \sqsubseteq y \quad y \sqsubseteq x \Rightarrow x=y
$$



$$
S=\{1, \div, 0,+, T\}\left[\begin{array}{r}
\Gamma^{\prime}=\begin{array}{r}
\{(1,1),(1, \div),(1,0),(1, T), \\
\\
(1, T),(\div, \div),(\div, T),(0,0), \\
(0, T),(1,+),(t, T),(T, T)\}
\end{array}
\end{array}\right.
$$

## - Exercise (Hasse Diagram)

- Given Hasse Diagram:

- Write down partial order ( $\boldsymbol{B}, \sqsubseteq$ ):
- Set $\boldsymbol{B}=\{\ldots\}$
- Relation '巨':
- Signature
- All elements of the relation (i.e., ${ }^{\prime}{ }^{\prime}=\{\ldots\}$ )
- Give example of element in ' $\sqsubset$ ' (w/ + w/o shorthand)
- Again, but for an element not in the relation


## $\vdash$ Example Partial-Orders

- Lattice Examples (as Hasse Diagrams):



- ...depending on what is analysed for!


## Least Upper Bound 'ப'

## - "Least Upper Bound"

- Upper bound:
- We say that ' $z$ ' is an upper bound for set ' $X$ '
- ...written $\mathrm{X} \mathrm{\sqsubseteq z}$ if $\forall x \in X: x \sqsubseteq z$
- Least upper bound:
- We say that ' $z$ ' is the least upper bound of set ' $X$ '

■ ...written |  |
| :---: |
| $z=\sqcup X$ |
| if $\underbrace{X \sqsubseteq z \wedge \forall z^{\prime}: X \sqsubseteq z^{\prime} \Rightarrow z \sqsubseteq z^{\prime}}_{\text {upper bound }}$ |

## - Example: Least upper bound

- Analyses use ' $\sqcup$ ' to combine information (at confluence points):


