Lecture Notes:
Control Flow Analysis for Functional Languages

17-355/17-665/17-819: Program Analysis (Spring 2020)

Claire Le Goues*
clegoues@cs.cmu.edu

We have made progress by expanding our dataflow analysis to handle programs with multiple procedures. However, the approach we've developed relies on a number of simplifying assumptions. Notably, in WHILE3ADDR with functions, it is always easy to tell which function is being called at any particular callsite. This is often *not* the case in real languages. Object-oriented languages (or any language with dynamic dispatch) and functional languages challenge this assumption: in both cases, it can be difficult to tell which function is being called, statically.

We therefore turn now to the general problem of statically analyzing functional languages. In doing so, we will see techniques for addressing this general question of determining control flow (or call graphs), and generalize several of our ideas about dataflow analysis (like the idea of a program point). Additionally, analyzing functional languages motivates and provides a good introduction to *constraint-based analyses*. We will additionally expand on a number of these ideas in subsequent classes.

1 A simple, labeled, functional language

Consider an idealized functional language based on the lambda calculus, similar to the core of Scheme or ML, with the additional property that we *label* all expressions:

e	$\in$	$Expressions$	...or labelled terms
t	$\in$	$Term$	...or unlabelled expressions
l	$\in$	$\mathcal{L}$	labels

e	::=	t^l
t	::=	$\lambda x.e$
	|	x
	|	$(e_1)\,(e_2)$
	|	let $x = e_1$ in e_2
	|	if e_0 then e_1 else e_2
	|	$n \mid e_1 + e_2 \mid \ldots$

The grammar includes a definition of an anonymous function $\lambda x.e$, where x is the function argument and e is the function body.[1] The function can include any of the other types of ex-

*These notes were developed together with Jonathan Aldrich

[1]The formulation in PPA also includes a syntactic construct for explicitly recursive functions. The ideas extend naturally, but we'll follow the simpler syntax for expository purposes.

pressions, such as variables x or function calls $(e_1^a)(e_2^b)$,[2] where e_1 is the function to be invoked and e_2 is passed to that function as an argument (labeled a and b respectively). We evaluate a function call $(\lambda x.e)(v)$ by substituting the argument v for all occurrences of x in e. For example, $((\lambda x.(x^a + 1^b)^c)^d(3)^e)^g$ evaluates to $3 + 1$, which of course evaluates to 4. A more interesting example is $(((\lambda f.(f^a\ 3^b)^c)^e(\lambda x.(x^g + 1^h)^i)^j)^k$, which first substitutes the argument for f, yielding $(\lambda x.x^g + 1^h)^i\ 3$. Then we invoke the function, getting $3 + 1$ which again evaluates to 4.

Note that this grammar associates each expression with a label $l \in \mathcal{L}$; this is important to keeping track of analysis information (analogous to program points in our imperative analysis), as we discuss next.

2 Simple Control Flow Analysis

Static analysis can be just as useful in this type of language as in imperative languages, but immediate complexities arise. For example: what is a *program point* in a language without obvious predecessors or successors? Computation is intrinsically nested. Second, because functions are first-class entities that can be passed around as variables, it's not obvious which function is being applied where. We need some way to figure this out, because the value a function returns (which we may hope to track, such as through constant propagation analysis) will inevitably depend on which function is called, as well as its arguments.

Control flow analysis (CFA)[3] seeks to statically determine which functions could be associated with which variables. Because functional languages are not based on statements but rather expressions, it is appropriate to reason about both the values of variables and the values expressions evaluate to.

2.1 0-CFA

We will start by discussing the simplest form of a CFA, called 0-CFA. This is the simplest form because it is context-insensitive (the "0-" label indicates no context is taken into account). We track analysis information for variables and labels, in lieu of the explicit program points in the control flow graphs we used before. Although this may feel like a big change, this approach actually connects directly to what we've been doing in imperative dataflow analysis so far. Dataflow analysis is a type of *abstract interpretation*, an overall framework or theory of sound approximation of program semantics. At a high level and separate from a particular program definition, abstract interpretation associates *labels* with *properties* by manipulating sets of states using monotonic functions over ordered sets as defined by lattices. In our formulation for imperative languages, we implicitly associated labels with the program points between nodes in a control flow graph.

That said, our analysis information σ maps each variable and label to a lattice value. 0-CFA analysis is only concerned with tracking which functions are possibly associated with each location or variable (we will add dataflow information later), and so the abstract domain is as follows:

$$\sigma \in \mathit{Var} \cup \mathcal{L} \to L \qquad L = \top + \mathcal{P}(\lambda x.e)$$

The analysis information at any given expression is the set of all functions that could be the result of evaluating that expression. As suggested above, expressions are identified by their labels

[2] In an imperative language this would more typically be written $e_1^a(e_2)^b$, but we follow the functional convention here, with parenthesis included when helpful syntactically.

[3] This nomenclature is confusing because it is also used to refer to analyses of control flow graphs in imperative languages; We usually abbreviate to CFA when discussing the analysis of functional languages.

l, and we track similar information for variables. We use $\top$ to denote all possible functions; if we know all the functions in the program, we could enumerate them, but a symbolic $\top$ representation is useful when we don't have the whole program available.

Question: what is the $\sqsubseteq$ relation on this dataflow state?

A 0-CFA is a *Constraint Based Analysis*: it is defined via inference rules that generate constraints over the possible dataflow values for each variable or labeled location; those constraints are then solved. We use the $\hookrightarrow$ to define constraint generation. The judgment $[\![e]\!]^l \hookrightarrow C$ can be read as "The analysis of expression e with label l generates constraints C over dataflow state σ." For our first CFA, we can define inference rules for this judgment as follows:

$$\frac{}{[\![x]\!]^l \hookrightarrow \sigma(x) \sqsubseteq \sigma(l)} \textit{ var}$$

In this rule, the variable value flows to the program location l. Although we didn't list it above (we generalize it below), a rule for constants produces the empty set, because this analysis is tracking only function values.

The rules for functions/calls is more complex:

$$\frac{[\![e]\!]^{l_0} \hookrightarrow C}{[\![\lambda x.e^{l_0}]\!]^l \hookrightarrow \{\lambda x.e\} \sqsubseteq \sigma(l) \cup C} \textit{ lambda}$$

$$\frac{[\![e_1]\!]^{l_1} \hookrightarrow C_1 \quad [\![e_2]\!]^{l_2} \hookrightarrow C_2}{[\![e_1^{l_1}\ e_2^{l_2}]\!]^l \hookrightarrow C_1 \cup C_2 \cup \textbf{fn}\ l_1 : l_2 \Rightarrow l} \textit{ apply}$$

The first rule just states that if a literal function is declared at a program location l, that function is part of the lattice value $\sigma(l)$ computed by the analysis for that location. Because we want to analyze the data flow inside the function, we also generate a set of constraints C from the function body and return those constraints as well.

The rule for application first analyzes the function and the argument to extract two sets of constraints C_1 and C_2. We then generate an abstract *function flow constraint* of the form $\textbf{fn}\ l_1 : l_2 \Rightarrow l$. This function flow constraint is interpreted by the constraint solver to generate additional concrete constraints using the following rule:

$$\frac{\lambda x.e_0^{l_0} \in \sigma(l_1)}{\textbf{fn}\ l_1 : l_2 \Rightarrow l \hookrightarrow \sigma(l_2) \sqsubseteq \sigma(x) \wedge \sigma(l_0) \sqsubseteq \sigma(l)} \textit{ function-flow}$$

This rule states that for every literal function $\lambda x.e_0^{l_0}$ that the analysis (eventually) determines the expression labeled l_1 may evaluate to, we must generate additional constraints that capture value flow from the actual argument expression l_2 to formal function argument x, and from the function result to the calling expression l.

Consider the first example program given above: $((\lambda x.(x^a + 1^b)^c)^d(3)^e)^g$. The first rule to use is *apply* (because that's the top-level program construct). We will work this out together, but the generated constraints could look like:

$$(\sigma(x) \sqsubseteq \sigma(a)) \cup (\{\lambda x.x + 1\} \sqsubseteq \sigma(d)) \cup (\sigma(e) \sqsubseteq \sigma(x)) \wedge (\sigma(c) \sqsubseteq \sigma(g))$$

There are many possible valid (typically referred to as *acceptable*) solutions to this constraint set. Eliding the formalities, it suffices to say that we would like the least solution to these constraints,

as that will be the most precise result. We will return to constraint solving properly later in the course; for now, we will simply assert that a σ that maps all variables and locations except d to $\varnothing$, and d to $\{\lambda x.x + 1\}$, satisfies this set of constraints.

Question: what might the rules for the if-then-else or arithmetic operator expressions look like?

2.2 0-CFA with dataflow information

The analysis in the previous subsection is interesting if all you're interested in is which functions can be called where, but doesn't solve the general problem of dataflow analysis of functional programs. Fortunately, extending that approach to a more general analysis space is straightforward: we simply add the abstract information we're tracking to the abstract domain defined above. For constant propagation, for example, we can extend the dataflow state as follows:

$$\sigma \in \mathit{Var} \cup \mathit{Lab} \to L \qquad L = \mathbb{Z} + \top + \mathcal{P}(\lambda x.e)$$

Now, the analysis information maps each program point (or variable) to an integer n, or $\top$, or a set of functions. This requires that we modify our inference rules slightly, but not as much as you might expect. Indeed, the rules mostly change for arithmetic operators (which we omitted above) and constants. We simply need to provide an abstraction over concrete values that captures the dataflow information in question. We get the following rules:

$$\frac{}{[\![n]\!]^l \hookrightarrow \alpha(n) \sqsubseteq \sigma(l)}\ \textit{const} \qquad \frac{}{[\![e_1^{l_1} + e_2^{l_2}]\!]^l \hookrightarrow (\sigma(l_1) +_\top \sigma(l_2)) \sqsubseteq \sigma(l)}\ \textit{plus}$$

Where α is defined as we discussed in abstract interpretation, and $+_\top$ is addition lifted to work over a domain that includes $\top$ (and simply ignores/drops any lambda values). There are similar rules for other arithmetic operations.

Consider the second example, above, properly labeled: $(((\lambda f.(f^a\ 3^b)^c)^e(\lambda x.(x^g + 1^h)^i)^j)^k$ A constant propagation analysis could produce the following results:

$\mathit{Var} \cup \mathit{Lab}$	L	by rule
e	$\lambda f.f\ 3$	lambda
j	$\lambda x.x + 1$	lambda
f	$\lambda x.x + 1$	apply
a	$\lambda x.x + 1$	var
b	3	const
x	3	apply
g	3	var
h	1	const
i	4	add
c	4	apply
k	4	apply

3 m-Calling Context Sensitive Control Flow Analysis (m-CFA)

The control flow analysis described above quickly becomes imprecise in more interesting programs that reuse functions in several calling contexts. This problem should seem familiar from interprocedural imperative program analysis, but the following code illustrates the problem in this new language:

$$\textbf{let } \mathit{add} = \lambda x.\ \lambda y.\ x + y$$
$$\textbf{let } \mathit{add5} = (\mathit{add}\ 5)^{a5}$$
$$\textbf{let } \mathit{add6} = (\mathit{add}\ 6)^{a6}$$
$$\textbf{let } \mathit{main} = (\mathit{add5}\ 2)^{m}$$

This example illustrates *currying*, in which a function such as *add* that takes two arguments x and y in sequence can be called with only one argument (e.g. 5 in the call labeled $a5$), resulting in a function that can later be called with the second argument (in this case, 2 at the call labeled m). The value 5 for the first argument in this example is stored with the function in the *closure add5*. Thus when the second argument is passed to *add5*, the closure holds the value of x so that the sum $x + y = 5 + 2 = 7$ can be computed.

In this case, we create two closures, *add5* and *add6*, binding 5 and 6 and the respective values for x. 0-CFA analysis cannot distinguish them, and because it only computes one value for x we learn only that x has the value $\top$. This is illustrated in the following analysis (we shorten the trace to focus only on the variables):

Var $\cup$ *Lab*	L	notes
add	$\lambda x.\ \lambda y.\ x + y$	
x	5	when analyzing first call
$add5$	$\lambda y.\ x + y$	
x	$\top$	when analyzing second call
$add6$	$\lambda y.\ x + y$	
$main$	$\top$	

We can add precision using a context-sensitive analysis. One could, in principle, use either the functional or call-string approach we discussed previously. In practice the call-string approach is more commonly used for control-flow analysis in functional programming languages, perhaps because functional programs will typically produced an intractable number of contexts per function, and it is easier to place a bound on the analysis in the call-string approach.

We add context sensitivity by making our analysis information σ track information separately for different call strings, denoted by Δ. Here a call string is a sequence of labels, each one denoting a function call site, where the sequence can be of any length between 0 and some bound m (in practice m will be in the range 0-2 for scalability reasons):

$$\sigma \in (\mathit{Var} \cup \mathit{Lab}) \times \Delta \to L \qquad \Delta = \mathit{Lab}^{n \leqslant m} \qquad L = \mathbb{Z} + \top + \mathcal{P}((\lambda x.e, \delta))$$

When a lambda expression is analyzed, we now consider as part of the lattice the call string context δ in which its free variables were captured. We can then define a set of rules that generate constraints which, when solved, provide an answer to control-flow analysis, as well as (in this case) constant propagation:

$$\frac{}{\delta \vdash [\![n]\!]^l \hookrightarrow \alpha(n) \sqsubseteq \sigma(l, \delta)}\ \mathit{const} \qquad \frac{}{\delta \vdash [\![x]\!]^l \hookrightarrow \sigma(x, \delta) \sqsubseteq \sigma(l, \delta)}\ \mathit{var}$$

$$\frac{}{\delta \vdash [\![\lambda x.e^{l_0}]\!]^l \hookrightarrow \{(\lambda x.e, \delta)\} \sqsubseteq \sigma(l, \delta)}\ \mathit{lambda}$$

$$\frac{\delta \vdash [\![e_1]\!]^{l_1} \hookrightarrow C_1 \qquad \delta \vdash [\![e_2]\!]^{l_2} \hookrightarrow C_2}{\delta \vdash [\![e_1^{l_1}\ e_2^{l_2}]\!]^l \hookrightarrow C_1 \cup C_2 \cup \mathbf{fn}_\delta\ l_1 : l_2 \Rightarrow l}\ \mathit{apply}$$

These rules contain a call string context δ in which the analysis of each line of code is done. The rules *const* and *var* are unchanged except for indexing σ by the current context δ. Similarly, the *apply* rule is the same except we index everything by δ and record δ as part of the function flow constraint. The *lambda* rule now captures the context δ along with the lambda expression, so that when the lambda expression is called the analysis knows in which context to look up the free variables. But the rule no longer analyzes inside the function; we want to delay that and do it for a new context δ' when the function is called.

$$\frac{\begin{array}{c}(\lambda x.e_0^{l_0}, \delta) \in \sigma(l_1) \qquad \delta' = \mathit{suffix}(\delta{+}{+}l, m) \\ C_1 = \sigma(l_2, \delta) \sqsubseteq \sigma(x, \delta') \wedge \sigma(l_0, \delta') \sqsubseteq \sigma(l, \delta) \\ C_2 = \{\sigma(y, \delta_0) \sqsubseteq \sigma(y, \delta') \mid y \in FV(\lambda x.e_0)\} \\ \delta' \vdash [\![e_0]\!]^{l_0} \hookrightarrow C_3\end{array}}{\mathbf{fn}_\delta\ l_1 : l_2 \Rightarrow l \hookrightarrow C_1 \cup C_2 \cup C_3} \textit{ function-flow-}\delta$$

The function flow constraint has gotten a bit more complicated. A new context δ' is formed by appending the current call site l to the old call string, then taking the suffix of length m (or less). For each function that may be called, we set up constraints between the actual and formal parameters and the function result, as before (C_1). We analyze the body of the function in the new context δ' (C_3). Finally, we produce constraints that bind the free variables in the new context: all free variables in the called function flow from the point δ_0 at which the closure was captured.

We can now reanalyze the earlier example, observing the benefit of context sensitivity. In the table below, $\bullet$ denotes the empty calling context (e.g. when analyzing the *main* procedure):

Var / Lab, δ	L	notes
add, $\bullet$	$(\lambda x.\ \lambda y.\ x + y,\ \bullet)$	
x, a5	5	
add5, $\bullet$	$(\lambda y.\ x + y,\ a5)$	
x, a6	6	
add6, $\bullet$	$(\lambda y.\ x + y,\ a6)$	
main, $\bullet$	7	

Note three points about this analysis. First, we can distinguish the values of x in the two calling contexts: x is 5 in the context a5 but it is 6 in the context a6. Second, the closures returned to the variables *add5* and *add6* record the scope in which the free variable x was bound when the closure was captured. This means, third, that when we invoke the closure *add5* at program point m, we will know that x was captured in calling context a5, and so when the analysis analyzes the addition, it knows that x holds the constant 5 in this context. This enables constant propagation to compute a precise answer, learning that the variable *main* holds the value 7.

Optional: Uniform k-Calling Context Sensitive Control Flow Analysis (k-CFA)

m-CFA was proposed recently by Might, Smaragdakis, and Van Horn as a more scalable version of the original k-CFA analysis developed by Shivers for Scheme. While m-CFA now seems to be a better tradeoff between scalability and precision, k-CFA is interesting both for historical reasons and because it illustrates a more precise approach to tracking the values of variables in a closure. The following example illustrates a situation in which m-CFA may be too imprecise:

$$
\begin{array}{ll}
\textbf{let } \mathit{adde} & = \lambda x. \\
& \quad \textbf{let } h = \lambda y.\ \lambda z.\ x + y + z \\
& \quad \textbf{let } r = h\ 8 \\
& \quad \textbf{in } r \\
\textbf{let } t & = (\mathit{adde}\ 2)^t \\
\textbf{let } f & = (\mathit{adde}\ 4)^f \\
\textbf{let } e & = (t\ 1)^e
\end{array}
$$

When we analyze it with m-CFA, we get the following results:

Var / Lab, δ	L	notes
adde, $\bullet$	$(\lambda x..., \bullet)$	
x, t	2	
y, r	8	
x, r	2	when analyzing first call
t, $\bullet$	$(\lambda z.\ x + y + z,\ r)$	
x, f	4	
x, r	$\top$	when analyzing second call
f, $\bullet$	$(\lambda z.\ x + y + z,\ r)$	
t, $\bullet$	$\top$	

The k-CFA analysis is like m-CFA, except that rather than keeping track of the scope in which a closure was captured, the analysis keeps track of the scope in which each variable captured in the closure was defined. We use an environment η to track this. Note that since η can represent a separate calling context for each variable, it has the potential to be more accurate, but also much more expensive. We can represent the analysis information as follows:

$$
\begin{array}{lcllcl}
\sigma & \in & (\mathit{Var} \cup \mathit{Lab}) \times \Delta \to L & \Delta & = & \mathit{Lab}^{n \leqslant k} \\
L & = & \mathbb{Z} + \top + \mathcal{P}(\lambda x.e, \eta) & \eta & \in & \mathit{Var} \to \Delta
\end{array}
$$

Let us briefly analyze the complexity of this analysis. In the worst case, if a closure captures n different variables, we may have a different call string for each of them. There are $O(n^k)$ different call strings for a program of size n, so if we keep track of one for each of n variables, we have $O(n^{n*k})$ different representations of the contexts for the variables captured in each closure. This exponential blowup is why k-CFA scales so badly. m-CFA is comparatively cheap—there are "only" $O(n^k)$ different contexts for the variables captured in each closure—still exponential in k, but polynomial in n for a fixed (and generally small) k.

We can now define the rules for k-CFA. They are similar to the rules for m-CFA, except that we now have two contexts: the calling context δ, and the environment context η tracking the context in which each variable is bound. When we analyze a variable x, we look it up not in the current context δ, but the context $\eta(x)$ in which it was bound. When a lambda is analyzed, we track the current environment η with the lambda, as this is the information necessary to determine where captured variables are bound. The function flow rule is actually somewhat simpler, because we do not copy bound variables into the context of the called procedure:

$$\frac{}{\delta, \eta \vdash [\![n]\!]^l \hookrightarrow \alpha(n) \sqsubseteq \sigma(l, \delta)} \textit{ const} \qquad \frac{}{\delta, \eta \vdash [\![x]\!]^l \hookrightarrow \sigma(x, \eta(x)) \sqsubseteq \sigma(l, \delta)} \textit{ var}$$

$$\frac{}{\delta, \eta \vdash [\![\lambda x.e^{l_0}]\!]^l \hookrightarrow \{(\lambda x.e, \eta)\} \sqsubseteq \sigma(l, \delta)} \textit{ lambda}$$

$$\frac{\delta, \eta \vdash [\![e_1]\!]^{l_1} \hookrightarrow C_1 \qquad \delta, \eta \vdash [\![e_2]\!]^{l_2} \hookrightarrow C_2}{\delta, \eta \vdash [\![e_1^{l_1}\ e_2^{l_2}]\!]^l \hookrightarrow C_1 \cup C_2 \cup \mathbf{fn}_\delta\ l_1 : l_2 \Rightarrow l} \textit{ apply}$$

$$\frac{\begin{array}{c}(\lambda x.e_0^{l_0}, \eta_0) \in \sigma(l_1) \qquad \delta' = \textit{suffix}(\delta{+}{+}l, m) \\ C_1 = \sigma(l_2, \delta) \sqsubseteq \sigma(x, \delta') \wedge \sigma(l_0, \delta') \sqsubseteq \sigma(l, \delta) \\ \delta', \eta_0 \vdash [\![e_0]\!]^{l_0} \hookrightarrow C_2\end{array}}{\mathbf{fn}_\delta\ l_1 : l_2 \Rightarrow l \hookrightarrow C_1 \cup C_2} \textit{ function-flow-}\delta$$

Now we can see how k-CFA analysis can more precisely analyze the latest example program. In the simulation below, we give two tables: one showing the order in which the functions are analyzed, along with the calling context δ and the environment η for each analysis, and the other as usual showing the analysis information computed for the variables in the program:

function	δ	η
main	$\bullet$	$\varnothing$
adde	t	$\{x \mapsto t\}$
h	r	$\{x \mapsto t,\ y \mapsto r\}$
adde	f	$\{x \mapsto f\}$
h	r	$\{x \mapsto f,\ y \mapsto r\}$
λz....	e	$\{x \mapsto t,\ y \mapsto r,\ z \mapsto e\}$

Var / Lab, δ	L	notes
adde, $\bullet$	$(\lambda x..., \bullet)$	
x, t	2	
y, r	8	
t, $\bullet$	$(\lambda z.\ x + y + z,\ \{x \mapsto t,\ y \mapsto r\})$	
x, f	4	
f, $\bullet$	$(\lambda z.\ x + y + z,\ \{x \mapsto f,\ y \mapsto r\})$	
z, e	1	
t, $\bullet$	11	

Tracking the definition point of each variable separately is enough to restore precision in this program. However, programs with this structure—in which analysis of the program depends on different calling contexts for bound variables even when the context is the same for the function eventually called—appear to be rare in practice. Might et al. observed no examples among the real programs they tested in which k-CFA was more accurate than m-CFA—but k-CFA was often far more costly. Thus at this point the m-CFA analysis seems to be a better tradeoff between efficiency and precision, compared to k-CFA.