Eötvös Loránd Tudományegyetem
Informatikai Kar
Programozási Nyelvek és
Fordítóprogramok Tanszék

Applying slicing algorithms on large code bases

Tibor Brunner
doktorandusz

Olivér Hechtl
programtervező informatikus MSc

Budapest, 2017

Contents

Chapter 1

Introduction

> Everyone knows that debugging is twice as hard as writing a program in the first place. So if you're as clever as you can be when you write it, how will you ever debug it?
>
> *Brian Kernighan*

Nowadays, there are a lot of tools available for debugging. A developer can examine call stacks, variable informations, and so on. There are a lot of times when a bug is when some variable does not behave like the writer of that code part would expect. When the programmer discovers it, he must follow back the path of assignments, and find out where did it get an unexpected value. Program slicing targets this kind of debugging. We can define a program slice as a part of the program which contains a selected statement, and all other statements which influences it, or gets influenced by it. These two types are respectively called backward and forward slicing. Basically this is what many programmers do intuitively, when facing with bugs. In this thesis, I'll introduce a few approaches for computing slices, and describe a working prototype tool application, which can analyze C++ programs, and compute slices of them. I've used the help of the Clang/LLVM compiler infrastructure, which builds the AST of the program and provides an interface to analyse it using C++.

Chapter 2

Slicing, methods and efficiency

2.1 About slicing

For better understanding programs, programmers organize code into structures. They write sub-problems into functions, and organize variables and data to structs. Also, with object oriented design, they put these into classes. These are all good for separating the data, and the procedures on data. But these are not helpful when we need to examine a flow of data in the program. Slicing gets useful in this scenario. This is a program analysis technique introduced by Mark Weiser[1]. In his paper, he wrote: "Program slicing is a decomposition based on data flow and control flow analysis". We define slicing as a subset of the program, which only includes the statements which have transitive control or data dependency regarding the selected statement.

2.2 Types of slicing

There is two different type of slicing known: static and dynamic. While dynamic slicing gets the statements which could affect the selected statement at a particular execution of the program with a fixed input, static slicing examines it statically, including all possible statements which could affect that selected statement. In this thesis, I'll focus on static slicing methods. There are two different subtypes of static slicing, backward and forward. They are indicating the relevant statements' direction from our selected statement.

2.3 Methods for slicing

We can construct slices via various methods on different representations of the program. All of these are using some kind of graph structures, which can be traversed through for searching the transitive data dependences.

2.3.1 Dependences

Before I describe the various methods currently available for slicing, we must get to know what dependences in programs are. There are two kinds of dependences: control and data. They can be defined by the control flow graph (CFG) of the program. Given the following example:

```
int main(){
  int sum = 0;
  int i = sum;
  while (i < 11){
    sum += i;
    i++;
  }
}
```

Figure 2.1: Example program

It's control-flow graph can be seen on 2.2. This type of graph is created from the program by grouping the statements into basic blocks. Each basic block consists a maximal amount of consecutive statements without any jumps. In a high-level language, such as C++, a jump can be either a branch statement, like an `if` and a `switch-case`, or a loop statement, like a `while`. In the example, we can see that it's CFG consists three basic blocks, and two special blocks, namely `entry` and `exit`, which represent the entry and exit points of the program, respectively.

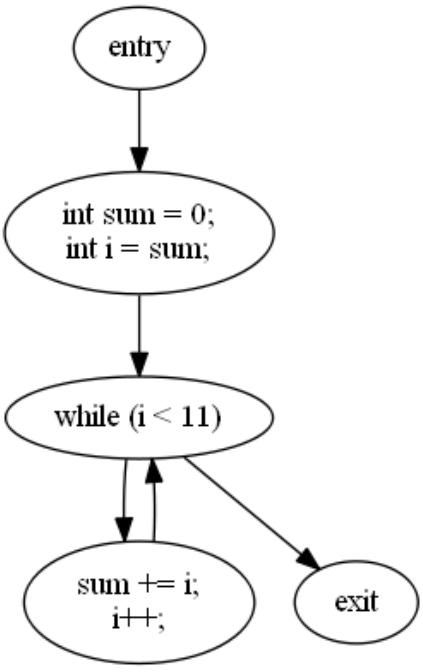

Figure 2.2: Control-flow graph

Therefore, control dependence can be defined in the knowing of post-dominance. As written in [2]:

"A node i in th CFG is *post-dominated* by

a node by j if all paths from i to STOP pass
through j. A node j is *control dependent* on a node i if (i) there exists a path P from i to j such that j post-dominates every node in P, excluding i and j, and (ii) is not post-dominated by j."

It basically means that every statement under a branch or loop are control dependent of the control statement's predicate. Excluding the unstructured and structured jump statements, `continue`,`break` and `goto`, drawing control dependences between statements creates a tree, with the root being the inspected function, and control statements form the branches. I discuss later in detail the control dependences created by structured jumps.

A data dependence between two statements means that if we change their order, the meaning of the program changes, or becomes ill-formed. There are two types of data dependences: *flow dependences* and *def-order dependences*. According to Horwitz[3], we can define these with the following rules:

There is a flow dependence between statement s1 and s2 if:

1. s1 defines variable x.
2. s2 uses x.
3. There is a path between s1 and s2 in program execution with no intervening definitions of x in between, and s2 can be reached by control from s1.

Definition in this context means a bit different than usual: it can be either an initial definition of a variable but an assignment where x is on the left side is also counts as a definition. In compiler theory, flow-dependence referred as reaching definition of a variable.

In the presence of loops, there are two further classification of flow-dependence: loop-carried and loop-independent. A flow-dependence is loop-carried if it satisfies all rules above and also:

1. includes a backedge to the predicate of the loop.
2. s1 and s2 are enclosed in the loop.

Backedge means that there is a flow dependence between the statement and the predicate of the loop, therefore the loop uses the variable which that statement defines. In the example above, the statement `i++` is like this.

Loop-independent flow-dependences can be described as the common ones which has no backedge to the loop predicate.

On the other hand, def-order dependence between statement s1 and s2 is different from loop-independent flow-dependences only in that s2 instead of using x, it also defines it. The other two rules stays the same.

These definitions are used in all slicing methods, but differently. I will elaborate these in each section.

2.3.2 Data flow equations

2.3.2.1 The algorithm

This method is created by Mark Weiser. It is based on the analysis of the CFG. As he wrote in [1], he defines a slice regarding a *slicing criterion*, which consists of a pair *(n, V)* where n is a statement of the program and V is a subset of the program's variables, which we are slicing on. He also writes that a slice of a program is an executable, which has only the relevant statements in it.

To calculate which statements should be included in the slice, he defines two sets of variables: *directly* and *indirectly relevant* variables. As written in [2], he provides the following equations for them:

For determining *directly* relevant variables and statements:

For each edge $i \rightarrow_{CFG} j$ in the CFG:

$$R_C^0(i) = R_C^0(i) \cup \{v | v \in R_C^0(j), v \notin \textsc{Def}(i)\} \cup \{v | v \in \textsc{Ref}(i), \textsc{Def}(i) \cap R_C^0(j) \neq \emptyset\}$$
$$S_C^0 = \{i | (\textsc{Def}(i) \cap R_C^0(j) \neq \emptyset, i \rightarrow_{CFG} j\}$$

And for determining *indirectly* relevant variables and statements:

$$B_C^k = \{b | \exists i \in S_C^k, i \in \textsc{Infl}(b)\}$$
$$R_C^{k+1}(i) = R_C^k(i) \cup \bigcup_{b \in B_C^k} R^0_{(b, \textsc{Ref}(b))}(i)$$
$$S_C^{k+1} = B_C^k \cup \{\textsc{Def}(i) \cap R_C^{k+1}(j) \neq \emptyset, i \rightarrow_{CFG} j\}$$

He also says that a slice is statement-minimal if it couldn't have less statements. The slice then is computed in two steps: First by determining *directly* relevant variables by the equation above. In it, $\textsc{Def}(i)$ and $\textsc{Ref}(i)$ means the variables that statement i defines or uses, respectively. $R_C^0(i)$ will contain the directly relevant variables of statement i. It starts with the values in the slicing criterion: $R_C^0(n) = V$, and for all other sets initialized as $\emptyset$. As it can be seen from the first equation above, to get the R_C^0 value for a node, it must use the computed values from every following node in the CFG. Weiser states in his work[1], that this is a simplification of the usual data flow information, which would use a PRE set to represent preservation

of variable values. Reading the first equation, we see that variable v is in the $R_C^0(i)$ set either if:

1. v is in the slicing criterion V,
2. or v is in REF(i) and the R_C^0 set of the following CFG node j has a value which is also in DEF(i)
3. or v is not in DEF(i) but v is in $R_C^0(j)$.

The last constraint ensures the propagation of the 'may be used' variables through statements which does not use nor reference the variables in the slicing criterion.

The second equation S_C^0 then contains the statements which are included in the slice. It goes through the CFG from start and collects all variables that may have an influence on the slicing criterion. I've listed the sets of variables which are in DEF, REF and R_C^0 of the statements of the example program above in the table below.

Node	DEF	REF	R_C^0
`int sum = 0;`	`sum`	$\emptyset$	$\emptyset$
`int i = sum;`	`i`	`sum`	`sum`
`while (i < 11)`	$\emptyset$	`i`	`i`
`sum += i;`	`sum`	`sum,i`	`i`
`i++`	`i`	`i`	`i`

Table 2.1: Results of the algorithm for the given example and criterion `(i++,{i})`

In this case, the set S_C^0 contains lines `2,3,6`. As we see, this slice is incomplete since it does not include the `while` statement. For this to be included, further iterations should be done. In the equations above, Weiser introduces the INFL set for every statement i, which consists all statements which are control dependent on i. In our example, it is the empty set for all statements except the `while`. Using this, he defines the B_C^k set, which includes all statements which have a statement in it's INFL set which is included in S_C^k.

Now using B_C^k, we can add the `while` statement since the slicing criterion is control dependent on it. This is defined in the last equation, which is almost the same as the second, but it also includes the B_C^k set in it. When facing with nested branch, or looping statements, multiple levels of indirect dependences must be considered. The R_C^{i+1} set contains these for each level. Because of this, Weiser's algorithm needs iterations for achieving the correct slice. The R_C^k and S_C^k sets are non-decreasing, and they reach a fixpoint which are called R_C and S_C, respectively. The final S_C set contains the slice for the criterion.

I must note that this method computes the *backward* slice for any criterion, and cannot be used for forward slicing, since it computes only the statements which have influence on the criterion.

2.3.2.2 Efficiency

Weiser states that the complexity for computing S_C is $O(neloge)$. He notes that this slice is not always the smallest, since it could include statements which cannot influence the slicing variable. He shows it in the following example program:

```
a = constant
while(p(k)){
  if (q(c)){
    b = a;
    x = 1;
  }else{
        c = b;
        y = 2;
  }
  k++;
}
z = x + y;
write(z)
```

For criterion `C = (write(z),{z})`, the set S_C will contain the first statement, but it does not have an influence on `z` since on any path by which `a` can influence `c` will execute `x = 1` and `y = 2`, therefore `z` becomes a constant value in this case.

However, computing slices with Weiser's algorithm needs little information from the program, since it requires only the CFG and the REF and DEF sets for every statement. The downside is that it only computes backwards slices, and since the criterion is fixed, it needs to compute every set in the equations if we need a slice for a different slicing criterion of the program.

2.3.3 Information flow relations

2.3.3.1 The algorithm

This algorithm is created by Bergeretti and Carré[4]. It has a different approach than Weiser's: it associates the program's statements with relations: that which expression may affect which variable. These relations has the definitions in the table below.

Empty statements

$D_S = \emptyset$

$\lambda_S = \emptyset$

$\rho_S = \iota$

$P_S = V$

$\mu_S = \emptyset$

Assignment Statements, in a form of `v = e`

$D_S = \{v\}$

$\lambda_S = \Gamma(e) \times \{e\}$

$\rho_S = \Gamma(e) \times \{v\} \cup (\iota - \{(v, v)\}$

$P_S = V - \{v\}$

$\mu_S = \{(e, v)\}$

Sequences of statements, in a form of A;B

$D_S = D_A \cup D_B$

$\lambda_S = \lambda_A \cup \rho_A \lambda_B$

$\rho_S = \rho_A \rho_B$

$P_S = P_A \cap P_B$

$\mu_S = \mu_A \rho_B \cup \mu_B$

Conditional statements, in a form: `if (e) A else B`

$D_S = D_A \cup D_B$

$\lambda_S = (\Gamma(e) \times \{e\}) \cup \lambda_A \cup \lambda_B$

$\rho_S = (\Gamma(e) \times (D_A \cup D_B)) \cup \rho_A \cup \rho_B$

$P_S = P_A \cup P_B$

$\mu_S = (\{e\} \times (D_A \cup D_B)) \cup \mu_A \cup \mu_B$

Conditional statements, without else: `if (e) A`

$D_S = D_A$

$\lambda_S = (\Gamma(e) \times \{e\}) \cup \lambda_A$

$\rho_S = (\Gamma(e) \times D_A) \cup \rho_A \cup \iota$

$P_S = V$

$\mu_S = (\{e\} \times D_A) \cup \mu_A$

Repetitive statements, in a form: `while (e) A`

$D_S = D_A$

$\lambda_S = \rho_A^*((\Gamma(e) \times \{e\}) \cup \lambda_A)$

$\rho_S = \rho_A^*((\Gamma(e) \times D_A) \cup \iota)$

$P_S = V$

$\mu_S = (\{e\} \times D_A) \cup \mu_A \rho_A^*((\Gamma(e) \times D_A) \cup \iota)$

Table 2.2: Table of information flow relations

The statements are classified into two primitive types: Empty and Assignment, and three hiearchical types: Sequence, Conditional with `else` branch, without `else` and the Repetitive statements. They are constructed from the control-flow graph. For analysis, they define *variables* and *expressions*.For Assignment, it's left side is the variable, and it's right side is the expression associated with it. Only this type of statement has a variable. For Conditional and Repetitive statements their conditional expression used, and Empty and Sequence has no variable or expression. There's also the definition of $\Gamma(e)$ which consists the variables which appear in e. We define E and V which contains all of the variables and expressions of the program. In D_S there are the variables which S *may define*. In P_S are the variables which S *may preserve*. The main three binary relations are the following:

1. λ_S, from V to E
2. μ_S, from E to V
3. ρ_S, from V to V

For a statement S, $v \in V, e \in E$: $v\lambda_S e$ means that 'the value of v on entry to S *may be used* in the evaluation of the expression e in S.' Taken the following program:

```
void fn(int x){
  read(n);              //1
  int i = 1;            //2
  while (i <= n){       //3
    if (i \% 2 = 0) {   //4
      x = 17;           //5
    }
    else{
      x = 18;           //6
    }
    i = i + 1;          //7
  }
  write(x);             //8
}
```

Figure 2.3: Example program 2

It can be seen that in case of the `while` statement, variable n and the predicate of `while` is in the $\lambda_{\texttt{while}}$ relation.

Taking a look at μ_S, it means that for expression e and variable v : $e\mu_S v$ 'a value of e in S *may be used* in obtaining the value of the variable v on exit from S'.

Looking at our example again, for the expression of `if` statement and variable x, $\mu_{\texttt{if}}$ holds, since the execution of assignments to x are control dependent on the `if` statement.

It is basically the dual relation of λ_S, which will be very useful in aspect of slicing.

The final relation ρ_S can be expressed in terms of λ_S,μ_S and P_S:

$$\rho_S = \lambda_S \mu_S \cup \Pi_S, \ \textit{where } \Pi_S = \{(v, v) \in V \times V \mid v \in P_S\}. \tag{2.1}$$

2.3.4 PDG based graph reachability

Chapter 3

LLVM/Clang infrastructure

3.1 About Clang

3.2 The Clang AST

3.3 AST Matchers

Chapter 4

Implementation and algorithm

4.1 The approach

4.2 Building the PDG

4.2.1 Control dependences

4.2.2 Data dependences

4.3 Implementing slicing

Bibliography

[1] M. Weiser, Program slicing, IEEE Transactions on Software Engineering, 10(4):352-357, 1984.

[2] Tip, Frank, A survey of program slicing techniques, Journal of programming languages 3.3, 121-189, 1995.

[3] Horwitz, Susan, Thomas Reps, and David Binkley, Interprocedural slicing using dependence graphs, ACM Transactions on Programming Languages and Systems, (TOPLAS) 12.1: 26-60, 1990.

[4] Bergeretti, Jean-Francois, and Bernard A. Carré, Information-flow and data-flow analysis of while-programs, ACM Transactions on Programming Languages and Systems, (TOPLAS) 7.1: 37-61, 1985