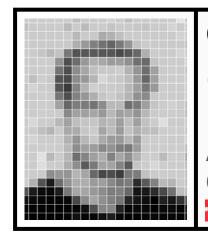


[HOW TO ANALYZE LANGUAGES AUTOMATICALLY]



Claus Brabrand

(((brabrand@itu.dk)))

Associate Professor, Ph.D. (((Programming, Logic, and Semantics))) IT University of Copenhagen

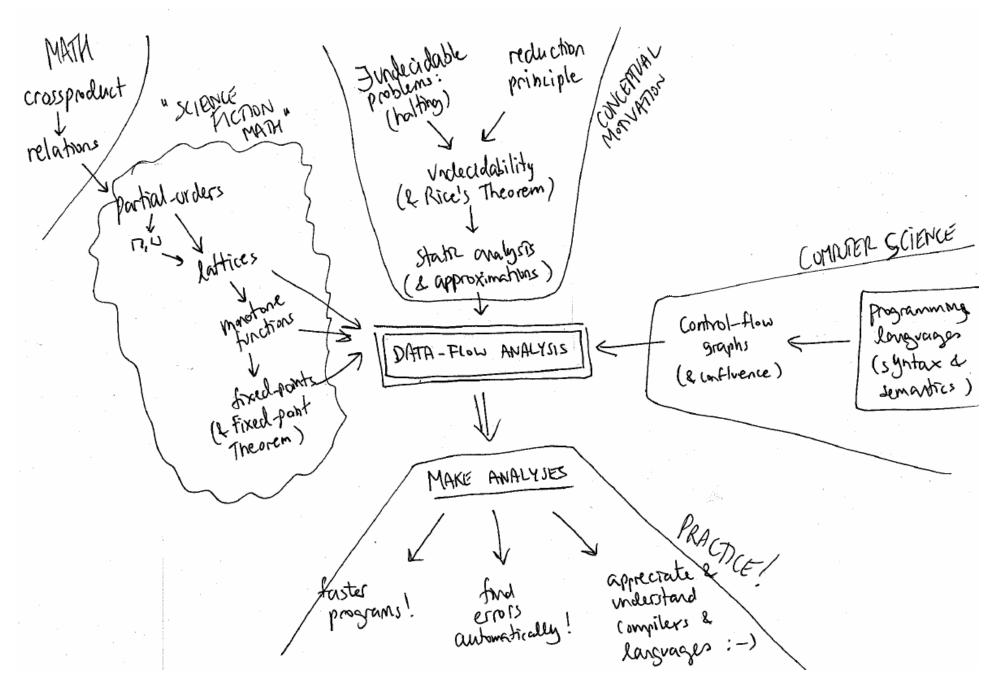
- Abstract

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.



H Agenda

Introduction:

Undecidability, Reduction, and Approximation

Data-flow Analysis:

- Quick tour of everything & running example
- Control-Flow Graphs:

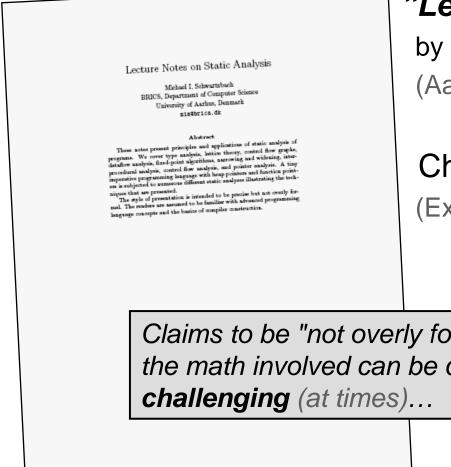
Control-flow, data-flow, and confluence

"Science-Fiction Math": (next monday)

Lattice theory, monotonicity, and fixed-points



Notes on Static Analysis



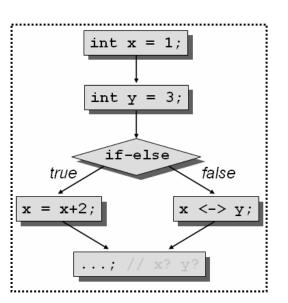
"Lecture Notes on Static Analysis"

by Michael I. Schwartzbach (Aarhus University)

Chapter 1, 2, 4, 5, 6 (until p. 19)

(Excl. "pointers")

Claims to be "not overly formal", but the math involved can be quite



Quiz: Optimization?

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



• 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

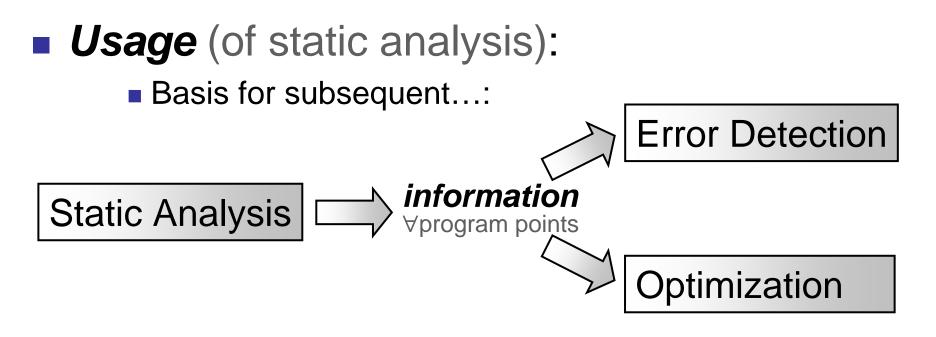
Π	#2 (ptr)	no opt	19.5	17.6	7.55
Ш	(1)				

- 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!

Flow Analysis

- **Purpose** (of Data-Flow Analysis):
 - Gather information (on running behavior of program)

■ "∀program points"



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"

-- Rice's Theorem [paraphrased] (1953)

*) written in a turing-complete language

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)
 - Finitely many books; i.e.:

- "The Bible" "War and Peace" "Programming Languages, An Interp.-Based Approach" ... The Book-of-all-Books
- 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. "<u>This</u> sentence is false"

 $(\mathbf{:})$

Formination Undecidable!

Assume termination is decidable (in Java);

• i.e. \exists some program, *halts*: Program \rightarrow bool

bool halts(Program p) { ... }

 $-- P_0.java --$

Program p₀ = read_program("P₀.java"); <u>if (halts(p₀)) loop();</u> else halt();

Q: Does P₀ loop or terminate...?

Hence: halts cannot exist!

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

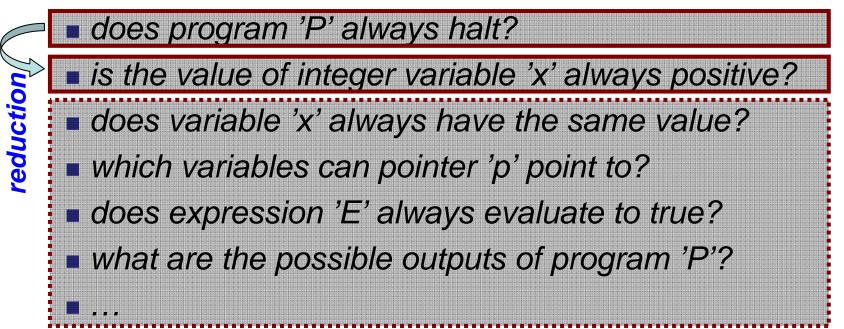
- Rice's Theorem (1953)

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

-- Rice's Theorem [paraphrased] (1953)

*) written in a turing-complete language

• Examples:



⊢ Reduction: solve *always-pos* ⇒ solve *halts*

- 1) Assume '*x-is-always-pos(P)*' is decidable
- 2) Given P (here's how we could solve 'halts(P)'):
- 3) Construct (veeeery clever) reduction program *R*:

```
-- R.java --

int x = 1;

P /* insert program P here :-) */

x = -1;
```

4) Run "supposedly decidable" analysis:

res = x-is-always-positive(R)

5) Deduce from result: if (res) then P loops!; else P halts :-)
6) <u>THUS</u>: 'x-is-always-pos(P)' must be undecidable!

Reduction Principle

Reduction principle (in short):

 $\frac{\phi(P) \text{ undecidable } \land [\text{solve } \psi(P) \Rightarrow \text{ solve } \phi(P)]}{\psi(P) \text{ undecidable}}$

Example:

reduction

'halts(P)' **undecidable** \land [**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 --
```

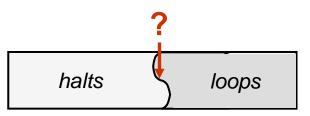
```
ptr q = 0xffff;
P /* insert program P (assume w/o 'q') */
q = null;
```

- 4) Run 'which-var-q-points-to(R)' = res
- 5) If (null \in res) *P* halts! else; *P* loops! :-)
- 6) <u>THUS</u>:

'which-var-q-points-to(P)' must be undecidable!

- Undecidability

Undecidability means that...:



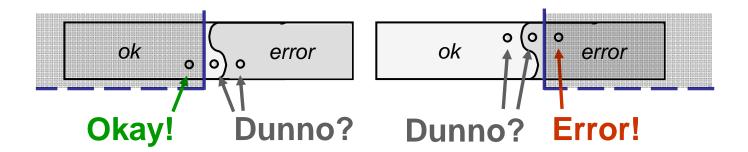
...no-one can decide this line (for all programs)!

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"*:

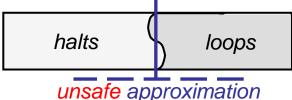
 Compilers use safe approximations (computed via "static analyses") such that:



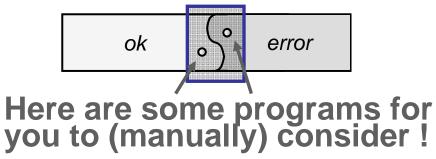
- "Side-Stepping Undecidability"

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:



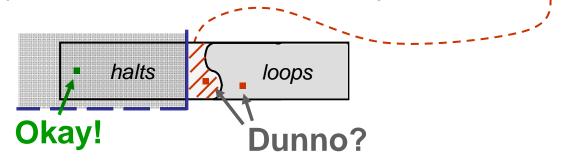
For testing it may be okay to "abandon" safety and use unsafe approximations:



DATA-FLOW ANALYSIS

- "Slack"

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



However, still useful:

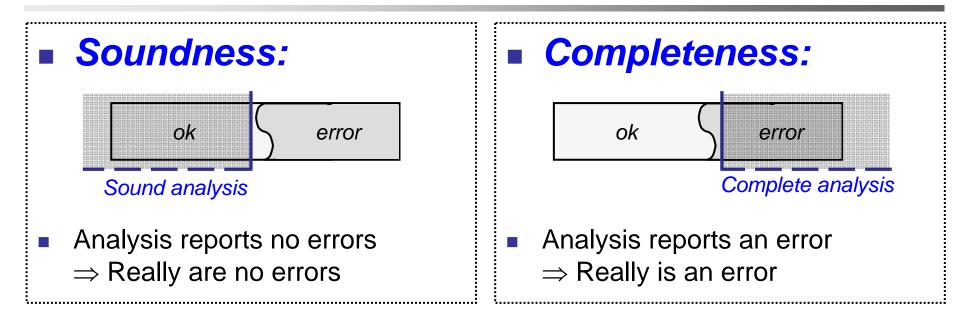
(possible interpretations of "Dunno?"):

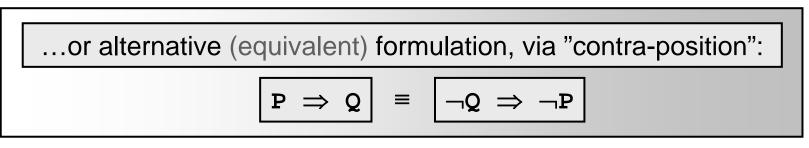
Treat as **error** (i.e., *reject program*):

- "Sorry, program not accepted!"
- Treat as warning (i.e., warn programmer):

• "Here are some **potential** problems: ..."

- Soundness & Completeness



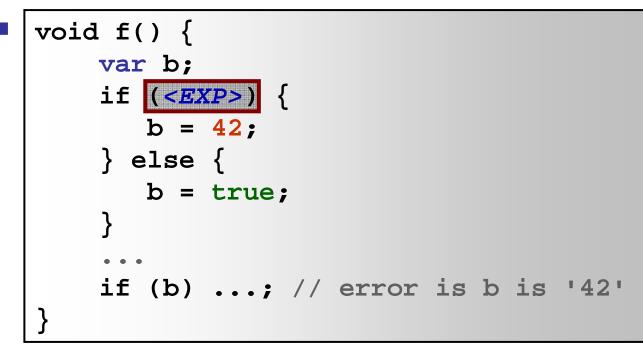


Really are error(s)
 Analysis reports error(s)

Really no error(s)
 Analysis reports no error(s)

Example: Type Checking

Will this program have type error (when run)?

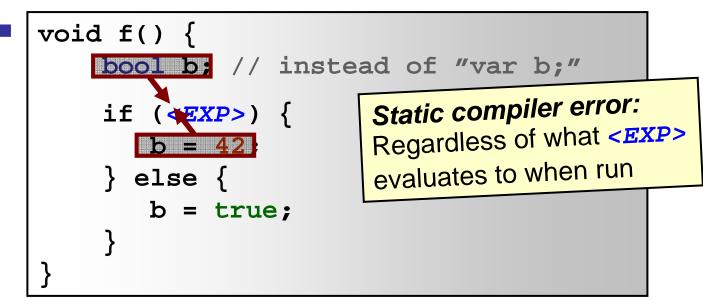


Undecidable (because of reduction):

■ Type error ⇔ <**EXP**> evaluates to true

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")

H Agenda

Introduction:

Undecidability, Reduction, and Approximation

Data-flow Analysis:

Quick tour & running example

Control-Flow Graphs:

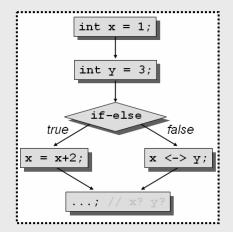
Control-flow, data-flow, and confluence

"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)))

Associate Professor, Ph.D. (((Programming, Logic, and Semantics))) IT University of Copenhagen

Claus Brabrand, UFPE, Brazil

DATA-FLOW ANALYSIS

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

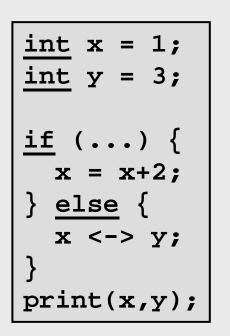
Example: "(integer) constant propagation"

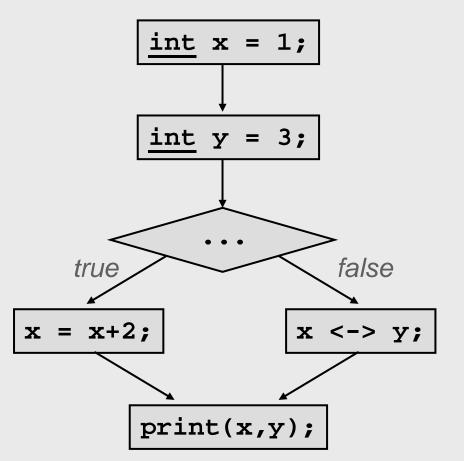
DATA-FLOW ANALYSIS



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

Given program:





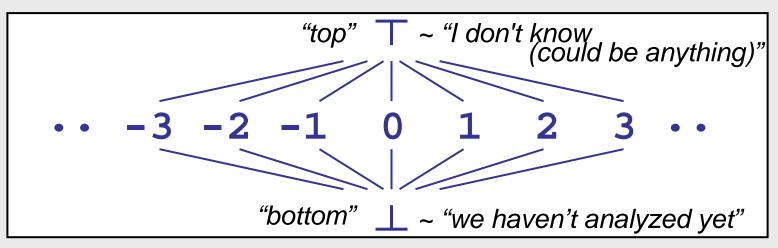
A Lattice

We (only) need 3 things: A *control-flow graph*

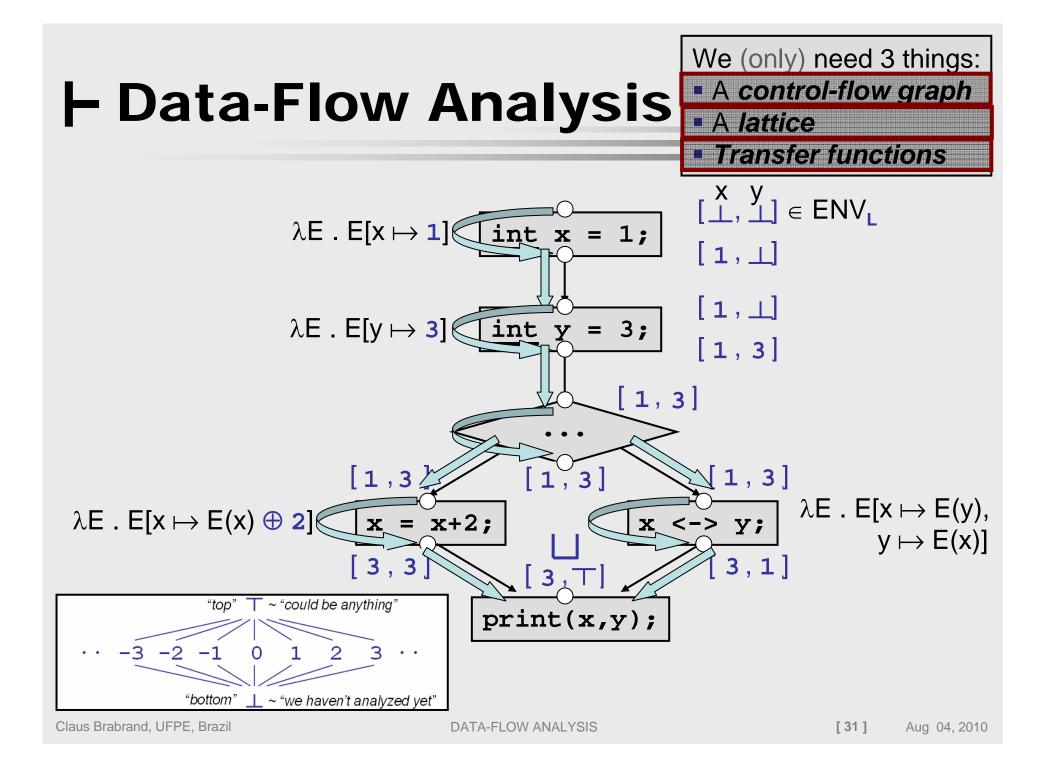
A lattice

Transfer functions

■ Lattice L of abstract values of interest and their relationships (i.e. ordering "≤"):



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



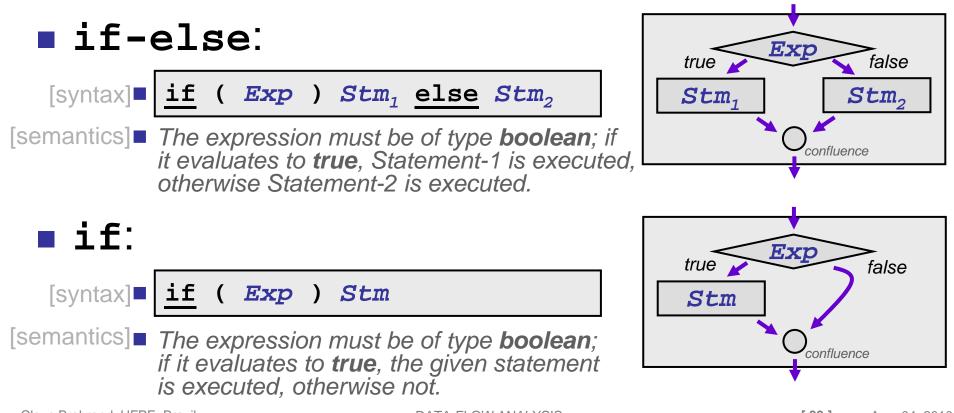
H Agenda

- 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

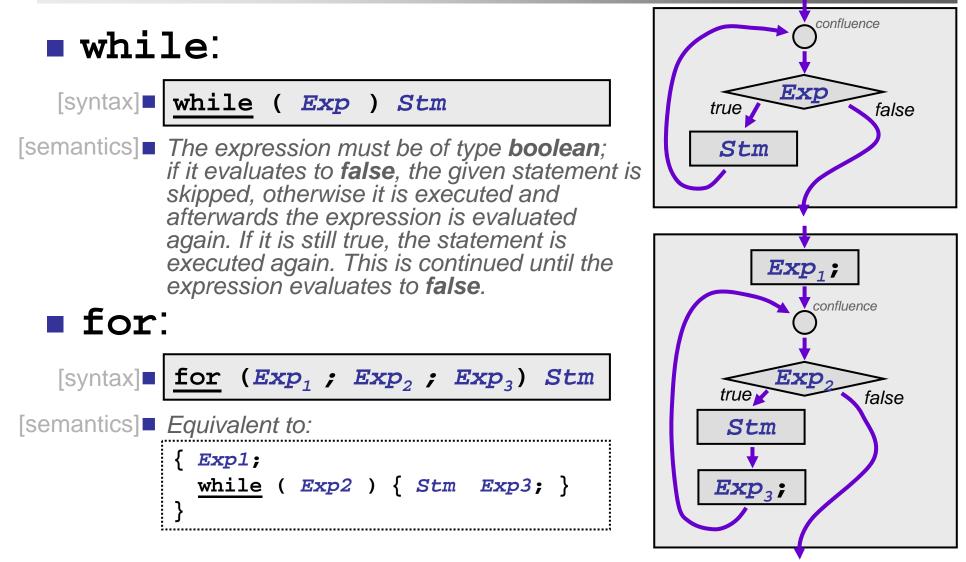
- Control Structures

Control Structures:

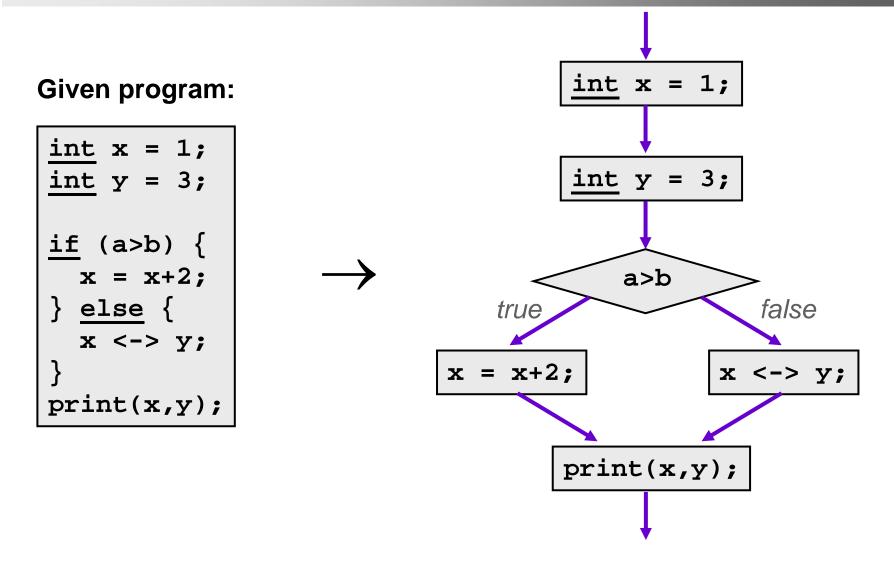
Statements (or Expr's) that affect "flow of control":



- Control Structures (cont'd)



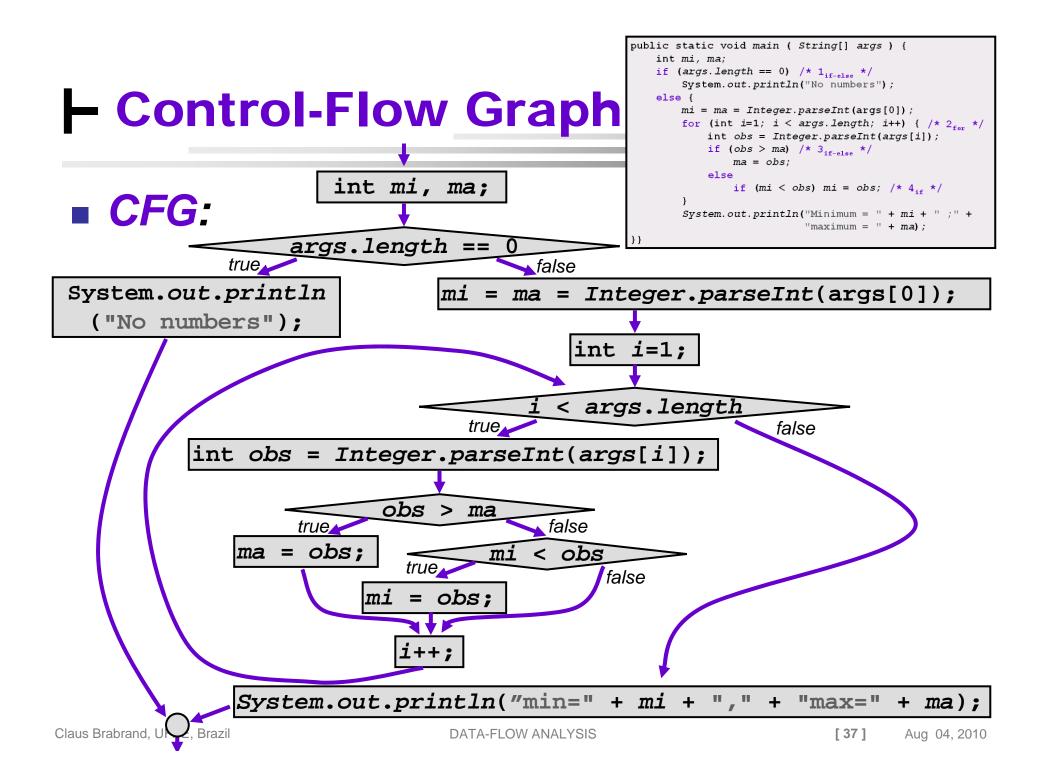
Control-flow graph



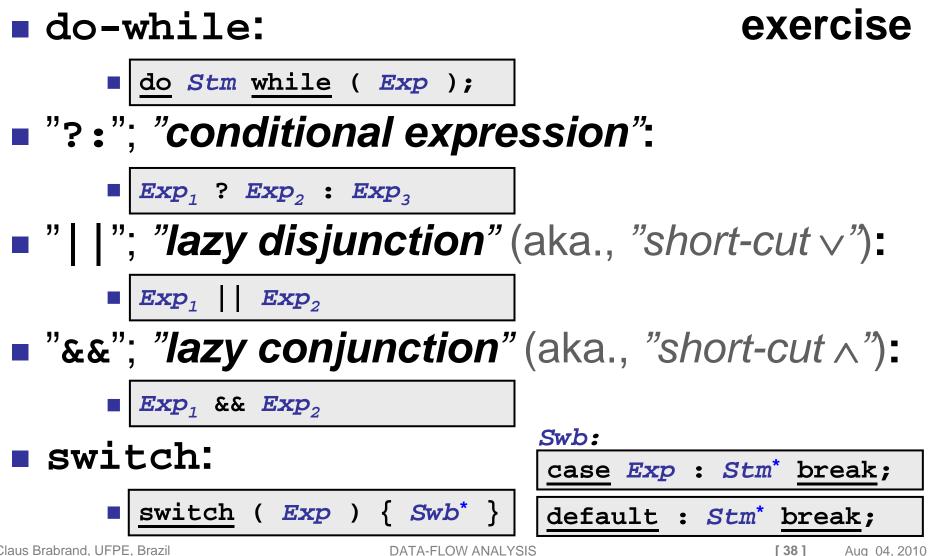
1

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++) {</pre>
            int obs = Integer.parseInt(args[i]);
             if (obs > ma)
                 ma = obs;
            else
                 if (mi < obs) mi = obs;
        System.out.println("min=" + mi + "," +
                             \max = \max + \max
} }
```



– Control Structures (cont'd²)



Claus Brabrand, UFPE, Brazil

DATA-FLOW ANALYSIS

- Control Structures (cont'd³)

try-catch-finally (exceptions):

 $\underline{try} Stm_1 \underline{catch}$ (\underline{Exp}) $Stm_2 \underline{finally} Stm_3$

return / break / continue:

return ; | return Exp ;

break ;

continue ;

"method invocation":

■ e.g.; **f(x)**

"recursive method invocation":

■ e.g.; **f(x)**

"virtual dispatching":

■ e.g.; f(x)

- Control Structures (cont'd⁴)

"function pointers":

■ e.g.; (*f)(x)

"higher-order functions":

■ e.g.; *λf.λx.(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

÷.

H 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

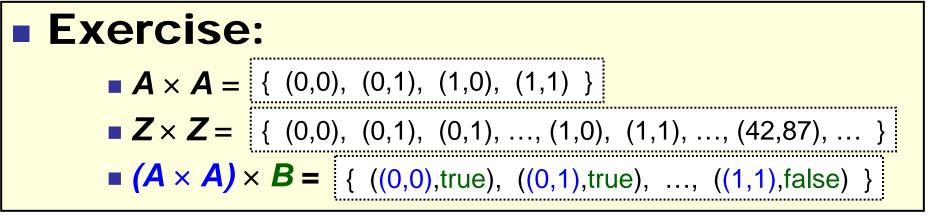
Crossproduct: 'X'

Crossproduct (binary operator on sets):

Given sets:

■ **B** = { true, false }

i.e., creates sets of pairs



Howersets : $'\mathcal{P}(S)'$

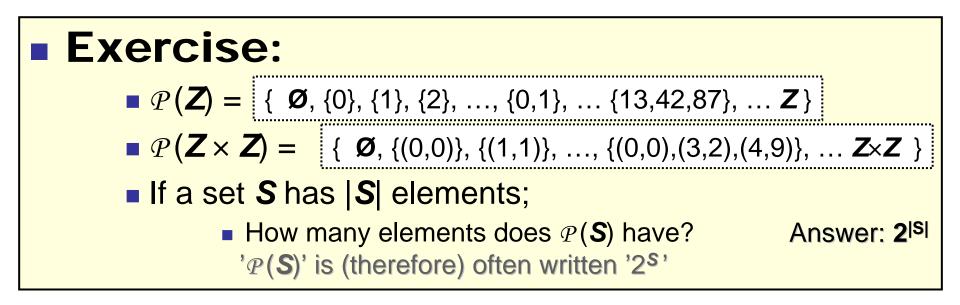
Powerset (unary operator on sets):

■ Given set "**S** = { A, B }";

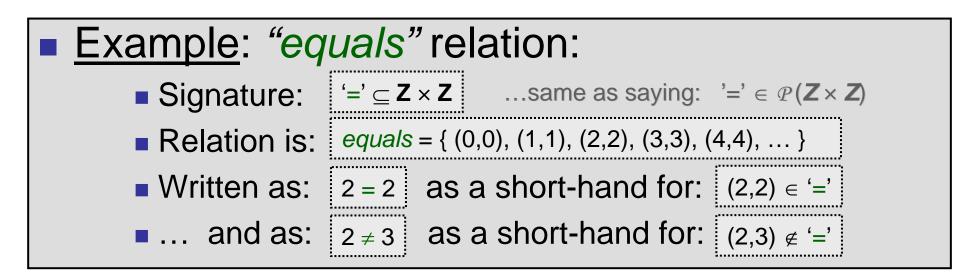
$$P(S) = \{ \emptyset, \{A\}, \{B\}, \{A,B\} = S \}$$

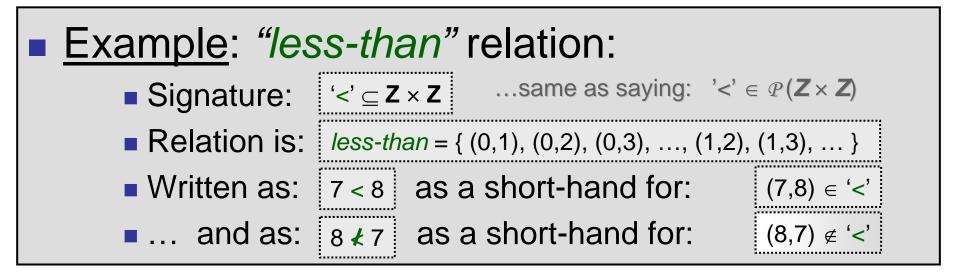
• i.e., creates the set of all subsets (of the set)

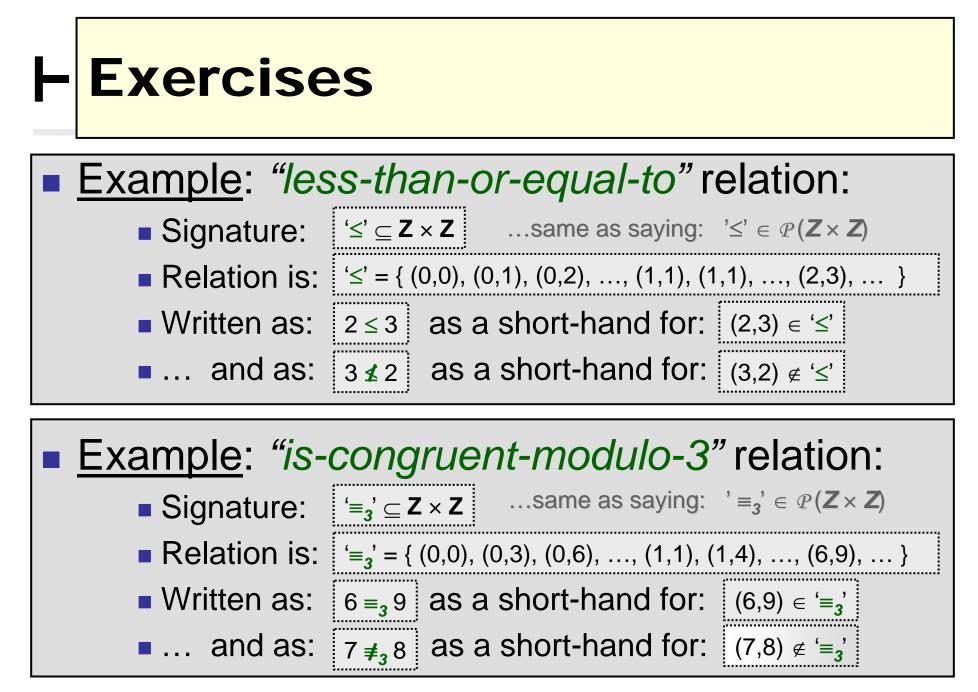
• Note:
$$X \subseteq S \iff X \in \mathcal{P}(S)$$



- Relations

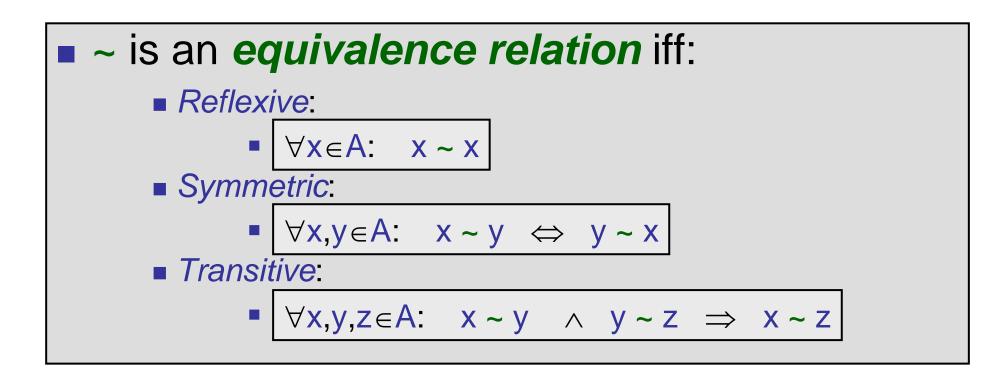






Equivalence Relation

Let '~' be a *binary relation* over set A: '~' ⊆ A × A



H 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...:

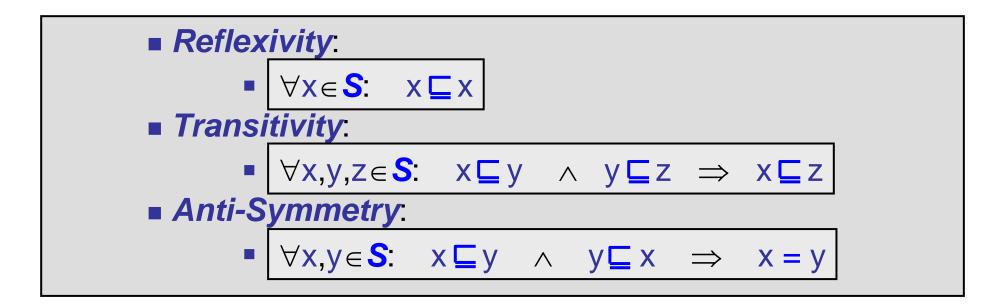
Example revisited

- Partial-Order

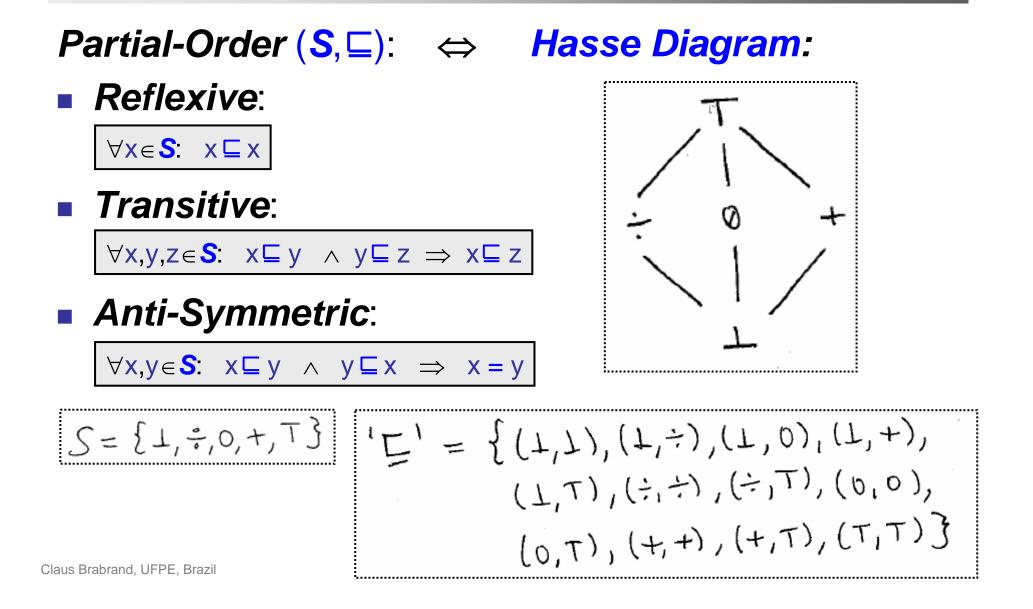
■ A *Partial-Order* is a structure (S, ⊆):

S is a set

• ' \sqsubseteq ' is a *binary relation* on **S** (i.e., ' \sqsubseteq ' \subseteq **S** × **S**) satisfying:

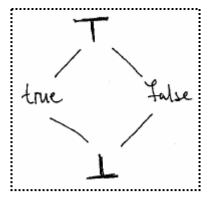


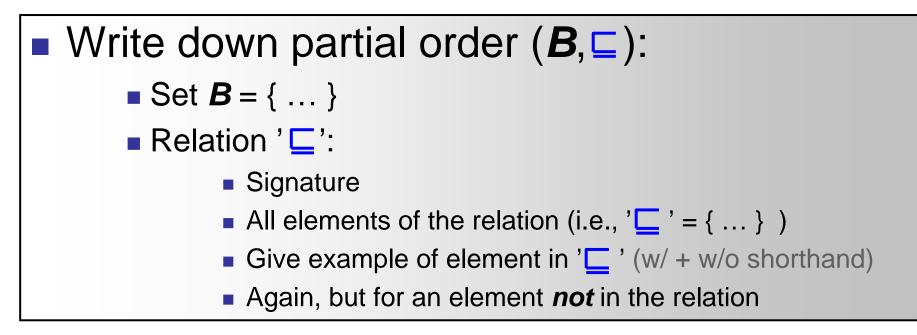
For the second secon

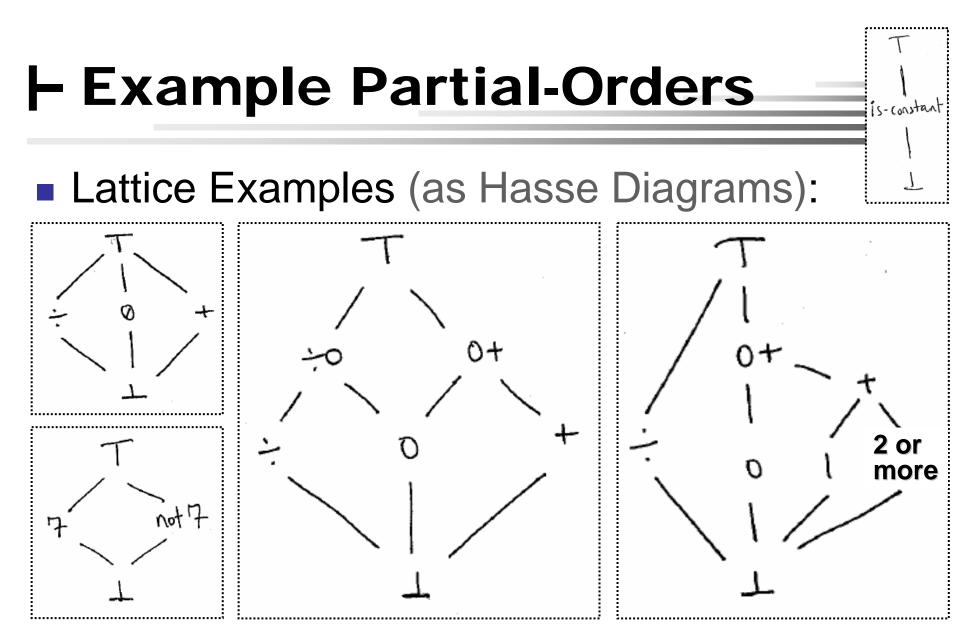


Exercise (Hasse Diagram)

Given Hasse Diagram:







...depending on what is analysed for!

Claus Brabrand, UFPE, Brazil

DATA-FLOW ANALYSIS

Least Upper Bound '

2

"Least Upper Bound"

Upper bound:

We say that 'z' is an upper bound for set 'X'

• ...written $|X \subseteq z|$ if $\forall x \in X$: $x \subseteq z$

Least upper bound:

We say that 'z' is the least upper bound of set 'X' $z = \mathbf{U}X$ if $X \sqsubseteq z \land \forall z': X \sqsubseteq z' \Rightarrow z \sqsubseteq z'$written upper bound least

Example: Least upper bound

Analyses use 'L' to combine information (at confluence points):

