

Course Script

Static analysis and all that

IN5440 / autumn 2024

Martin Steffen

Contents

1 Introduction **1**
1.1 Data-flow analysis . . . 1
1.1.1 A simplistic while-language . . . 3
1.1.2 Example analysis: Reaching definitions . . . 9
1.1.3 Equational approach . . . 13
1.1.4 Constraint-based approach . . . 20
1.2 Control-flow analysis (constraint-based) . . . 22
1.3 Type and effect systems . . . 29
1.3.1 Introduction . . . 29
1.3.2 Annotated type systems . . . 35
1.3.3 Annotated type constructors . . . 38
1.3.4 Effect systems . . . 39
1.4 Algorithms . . . 44
1.5 Conclusion . . . 46

Chapter 1 Introduction

What is it about?

Learning Targets of this Chapter

Apart from a motivational introduction, the chapter gives a high-level overview over larger topics covered in the lecture. They are treated here just as a teaser and in less depth compared to later but there is already technical content.

Contents

1.1 Data-flow analysis 1
1.1.1 A simplistic while-language 3
1.1.2 Example analysis: Reaching definitions . 9
1.1.3 Equational approach . 13
1.1.4 Constraint-based approach 20
1.2 Control-flow analysis (constraint-based) 22
1.3 Type and effect systems . . . 29
1.3.1 Introduction 29
1.3.2 Annotated type systems 35
1.3.3 Annotated type constructors 38
1.3.4 Effect systems 39
1.4 Algorithms 44
1.5 Conclusion 46

1.1 Data-flow analysis

Data-flow analysis is one well-established quite standard, general form of program analysis. It is routinely used for many purposes in compilers. It can take different forms (and can be done for different purposes), but all those are based on common notions and principles. We will look at those principles, here first more cursory, later in more depth.

It is a static analysis technique par excellence, in many aspects. It's widely used and widely applicable with different goals, and it's based on well-understood principles. It's pretty ancient, as well. Indeed, a form of data-flow analysis was done in one of the very first compilers overall, for Fortran. The Fortran compiler was not the absolutely first compiler, but it has been dubbed the first complete and commercially available compiler. Actually, it was not even called a compiler in the paper **?**] from **?** , giving a short overview over the status of the FORTRAN project and the resulting compiler. In the

title, it's called "automatic coding system" and the text talks about the *translator*, the word "compiler" does not show up, though it's mentioned that the Fortran translator *compiles* Fortran programs to IBM704 programs. Anyway, the document speaks about **flow analysis** (with the purpose of register allocation). This is done by "section 4 and 5" of that compiler (the phases of the compiler were called "sections"). The compiler also worked with the notion of **basic blocks** and **control-flow graphs** (though the latter word is not used). We will cover those concepts in this lecture, as well.

Another point why one should know about data-flow analysis is that it illustrates in a straightforward setting a number of general concepts or mind-sets underlying basically all static analyses. In particular a clear notion of working on **approximations**, central to basically all static analyses. Technically one can understand data-flow analysis to work with **constraints**, another central concept. For basic data-flow analysis, the involved constraints are likewise "basic". They are known as data-flow constraints and data-flow analysis is one simple example of **constraint solving**. If one sticks to basic settings (as we do for the beginning, especially in this warm-up section), data-flow analysis is typically manageable, complexity-wise. And, as mentioned, it was always around in all compilers, ever since basically the first one for Fortran.

1.1.0.1 About abstract syntax

This warm-up section will also give a first glimpse of what we will repeatedly see in the following. If one reads other literature, like publications explaining new forms of analyses, the way things are presented there often also resemble the presentation here (basically taken from [2]). Static analysis normally does not work directly on program text (in the form of a string of characters in concrete syntax), but on more reasonable, structured representations of the program. One such intermediate representation is known as the **abstract syntax** of a programming language, and one particular program is represented by one concrete element of the abstract syntax, which is a tree, namely the **abstract syntax tree** of the given program. In a concrete implementation of a compiler, it's an important data structure as output from the parser, perhaps implemented using inductive data types (when working with a functional language), or with interfaces and sub-classes in a language like Java (and there are other ways how one can implement such trees).

How will we talk about abstract syntax? We (as everybody else) will use something called **context-free grammars** for that, more specifically, a *notation* for that kind of grammars known as BNF resp. EBNF ((extended) Backus-Naur form). Also this concept originated from the Fortran language, at least their use for programming languages (the formal notion of grammars originated in linguistics). The Fortran folks did indeed some truly pioneering stuff. We will see BNFs describing abstract syntax variously in this lecture, starting in the next Section 1.1.1 with the abstract syntax of a small while-language.

Now, if one happens to know a bit about the parsing phase of a compiler, one knows, that parsing is obsessed with context-free grammars: What kind of grammar one can use for what kind of parsing machinery, how much look-ahead is supported, is the grammar left-recursive, ambiguous, normalized, and what not. We, in contrast, are not obsessed with that and we are not doing parsing, which is done for (a token stream of) *concrete syntax*. So we don't care about those issues. We use grammars in BNF is to describe

the structure of trees, i.e., a means to concisely *communicate* the (abstract) syntax of the language at hand (for instance, the "while-language" coming up next). Therefore, we will not formally introduce grammars, BNFs etc, we just present the notation relying on that it's not only precise (which it is) but also clear enough. Or at least becomes clear enough after a while, when after seeing different examples throughout the course, one just learns by "getting used to it".

Also typical for many textbook or similar presentations is, that one often chooses to work with a "simplified" version of the abstract syntax. The simplification refers to different aspects. One is that the abstract syntax is presented mathematically and compactly, omitting things that are included in abstract syntax tree data structures in concrete implementation (which may include line numbers, types and what not). There is another, more relevant way the abstract syntax is "simplified". We *focus* on aspects we are interested in. For example, the while-language is unrealistically simple for modern languages. There is no scoping, no procedures, no pointers, etc. That's partly pedagogical: one learns the basics one some cut-down focused core language, knowing that in a real language, there are other things to be dealt with as well. Actually, later in the lecture we may deal with some of those (like adding procedures, pointers, etc., thereby extending the abstract syntax more and more). However, still we won't cover the abstract syntax of a full realistic language.

But a simplified and lean abstract syntax is not only "pedagogical". Properly designed, it should consist of a carefully chosen selection of core concepts or features, each of which might only be represented once syntactically. Note that we are talking about the abstract syntax of the core language itself. On top of that language, many fancy things can be supported in libraries, which may indeed be massive. Also, in case the designer chooses to support different syntactic ways to express similar things in the language itself, there is often no need to clutter the *abstract* syntax with this. For illustration, if we had a simple imperative language (like the while language), the feature is looping and that loops can be nested. For the user, i.e., in concrete syntax, the language may support different ways of expressing that concept, like while-loops and repeat-until-loops and other variants. However, in the abstract syntax, it may be a good idea that just one form of loop is represented (let's say corresponding to while-loops). "Syntactic" variations of that sort, that are "compiled away" early on, before the real work starts, are also know as **syntactic sugar ?**] and the process of getting rid of those and working with a "simplified" or cut-down version of the user syntax is called *desugaring*.

A last word about abstract syntax. We talked about it in a way that seem to indicate that abstract syntax captures the syntactical essence of the language the user sees (in the form of concrete syntax). Fair enough, but as stressed, BNF is a way to describe trees ("syntax" tree). Therefore, abstract syntax and corresponding abstract syntax trees may well apply to the syntax of various forms of intermediate code, even if the intermediate code is internal and without "concrete syntax" attached to it.

1.1.1 A simplistic while-language

Abstract syntax The given while-language here (and other languages later) is rather simplistic. Nonetheless, if we don't assume an upper bound on the memory, it is Turing

complete.

This fact of being simple is, however, not the reason why we call the syntax here *abstract* (not like "it's a very abstract language, it has only 5 constructs").

This program "written" in abstract syntax are thought of as **trees**. Since trees are notationally not easy to write down, we don't depict them as actual trees (even if that would be accurate). Instead, we use textual notations and to appeal to the understanding in which way that represents trees, and allow ourselves grouping constructs like parentheses (..) to disambiguate the tree structure, even if of course parentheses and similar constructs are general not part of abstract syntax trees; they are just needed to help the human reader here to understand the underlying tree structure.

Furthermore, and also as a general remark: the while language here (even in its restricted syntactic capabilities) resembles a "high-level" language, at least insofar that it supports "structured programming" in the form of while-loops, as opposed to conditional jumps. As mentioned, later we might extend it also with further capabilities and programming abstractions, for instance, procedure calls. Anyway, techniques similar to the ones we cover, can also be applied to lower-level intermediate languages, not just abstract syntax of the source language. Also in that case, in a more formal account, one might start fixing the abstract syntax of the intermediate language for the programs one intends to analyze.

Types As mentioned, working directly with abstract syntax, we very much ignore syntactical questions. Besides that, and in particular for the while-language, we more or less ignore another issue, that of typing. Typing and type checking is one very important form of static analysis and will be covered in the course, just not really in connection with the while-language and its derivatives. The "type system" here is so impoverished, supporting basically only integers and booleans, that it's not much worth introducing a type system for doing that. Instead, sidestepping the question, we deal with it "syntactically", i.e., *within* the grammar. *Arithmetic* expressions are represented by the non-terminal a, boolean expressions by b, and, as it happens, we have only variables of arithmetic type. In that sense the language *is* typed, only not very interestingly. Since we have only two types (and no polymorphism or other complications), we can bake the very rudimentary type directly into the grammar. For functional languages, resp. calculi (variations of the λ-calculus), we will take care of type checking and extensions of traditional type checking, as it becomes more challenging.

1.1.1.1 Abstract syntax and labeled abstract syntax

There will be *two versions* of the abstract syntax. The first one can be seen as *the* abstract syntax of the while language. The other one will be an augmented version, decorated with additional information. But let's start with the standard version from Table 1.1.

The abstract syntax is written in BNF notation, using the "meta"-symbols ::= and |. The ::= is "definitorial", defining the entity of the left is defined by the construction(s) on the right. The | reads "or" i.e., represents alternatives. The last column is just added for human readability, as a kind of comment, in that sense it does not belong to the

a	::=	$x \mid n \mid a\ \mathsf{op}_a\ a$	arithm. expressions
b	::=	$\mathsf{true} \mid \mathsf{false} \mid \mathsf{not}\ b \mid b\ \mathsf{op}_b\ b \mid a\ \mathsf{op}_r\ a$	boolean expr.
S	::=	$x := a \mid \mathsf{skip} \mid S; S$	statements
	$\mid$	$\mathtt{if}\ b\ \mathtt{then}\ S\ \mathtt{else}\ S \mid \mathtt{while}\ b\ \mathtt{do}\ S$	

Table 1.1: Abstract syntax

mathematical content of the grammar. For instance, the symbol S, which is supposed to represent "statements" can be given or *constructed* in exactly 5 different ways; the different cases, separated by | stretch out over 2 lines. The first two alternatives could call **basic** (or elementary or atomic) statements, they represent assignments and an "empty" statement called skip. The last 3 cases represent **composite** statements, statements composed from simpler ones; what they intend to represent should be self-explanatory.

As for notational or typographic conventions we try to follow (though not slavishly): The "constructors" of the abstract syntax, we use some specific `font`, like for `if`, `then`, etc. They can be seen as "keywords" of the language *terminals* of the grammars. That's perhaps a helpful way to see it, though the notion keywords (and tokens) belong rather to the use of grammars for parsing and thus for *concrete* syntax. We are using the BNF as a grammar to describe abstract syntax trees. So the `if_then_else` is a construction that can build a statement "node" when given 3 arguments, a boolean expression (represented by b) and two statements, both of type S. It's one single constructor, not three individual "keywords". We might have chosen to write it `ITE` $b\ S\ S$ to stress it's conceptually better seen as one three-argument constructor, but it makes it harder to read as human. The letters (or sometimes strings) written in *italics* like S, x, etc. are in the terminology of grammars *non-terminals*.

Adding "labels" The second version of the abstract syntax from Table 1.2 contains additional information, called labels l (written as superscript and with some [bracket] notation, which has no purpose except for clarity or readability). We don't explain the role of labels and how they can be "implemented" here much, since we are in the overview, and we pick up on the syntax, the labeling and other issues in a later chapter. The labels can be understood as being able to identify different points in a program.

a	::=	$x \mid n \mid a\ \mathsf{op}_a\ a$	arithm. expressions
b	::=	$\mathsf{true} \mid \mathsf{false} \mid \mathsf{not}\ b \mid b\ \mathsf{op}_b\ b \mid a\ \mathsf{op}_r\ a$	boolean expr.
S	::=	$[x := a]^l \mid [\mathsf{skip}]^l \mid S; S$	statements
	$\mid$	$\mathtt{if}[b]^l\ \mathtt{then}\ S\ \mathtt{else}\ S \mid \mathtt{while}[b]^l\ \mathtt{do}\ S$	

Table 1.2: Labeled abstract syntax

Remark 1.1.1 (Labels)**.** *People who went to the compiler course will have encountered the notion of labels, those were basically the same concept. In more low-level languages*

(including certain intermediate code languages or also assembler) or in a higher-level language with goto's, labels (like here) allow to refer to specific points in the program. That allows to mark a line with a symbolic label like `label_1` *or* `error-handling` *and then do things like* `goto error-handling`*. Used this way, labels are symbolic representations of line numbers in the code. It's a slightly better and less brittle way of programming than doing things like* `goto 137` *where 137 is the line number (some versions of BASIC supported that). That's a perhaps helpful analogy, though again, we are not working with "lines of code", we are working with abstract syntax* trees*, described by the grammar, so there are no lines at all. Still the purpose of the labels is analogous. The labels will not be part of syntax the user can provide, the labeling is added* automatically*. In an actual implementation, one perhaps would use integers and a counter, to attach them to a given abstract syntax tree in a manner without handing out a label twice. In this lecture, we will use labels abstractly, for instance, the labeled version of the factorial from equation (1.2) uses labels* l_0*,* l_1*,* l_2*. It is implicitly assumed that the* $l_0, l_1, \ldots$ *are all different. But otherwise, labels are treated abstract and it's not intended to imply that* l_0 *is the integer* 0 *etc, or even that internally it's an integer, though in a concrete implementation, integers may be a valid representation.* □

The following is a very simple example. What it does should be clear, but it's not too important anyway. The example uses most constructs. The skip-construct is missing (which is a boring one) as is the conditional, but the `while` works similar to the conditional, like a conditional with the addition complication that it leads to a cycle. The notions of input and output variables are not officially part of the abstract syntax, it's more like, in this example, x is seen as the input and z the output.

Afterwards, we see the same program, but being properly labeled. We just use 0, 1, 2, not l_0, l_1, not to overdo it with subscripts or superscripts. The subsequent control flow graph is decorated with l's.

Example 1.1.2 (Factorial). The code of the factorial is shown in equation (1.1), with *input variable* x and *output variable* y.

$$y := x; z := 1; \texttt{while}\ y > 1\ \texttt{do}\ (z := z * y; y := y - 1); y := 0 \tag{1.1}$$

The **labeled** version of the code is given in equation (1.2), and the corresponding control-flow graph in Figure 1.1.

$$\begin{array}{l} [y := x]^0; \\ [z := 1]^1; \\ \ \texttt{while}\ [y > 1]^2 \\ \ \texttt{do}([z := z * y]^3; [y := y - 1]^4); \\ \ [y := 0]^5 \end{array} \tag{1.2}$$

□

Note that x is not initialized in the program, and not assigned to. That's common for input variables (and in a more realistic language, which supports the concepts of input and

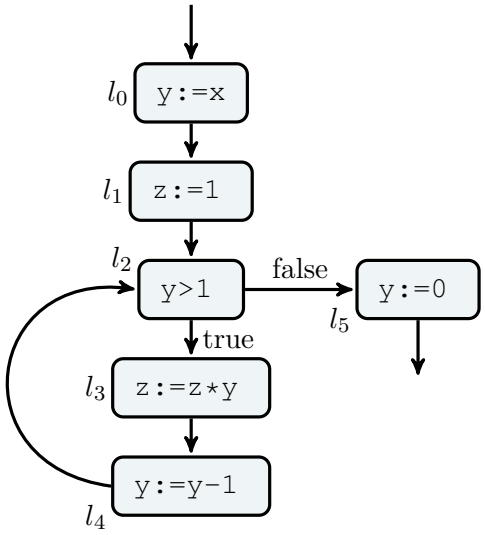


Figure 1.1: CFG factorial

output variables explicitly, it may be enforced by a (trivial) analysis.[1] That all non-input variables are properly initialized could likewise be covered by a non-trivial but still basic static analysis, in fact by a standard data-flow analysis like the one were are about to look at, "reaching definition". We will later comment on that again, in the context of that particular analysis.

Control-flow graphs The factorial used to illustrate the important concept of *control flow graphs*. Later we will look in slightly more detail how to actually *calculate* a control-flow graph, given a program in abstract syntax (i.e., in the form of an abstract syntax tree). Actually, it's not very complex; basically, one has to traverse the tree, "label" the nodes appropriately, which typically means, creating one new node of the graph for each encountered basic construct. In the illustrations here, the nodes are identified by unique labels $l_0, l_1, l_2 \ldots$ (in some way, the numbers themselves serve as identification). The "labeled" abstract syntax from above is a different notation for the *nodes* of the control flow graph + the basic expression "contained in" or "associated with" the nodes (and the nodes of the graph are the same as the labels in the syntax notation). Missing in the labeled program from equation (1.2) obviously are the **edges** of the graph, but, as said, it's not too hard to calculate them as well (while labeling the statements and traversing the syntax tree).

Remark 1.1.3 (Doing a control-flow graph from assembler code or intermediate code)**.** *For participants of the course* compiler construction *(INF5110): the definition of the control flow graph was done in that course at a lower level, i.e., at some* intermediate code *(like three-address-code) without looping constructs, but conditional jumps instead. That intermediate code can be seen as a kind of "machine-independent machine language", i.e., an intermediate language already rather close to actual machine language, but not yet*

[1] One would not bother a data-flow analysis for that, a simply syntactic check seems more adequate.

quite there. In any case, that intermediate language language had officially *labels, as part of the syntax, that could be used for conditional or unconditional* jumps *(which are like goto's). In other words, that intermediate language was "officially labeled" via a specific* `label`*-command (which was counted among the so-called* pseudo-instructions*). Because of that, that form of language almost immediately contains its control-flow graph, since the jumps (and fall-throughs) correspond obviously to the edges. More precisely, for conditional jumps, one of the edges is goes from the conditional jump-statement to the target label, the other edge is the "fall-through". Unconditional jumps correspond to one single edge. In principle, the notion of control-flow graph there is analogous to the one used here (and elsewhere).* □

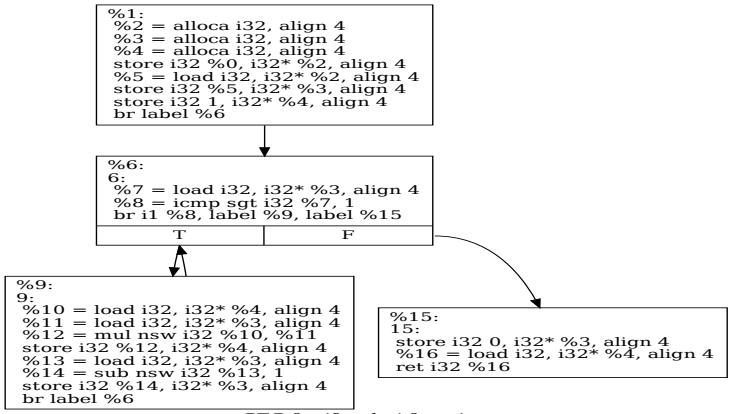
Basic blocks A final remark on the nodes and the code contained therein. In the lecture here, we make the assumption for the imperative language: "one node, one basic statement". In practice, it's often the case that one groups together sequences of statements like assignments together into larger blocks which contains no branching or jumps into it or out of it. Those larger blocks are sometimes called *basic blocks* and the grouping is done for efficiency in an implementation, when running some analysis like reaching definitions on the graph. In the factorial example, l_0 and l_1 could be lumped together (as well as l_3 and l_4). While making larger basic blocks is a good idea in practice for efficiency, we don't care much here, as the principles of data-flow would not change under this refinement.

We may also have a glimpse of how the CFG for the factorial looks in reality. For that, I used a C version of the factorial function. The input variable z is now the formal parameter of the procedure, and the output variable is used in the return-statement.

```
int faculty (int x) {
  int  y,z;
  y = x;
  z = 1;
  while(y > 1) {
    z = z*y;
    y = y-1;
  };
  y=0;
  return z;
};
```

Listing 1.1: Factorial, as C program

The CFG is produced with **LLVM** and `clang` (and rendered with `graphviz`). It's more complicated than our class-room version, though it has lesser nodes: the CFG here uses basic blocks, containing maximal stretches of straight-line code.



CFG for 'faculty' function

When saying, "realistic", on the one hand, that's true, being LLVM and all. On the other hand, one should keep in mind shown is of course not the *actual* CFG data structure, but a graphical representation of it. It is generated with the support of `clang` and `opt` into half-way human-readable textual form (a print-out of the graph, so to say), which in turn is visualized by `graphviz`. But one can see the labels, represented here as numbers and how they are used by the jump-commands ("branch"). More could be said about CFGs in LLVM, including that it makes use of the so-called SSA format, but we can keep that for later).

If one wants to experiment with CFGs oneself, that's the commands I used

```
clang factorial.c  -O0 -S -emit-llvm  -o factorial.o
opt -dot-cfg factorial.o
dot -Tpdf .factorial.dot -o factorial.pdf
```

1.1.2 Example analysis: Reaching definitions

In the following, we look at a first instance of a data-flow analysis. The reaching definitions analysis is just one of a family of similar analyses the lecture covers (and which themselves are only a small selection of many much more). It's a typical *data-flow problem* and the techniques to solve this problem can be used analogously to a wide range of problem. This class of problems is known as *monotone frameworks* and goes back to **?**].

The data which is imagined to "flow" through the program, resp. rather flow through the control flow graphs is not so much the *actual* program data (i.e., here, the integer values stored in the variables). It's rather *information about the data* and the program that flows through the graph. This information may be seen as an *abstraction*. Of course, there may be different kinds of information one could be interested in, and the choice determines the analysis. In case of the reaching definition analysis here, we are interested in the places, where variables are "defined", i.e., the nodes or labels where variables are being assigned to. We will illustrate the analysis with the example program from before, the factorial.

Definition 1.1.4 (Reaching definition)**.** An *assignment* (= "definition") $[x := a]^l$ **may** reach a program point, if there **exists** an execution where x was *last assigned to* at l, when the mentioned program point is reached.

In standard terminology in the context of data-flow analysis, a "definition" of a variable x is an assignment to x, here written $x := a$. The **data of interest** for reaching definitions sets of pairs of variable names $\times$ and labels.

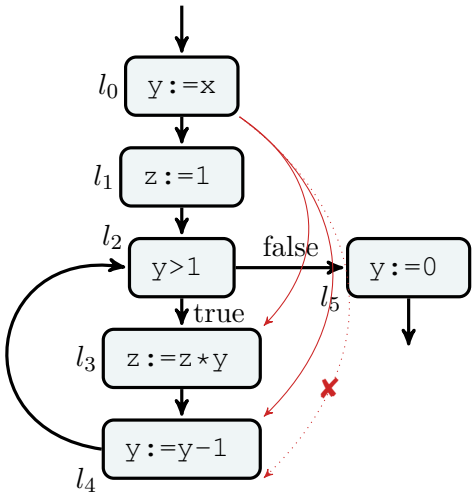


Figure 1.2: Factorial: reaching definitions

Figure 1.2 illustrates, which points are reachable from the choice of origin $(y, 0)$, but, to keep it readable, not all reachable such points are indicated via red arrows. Note also, that by "points" in the program, one does not just say, the node labeled l_x is reachable. It's more fine-grained in that one **distinguishes between the entry and the exit** of a node. For instance, the *exit* of the node l_4 cannot be reached by the assignment to y in the node l_1, whereas the entry of that node l_4 may well be reached. That's indicated by the dotted, resp. the solid arrows, in the illustration.

For the given CFG of the factorial with 6 nodes, that gives altogether 12 "points" of interests. Since there are 6 nodes, there are 6×12 combinations for the question: can the assignment from a node reach the entry or exit of another node. Actually, in the program, l_2 is as starting point is uninteresting insofar it does not contain an assignment. Additionally, for reaching definitions, there are definitions or assignments to the variables, which play a role, which are *not* shown in the graph. Conceptually, it's assumed that all variables have some value "assigned" *prior* to the execution of the program, they have some initial values, though it's unclear what they are. Some languages may guarantee, that variables have specific initial values (0, or a nil-pointer), some don't. The analysis here uses ? **as specific label or node for that.**

It should be also noted: when accepting that the analysis works on tuples (x, l), we immediately see that the problem intuitively is decidable: there are only finitely many variables and finitely many nodes (= labels) in the control flow graph. One could analyze thereby the problem, by making some graph search for all combinations of variables with

start nodes and end nodes. That would be a very **wasteful** approach, and *data-flow analysis* is about more **efficient** techniques.

Table 1.3 shows the complete reaching definition information in tabular form or an array. It can be seen as a pair of functions

$$\mathsf{RD} = (\mathsf{RD}_{entry}, \mathsf{RD}_{exit}) \tag{1.3}$$

assigning to each entry and exit in the control-flow graph the information we are interested in (sets of pairs). For the factorial, there are 12 slots in the table as there are 12 "**points** " in the program for which the information is calculated, namely the *entry* and *exit* to each elementary block or label.

l	RD_{entry}	RD_{exit}
0	$(x,?),(y,?),(z,?)$	$(x,?),(y,0),(z,?)$
1	$(x,?),(y,0),(z,?)$	$(x,?),(y,0),(z,1)$
2	$(x,?),(y,0),(y,4),(z,1),(z,3)$	$(x,?),(y,0),(y,4),(z,1),(z,3)$
3	$(x,?),(y,0),(y,4),(z,1),(z,3)$	$(x,?),(y,0),(y,4),(z,3)$
4	$(x,?),(y,0),(y,4),(z,3)$	$(x,?),(y,4),(z,3)$
5	$(x,?),(y,0),(y,4),(z,1),(z,3)$	$(x,?),(y,5),(z,1),(z,3)$

Table 1.3: Factorial: reaching assignments solution in tabular form

The table shows the *result* of the reaching definition analysis. For the entries and exits for each node, they show, by which definitions from which nodes they can be reached.

Now for the entry points: one needs to look at the graph, they are determined by the inter-block relations, i.e., the *edges*. In the easiest case (for instance $l = 1$), the set just coincides with the post-set of the predecessor block. Interesting is, of course, label $l = 2$, where the graph (interpreted forwardly) joins two arrows, i.e., the block 2 has *two predecessors*, 1 and 4. The question then is: is it the *union* or the *intersection* of the information. Intuitively: it must be *union*. That's due to the nature of the analysis here: the question is, *may* an assignment reach a point in question.

Actually, the table actually does not show *one* possible result of the reaching definition analysis, i.e., not the only **safe** approximation. But it's the "canonical" one insofar that it is the **smallest** over-approximation, i.e., the **most precise** but still **safe**, and thereby the best solution one can (at the given level of abstraction).

Remark 1.1.5 (Input variables and uninitialized variables)**.** *We take up again the notion of input variables (which we don't bother to introduce explicitly). As a result of the analysis, variable x is marked as ?. That can be interpreted as a sign that it's not assigned to. More precisely, the fact that there is the tuple $(x,?)$ at the program exist* together *with the fact that there is* no other *tuple containing x indicates that x is never assigned to (since we are dealing with "may" information"). That's the reason why we can interpret it as* input variable*: It is often a convention that input variables are* not *assigned to inside the program, they are assumed to be given a value up front (indicated by ?).*

If one analyses a function body —each function conventionally is represented by its own control flow graph capturing the control-flow of its body— the formal parameters are given their initial value from "outside" via parameter passing. It's often considered as bad style, at least when passing parameters by value, to assign to the formal parameters.

Of course, non-input variables which are not assigned to make no sense. Actually, if such variables occur, it may indicate a problem (initialized variables) which may lead to random results or errors like nil-pointer exceptions. □

Optimal solution Currently it may not yet be completely clear in what sense the solution is minimal/optimal. Much of the underlying theory (covered later) is about assuring that such a minimal solution demonstrably and uniquely exists (it will be based on the notion of lattices).

Intuitively, it can be compared to the picture from before (static analysis as approximation). An over-approximation is an area fully "covering" the actually, irregularly shaped behavior. Restricting to a certain class of analysis (for instance simple data-flow, specifically say simple reaching definition) can be "visualized" by allowing over-approximation only in the form of a *circle*, let's say. The optimal solution is then the smallest such covering circle (which still is an approximation). If one invests in a more detailed analysis, that might correspond to having, let's say, ellipses or six-edged polygons as (illustration of) allowed results of the analysis. In that case, potentially better and more precise approximations may be doable (typically at higher computational costs). It may also be the case, that no longer a single best solution exists in general. For instance, there may be *two* covering polygons, both of which cannot be made smaller without becoming unsound, but are on the other hand **incomparable** (neither is contained in the other). In this lecture, we typically deal with settings where *optimal* solutions do exists, though.

So far we have a impression of the information we are after, at least in the reaching definition case. We have a feeling when the information is correct (i.e. when it's a safe approximation). Missing is: how to obtain that information? The way we proceed is: first show how to *specify* the data-flow problem for a given program, and afterwards show how to *solve* the data-flow problem.

A program gives rise to a data-flow problem in the form of **constraints**. The notion of constraints is a very general one, there are may different forms of constraints, also many in use for static analysis. Consequently there are many different algorithms to tackle specific classes of constraints and may different constraint solvers.

A constraint is nothing else then a formula of some form containing some (free) variables or unknowns. Solving a constraint is nothing else that filling out the unknowns with values in such a way that the formula becomes true. So the concept as such is very familiar, and one will have it encountered it already in early math classes in elementary school, when one has to solve things like $x + 6 = 14$ or what not. Some folks solve constraints for recreational purpose, filling out crossword puzzles or sudokus (the white squares are the unknowns or "variables").

For us: Data-flow problems are straightforwardly cast as a simple form of constraint system. We illustrate that in the following, presenting two variations, which are slightly different ways to do that. The difference is **non-essential** in that both representations

give basically the same results and both representations will be solved algorithmically exactly the same way. Still we go through both variants, since the second one will be more flexible. The flexibility is irrelevant for the simple form of constraints right now (reaching definitions), but better suitable for more complex setting (like analyzing functional languages).

The first approach phrases the data-flow problem as a set of **equations**, the *equational* approach. Afterward we generalize the equations by **inequations**, i.e. subset-constraints, using $\subseteq$. That's the only difference.

For precision's sake: we have be careful with our statement claiming that both approaches "give the same results" because literally taken, it's not true that both versions have the same set of solution: inequations allow (naturally) more solutions. As analogy, $x + 4 \leq y$ has more solutions than $x + 4 = y$ (those are not data-flow constraints, but constraints involving equations and inequations nonetheless). But as far as the intended solution is concerned, both variants give the same in the particular mathematical setting of data-flow constraints (more later).

1.1.3 Equational approach

The constraint system takes a particularly simple form. Actually, it's not so much the fact that there are equations, and we don't want to imply, equations in the first approach are simple whereas the "inequations" cover later, using $\subseteq$ instead, are complex. They are equally simple.

What makes the constraint problem simple when compared to other forms of constraints is perhaps two facts. First is the *form* of constraints. Constraints are formulas with variables. In this case, the "formulas" are sets of equations (or inequations later), where the set can be seen as *conjunctions*, insofar that a solution must satisfy all (in-)equations. What makes it simple is, that one side of all questions, the variables are "isolated". As illustration: in an (numerical) equation system like

$$x = y - 16 \tag{1.4}$$
$$y = 4 \tag{1.5}$$

in the first line, x is "isolated" on the left, but not y, and in the second line y is "isolated". For inequations, one would require, that both lines would either be both $\leq$ or both $\geq$, a mixture would be no longer in the simple form characteristic of data-flow equations (and the techniques for solving data-flow equation would no longer apply). This specific restriction on the form of the (in-)equations is only aspect what makes data-flow constraint system simple (actually, in some presentations, the word "simple constraint system" is meant as technical term to denote said syntactic restriction). The other aspect that makes the data-flow setting quite straightforward is not visible in the form of the constraints. It has to do with the **domain** over which the constraints are solved to fill in the unknowns.

Before we go there, let's have a second look at the constraint system from (the non-data-flow) equations (1.4) and (1.5). A plausible domain over which the system is to be solved is

integers. Another one could be the natural numbers, in which case there'd be no solution. So, let's say it's integers. In that case, to solve the equation system is not just easy, but super-easy. Everyone (even in elementary school) knows how to proceed. That it is so trivial has to do with the fact that the equation are simple, with variables isolated on the left. For y, the solution is directly given, and that we just have to set into the first equation, do a simple numerical calculation, and that gives the result -12 for x, and we are done.

Now, that was not just simple and easy, but outright trivial. Data-flow constraints are still simple in the way described. However, they involve typically **recursion**. We could change the above equation as follows, making the two equations *mutually recursive*.

$$x = y + 12 \tag{1.6}$$
$$y = 5 + x \tag{1.7}$$

A solution for x depends on a solution for y, and vice versa. Such equations are not typically encountered in elementary school... Adding recursion catapults the constraint system to a whole new level of complication (thought the form is still called "simple" since at least the variables on one side are "isolated").

The particular recursive equation system has *no* solution over integers nor natural numbers (nor in any standard numerical type).[2]

However, data-flow equations are seldom solved over natural numbers or integers and equations would typically not involve plus and minus on numbers. Instead, they are solved over domains that have nicer mathematical properties. What is nice or not, of course, may be a matter of taste, but nice in the sense that the properties of the domain fit *exactly* to the needs that allow the data-flow problem to be solved. We will shortly see that in the example of reaching definitions, where the domain are **sets** of elements containing RD information (the pairs of the form $(x, ?)$ from before), ordered by $\subseteq$. Calculations over such kind of non-numerical domains are consequently also not plus or minus or things like that, but set union and set intersection. Later we discuss more in detail what properties such domains need to have (they are known as complete lattices) to guarantee that simple recursive equations like the one before do have solutions. Actually, not only that, working with such solution domains guarantees **unique best solutions** and that they can be **efficiently** calculated, at least in the sense that no backtracking is required.

We will cover the necessary background later. For now: the reaching definitions (and other data-flow questions) are solved over a domain which is a complete lattice. In a way, that's what basic data-flow analysis is all about:

> *data-flow analysis = solving simple (but typically recursive) equations over a lattice.*

As said, sets (ordered by $\subseteq$) are a lattice, the natural numbers (ordered by $\leq$) are *not* a lattice.

[2] I say standard, since if one considered "infinity" a some non-standard number, one could reasonably say $x = \infty$ and $y = \infty$ is a valid solution.

By the fact alone, that equations (1.6) and (1.7) have no solution in $\mathbb{N}$ shows, that it can't be a lattice, because being a lattice would *guarantee* that a solution exist. Well, to be really precise: to really guarantee that the system has a (best) solution, we have also take into account that + is a *monotone* function (which it is), otherwise, a solution may exist or may not. Indeed, if we would use $\mathbb{N}^\infty$ instead of the standard natural numbers, with and additional value ∞ representing an element $\geq$ than all other numbers, that would turn the domain into a (complete) lattice. Since +, properly extended, is still monotone for $\mathbb{N}^\infty$, the above equation system does have a (unique smallest) solution. Namely ∞, as we already mentioned.

1.1.3.1 From flow graphs to equations

The next task is to turn the control flow graph into an equation system or equational constraint system. That's actually straightforward. The equation system describes the "flow of information", here: the information related to reaching definitions, and the information imagined to flow forwards. Equation system resp. constraint systems contain variables. These are not the variables of course of the program we analyze (here the factorial) but are "meta" variables of the constraint system, intended to contain the solution to the flow problem. So those variables represent the of interests, for the factorial in the example (or the slots in the Table 1.3 and the content of the slot is the solution).

As mentioned, the constraints serve to capture the flow, more precisely how reaching definition pairs at one point influence the reaching definition information at other points. Or conversely, how reaching definition information at one point **depends on** the reaching definition information at other points.

Example 1.1.6 (Equational constraint system for the factorial). The reaching definition constraint system for the control-flow graph from Figure 1.1 is given by the 6 equations from (1.8) and the 5+1 equations from 1.9.

$$\begin{array}{lcl}
\mathsf{RD}_{exit}(0) & = & \mathsf{RD}_{entry}(0) \setminus \{(y,l) \mid l \in \mathbf{Lab}\} \cup \{(y,0)\} \\
\mathsf{RD}_{exit}(1) & = & \mathsf{RD}_{entry}(1) \setminus \{(z,l) \mid l \in \mathbf{Lab}\} \cup \{(z,1)\} \\
\mathsf{RD}_{exit}(2) & = & \mathsf{RD}_{entry}(2) \\
\mathsf{RD}_{exit}(3) & = & \mathsf{RD}_{entry}(3) \setminus \{(z,l) \mid l \in \mathbf{Lab}\} \cup \{(z,3)\} \\
\mathsf{RD}_{exit}(4) & = & \mathsf{RD}_{entry}(4) \setminus \{(y,l) \mid l \in \mathbf{Lab}\} \cup \{(y,4)\} \\
\mathsf{RD}_{exit}(5) & = & \mathsf{RD}_{entry}(5) \setminus \{(y,l) \mid l \in \mathbf{Lab}\} \cup \{(y,5)\}
\end{array} \tag{1.8}$$

$$\begin{array}{lcl}
\mathsf{RD}_{entry}(1) & = & \mathsf{RD}_{exit}(0) \\
\mathsf{RD}_{entry}(2) & = & \mathsf{RD}_{exit}(1) \cup \mathsf{RD}_{exit}(4) \\
\mathsf{RD}_{entry}(3) & = & \mathsf{RD}_{exit}(2) \\
\mathsf{RD}_{entry}(4) & = & \mathsf{RD}_{exit}(3) \\
\mathsf{RD}_{entry}(5) & = & \mathsf{RD}_{exit}(2) \\
\\
\mathsf{RD}_{entry}(0) & = & \{(x,?),(y,?),(z,?)\}
\end{array} \tag{1.9}$$

The first set of equations (1.8) capture the **intra-block** flow, the ones from (1.9) capture the **inter-block** flow.

The fact that there are altogether 12=6+6 equations and there are 12=6+6 constraint variables is a coincidence. The equations, as said capture how the flow information at different points in the control flow graph depend on the information on other points. There are two types of such dependencies. One is that the reaching definition information at the **exit** of a block depends on the flow information at the `entry` of the same block or label. Since there are 6 labels resp. 6 nodes in the CFG, there are 6 intra-block flow equations (equation (1.8)). They describe the "flow" of analysis information through each block.

The second type of equations deal with how information flows **between** the blocks, and thus correspond to the **edges** of the graph. Note that it's not the case that there is one equation per edge! There are 6 "internal" edges in the CFG.[3] However, they are represented by *5* equations, since the equation for the entry of node 2 reads $\mathsf{RD}_{entry}(2) = \mathsf{RD}_{exit}(1) \cup \mathsf{RD}_{exit}(4)$ and covers the two incoming edges.

The last equation from (1.9) covers the incoming pointing to the first node. □

Note that the intra-block equations all start with with RD_{exit} as left-hand side, whereas inter-block nodes start with RD_{entry}. This reflect the fact that the information flows **forwards.**

Remark 1.1.7 (One isolated entry node)**.** *The entry node has no internally incoming edge in this example (only one that is shown in the picture to come from "outside" which is therefore not part of the graph nor an edge, it's just a conventional, graphical indication of the initial node). This is not a coincidence, insofar that one generally assumes that the control-flow graph does not have such an* initial loop*. In the unlikely event that the program would start immediately with a loop or similar, the compiler would arrange it so that there is an extra "skip" entry node (a* sentinel*). Alternatively, the generation of the control-flow graph might simply automatically add some extra sentinel, just in case.*

Technically, nothing would go wrong allowing such initial loops, it's just that the technical representation later (and the algorithms) are slightly simpler. Sure, not fundamentally simpler, only avoiding some extra corner cases, it just gets a tiny bit more smooth. □

Remark 1.1.8. *For participants of "compiler construction" (INF5110): that simplifying "restriction" from the previous remark is mildly reminiscent of the treatment of context-free grammars for some constructions (like the LR(0)-DFA construction). For some constructions on grammars, it was assumed that the grammar's start symbol, say S, does not show up on the right-hand side of any production. That would correspond to a loop back to the "initial state" here. Since also there, one wanted to slightly simplify the treatment avoiding that case, the standard construction routinely simply added "another" start symbol S' and a production $S' ::= S$.*

[3] Not counting the "incoming" edge indicating the entry of the factorial program and the outgoing edge indicating the end of the program.

1.1.3.2 General scheme (for RD)

The following equations show the scheme of data-flow equations of a while-program, more precisely the flow equations given the control-flow graph of a while-program. Besides equations for the initial node, there are equations *per node* of the graph, and equations in connection with the *edges*. Since the nodes represent typically *basic blocks*, one speaks of equations representing *intra-block data-flow* (inside a block), and equations, representing *inter-block data-flow* (between blocks).

As said earlier, we allow ourselves to make nodes that contain only one statement, where typically in implementations, it would be maximal stretches of straight-line code. Nothing much would change though as far as the principles are concerned. The only complication would be: currently, for intra-block flow, there are exactly three cases to consider: conditions coming from loops or conditionals) and the two possible cases for statements, assignments or skip-statements. Those are covered by equations (1.10) and (1.11). In the realistic setting that nodes contains sequences of assignments, the nodes themselves cannot be by a simple finite case distinction, they require an analysis of their own. But that does not change anything, it's just that per a node, one would have a number of equations, which correspond to the equations we would get here, when stretching out the statements across multiple nodes.

If either representation, lumping straight-line code together into one large node or stretching them out to a linear sequence of nodes does no make a difference, why do real compilers bother to work with basic blocks? It's largely a matter of efficiency. Soon we will get an impression of how such equations can be solved. In any rate, if the program contains loops, and most programs do, the analysis will handle equations *repeatedly*. That's caused by the recursive nature of the equations, which is caused by loops in the program code. Because of that, one gains efficiency, if one calculates the accumulated effect of stretches of linear code in a first stage. Of course, if a large basic block is part of a loop, the overall effect of the basic block needs still may be repeatedly taken into account, but one does not repeatedly accumulate the effect of the block's individual elementary steps one over and over again, since one has *pre-computed* their summary effect.

Intra

- for assignments $[x := a]^l$

$$\mathsf{RD}_{exit}(l) = \mathsf{RD}_{entry}(l) \setminus \{(x, l') \mid l' \in \mathbf{Lab}\} \cup \{(x, l)\} \tag{1.10}$$

- for other blocks ($[b]^l$ or $[\mathsf{skip}]^l$), side-effect free:

$$\mathsf{RD}_{exit}(l) = \mathsf{RD}_{entry}(l) \tag{1.11}$$

Inter

$$\mathsf{RD}_{entry}(l) = \bigcup_{l' \to l} \mathsf{RD}_{exit}(l') \tag{1.12}$$

Initial l: label of the initial block (isolated entry)

$$\mathsf{RD}_{entry}(l) = \{(x, ?) \mid x \text{ is a program variable}\} \tag{1.13}$$

The formulas in equations (1.11), (1.12), and (1.13) are of course *schemas* of equational constraints, so they represent *sets* of equations. The bigger the control-flow graph, the larger the number of equational constraints, with variables of the form $\mathsf{RD}_{entry}l$ and $\mathsf{RD}_{exit}l$.

So far we have only talked about the equations, which can be seen as a **specification** of the flow problem. But we have not solved it. That comes next, where we speak about solving an equation. One needs also think about *how many* solutions such equations have. The equations seen so far are already in a form, which is rather restricted: the left-hand side of the equations are "isolated". That's almost solved, except for one (important) thing: recursion ... Both facts —"isolation" and recursion— will show up more clearly in that we turn it into a *fix-point* equation.

1.1.3.3 Fix point equations

Earlier, we commented on the specific nature of data-flow equations ("isolated" variables on one side, but recursive). We will soon see how to approach them, in particular how to tackle the phenomenon of recursion. As a first step towards that, we take a second look at those equation, to understand them as what is known as **fix-point** equations. What a fix-point is, is actually very straightforward. Assume one has some kind of function f of type $A \to A$ (where for the time being, A does not matter). Now, a *fix-point* of f is an element a from A such that

$$f(a) = a$$

In other words:

> A fix-point for f is a solution to the *fix-point equation*
>
> $$f(x) = x$$

i.e., finding a value for x in this constraint. The fix-point equation corresponds directly to the form of the data-flow constraints as we will explain, with the only difference that the data-flow equations typically involve more than one variable; they are of the form $f(\vec{x}) = \vec{x}$. That's actually not a very significant difference, one can see x also as a variable being of a type of a tuple; that's the same of having multiple variables $\vec{x}$.

Now that we know that data-flow equations are nothing else than fix-point equations, we will need to figure out how to solve those. Let's see first in some familiar mathematical setting, like "numbers". Assume as example the square function f_{square} defined as $\lambda y.y^2$ of type $\mathbb{R} \to \mathbb{R}$. So, what about fix-points of that function, i.e., what are solutions of the fix-point equation $f_{square}(x) = x$, i.e., solutions of

$$x^2 = x$$

In this case, it's easy, there are exactly two fix-points, 0 and 1. Those can be visualized by plotting the function x^2 in the standard manner, and the fix-points are where the graph intersects with the 45°-line (representing the identity function), see Figure 1.3.

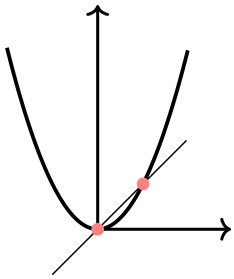


Figure 1.3: Two fix-points of x^2

If we modify the equation a bit to $x^2 + 2 = x$, then there are no fix-point solutions (at least none from $\mathbb{R}$). Other kind of numerical equations over $\mathbb{R}$ involving polynomials may have exactly one, or more than 2 solutions. So, the solution space can consist of 0, 1, 2, etc. solutions.

Data-flow analysis is typically facing fix-point constraints over *non-numerical* domains. We have seen an example characteristic for such domains in the reaching definition example. For such domains, the situation is nicer. For those it is *guaranteed* that the data-flow equations do have a solution, there cannot be *no* solution. Typically, there are many solutions. That, however, does not cause headache. Multiple solutions could be problematic, because in this case, one might not say what *the* solution of the data-flow problem is. However, the solution domains is of a form, that the constraint system has **exactly one most precise solution**. That's the one the analysis returns. That's good news because in this way, one can speak about *the* solution of the data-flow problem. But that's not all. The structure of the solution domain does not only guarantee the existence of a unique best solution. It also assures, on top of that, that it can be calculated without backtracking, i.e., **efficiently.**

1.1.3.4 The reaching definition equation system for the factorial function as fix point

The RD equation system for for the factorial has 12 variables $\mathsf{RD}_{entry}(0), \ldots, \mathsf{RD}_{exit}(5)$, supposed to contain sets of pairs of the form (x, l). A solution of the equation system was shown in tabular form in Table 1.3. We can write the twelve-tuple of variables in "vector" form:

$$\vec{\mathsf{RD}} = F(\vec{\mathsf{RD}}) \tag{1.14}$$

where F represent the right-hand sides of the individual equations. This is just a more concise representation of the equation system consisting of 12 individual equations, where instead of individual functions representing the right-hand sides as shown in equations (1.8)) and (1.9). So F from equation (1.14) is a function on of type

$$F \ : \ (2^{\mathbf{Var}_* \times \mathbf{Lab}_*})^{12} \to (2^{\mathbf{Var}_* \times \mathbf{Lab}_*})^{12} \tag{1.15}$$

i.e., the solution domain is a 12-tuple of the information of interest, namely sets of pairs of variables and labels. One 12-tuple contains the RD-information for each individual point in the factorial. Table (1.3) is such a 12-tuple, arranged in more readable form. As for the

used symbols: $\mathbf{Var}_*$ represents the variables "of interest" (i.e., all variables occurring in the program). For the factorial example, that is x, y, and z. The set $\mathbf{Lab}_*$ are the labels of interest (i.e.,all labels occuring on the program. For the factorial example, the two sets are

$$\mathbf{Var}_* = \{x, y, z\} \quad \text{and} \quad \mathbf{Lab}_* = \{?, 0, \ldots, 5\}$$

As indicated we want the best, in this case the least or smallest solution (and the underlying theory guarantees that it exists and it exists uniquely. Solutions naturally ordered (by subset ordering, a subset is better), and the order is a so-called *partial order*. The (12-)-tuples are ordered accordingly. One says they are ordered *pointwise*:

$$\vec{\mathsf{RD}} \sqsubseteq \vec{\mathsf{RD}}' \text{ iff } \forall i.\ \mathsf{RD}_i \subseteq \mathsf{RD}'_i \tag{1.16}$$

The solution domain is not just partially ordered as given in equation (1.16), it satisfies some extra conditions, which makes it into a specialization of a partial order called a (complete) **lattice**, which guarantees the unique existance of smallest solutions. The relevant information form a lattice not just for RD but for quite many other analyses.

1.1.4 Constraint-based approach

This section is basically a **rerun** of the previous one: the constraint-based approach here is nothing much more than a variation of the equational approach. As for terminology, the distinction between "equational" vs. "constraint-based" is slightly misleading. It's misleading insofar, as also the equational approach is based on constraints, namely **equational constraints**. In the following, instead of equations, we will use "inequations". In the current RD analysis, the solutions are sets (of pairs (x, l)) and the inequations will be **subset** constraints $\subseteq$ on sets, instead of equality $=$ on sets.

In the current setting, a basic data-flow problem for a simplistic imperative language, there's not much difference between the two approaches ($=$ vs. $\subseteq$). Not only is it straightforward to replace the $=$ by $\subseteq$. Also on a deeper level, there are technical reasons resulting in the following fact: the best (here the smallest) solution of the equational approach *coincides* with the smallest solution of the constraint-based approach. That means, for the purpose of program analysis, which is after the best, safe approximation, **both approaches are the same**!

The setting here with simple control-flow graphs, however, *is* indeed simple. In more complex settings (for instance functional languages with much more flexible control flow) it's not always the case that equational approaches and constraint based approaches result in the same formulation of the problem (and as a result give the same (best) analysis result). In general, the subset-constraint-based approach offers more flexibility (thus sometimes more precise analysis) in more complex settings. But, as said, not right now, here it does not really matter. In the context of type and effect systems, we will encounter more complex settings, where $\subseteq$-constraints come in more handy.

When stating that (here) the step from an equational to a constraint-based approach amounts to a trivial reformulation of the equations leaves, however, one question still remains to be answered:

Should $=$ be replaced by $\subseteq$ or $\supseteq$?

We will see that the answer to that question depends on the analysis we are doing (basically whether we are interested in safe over-approximations or under-approximation). In the case of the reaching definitions ("... may reach ..."), a solution *larger* than a safe one is safe, as well. As a consequence, the "left-hand sides" of the previous equations need to be $\supseteq$-larger than the "right-hand sides".

A final remark: the fact that we move from equations to $\supseteq$ allows for a simple *rearrangement* of the constraints. Basically, instead of replacing $s_1 = s_2 \cup s_3$ by $s_1 \supseteq s_2 \cup s_3$, the latter is split into

$$\begin{aligned} s_1 &\supseteq s_2 \\ s_1 &\supseteq s_3 \end{aligned}$$

With this, the reformulation of the equations is straightforward.

Example 1.1.9 (Factorial program: Intra- and inter-block constraints). Again, we group them into intra-block and inter-block equations. Cf. also the equation sets (1.8) and (1.9) from before.

$$\begin{aligned}
&\mathsf{RD}_{exit}(0) \supseteq \mathsf{RD}_{entry}(0) \setminus \{(y,l) \mid l \in \mathbf{Lab}\} \\
&\mathsf{RD}_{exit}(0) \supseteq \{(y,0)\} \\
&\mathsf{RD}_{exit}(1) \supseteq \mathsf{RD}_{entry}(1) \setminus \{(z,l) \mid l \in \mathbf{Lab}\} \\
&\mathsf{RD}_{exit}(1) \supseteq \{(z,1)\} \\
&\mathsf{RD}_{exit}(2) \supseteq \mathsf{RD}_{entry}(2) \\
&\mathsf{RD}_{exit}(3) \supseteq \mathsf{RD}_{entry}(3) \setminus \{(z,l) \mid l \in \mathbf{Lab}\} \\
&\mathsf{RD}_{exit}(3) \supseteq \{(z,3)\} \\
&\mathsf{RD}_{exit}(4) \supseteq \mathsf{RD}_{entry}(4) \setminus \{(y,l) \mid l \in \mathbf{Lab}\} \\
&\mathsf{RD}_{exit}(4) \supseteq \{(y,4)\} \\
&\mathsf{RD}_{exit}(5) \supseteq \mathsf{RD}_{entry}(5) \setminus \{(y,l) \mid l \in \mathbf{Lab}\} \\
&\mathsf{RD}_{exit}(5) \supseteq \{(y,5)\}
\end{aligned} \tag{1.17}$$

$$\begin{aligned}
\mathsf{RD}_{entry}(1) &\supseteq \mathsf{RD}_{exit}(0) \\
\mathsf{RD}_{entry}(2) &\supseteq \mathsf{RD}_{exit}(1) \\
\mathsf{RD}_{entry}(2) &\supseteq \mathsf{RD}_{exit}(4) \\
\mathsf{RD}_{entry}(3) &\supseteq \mathsf{RD}_{exit}(2) \\
\mathsf{RD}_{entry}(4) &\supseteq \mathsf{RD}_{exit}(3) \\
\mathsf{RD}_{entry}(5) &\supseteq \mathsf{RD}_{exit}(2)
\end{aligned}$$

$$\mathsf{RD}_{entry}(0) \supseteq \{(x,?),(y,?),(z,?)\} \tag{1.18}$$

Both constraint systems (the one with equational constraints and the one with subset constraints) look very similar. One adaptation is that with the subset constraints, we can split constraints in the equational formulation into separate $\subseteq$-constraints; of course *not* splitting them and just replacing $=$ by $\subseteq$ would also be fine, especially since the greater "flexibility" the splitting allows does not matter for the simple setting of the while-language and RD analysis. □

Now, let's revisit the question of solutions of the constraint system, in particular let's **revisit the least solution.** Instead of solving the fix-point constraint system $F(\vec{\mathsf{RD}}) = \vec{\mathsf{RD}}$ from equation 1.14, we are tasked to solve the following subset constraints:

$$F(\vec{\mathsf{RD}}) \sqsubseteq \vec{\mathsf{RD}} \tag{1.19}$$

Solutions to that are also called **pre-fixpoints** as opposed to **fixpoints** which are solutions of the equational formulation.[4] It should be clear that solution to the equation system are also solution to the $\subseteq$-constraint system, but not necessarily vice versa: the $\subset$-formulation has in general more solutions, as the constraints are more "lax". In that sense, the two constraint system are different, but in one aspect, and that's the crucial one, they are "the same". Both have **the identical smallest solution**.

1.2 Control-flow analysis (constraint-based)

As remarked in the data-flow analysis part, both the approach with data-flow **equations** as well as the one based on $\subseteq$ are data-flow **constraints**. It was stressed *there* (in that simple setting) that the equational approach and the approach based on subset constraints give the same result (in that the best solution coincides in both representations).

Another thing that was simpler for the while-language is the control flow: determining the CFG was so simple that one would not even bother to call it control-flow analysis (even if it is). Here, we are facing a non-trivial **control-flow analysis** problem and we use a $\subseteq$-constraint-based approach (not an equational) to tackle it.

Now things change, insofar that we are dealing with a **functional language**. As sketched in the previous section, determining the control-flow for a very simple language like the while-language is really straightforward. On top of the control-flow graph of a given program, we determined some data-flow (reaching definitions for instance). So, this was clearly a **two-stage** approach, **first** determine the control-flow in the form of the control-flow graph, and **afterwards**, on top of that, the data-flow analysis. For precision's sake: of course the control-flow graph, strictly speaking, *does not* represent *the* control-flow. It represents a *static abstraction* of the control flow.

When working now with *functional* languages, things get more involved. The simple picture of control-flow on the one hand, that can easily visualized by a graph, and data on the other hand, that "flows" through that graph does no longer work. A reason for that is: characteristic for functional languages is that the line between "control" on one hand and "data" on the other gets blurred. Basically, all is data *and* control in the following sense: the central concept in functional languages is the notion of *functions*. Those are the things being executed, in this sense they are "code". On the other hand, they can be passed around as arguments of other functions and returned as results of functions. This, together with the possibility of nested, lexical scopes for functions, is known as **higher-order functions**. Functions as **first-class citizens** of the language, which means there is no longer a strict separation between "data" on the one hand, and functions on the other. Some (mostly non-functional) languages have restrictions and don't treat functions

[4] We will later,in the main part, also encounter **post-fixpoints** (which are obtained when replacing $\sqsubseteq$ by $\sqsupseteq$). It corresponds to the switch from **over**-approximations to **under**-approximations.

as first-class citizens, for instance by regulating: the language has lexical scope, one can have local data inside a function, but unfortunately no local function definition. Or, the language has functions, one can pass all kinds of data as argument and as return value, but unfortunately not functions themselves etc.

In functional languages, there are typically no such restrictions. As a consequence, it's no surprise that static program analysis gets more involved. That's even true taking into account that the setting for data-flow analysis in the introduction so far was really simple in that we did not even consider (non-higher-order) functions there. Adding first-order procedures would complicate things —later in the lecture we will cover that to some extent— but still the 2 stages of determining the control-flow graph first, and afterwards doing data-flow constraints are clearly discernible. When saying that analysis for languages with higher-order functions get more involved, then that's fair to say; for instance, control-flow analysis becomes more involved.

Though it may be the case that static analysis becomes more simpler in some aspects, as well. Functional languages often downplay the role of **mutable state** (and compensate by the added flexibility of higher-order functions). Still most functional languages support mutable state, it' just not encouraged; only purely functional or declarative languages put a total ban on mutable state. Mutable state poses problems for static analysis, so avoiding them makes some aspects of some analyses easier. In some languages that do *not* support higher-order functions, there are ways to recover some expressivity of that feature. Like "patterns" that (almost) work as if one had such functions, perhaps involving function pointers. Relying on (in principle mutable) pointers opens another box of Padora, the flexibility of mutable pointers are notoriosly hard to analyse (and are a notorious source of errors).

As for the importance of functional or decrarative ways of programming: Indeed, some intermediate representations for imperative languages turn the imperative code in some form of "functional representation" known as SSA-format (static single assignment). For instance, LLVM uses CFG+SSA as intermediate representation, because further analyses profit from such a "quasi-functional" intermediate representation. It does not mean, LLVM uses higher-order functions in that intermediate representation, it just treats variables in a quasi-declarative way, which makes things easier afterwards (of course turning imperative code to SSA takes some effort in itself, but the fact that SSA is very popular shows, it's often worth it).

For the lecture, similar as what we did for the imperative while-language, we work with a stripped down core language capturing (some) core aspects of functional languages

> That's "the" λ-calculus.

There are many λ-calculi (especially, when it comes to typed versions, different type systems drastically change expressivity). Here, in this part of the intro, we restrict ourselves to the bare-bones version: the *untyped λ-calculus*.[5] We use that to show, how one does labels a program in that language. The labelling is more "complex", though not more

[5] Later in the intro, we also present a typed version. But also when being untyped, there are different untyped λ-calculi, we take a very stripped down version, in that sense "the" untyped λ-calculus, and later we assume call-by-value (as opposed to lazy evaluation).

complex in that it's harder to understand. But the labels can in a way be "nested". In the CFG for the while language, the labels corresponded to the nodes of the CFG. In that sense, one could consider that now nodes can be contained in nodes. On top of that, running the program would change the nesting structure, so the "nested-graph" would somehow dynamically change. In short, there is no good way to visualize the control-flow with some static graph structure. Still, we can capture the relationships between the labels by *constraints*. Solving those constraints then is a form of *control-flow analysis*.

Another way of seeing it is as follows: the control-flow problem here (which is inseparable from data-flow aspects as functions are higher-order), will not even result in a control-flow graph. For a non-higher order language, each function body has its own CFG, all clearly separated, but such a depiction makes no longer sense. On the other hand: in the data-flow section, we slightly touched upon the close connection between control-flow graph and the constraints (equational or otherwise). We did not explore it to the end, but one can view each edge of a control-flow graph as constraints connecting (in a forward analysis) the solutions of a target node dependent (via a constraint) on the solution on the source nodes and the nodes (= labels) play the role of *variables* in the constraint system. The connection is so close that a constraint systems in the form we covered in the data-flow section *is nothing else* than some representation of the control-flow graph (together with constraints connecting the nodes). As illustration, an inequation

$$x_1 \supseteq f(x_2, x_3)$$

represents the dependence of (the solution for) x_1 on the solutions of x_2 and x_3 via some f, thereby representing two incoming edges in x_1, starting from x_2 resp. x_3. In Section 1.1, this representation was not made explicit, and we did not call the variables x_i but RD_i (and split between entries and exits), but in general the connection is between variables of the constraint system and the nodes of the control-flow graph.

Now, in the more complex situation of higher-order functions, we operate directly with constraints; since they are now more complicated, they can no longer be visualized as expressing relations on values in some graph as we said before.

1.2.0.1 Control-flow analysis

> The **general goal for CFA** (here) is to determine which elememtary blocks lead (or may lead) to which other elementary blocks

Example 1.2.1. Let's look at a simple functional program. In the program, the function f takes a functional argument. The example does not involve a situation where a function *returns* a function.

```
let f = fn x => x 1;
    g = fn y => y + 2;
    h = fn z => z + 3;
in (f g) + (f h)
```

Listing 1.2: Simple example

We assume the code to be untyped (at least statically untyped). Though it's clear from the code snippet that g and h takes numbers as argument, and f takes as argument functions, as the formal parameter x is applied to 1 in the body. This application is of course the interesting part of the code, as it requires analysis to figure out, which function(s) are actually called. In the example, it's of course immedately clear by looking at it, namely g and h, but in general it might be more involved.

The code also makes use of definitions via let-in, even though the current calculus won't have that contstruct. □

The **goal more specifically here is** to determine for each function application, **which function** may be applied.

The code snippet is a simple example of a functional program, covering in particular the possibility of having functions as argument. What is written here as $\texttt{fn}\, x \Rightarrow \ldots$ might be syntax one could find in a functional language. In λ-calculus notation, it would correspond to $\lambda x. \ldots$. Make sure, even if not familiar with functional programmig, that the short piece of code makes intuitive sense, i.e., one should have a feeling what it does when executed. There is no point in trying to get a feeling what the following analysis does when approximating the behavior of the program, if the behavior of the program is not at least half-way clear for a start.

Local definitions with let The above example uses `let` for local definition. That is slighty different than assignments. For once, it's "single assignment", secondly it has a local scope associate with it. In later chapters, the functional language may support that construct, for the introduction we ignore it.

1.2.0.2 Labeling

- more complex language ⇒ more *complex* **labelling**
- "elem. blocks" can be *nested*
- *all* syntactic constructs (expressions) are labelled
- consider:

Unlabelled abstract syntax

$$(\texttt{fn}\, x \Rightarrow x)\ (\texttt{fn}\, y \Rightarrow y)$$

Full labelling

$$[\ [\texttt{fn}\, x \Rightarrow [x]^1]^2\ [\texttt{fn}\, y \Rightarrow [y]^3]^4\ ]^5$$

- functional language: side-effect free

$\Rightarrow$ *no* need to distinguish *entry* and *exit* of labelled blocks.

As promised at the beginning of the section: the labelling may be conceptually be more complex (allowing labels "inside" labels, so to say) but actually not more difficult to understand how the labelling is done. compared to the while-language.

1.2.0.3 Data of the analysis

Pairs $(\hat{C}, \hat{\rho})$ of mappings:

abstract cache: $\hat{C}(l)$: set of values/function abstractions, the subexpression labelled l may evaluate to

abstract env.: $\hat{\rho}$: values, x may be bound to

1.2.0.4 The constraint system

- ignoring "`let`" here: *three* syntactic constructs $\Rightarrow$ *three* kinds of constraints
- relating $\hat{C}$, $\hat{\rho}$, and the program in form of subset constraints (subsets, order-relation)

3 syntactic classes

- function abstraction: $[\texttt{fn}\, x \Rightarrow x]^l$
- variables: $[x]^l$
- application: $[f\ g]^l$

1.2.0.5 Constraint system for the small example

Labelled example

$$[\ [\texttt{fn}\, x \Rightarrow [x]^1]^2\ [\texttt{fn}\, y \Rightarrow [y]^3]^4\ ]^5$$

- application: connecting function entry and (body) exit with the argument but:
- also $[\texttt{fn}\, y \Rightarrow [y]^3]^4$ is a candidate at 2! (according to $\hat{C}(2)$)
- function abstractions

- variables (occurrences of use)

$$
\begin{array}{rcl}
\{\texttt{fn}\, x \Rightarrow [x]^1\} & \subseteq & \hat{C}(2) \\
\{\texttt{fn}\, y \Rightarrow [y]^3\} & \subseteq & \hat{C}(4) \\
\hat{\rho}(x) & \subseteq & \hat{C}(1) \\
\hat{\rho}(y) & \subseteq & \hat{C}(3) \\
\{\texttt{fn}\, x \Rightarrow [x]^1\} \subseteq \hat{C}(2) \quad \Rightarrow & \hat{C}(4) & \subseteq \hat{\rho}(x) \\
\{\texttt{fn}\, x \Rightarrow [x]^1\} \subseteq \hat{C}(2) \quad \Rightarrow & \hat{C}(1) & \subseteq \hat{C}(5) \\
\{\texttt{fn}\, y \Rightarrow [y]^3\} \subseteq \hat{C}(2) \quad \Rightarrow & \hat{\mathbf{C}}(\mathbf{4}) & \subseteq \hat{\rho}(y) \\
\{\texttt{fn}\, y \Rightarrow [y]^3\} \subseteq \hat{C}(2) \quad \Rightarrow & \hat{C}(3) & \subseteq \hat{\mathbf{C}}(\mathbf{5})
\end{array}
$$

The example is rather small, but contains all constructs of the language (ignoring `let`), therefore it illustrates how a labeled program gives rise to a constraint system, in this case for control-flow analysis. It's still a bit of an illustration only, as we don't give the actual construction, we just show how it works in the given example.

The slides (but not the script) present the construction in *2 stages* (for didactic reasons). The first stage gives "standard", **unconditional** constraints. The resulting system is fine in the sense that all solutions to it are *safe approximations* of the problem. The second stage is an improvement in that it gives more precise solutions. It is formulated with *conditional* constraints.

For programmes in the given calculus, there are **3 different kinds** of constraints, since there are three classes of syntactial constructs. The classification is done according to the top-level construct at the label being considered. Remember that abstract syntax represents trees, and now the labels indentify *subtrees*. The three cases are the following:

Abstractions: We have two abstractions, at label 2 and 4. The constraints in this category are relevant for the $\hat{C}$, only, as there are no variables "involved" (only the bound variables inside the function definition). Therefore, the abstract environment is not mentioned here. The inequations are easy. They just state that the expression at the given label (which labels, as said, an abstraction), *at least* evaluates to that abstraction. That's pretty obvious.

Variables: The constraints for variables *connects* the $\hat{C}$ and the abstract environment. Basically it relates directly the $\hat{C}$ for the label with the $\hat{\rho}$ variable carrying that label. We do not write to the variable, but **read** from it and get the value from the variable into the program (at the given label). Consequently the information flows *from* the variable *to* the "surrounding" label. As we have a forward-may-analysis and go for the smallest value, the *"post-state" (at l)* must be equal or **larger** than the *"pre-state" (at x)*.

Applications: We have only one application (labelled 5) in the example. So we have to think of the flow there. One is that the argument "flows into" the variable which constitutes the formal argument of the function. That coversthe **input** into the function. The same reasoning (intuitive direction of the flow of information) determines the direction of the inequation. The *second* flow is the **output:** what comes *out of the body of the function resp. out of the application.* That relates the $\hat{C}$ of the corresponding labels. This explains the first two inequations (which are not yet complete, there will be 2 steps to capture the flow (more precisely)). The first step deals with the fact

that we don't really know which function a variable represents (higher-order). The second one is an optimization, introducing *conditional* constraints.

However, that is **not all** for applications. So far we have just looked "intuitively" at the code, but we need to relate to the analysis/an algorithm. This means, we should not just look at what function there "is" according to our intuition at label 2 (the place for the function in the application) but formally we need to think, what functions location 2 evaluates **according to the analysis**. In other words, we need to consult $\hat{C}(2)$. We have only a **lower bound** for $\hat{C}(2)$, therefore, in principle *all* possible abstractions (there are 2) are candidates to be at the applicator position in the application. I.e., the one *missing* is $[\texttt{fn}\, y \Rightarrow [\mathbf{y}]^{\mathbf{3}}]^4$. This adds *two more* inequations, again one for input and the other for output, this time for the second abstraction.

What we have so far *would work* in the sense of leading to a sound analysis. As mentioned before, we can do better, however, in the category for applications. Namely that we formulate the inequations **conditionally** making the information about which function for $l = 2$ evaluates to a *more precise* (smaller) solution. In the example, we will see that the analysis finds out that $\hat{C}(2)$ indeed only contains *one* function, the obvious one. This additional precision gives indeed a smaller solution.

The refinement can be explained as follows: it concerns the category for applications (location 5). Each "instance" of an application leads to *two* constraints "in" and "out" (for parameter passing and returning the data). The question only is (as discussed): *which* function definition is actually meant where the data-flows "in" and "out". What we know is that the function, whichever it is, is placed at location 2. Therefore the condition in the conditional refinement of the constraints expresses the following

> "if such-and-such function occurs at 2, then consider the "in" constraint (and "out" constraint) for that function"

1.2.0.6 The least (= best) solution

$$
\begin{array}{rcl}
\hat{C}(1) & = & \{\texttt{fn}\, y \Rightarrow [y]^3\} \\
\hat{C}(2) & = & \{\texttt{fn}\, x \Rightarrow [x]^1\} \\
\hat{C}(3) & = & \emptyset \\
\hat{C}(4) & = & \{\texttt{fn}\, y \Rightarrow [y]^3\} \\
\hat{C}(5) & = & \{\texttt{fn}\, y \Rightarrow [y]^3\} \\
\hline
\hat{\rho}(x) & = & \{\texttt{fn}\, y \Rightarrow [y]^3\} \\
\hat{\rho}(y) & = & \emptyset
\end{array}
$$

One interesting bit here in the solution is: $\hat{\rho}(y) = \emptyset$: that means, the variable y is never evaluated, i.e., the function is not applied at all.

1.3 Type and effect systems

1.3.1 Introduction

1.3.1.1 Standard type systems and others

In this section, we are mainly dealing with "non-standard" type systems, but let's have a look at *standard* type systems first.

Types and type systems are a well-established and central part of static analysis. Before continuing further: As far as the lecture is concerned, we are dealing mainly with **static** type systems. The typing part of the semantic analysis phase is also kind of familiar, as they are often visible in the programming language and the programmer has to learn to deal with them. For languages being typed at run-time (for instance some scripting languages), types may be less visible, even if they are still there (but dynamically typed languages are mostly out of scope for this lecture anyway). In any case, even novice programmers are aware that somewhere under the hood, the compiler checks whether the given program adheres to the language's typing discipline; that's the task of the type checker.

The "standard" role of types is to describe "data values" and type checking assures that "meaningful" use is made of the data. At the lowest level, it serves the compiler, for instance, to know the "size" of data so as to allocate an adequate amount of memory for storing and accessing it. With the memory places of the right size and the bit patterns fitting to the ones chosen to represent data of the type, there are no run-time problems when working with the bits.

The lecture here covers types mostly on a higher abstraction level, and types are seen as *specifications* of allowed uses of data. For instance, the type system may allow to use the logical operations `and` and `or` on boolean typed values, but not addition. On that level, type and type systems are very rudimentary and potentially very rigid (but still important). Modern (standard) type system are far more complex, basically due to the wish or need to add flexibility, for instance, offering different forms of so-called **polymorphism**, but still maintaining efficient, safe, scalable, and decidable static type checking.

Type theory —the study of type systems, their expressiveness, efficiency, etc.— often deals with functional languages, one reason being that functional languages feature one particular expressive form of data, namely functions. There are more fundamental reasons why type systems for functional languages have been widely studied, but that's perhaps outside the scope of the lecture and we leave it at that.

At any rate, whether it's the more down to earth "size-of-memory" basic types for code generation, or advanced polymorphic type systems for some λ-calculus or other, standard type systems are always concerned with specifying sets of *values* (and functional languages, functions *are* also data), all with the purpose of regulating their usage.

So far for *standard* type systems, what about **non-standard** ones? The latter term here is used for all aspects *different* from describing *values* in a programming language. A value of a program is "what comes out at the end", that's why executing a program is sometimes

called **evaluation**... Another name we will encounter is **reduction**, the intuition being that the program is "reduced" to its final *value*. Anyhow, non-standard type systems, in contrast, specify aspects that are not related to the final value. For instance, data flow or control flow information is certainly something that typically is not covered by type systems. *Effects*, in particular, cover aspects that happen **"during"** the execution. This is a very broad characterization. Most things that need semantic analysis refer to "what happens when the program runs" (as opposed to "what's the final value" for which there are standard type systems). And so, most things can be described effect systems (and the rest by standard type system ...) Often, the description is combined with a type system, in which case it's called a *type and effect system*. The lecture will probably cover different effects (exceptions, communication, ...).

For Haskell progammers, the separation between the "pure", functional part and all "the other aspects" could sound familiar. Haskell goes to great lengths to separate both aspects, encapsulating all impure effects in so-called *monads*. So one might call standard types all that captures the pure (effect-free) core of the language, and effects correspond to those covered by monads. The connection between effects and monads can even be made formal (giving a monadic interpretation to effects), but that's beyond this lecture.

Let's look at a very general level at type systems (first standard ones). They are, as said, a "classical" static analysis and they are systems to check or derive types. We write (in this section) often sometimes t for "programs", or e ("expressions"), or S ("statements" when talking about imperative programs. To express that a term t is of type T, we use symbol : ("is of type"or "has type"). This relationship between programs and types will be formalized by a **type system**. However, stipulating $t : T$ is not really meaningful in isolation.

Fragments and sub-parts of a programs generally contain **free** variable, whose content and type are determined **elsewhere** in the overall program. Maybe variables are declared before they are used, or the definition is found in a surrounding scope, or in some other way. When focusing on a part t of a program (for instance trying to determine its type), the **rest** of the program is called its **context**. The type of t resp., the fact whether or not t is well-typed depends on its context. For instance, the term `x + 14` is well-typed (with type integer) if and only if `x` is typed by an integer. And the type of `x` is introduced elsewhere (often it's required to be declare before, not afterwards).

Consequently, type checking is not a **context-free** (or syntactic) analysis, but a **context-sensitive**.[6]

So type systems derive claims like *"program t is of type T given a context Γ"*. That is notationally often written as follows

$$\Gamma \vdash t : T \tag{1.20}$$

It's also called a **judgement** or **statement**, and the **context** is often also called **environment.**

[6] It leaves open the possibility that a type system is undecidable, and hence no longer context-sensitive, but typically one tries to avoid static type systems that are this complex. And we certainly do.

We called the context of a program fragment t it's context. However, type systems won't make use of all of t's context as appearing in the program, it just remembers in Γ information **relevant** for the type checking task at hand. For instance, if the type system is based on a "declare-before-use" discipline, the relevant part of the context need to contain only information that appers *before* t, obviously. And also for that, not all is needed. Normally, what's needed for Γ is information which types various **variables** have. So the relevant context information is "which variable has which type". Γ is thus often also called **type** context or **type** environment (as opposed to just context or environment). If one wants to implement such a type system, one needs a data structure that realizes the association of variable names with information (here type information). In concrete implementations, that data structure is known as **symbol table**. That's a data structure, often realized as hash table or similar, used to keep relevant information about syntactic entities during the semantical phase (and also for code generation), in particular about variables or other "symbols". One prominent piece of information is about the types, and that's what Γ contains. When coming to effect systems and some non-standard types, also other kind of information may be stored in the types, and therefore in Γ, as well.

- Γ: *assumption* or *context* or type environment
- here: *"non-standard"* type systems: effects and annotations
- natural setting: typed languages, here: *trivial!* setting (while-language)

A type system is a classical **context sensitive** analysis, following parsing, which is the classical context-free analysis. It's thus not a coincidence, that the component Γ in the judgements is called "context". In implementations, it directly corresponds to the so-called *symbol table*.

Remark 1.3.1. *For the participants for the compiler construction course (INF5110): information attached to syntactic elements were there also called* attributes*. In that parlance, a type is an attribute of a variable (or an expression etc.). Note in passing, that also the labelling of expressions can well be seen as attribuation.*

Context-sensitive information assosciated with syntactic entities was called attributes basically because in the static analysis part, the lecture was working with so-called attribute grammars*. In this lecture we don't bother with that general formalism. The type and effect systems will be represented in the form of* derivation systems*, but in priciple that can be viewed as some special notation for specific attribute grammars, as well.* □

1.3.1.2 "Trival" type system

- setting: while-language
- each statement maps: state to states
- Σ: type of *states*

judgement

$$\vdash S : \Sigma \to \Sigma \tag{1.21}$$

- specified as a *derivation* system

- note: *partial* correctness assertion

The "type system" here is *trivial* in a technical sense (not just meant to say the type system is "super-easy", which it also is). Type systems are used to *distinguish* well-typed programs from *ill-typed* ones, and reject the latter. This is generally not a context-free task, and therefore cannot be done by the parser alone, even though, for simple languages, type checking can be done *while* parsing. Here, the "type system" (shown below) accepts *all* programs, all syntactically correct programs are already well-typed (remember that we are dealing with abstract syntax, i.e., trees, which represent syntactically correct programs). Since all programs are well-typed, there is actually no need for a type system whatsoever. It's only used here to illustrate how to *extend* a (in this case trivial) type system with extra information, leading to *annotated* type system or an *effect type system*, which then yields useful information. Note in passing that the above judgment from equation (1.21) has no context Γ, reflecting the fact that it's technically context-free and not context-sensitive, and thereby trivial.

The typing judgment, as is the case in general for standard type system, is intended to be a *partial correctness* assertion. The trivial type system here is a bad illustation for that fact (being so trivial), but in general, the meaning of a judgment $e : T$ is:

> *if* the statement, expression or progam e terminates, then the resulting value conforms to the type T.

Type systems typically don't attempt have an opinion about *termination*, they only guarantee that the data at the end is ok, *provided* the program terminates thereby yielding said value. This restricted form of assertion is known as *partial correctness* (as opposed to total correctness). Another word used for such specifications is, that typing is conventionally concerned with *safety* properties (as opposed to *liveness* properties).

This is not intended to say, that it's impossible to devise type systems that try to capture "total correctness" in that they would guarantee termination or would warn against possible non-termination. In general, that would require "non-standard" augmented information.

Remark 1.3.2 (Termination)**.** *As a final side remark: there is a connection between termination and standard type system in the following way: for typed λ-calculi (without recursion), well-typed programs guarantee termination, which requires non-trivial techniques for proving that, and furthermore makes pure, typed λ-calculi (without recursion) no real programming languages insofar as they are not Turing-complete. The λ-calculus we will encounter later does have a recursion operator for that reason.* □

As mentioned, the "type system" does not do anything useful. It shows, however, the general *style* of writing down type systems, namely in the form of derivation rules. In the current version, there are five rules, one for each construct. Each rule has a set of *premises* and one *conclusion*. In case, there are *no* premises, a rule is also called *axiom*. That's the case for Ass and Seq, which are dealing with the *basic* constructs of the language. The compound statement are treated by the rules with a non-empty set of premises, where the premises deal with the sub-constructs.

$$\vdash [x := a]^l : \Sigma \to \Sigma \qquad \textsc{Ass}$$

$$[\mathsf{skip}]^l : \Sigma \to \Sigma \qquad \textsc{Skip}$$

$$\frac{\vdash S_1 : \Sigma \to \Sigma \qquad S_2 : \Sigma \to \Sigma}{\vdash S_1; S_2 : \Sigma \to \Sigma}\ \textsc{Seq} \qquad ,$$

$$\frac{\vdash S : \Sigma \to \Sigma}{\vdash \mathtt{while}[b]^l\ \mathtt{do}\ S : \Sigma \to \Sigma}\ \textsc{While}$$

$$\frac{\vdash S_1 : \Sigma \to \Sigma \qquad \vdash S_2 : \Sigma \to \Sigma}{\vdash \mathtt{if}[b]^l\ \mathtt{then}\ S_1\ \mathtt{else}\ S_2 : \Sigma \to \Sigma}\ \textsc{Cond}$$

Table 1.4: "Trivial" type system: rules

The rules can be viewed a logical "implications", reading them from top to bottom: If S_1 is well-typed and S_2 is well-typed, then so is the sequential $S_1; S_2$ composition (in rule Seq). A program S is well-typed, if there *exists* a derivation tree using the givem rules such that $\vdash S : \Sigma \to \Sigma$ is *derivable*, i.e., is the root of the tree (derivation trees have their roots at the bottom and their leaves, corresponding to axioms, at the top...)

One can view the rules also as the specification of a recursive procedure: in order to establish that $S_1; S_2$ is well-typed, one has to recursively check S_1 and S_2 for well-typedness. It's of course just a different angle on the same thing. This "reading" of the rules is sometimes called "goal-directed": in order to establish the conclusion, establish the premises first.

Seen in this goal-directed manner, the rules directly can be seen as a recursive procedure (corresponding to a tree-traversal of the abstract syntax tree). It's a very straight-forward divide and conquer strategy. Note in this context: the language has 5 syntactic constructs (2 basic ones and 3 compound ones) and the derivation system has 5 rules (including 2 axioms), exactly one for each contruct.

That entails that the top-level constuct of a subprogram *determines* which rule to apply recursively (in the goal-directed reading of the rules). Furthermore, the premises always deals with proper *subterms* compared to the term or statement in the conclusion, which in particular guarantees termination. Type systems, or derivation systems in general with these "one-rule-per-construct" and "in-the-premises-subterms-only" properties are called *syntax-directed*. In the terminology of attribute grammars: the types here, using the rules in a goal-directed manner, correspond to *synthesized* attributes (in their simplest form). But keep in mind that the "type system" here is trival to the point of being meaningless. More realistic type systems do *not* correspond to synthesized attributes, they may be more complex (and they may not be syntaxt-directed).

Not all type systems are given in a syntax-directed manner. That means, not all type system specifications can be directly understood as an algorithm. Sometimes, that is

for fundamental reasons: the rules describe a type system so complex that the question whether $\Gamma \vdash e : T$ is derivable or not is *undecidable*. We won't see much of that in this lecture. A more common reason is: the rules of the type system are not a priori *intended* as an *algorithm*, they are rather intended as *specification* of the typing discipline, and showing an algorithm would obscure that specification (normally the specification of the type system reads simpler and clearer). A non-syntax directed specification for instance will be used when dealing with *subtyping* or other forms of polymorphism, resp. *subeffecting*. To illustrate the point, let's have a look at subtypong. One could for instance specify the following:

$$\frac{\Gamma \vdash T_1 \leq T_2 \qquad \Gamma \vdash T_2 \leq T_3}{\Gamma \vdash T_1 \leq T_3}\ \text{S-Trans}$$

That's a crystal clear specification requiring that subtyping is supposed to be *transitive*: If T_1 is a subtype of T_2, which in turn is a subtype of T_3, then T_1 is indeed a subtype of T_3. That's what the rule unmistakingly stipulates when reading it as "implication" from the premises to the conclusion. When trying to interpret in a goal-directed manner, the reading goes: To establish that T_1 is a subtype of T_3, one can "break down" the problem to the subproblems to check that T_1 is a subtype of T_2 and T_2 a subtype of T_3. The problem just is: what has been said, going from the goal in the conclusion to the subgoals in the premises cannot really be called an act of "breaking down" the problems. It's not clear in which way the problem has become smaller in the premises, they are not really sub-problems, and it's equally unclear which T_2 to take. Actually, interpreting the rule in a goal directed manner just say "find such a T_2 in the middle" to establish the goal. Typically, systems have infinitely many types, just randomly trying out all candidates is a no-go.

We will not dig deeper into the issue here, it's meant to illustrate the point that there is often a gap between the type system, and its algorithmic realization. Later in the lecture we touch upon this question again: how can one turn a type system specification it into a syntax-directed, thus algorithmic version. Sometimes it's far from trivial, sometimes impossible (the type system may be undecidable), but there is a standard route one can try, which we may look at, —and we don't go into too complex systems, anyway.

1.3.1.3 Types, effects, and annotations

$$\vdash S : \Sigma_1 \to \Sigma_2 \qquad (1.22) \qquad\qquad \vdash S : \Sigma \xrightarrow{\varphi} \Sigma \qquad (1.23)$$

type and effect system (TES)

- *effect* system + *annotated* type system
- borderline fuzzy
- **annotated type system**
  - Σ_i: property of state ("$\Sigma_i \subseteq \Sigma$")
  - "abstract" properties: invariants, a variable is positive, etc.
- **effect system**
  - "statement S maps state to state, with (potential ...) effect φ"

– *effect* φ: e.g.: errors, exceptions, file/resource access, ...

1.3.2 Annotated type systems

In this section we use again the **reaching definitions** analysis for the while-language. We formulate the analysis in two different flavors (remember equations (1.22) and (1.23).

$$\vdash S : \mathsf{RD}_1 \to \mathsf{RD}_2 \qquad (1.24) \qquad\qquad \vdash S : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma \qquad (1.25)$$

Here we see that the border line between annotated type systems and effect system is fuzzy. The first sub-flavor corresponds to the intuition we have used so far: the states are restricted. The second one, if we think of it as functional type, can be seen as if the "functional" type is annotated. However, the annotation is like an effect (as we will see later). Anyway,

1.3.2.1 RD with annotated base types

The information of interest, as before, consists of pairs of variables and labels, i.e., sets $\mathsf{RD} \subseteq 2^{\mathbf{Var}\times\mathbf{Lab}}$. We use $\mathit{init}(S)$ to denote the (unique) label at the entry of S and $\mathit{final}(S)$ for the set of labels at the exit(s) of S. And the judgements from the derivation system are of the form $\vdash S : \mathsf{RD}_1 \to \mathsf{RD}_2$ (see equation (1.24)).

Intended"meaning" of a judgment $\vdash S : \mathsf{RD}_1 \to \mathsf{RD}_2$
"RD_1 is the set of var/label reaching the entry of S and RD_2 the corresponding set at the exit(s) of S":

$$\begin{aligned} \mathsf{RD}_1 &= \mathsf{RD}_{\mathit{entry}}(\mathit{init}(S)) \\ \mathsf{RD}_2 &= \bigcup\{\mathsf{RD}_{\mathit{exit}}(l) \mid l \in \mathit{final}(S)\} \end{aligned} \qquad (1.26)$$

Concerning the "meaning" of the judgment: the formulation is **not 100% correct** (to say it more bluntly, it's incorrect when taking things precisely ...), it's too strict as we will see.[7] The problem is the claim that RD is **the** set of ..... As in the data flow section, there is not just one single *safe* solution, but many. There is (as before) exactly one *minimal*, i.e., best one. However, the effect system is "lax" in that it specifies all safe ones. That is completely in analogy to the previous constraint system approaches.

One may compare the sets RD to the analysis data used in the "original" reaching definitions analysis (in the equational or constraint based approach, it does not matter which). The "functional" type here expresses the pre- and post-states. For elementary blocks, that corresponds to the entry and the exit point of the node or label. As one rule (coming next) treats one block/statement (elementary or not) at a time, it's not the *12-tuple*, of course, (taking the *concrete* factorial example) but just 2 generic slots of it: the pre- and the post-state.

[7] The corrected version will be shown in equation (1.3.2.2).

Concerning the auxiliary functions (initial and final): They calculate *implicitly* the *control-flow graph*.

$$\vdash [x := a]^{l'} : \mathsf{RD} \to \mathsf{RD} \setminus \{(x, l) \mid l \in \mathbf{Lab}\} \cup \{(x, l')\} \qquad \textsc{ass}$$

$$\vdash [\mathsf{skip}]^{l} : \mathsf{RD} \to \mathsf{RD} \qquad \textsc{skip}$$

$$\frac{\vdash S_1 : \mathsf{RD}_1 \to \mathsf{RD}_2 \qquad \vdash S_2 : \mathsf{RD}_2 \to \mathsf{RD}_3}{\vdash S_1; S_2 : \mathsf{RD}_1 \to \mathsf{RD}_3} \textsc{Seq}$$

$$\frac{\vdash S_1 : \mathsf{RD}_1 \to \mathsf{RD}_2 \qquad \vdash S_2 : \mathsf{RD}_1 \to \mathsf{RD}_2}{\vdash \mathtt{if}[b]^{l}\,\mathtt{then}\,S_1\,\mathtt{else}\,S_2 : \mathsf{RD}_1 \to \mathsf{RD}_2} \textsc{If}$$

$$\frac{\vdash S : \mathsf{RD} \to \mathsf{RD}}{\vdash \mathtt{while}[b]^{l}\,\mathtt{do}\,S : \mathsf{RD} \to \mathsf{RD}} \textsc{While}$$

$$\frac{\vdash S : \mathsf{RD}'_1 \to \mathsf{RD}'_2 \qquad \mathsf{RD}_1 \subseteq \mathsf{RD}'_1 \qquad \mathsf{RD}'_2 \subseteq \mathsf{RD}_2}{\vdash S : \mathsf{RD}_1 \to \mathsf{RD}_2} \textsc{Sub}$$

Table 1.5: Rules for RD with annotated base types

The rules may be compared with the constraint-based formulation of the reaching definitions early, which was based on the control-flow graph. The **intra-block** equations are covered by the axioms here, and the **inter-block** equations are the compositional part, the rules.

Worth mentioning is also the fact that in the annotated type system now, we have **6 rules** (for 5 syntactic constructs). So, there goes the syntax-directedness out of the window... As a consequence, the type system does not (directly) describe an algorithm.

The culprit is rule Sub (for *subsumption*). We will encounter subsumption many times (for subeffecting, and also subtyping), and it's one classical reason why type systems are not syntax directed. It breaks syntax-directedness in two ways in two ways: first, for each construct, there are now *two rules* to choose from (if one thinks in a goal-directed manner), as subsumption is *always* possible. as well. Secondly, the core judgment in the premise does not assert well-typedness for a proper *sub-term* of S in the conclusion. So, going from conclusion to the premise, S does not get "smaller". Suddenly, termination of type checking may become a non-trivial issue.

1.3.2.2 Meaning of annotated judgments

Let's go back to the point where it was said equation (1.26) that the meaning of the judgement $\vdash S : \mathsf{RD} \to \mathsf{RD}'$. Let's looking at conditionals, for instance in the following simple example:

$$\mathtt{if}[b]^{l}\,\mathtt{then}\,[x := y]^{l_1}\,\mathtt{else}\,[y := x]^{l_2} \tag{1.27}$$

Now, that the two branches definitely have different results: one branch adds $(x, \{l_1\})$ and the other one $(y, \{l_2\})$ (and both potentially remove stuff) and in particular delete potential entries concerning the variables that the other *does not delete*. Even if the statement is embedded in some larger context (loop), the example makes sure, that at the end of the conditional, the RD-sets are different (if RD represents the intended **smallest** solution). The conditional rule, however, insists in its premises that the to branches have *the same* pre/post information. So, especially for the premises, covering the two branches, the given interpretation, namely that RD represents "the" reaching definition" at the beginning and the end, is not correct, if we wish to have that IF-rule as given. And indeed, we should should have been warned. There is not something as "the" RD information. There is a thing like *the smallest* RD *information*, i.e., smallest safe approximation of the corresponding data flow constraints. The discussion here hightlights the fact, that the type and effect system does *not* capture that, it captures *all* safe solutions.

That means, the interpretation of the two branches could be

$$\vdash \mathsf{RD}_1\ [x := y]^{l_1}\ \{(x, l_1), (y, ?), (z, ?)\} \text{ and } \vdash \mathsf{RD}_2\ [y := x]^{l_2}\ \{(x, ?), (y, l_2), (z, ?)\} \quad (1.28)$$

with $\mathsf{RD}_1 = \{(x, ?), (y, ?), (z, ?)\}$, then RD_2', when assuming that we are dealing with x, y, and z, all yet "uninitialized". Both seems reasonable in isolation, and $\{(x, ?), (y, l_2), (z, ?)\}$ and $\{(x, l_1), (y, ?), (z, ?)\}$ presents "the" flow information at the end of each branch, in the sense that it's the *smallest* safe one. But the derivation system is not designed to derive the smallest one, it derives *all* safe ones. It thus corresponds to the setting with control-flow graphs and constraints; also there, the corresponding contraint systems (both the equational formulation and the one based on $\subseteq$) had multiple solutions, but unique best ones. What the derivation system thus allows to derive is therefore safe approximations of one we are after, which is the smallest, and actually a bit more, the derivation system allows to derive *all* safe approximation which implies that since all such safe solutions are derivable, the smallest one is derivable as well, it's only a matter how to find it.[8] The laxness of the derivation system can be phrased as follows: If the system derives a solution, then it derives a **larger one as well**. And that's exactly what is captured by the **subsumption** or **weakening** rule SUB. The rule SUB is needed in the xample to **"reconcile"** the two different outcomes from the branches in the two judgements from equation (1.28) to a joint result for the conditional from equation (1.27), taking the union of the two outcomes at the end

$$\{(x, l_1), (y_1, l_2), (z, ?)\} \quad (1.29)$$

In the discussion we focused on the reaching definitions at the end (one could also call it *post-condition*). However, what about the *pre-condition?* In this particular small example: what justifies the presence of z? Nothing, really, when looking at the conditional of the conditional in isolation in the context of a larger program. So the pre-condition RD_1 is to be interpreted as *assumption*: **if** the execution of the conditional from equation 81.27) starts with reaching definition information $\mathsf{RD}_1 = \{(x, ?), (y, ?), (z, ?)\}$ as precondition, **then** the reaching definition information from (1.29) is a possible, i.e., safe post-condition. Actually, it's the smallest safe one, *given* the assumed precondition.

[8] That assumes that a smallest solution exists, but the underlying theory guarantees that.

After all this discussion, we can reformulate the first attempt from equation from 1.26 into the following more "lax" and conditional formulation:

Meaning of a judgment $\vdash S : \mathsf{RD}_1 \to \mathsf{RD}_2$
if RD_1 is a set of var/label reaching the entry of S, then RD_2 is a corresponding set at the exit(s) of S:

$$\begin{array}{rrcl} & \text{if} \quad \mathsf{RD}_1 & \subseteq & \mathsf{RD}_{entry}(init(S)) \\ \text{then} & \forall l \in final(S).\ \mathsf{RD}_{exit}(l) & \subseteq & \mathsf{RD}_2 \end{array} \tag{1.30}$$

The role of the subsumption or weakening rule is shown also in the following example, in connection with a loop:

Example 1.3.3 (Factorial). Let's revisit the factorial from Example 1.1.2. One may also compare with the control-flow graph from Figure 1.1.

To fit the derivation on a page, we use the following abbreviation: $f_2 = \texttt{while}\ldots;[y := 0]^5$. Since we consider the we program as complete, we use as global and initial precondition $\mathsf{RD}_0 = \{?_x, ?_y, ?_z\}$. The smallest derivable postcondition for the program is $\mathsf{RD}_{final} = \{?_x, 5, 1, 3\}$. It is the smallest (given the precondition), because we used the subsumption rule as disciplined as possible, and not overly generous.

$$\dfrac{[y := x]^0 : \mathsf{RD}_0 \to \{?_x, 1, ?_z\} \qquad \dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[z := _]^3 : \mathsf{RD}_{body} \to \{?_x, 0, 4, 3, \cancel{1}\}}{[y := _]^4 : \{?_x, 1, 5, 4\} \to \{?_x, 4, 3\}}}{f_{body} : \mathsf{RD}_{body} \to \{?_x, 4, 3\}}}{f_{body} : \mathsf{RD}_{body} \to \mathsf{RD}_{body}}\ \textsc{Sub}}{f_{while} : \mathsf{RD}_{body} \to \mathsf{RD}_{body}}}{f_{while} : \{?_x, 0, 1\} \to \mathsf{RD}_{body}}\ \textsc{Sub} \qquad [y := 0]^5 : \mathsf{RD}_{body} \to \mathsf{RD}_{final}}{f_2 : \{?_x, 0, 1\} \to \mathsf{RD}_{final}}}{f_1 : \{?_x, 0, ?_z\} \to \mathsf{RD}_{final}}}{f : \mathsf{RD}_0 \to \mathsf{RD}_{final}}$$

The example needs subsumption for the internal loop. Rule While requires that the pre- and the post-condition of the body of the loop is the same. In other words one has to **find** a **loop invariant!**

$$\mathsf{RD}_{body} = \{?_x, 0, 4, 1, 3\}$$

□

1.3.3 Annotated type constructors

The section covers an alternative to the approach from the previous Section 1.3.2. Instead of adding information to the base "types" Σ, the arrow type constructor itself is annotated, writing $\xrightarrow{\varphi}$. The annotation φ is the **effect**, i.e., a description of what happens during the execution of the statement. The two formulation are rather similar: The effect-annotation

on the $\rightarrow$-constructor directly captures the effect, whereas in the previous formulation, the effect is implicit in the **change** of the state-information attached.

For the reaching definitions, the effect is split into to parts, RD, which is the reaching definition information that reaches the end of the statement. Unlike in the version with annotated types, it's only the information the statement itself is responsible for, it's not relative to assuming a flow as pre-condition. This information is not enough, one needs also "negative"information which flow is taken away. This corresponds to the set of programming variables that are assigned to or defined in the statement. These to parts of the annotation will later be called the **generate** and the **kill** part of data flow analyses. This will be typical for data-flow analysis for imperative languages and is not specific for reaching definitions.

For no particular reasons, the two parts of the reaching definition analysis are written (at least here) on top and below the arrow, leading to **judgements** of the form

$$S : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma \ . \tag{1.31}$$

For the intented meaning, again it's not an exact statement ("the effect"), but lax, describing all possible safe approximations:

> **Intended meaning of $S : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma$:** "$S$ maps states to states, where RD is the set of reaching definitions, S **may** produce (= generate) and X the set of var's S **must** unavoidably assign (= kill)."

The rules for the analysis are shown in Table 1.6. Again, there is an subsumption rule and the rule for conditionals requires that both branches have the same effect, resp. the rule for loops requires that the information in the loop body is invariant.

Note that the two pieces of information, the RD and the X-parts work in opposite direction. It's safe to **overapproximate** RD and to **underapproximate** X. That corresponds also to the formulation of the intended meaning, where RD is the information the statement **may** generate, and X the information the statement **must** kill. In connection with that: the set above the arrow is $\emptyset$. That's because the while-loop may not be taken at all, so there is no variable that is unavoidable defined.

1.3.4 Effect systems

Now we switch back to a functional language, some variant of the λ-calculus. Unlike for the while-language, we will have a real underlying type system, not a trivial one. The underlying type system uses bool, int, and $T \rightarrow T$ as **types**, and **judgements** of the form:

$$\Gamma \vdash e : \tau \ . \tag{1.32}$$

$$[x := a]^l : \Sigma \xrightarrow[\{(x,l)\}]{\{x\}} \Sigma \quad \textsc{Ass} \qquad [\mathsf{skip}]^l : \Sigma \xrightarrow[\emptyset]{\emptyset} \Sigma \quad \textsc{Skip}$$

$$\frac{S_1 : \Sigma \xrightarrow[\mathsf{RD}_1]{X_1} \Sigma \qquad S_2 : \Sigma \xrightarrow[\mathsf{RD}_2]{X_2} \Sigma}{S_1; S_2 : \Sigma \xrightarrow[\mathsf{RD}_1 \setminus X_2 \cup \mathsf{RD}_2]{X_1 \cup X_2} \Sigma} \textsc{Seq}$$

$$\frac{S_1 : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma \qquad S_2 : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma}{\mathtt{if}[b]^l \,\mathtt{then}\, S_1 \,\mathtt{else}\, S_2 : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma} \textsc{If}$$

$$\frac{S : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma}{\mathtt{while}[b]^l \,\mathtt{do}\, S : \Sigma \xrightarrow[\mathsf{RD}]{\emptyset} \Sigma} \textsc{While}$$

$$\frac{S : \Sigma \xrightarrow[\mathsf{RD}']{X'} \Sigma \qquad X \subseteq X' \qquad \mathsf{RD}' \subseteq \mathsf{RD}}{S : \Sigma \xrightarrow[\mathsf{RD}]{X} \Sigma} \textsc{Sub}$$

Table 1.6: Derivation rules

The part Γ from the judgement is the *(type) environment* (or (type) context, a mapping from variables to types; in corresponds to the symbol table in practical implementation, as mentioned earlier.

The language here basically is known as the **simply typed λ-calculus**. It is a variant which is known as *Curry-style* formulation of the type system. That refers to the fact that abstractions are written as $\lambda x.e$ instead of $\lambda x{:}T.e$. In the latter case, it's know as "Church style". In the less explicit Curry-style, one is facing a problem of so-called **type inference**, i.e., figuring out what the (omitted) type of x should be. That in itself is an important and interesting problem, which we might cover later, but in the introduction, we focus on the *effect* part, not the typing.

As far as the syntax is concerned: The π-subscript of the abstraction is *non-standard* as far as the typed λ-calculus and its typing is concerned. It's added for the purpose of the effect system later, only.

In comparison to the earlier "type-system" we used a starting point for the while language, now the system is non-trivial (even if completely standard and simple).

As mentioned, the flavor of the (simply typed) λ-calculus here is Curry-style. As before, an easy way to interpret the rules is as derivation rules, where the premises imply the conclusion.

It may be also instructive to look at them in a goal-directed manner, as if it were a recursive procedure: what is needed in order to derive a conclusion.

$$\frac{\Gamma(x) = \tau}{\Gamma \vdash x : \tau}\ \textsc{Var}$$

$$\frac{\Gamma, x{:}\tau_1 \vdash e : \tau_2}{\Gamma \vdash \ \mathtt{fn}_{\,\pi} x \Rightarrow e : \tau_1 \to \tau_2}\ \textsc{Abs}$$

$$\frac{\Gamma \vdash e_1 : \tau_1 \to \tau_2 \qquad \Gamma \vdash e_2 : \tau_1}{\Gamma \vdash e_1\ e_2 : \tau_2}\ \textsc{App}$$

Table 1.7: Simply typed λ-calculus: type system

But even more instructive is to think what such a recursive interpretation is supposed to achieve. To be useful as conventional type checker, it's not meant as to **check** if the following is true

$$\Gamma \vdash e : T\ ? \tag{1.33}$$

like confirming the user's guess whether or not it's true that e has type T (given the assumptions Γ). Instead, more useful is the interpretation

$$\Gamma \vdash e : ? \tag{1.34}$$

i.e., an answer to the question:

> is e well-typed under the given assumption, and if so, what's its type?

In other words, when interpreting the rules, one useful reading is to see Γ and e as **input** of a recursive procedure, and T as the **output**.

With this interpretation in mind, especially rule Abs is interesting. Here, in the Curry-style formulation, the formal parameter x is mentioned in the conclusion **with out type**. That means, that the "recursive call" corresponding to the premise **guesses** a type T_1 for it. This guess may be right in that it leads to a successful type check of the premise and thus a successful type check and returning $T_1 \to T_2$ for the abstraction in the conclusion. Or it may also fail. The form of abstraction may determine that a recursive call according rule App is in place; however, not enough information is given in the input Γ and $\lambda x.e$ to determine the subsequent recusive call (as T_1 has to be guessed). Note that there are *infinitely* many types to choose from.... In other words, the rules (in the interpretation given) do **not really qualify as algorithm**, and they would not qualify as syntax-directed (if we interpret the rules as describing a problem as in equation (1.34) and not as *confirming* the type as in equation (1.33).

Solving the "guessing problem" is known as **type inference** or **type reconstruction** and we might return to it later. Also note: if the abstraction would be of the form $\lambda x{:}T.e$, the guessing problem would go away as well. This form is also known a *Church-style typing*, as opposed to Curry-style.

1.3.4.1 Effects: Call tracking analysis

The underlying type system is concerned with "classical" types, providing information of the *value*. Indeed, being purely functional, expressions in λ-calculus have no side effects. Still, one can consider *effects* in functional languages. Effects are generally things that occur *during* execution, as opposed to the resulting values that is officially returned upon termination. Side effects are classical effects (for not purely functional settings): executing a program has the effect of changing the contents of variables, resp. the memory. Another effect that is even sometimes part of the official, usual-visible type system are **exceptions**. For instance, when a method in Java may throw an exception, that may have to be declared as part of its type information. In that sense, it's a type-and-effect system. Functional languages also feature *exceptions*. When a procedure "gives back" an exception, that's not considered that it's a standard return value, and hence it's considered an effect (and thus a functional language or λ-calculus with effect can be counted as no longer purely functional).

However, even in a purely functional setting without any effects, one can analyze what happens during execution, i.e., one can have an effect analysis. Here, we are doing **call tracking** which means

> Determine for each subexpression: which function abstractions may be applied, i.e., called, **during** the subexpression's evaluation.

It's slightly different (but related) to the constraint-based control-flow analysis we had before (in Section 1.2), where the question was for each application, which function it could potentially be applied to.

Functions are labelled or annotated (for the purpose of tracking them), so the effect information is a *set* of such function labels or names. **Judgements** of the analysis are of the form

$$\hat{\Gamma} \vdash e : \hat{\tau} :: \varphi \tag{1.35}$$

Remark 1.3.4. *The types are now written $\hat{T}$, not T, and contexts now $\hat{\Gamma}$ instead of Γ. Maybe that's overdoing it notationally, but it's meant to highlight, that we are no longer dealing with the standard types, but annotated types, and analogously with context that contain annotated types. Sometimes we just write "types" and "context" even if we mean the annotated version.*

The annotation is attached to the $\rightarrow$-constructor, i.e., annotated types look as follows:

$$\hat{T}_1 \xrightarrow{\varphi} \hat{T}_2 \tag{1.36}$$

The base types int and bool are not annotated, as the are not "executed" and thus cannot call procedures. The φ in the judgment from equation (1.35) is the effect of e, i.e., describing functions that are potentially called when evaluating e. What about the annotation on the arrow type from equation (1.36)? It's analogous effect information, bit with a slight twist. Let's look at a bit more concrete example:

$$\hat{\Gamma} \vdash e : \mathsf{int} \xrightarrow{\varphi_1} \mathsf{bool} :: \varphi_2 \tag{1.37}$$

The type describes a function from booleans to integers; the expression e is an expression that evaluates (if it terminates) to a λ-expression and during the evaluation from e to $\mathtt{fn}_\pi x \Rightarrow e$, it may call procedures in φ_1. Expression $\mathtt{fn}_\pi x \Rightarrow e$ is also of type $\mathsf{int} \xrightarrow{\varphi_1} \mathsf{bool}$ of course:

$$\hat{\Gamma} \vdash \mathtt{fn}_\pi x \Rightarrow e : \mathsf{int} \xrightarrow{\varphi_1} \mathsf{bool} :: \varphi_2 \tag{1.38}$$

But $\mathtt{fn}_\pi x \Rightarrow e$ is a value (i.e., the value that e has evaluated to). That means that evaluating it has **no effect**, no function are being called. So instead of the type and effect information from the previous equation (1.38), we could derive the more precise and thus better information

$$\hat{\Gamma} \vdash \mathtt{fn}_\pi x \Rightarrow e : \mathsf{int} \xrightarrow{\varphi_1} \mathsf{bool} \ :: \emptyset \tag{1.39}$$

But what about φ_1? That's the procedures that are potentially call in the case (if ever) the procedure π will be called. This is also called the procedures **latent effect**.

Remark 1.3.5 (Preservation of typing under evalution)**.** *The text previously remarked that the effect type of e is "of course" also the effect type of the λ-abstraction in equation (1.38).*

*It's indeed an **essential** property of static type systems that a **reduction preserves typing**. That's what "static" means: one determines statically the type, i.e., figures out what the type is (or the types are) that an expression can evaluate or reduce to. Then when reducing the expression (till the end or just partially) must lead to an expression which still has this type, otherwise the type originally determined had been wrong (i.e.. unsafe), and the type system incorrect. So having a type system that enjoys type preservation is actually synonymus with having a static type system (a correct one, that is).*

But what about effects? Did we not say that the effect of the final value is smaller*, when fully evaluated as in equation (1.39) the effect is empty? That's correct, but as we stressed earlier (in connection with subsumption) the analysis allows to derive* all *save static approximations, and larger ones of a safe one are also safe (when doing overapproximation as for the analysis here). So while the result of the analysis is preserved, it's common that the* best *solution improves when the program evolves. That's to be expected. Static analysis makes a prognosis of that the program will do when run or what it will evaluate into, but when it has partially run already the prognosis can become more clear, resp. the information more precise.*

Actually, the fact that the the static information can become more precise when the program evolves is not only characteristic for effects, it applies often also to standard, conventional types, at least in type system when programs can legitimately have multiple types. That's in particular the case for type systems supporting subtyping*. For instance, an expression* $\mathtt{if}\, x\, \mathtt{then}\, 1\, \mathtt{else}\, 2.4$ *could have a type* `float`*, but if it evolves (when x is true) to* 1*, it's best = smallest type would be* `int` *(and still* `float` *as well), if one has a type system with* $\mathtt{int} \leq \mathtt{float}$*.* □

It may be worthwhile to reflect what the connection is between the call-tracking analysis here and the constraint-based analysis we had before. Both analyses are not the same. In the constraint-based analysis from earlier, we wanted to know *at each point* to which

functions it evaluates to answer the question for applications $f\ a$, which functions are actually called. Here, the perspective is different: we take an expression and think of it as being evaluated and ask, which functions are being called *during evaluation*. Both questions, however, are related. Remember that the analysis earlier was *not* formulated as a type system.

Later in the lecture, we may revisit the control-flow analysis in a type-based formulation. This will be easier to compare to the call-tracking analysis from here, basically because it's formulated similarly, namely as a type-based derivation system quite similar to the one here. This allows us to see the differences and similarities more clearly. The earlier control-flow analysis was *not* formalized as type system, but with constraints. It's not, however, meant to mean that constraint based systems are fundamentally different from type and effect systems. After all they may tackle the same problems. Again it's more a stylistic question and type systems as the one here can also be seen as a particular way (using logical derivation rules as used for type system) to specify the constraint system.

$$\frac{\hat{\Gamma}(x) = \hat{\tau}}{\hat{\Gamma} \vdash x : \hat{\tau} :: \emptyset}\ \textsc{Var}$$

$$\frac{\Gamma, x{:}\hat{\tau}_1 \vdash e : \hat{\tau}_2 :: \varphi}{\Gamma \vdash \mathtt{fn}_\pi x \Rightarrow e : \hat{\tau}_1 \xrightarrow{\varphi \cup \{\pi\}} \hat{\tau}_2 :: \emptyset}\ \textsc{Abs}$$

$$\frac{\hat{\Gamma} \vdash e_1 : \hat{\tau}_1 \xrightarrow{\varphi} \hat{\tau}_2 :: \varphi_1 \qquad \hat{\Gamma} \vdash e_2 : \hat{\tau}_1 :: \varphi_2}{\hat{\Gamma} \vdash e_1\ e_2 : \hat{\tau}_2 :: \varphi \cup \varphi_1 \cup \varphi_2}\ \textsc{App}$$

Table 1.8: Call tracking rules

Example 1.3.6 (Call tracking). Let's use as a very simple example the application of the identity function to another instance of the idendity function.

$$\dfrac{\dfrac{x{:}\mathsf{int} \xrightarrow{\{Y\}} \mathsf{int} \vdash x{:}\mathsf{int} \xrightarrow{\{Y\}} \mathsf{int} :: \emptyset}{\vdash (\mathtt{fn}_X x \Rightarrow x) : (\mathsf{int} \xrightarrow{\{Y\}} \mathsf{int}) \xrightarrow{\{X\}} (\mathsf{int} \xrightarrow{\{Y\}} \mathsf{int}) :: \emptyset} \qquad \vdash (\mathtt{fn}_Y y \Rightarrow y) : \mathsf{int} \xrightarrow{\{Y\}} \mathsf{int} :: \emptyset}{\vdash (\mathtt{fn}_X x \Rightarrow x)\ (\mathtt{fn}_Y y \Rightarrow y) : \mathsf{int} \xrightarrow{\{Y\}} \mathsf{int} :: \{X\}}$$

The small derivation shows a small concrete deriviation. It's worth to look at the ultimate effect, the set $\{X\}$. The "important" information there is that Y is *not* mentioned, which means that the correspondingly-labelled function is *not* called. □

1.4 Algorithms

This part is rather short. So far, we touched upon here and there on issues how to algorithmically treat the problems we encountered. In particular in the context of the type systems, some remarks tried to raise awareness under which circumstances the derivation rules could straightforwardly be interpreted as a recursive procedure, and when not

(basically in all interesting cases). But we never really tackled the issue of obtaining an algorithm, especially not for the "flow problems" (data flow, effects etc.).

In particular: we saw how data flow problems such as reaching definitions can be described or specified as constraint systems (equational or otherwise). What we did *not* do so far is giving hints of how to actually *solve* such constraint system, i.e., how do do *constraint solving*.

The notion of "constraint system" is extremely broad and can capture all sorts of problems. The lecture is mostly concerned with particular forms of constraint systems which capture flow problems (or program analysis problem in general). In addition, we discuss under which circumstances those kind of constraint systems have solutions, in particular under which circumstances they have *unique best* solutions. Having that is a very welcome sitation and deserve a thorough treatment.

Here, we just hint at a very simple strategy (also only sketched on a very high-level), which is non-deterministic. Being *non-deterministic* makes it not really directly an implementation unless one wishes to make use of some random-generator which helps to implement the non-determinism. There would be in practice no point in doing so. Instead one would aim for *deterministic* solution, perhaps realizing specific strategies or heuristics.

Why then bother with a non-deterministic description at all? Basically it's to separate the question of *correctness* of the algorithm from the question of *efficiency*. As it turns out, the kind of constraints we are considering and the solution domains have the *very, very* desirable property that

> when facing a non-deterministic, choice, no matter how one resolve it, the one never make a "wrong choice".

So the choice does not really matter except for how fast the algorithm terminates, i.e., how efficient the algo runs. Even there are perhaps smarter and less smart choices at each point, there are not real *wrong* ones. As a consequence, there is no **backtracking**. And that's crucial for being usable. As a side remark: it's well-known that many forms of constraints, even simple ones as boolean constraints ("SAT-solving") don't have this favorable property and there is basically no way around solving those by brute force *combinatorial exploration*. Indeed, SAT (boolean satisfiability) one of the most famous NP complete problems there is (and the first one for which that was proven). Fortunately, for simple data flow equations, things look much brighter.

1.4.0.1 Chaotic iteration

- back to data flow/reaching def's
- goal: **solve**

$$\vec{RD} = F(\mathsf{RD}) \quad \text{or} \quad \vec{RD} \sqsubseteq F(\vec{\mathsf{RD}})$$

- F: monotone, finite domain

straightforward approach

init $\vec{\mathsf{RD}}_0 = F^0(\emptyset)$

iterate $\vec{\mathsf{RD}}_{n+1} = F(\vec{\mathsf{RD}}_n) = F^{n+1}(\emptyset)$ until stabilization

- approach to implement that: **chaotic iteration**
- non-deterministic stategy
- abbreviate:

$\vec{\mathsf{RD}} = (\mathsf{RD}_1, \ldots, \mathsf{RD}_{12})$

1.4.0.2 Chaotic iteration (for RD)

Input: equations for reaching defs
for the given program
Output: least solution: $\vec{\mathsf{RD}} = (\mathsf{RD}_1, \ldots, \mathsf{RD}_{12})$

Initialization:
$\mathsf{RD}_1 := \emptyset; \ldots; \mathsf{RD}_{12} := \emptyset$
Iteration:
while $\mathsf{RD}_j \neq F_j(\mathsf{RD}_1, \ldots, \mathsf{RD}_{12})$ for some j
do
$\mathsf{RD}_j := F_j(\mathsf{RD}_1, \ldots, \mathsf{RD}_{12})$

1.5 Conclusion

The introductory part touched upon different topics. The approaches are also related, i.e., it's sometimes a bit of a matter of preference if one represents a problem directly as flow equations, type systems etc. Things not covered in the introduction (but probably later are complications in the while language, like procedure calls or pointers). Also, there will be other analyses besides reaching definitions (and a systematic common overview over similar analyses (knows as monotone framework). Also we go into the underlying theory (lattices) as well has considering in which way to establish that the various analysis are *actually* a sound overapproximation of the program behavior ("soundness", "correctness", "safe approximation", all mean the same).

Bibliography

[1] Cardelli, L. and Wegner, P. (1985). On understanding types, data abstraction and polymorphism. *Computing Surveys*, 17(4):471–522.

[2] Nielson, F., Nielson, H.-R., and Hankin, C. L. (1999). *Principles of Program Analysis*. Springer Verlag.

[3] Robinson, J. A. (1965). A machine-oriented logic based on the resolution principle. *Journal of the ACM*, 12:23–41.

[4] Vytiniotis, D., Peyton Jones, S., and Schrijvers, T. (2010). Let should not be generalised. In *Proceedings of TDLI'10*.

Index

λ-calculus, 4
    typed, 32
$\Gamma \vdash t : T$, 31

abstract cache, 26
abstract environment, 26
algorithms, 44
analysis
    control-flow, 22
attribute, 31
attribute grammar, 31
axiom, 32

call tracking analysis, 42
chaotic iteration, 45
Church-style typing, 41
constraint solving, 45
constraint system, 45
context-sensitive, 31
control flow graph, 7
control-flow, 22
control-flow analysis, 22
control-flow graph, 36
correctness
    partial, 32
Curry-style typing, 41

derivation rule, 32
desugaring, 3

effect, 30
effect system, 34
evaluation, 29

factorial, 38
fix-point, 18
FORTRAN, 2
functional language, 22

Haskell, 30
higher-order function, 22

inference
    type, 40

judgment, 31

labelling, 25
lattice, 12
    complete, 20

monad, 30

partial correctness, 32
polymorphism, 29
pre-fixpoint, 22

reaching definitions, 9
reduction, 29

static type system, 29
subsumption, 36
symbol table, 31
syntactic sugar, 3
syntax-directed, 33

total correctness, 32
type and effect system, 34
type inference, 40
type system, 29, 32
    static, 29
type theory, 29
type vs. effect, 29

value, 29