Static Program Analysis
Part 1 – the TIP language

http://cs.au.dk/~amoeller/spa/

Anders Møller & Michael I. Schwartzbach
Computer Science, Aarhus University
Questions about programs

• Does the program terminate on all inputs?
• How large can the heap become during execution?
• Can sensitive information leak to non-trusted users?
• Can non-trusted users affect sensitive information?
• Are buffer-overflow runs possible?
• Data races?
• SQL injections?
• XSS?
• …
Program points

```javascript
foo(p,x) {
    var f,q;
    if (*p==0) { f=1; }
    else {
        q = alloc 10;
        *q = (*p)-1;
        f=(*p)*(x(q,x));
    }
    return f;
}
```

Invariants:

A property holds at a program point if it holds in any such state for any execution with any input
Questions about program points

- Will the value of x be read in the future?
- Can the pointer p be null?
- Which variables can p point to?
- Is the variable x initialized before it is read?
- What is a lower and upper bound on the value of the integer variable x?
- At which program points could x be assigned its current value?
- Do p and q point to disjoint structures in the heap?
- Can this assert statement fail?
Why are the answers interesting?

• Increase efficiency
  – resource usage
  – compiler optimizations

• Ensure correctness
  – verify behavior
  – catch bugs early

• Support program understanding
• Enable refactorings
“Program testing can be used to show the presence of bugs, but never to show their absence.”

[Dijkstra, 1972]

Nevertheless, testing often takes 50% of the development cost
Programs that reason about programs

a program analyzer $A$

a program $P$

$P$ always works correctly

$P$ fails for some inputs
Requirements to the perfect program analyzer

- **SOUNDNESS** (don’t miss any errors)
- **COMPLETENESS** (don’t raise false alarms)
- **TERMINATION** (always give an answer)
CLASSES OF RECURSIVELY ENUMERABLE SETS AND THEIR DECISION PROBLEMS

BY

H. G. RICE

1. Introduction. In this paper we consider classes whose elements are recursively enumerable sets of non-negative integers. No discussion of recursively enumerable sets can avoid the use of such classes, so that it seems desirable to know some of their properties. We give our attention here to the properties of complete recursive enumerability and complete recursiveness (which may be intuitively interpreted as decidability). Perhaps our most interesting result (and the one which gives this paper its name) is the fact that no nontrivial class is completely recursive.

We assume familiarity with a paper of Kleene [5]({2}), and with ideas which are well summarized in the first sections of a paper of Post [7].

I. Fundamental definitions

2. Partial recursive functions. We shall characterize recursively enumer-

COROLLARY B. There are no nontrivial c.r. classes by the strong definition.
Rice’s theorem

Any non-trivial property of the behavior of programs in a Turing-complete language is undecidable!
Reduction to the halting problem

• Can we decide if a variable has a constant value?

\[
x = 17; \text{if (TM}(j)\text{)} x = 18;
\]

• Here, \(x\) is constant if and only if the \(j\)’th Turing machine does not halt on empty input
Build $e(S_T)$ from $e(T)$

Does $M$ accept input $e(M)$?

- Simulate $T$ on input $e(T)$ (ignoring input $w$)
  - If simulation reaches ACCEPT, then goto FAIL
  - Otherwise, just terminate or loop forever (without reaching FAIL)

$P$ Is the FAIL state unreachable in the given program (for any input)?

$M$

- $e(T)$
- $e(S_T)$

$S_T$

- $w$

ACCEPT

REJECT

yes

no
Undecidability of program correctness

Does $M$ accept input $e(M)$?

(Note: this proof works even if we only consider programs that always terminate!)
Approximation

• **Approximate** answers may be decidable!

• The approximation must be *conservative*:
  – i.e. only err on “the safe side”
  – which direction depends on the *client application*

• We'll focus on decision problems

• More subtle approximations if not only “yes”/“no”
  – e.g. memory usage, pointer targets
False positives and false negatives
Example approximations

• Decide if a given function is ever called at runtime:
  – if “no”, remove the function from the code
  – if “yes”, don’t do anything
  – the “no” answer must always be correct if given

• Decide if a cast \( (A)x \) will always succeed:
  – if “yes”, don’t generate a runtime check
  – if “no”, generate code for the cast
  – the “yes” answer must always be correct if given
Beyond “yes”/“no” problems

• How much memory / time may be used in any execution?

• Which variables may be the targets of a pointer variable p?
The engineering challenge

• A correct but trivial approximation algorithm may just give the useless answer every time

• The *engineering challenge* is to give the useful answer often enough to fuel the client application

• ... and to do so within reasonable time and space

• This is the hard (and fun) part of static analysis!
int main() {
    char *p,*q;
    p = NULL;
    printf("%s",p);
    q = (char *)malloc(100);
    p = q;
    free(q);
    *p = 'x';
    free(p);
    p = (char *)malloc(100);
    p = (char *)malloc(100);
    q = p;
    strcat(p,q);
}

https://en.wikipedia.org/wiki/Lint_(software)
Does anyone use static program analysis?

For optimization:
•

For verification or error detection:
•

The capabilities of static analyzers, with our work on the Infer static analyzer, can be used on source analysis tools like Findbugs, to find procedural bugs, or bugs that involve arrays.

We’ll take a look at two examples of open source DuckDuckGo Android app, as part of the tools mentioned above, which use Infer’s static code analysis tool. Infer provides both a static code analysis tool — only intra-file analysis (per function, per unit, a file-with-includes).

Inter-procedural bugs are significant. Facebook developers have fixed those bugs, which can have a large impact; we included a few from Facebook. As we have found, inter-procedural bugs are in codebases that consist of millions of lines of code.
A constraint-based approach

Conceptually separates the analysis specification from algorithmic aspects and implementation details

```java
public class Matrix {
    public static void main(String[] args) {
        int arr[][] = new int[3][3];
        System.out.println("Enter nine elements");
        Scanner sc = new Scanner(System.in);
        for (int i = 0; i < arr.length; i++) {
            for (int j = 0; j < arr.length; j++) {
                arr[i][j] = sc.nextInt();
            }
        }
        int sum = 0;
        for (int i = 0; i < arr.length; i++) {
            for (int j = 0; j < arr.length; j++) {
                if (i == j)
                    sum = sum + arr[i][j];
            }
        }
        System.out.println(sum);
    }
}
```

program to analyze

mathematical constraints

constraint solver

solution
Challenging features in modern programming language

- Higher-order functions
- Mutable records or objects, arrays
- Integer or floating-point computations
- Dynamic dispatching
- Inheritance
- Exceptions
- Reflection
- ...

The TIP language

• Tiny Imperative Programming language

• Example language used in this course:
  – minimal C-style syntax
  – cut down as much as possible
  – enough features to make static analysis challenging and fun

• Scala implementation available
Expressions

Expresssions grammar:

\[ \text{Exp} \rightarrow \text{Int} \]

\[ | \quad \text{Id} \]

\[ | \quad \text{Exp} + \text{Exp} \quad | \quad \text{Exp} - \text{Exp} \quad | \quad \text{Exp} \times \text{Exp} \quad | \quad \text{Exp} / \text{Exp} \]

\[ | \quad \text{Exp} > \text{Exp} \quad | \quad \text{Exp} \equiv \text{Exp} \]

\[ | \quad ( \text{Exp} ) \]

\[ | \quad \text{input} \]

- \( I \in \text{Int} \) represents an integer literal
- \( X \in \text{Id} \) represents an identifier (\( x, y, z, ... \))
- \text{input expression} reads an integer from the input stream
- comparison operators yield 0 (false) or 1 (true)
• In conditions, 0 is false, all other values are true
• The output statement writes an integer value to the output stream
Functions

• Functions take any number of arguments and return a single value:

\[
\text{Fun} \rightarrow \text{Id} ( \text{Id}, \ldots, \text{Id} ) \{ \\
\quad [\text{var Id}, \ldots, \text{Id};]? \\
\quad \text{Stm} \\
\quad \text{return Exp;} \\
\}
\]

• The optional \texttt{var} block declares a collection of uninitialized variables

• Function calls are an extra kind of expressions:

\[
\text{Exp} \rightarrow \ldots \mid \text{Id} ( \text{Exp}, \ldots, \text{Exp} )
\]
Pointers

\[ Exp \rightarrow \ldots \]
\[ \quad | \quad \text{alloc} \ Exp \]
\[ \quad | \quad \& \ id \]
\[ \quad | \quad * \ Exp \]
\[ \quad | \quad \text{null} \]

\[ Stm \rightarrow \ldots \quad | \quad * \ Exp = \ Exp ; \]

(No pointer arithmetic)
Records

\[
\text{Exp} \rightarrow \ldots \\
| \{ \text{Id} : \text{Exp}, \ldots, \text{Id} : \text{Exp} \} \\
| \text{Exp} \cdot \text{Id}
\]

\[
\text{Stm} \rightarrow \ldots \\
| \text{Id} \cdot \text{Id} = \text{Exp} ; \\
| (\ast \text{Exp}) \cdot \text{Id} = \text{Exp} ;
\]

Records are passed by value (like structs in C)
For simplicity, values of record fields cannot themselves be records
Functions as values

• Functions are first-class values
• The name of a function is like a variable that refers to that function

• Generalized function calls:

\[
\text{Exp} \rightarrow \ldots \mid \text{Exp}(\text{Exp}, \ldots, \text{Exp})
\]

• Function values suffice to illustrate the main challenges with methods (in object-oriented languages) and higher-order functions (in functional languages)
Programs

• A program is a collection of functions
• The function named **main** initiates execution
  – its arguments are taken from the input stream
  – its result is placed on the output stream
• We assume that all declared identifiers are unique

\[ \text{Prog} \rightarrow \text{Fun} \ldots \text{Fun} \]
An iterative factorial function

ite(n) {
    var f;
    f = 1;
    while (n>0) {
        f = f*n;
        n = n-1;
    }
    return f;
}
A recursive factorial function

```javascript
rec(n) {
    var f;
    if (n==0) {
        f=1;
    } else {
        f=n*rec(n-1);
    }
    return f;
}
```
An unnecessarily complicated function

foo(p, x) {
  var f, q;
  if (*p == 0) {
    f = 1;
  } else {
    q = alloc 0;
    *q = (*p) - 1;
    f = (*p) * (x(q, x));
  }
  return f;
}

main() {
  var n;
  n = input;
  return foo(&n, foo);
}
Beyond TIP

Other common language features in mainstream languages:

• global variables
• objects
• nested functions
• ...

ite(n) {
    var f;
    f = 1;
    while (n>0) {
        f = f*n;
        n = n-1;
    }
    return f;
}
Control flow graphs

• A control flow graph (CFG) is a directed graph:
  – nodes correspond to program points
    (either immediately before or after statements)
  – edges represent possible flow of control

• A CFG always has
  – a single point of entry
  – a single point of exit
    (think of them as no-op statements)

• Let v be a node in a CFG
  – pred(v) is the set of predecessor nodes
  – succ(v) is the set of successor nodes
For the simple \texttt{while} fragment of TIP, CFGs are constructed inductively.

CFGs for simple statements etc.:
CFG construction (2/3)

For a statement sequence $S_1 S_2$:

– eliminate the exit node of $S_1$ and the entry node of $S_2$

– glue the statements together
Similarly for the other control structures:
Normalization

• Sometimes convenient to ensure that each CFG node performs only one operation

• *Normalization*: flatten nested expressions, using fresh variables

```
x = f(y+3)*5;
```

```
t1 = y+3;
t2 = f(t1);
x = t2*5;
```