



**Faculty of Media Engineering and Technology
German University in Cairo**



**Faculty of Computer Science
Duale Hochschule Baden-Württemberg**

Static Program Analysis for SetlX

Bachelor Thesis

Author: Ahmed AboulAzm
Supervisor: Prof. Dr. Karl Stroetmann
Submission Date: August 4, 2013

I hereby certify that the thesis comprises only my original work towards my Bachelor Degree.

Ahmed AboulAzm
August 4, 2013

ACKNOWLEDGMENT

Foremost I would like to express my sincere gratitude to my advisor Prof. Dr. Karl Stroetmann for his continuous support in my bachelor project, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of implementation and writing this thesis. I could not have imagined having a better advisor and mentor for my bachelor project.

I would like also to thank my mother Amaal Morsi for her continuous support and encouragement, and for pushing me to do my bachelor project abroad. I wouldn't have made it without her.

I would also like to express my sincere thanks to Heidy Heweidy for her continuous support throughout my bachelor project.

Last but not least, I would like to thank my sister Deena Aboulazm and my friends for always being there for me when I needed support.

Contents

1	Introduction	4
2	Introducing setlX	6
2.1	SETLX Data Types	6
2.2	Statements in SETLX	12
2.3	Regular Expressions	18
2.4	Functional Programming	19
2.5	Exceptions and Backtracking	22
2.6	Objects and Classes	22
2.7	Predefined Functions	25
3	Implementation	28
3.1	main Function	28
3.2	checkCode Function	30
3.3	addDefinedVariables Function	36
3.4	getUsedVariables Function	38
3.5	identifyAllClasses Function	42
3.6	identifyAllProcedures Function	42
3.7	checkProcedure Function	43
3.8	getProcedureParameters Function	43
3.9	analyzeAllLoadedFiles Function	43
3.10	getMatchedVariables Function	43
3.11	Helper Functions	44
4	Regression Tests	45
5	Conclusion	47
5.1	Sample Run	47
5.2	Future Work	50

Chapter 1

Introduction

As SETLX is an untyped language, type errors are only caught at runtime. This is the same as in untyped languages like *Prolog*[2] or *Python*[6]. However, current *Prolog* implementations check for *singleton variables*, i.e. variables that are used only once. The reason is that many typos result in singleton variables. Therefore, checking singleton variables can uncover mistyped variable names. Our intention is to implement similar static analysis techniques for SETLX. In contrast to *Prolog* programs, where a predicate can both read a variable and define it, the direction of the data flow in SETLX programs is well defined. Therefore, it is possible to implement a static analysis for SETLX programs that is more precise than the corresponding analysis for *Prolog*: In particular, six main different checks are possible.

1. We can check whether a variable is defined before it is used. This is commonly referred to as definite assignment analysis[4].
2. We can check whether a variable is read after it is written. If a variable is written but never read afterwards, the assignment to this variable is useless. This kind of data flow analysis is known as live variable analysis[7].
3. Checking if any used procedure or class are defined in the code or not.
4. Checking that any procedure or class are called with the correct number of parameters.
5. Checking if all member variables and procedures called for objects actually do exist or not.
6. Giving a warning in classes in case the user defined a new variable with the same name of a class variable due to not adding a prefix `"this."` before the variable's name.

The following chapters are going to be as follows

1. Chapter 2 is an introduction about the main language SETLX.
2. Chapter 3 discusses the implementation of the analyzer.

3. Chapter 4 covers regression tests.
4. Chapter 5 concludes this thesis.

Chapter 2

Introducing setlX

SETLX has been upgraded lately from the old untyped language *SETL*. It has been invented mainly for educational purposes as it is well suited to familiarize students with set theory, terms and functional programming.

In the following sections we are going to introduce the most important features of this language. If you wish to see the official full tutorial and instructions to download SETLX kindly visit this following link :

<http://www.lehre.dhbw-stuttgart.de/~stroetma/SetlX/setlX.php>

2.1 SetlX Data Types

SETLX is an interpreter that can be used to evaluate simple expressions. For example by typing

```
1/3 + 2/5;
```

and hitting return yields to this response

```
~< Result:  11/15 >~
```

This example shows that SETLX supports rational numbers, it also makes sure that the answer is in the simplest form. So after typing

```
1/3 + 2/3;
```

the response would be

```
~< Result:  1 >~
```

as in this case, the simplest form of the result will have a denominator of 1, thus SETLX only prints the nominator.

The precision of any result computed in SETLX is unlimited unless there is no more memory space to take it, so if we compute the factorial of 50

```
50!;
```

the result would be

```
~< Result:  3041409320171337804361260816606476884437764156896051200000000 >~
```

To calculate floating point values in SETLX the simplest way is to add 0.0 to the calculated expression as in

```
1/3 + 2/5 + 0.0;
```

which will yield to a result in this case as shown

```
~< Result: 0.7333333333333333 >~
```

To create a string in SETLX, you just need to put some characters between double quotations. Which makes the *hello world program* in SETLX's interactive mode as simple as

```
"Hello world!";
```

which outputs

```
~< Result: "Hello world!" >~
```

To assign a value to a variable in SETLX, we use the ':= ' operator which is the major difference in the syntax between SETLX and the programming language *C* so writing

```
x := 2;
```

assigns the value '2' to the variable 'x'. Always keep in mind that in SETLX variables always have to start with a lower case character.

SETLX provides some operators to combine boolean expressions. Which are

1. "&&" as the logical and.
2. "||" as the logical or.
3. "!" as the logical not.
4. "=>" as the logical implication.
5. "<==>" as the logical equivalence which can be also replaced with "==".
6. "<!=>" as the logical antivalence which can be also replaced with "!=".

It provides some operators to create boolean expressions as well such as ">", "<", "==", "!=", "<=", ">=".

It also supports the universal and existential quantifiers "*forall*" and "*exists*". So in order to evaluate the formula

$$\forall x \in \{1, \dots, 10\} : x^2 \leq 2^x$$

we simply write

```
forall (x in {1..10} | x ** 2 <= 2 ** x);
```

and to evaluate

$$\exists x \in \{1, \dots, 10\} : 2^x < x^2$$

we simply write

```
exists (x in {1..10} | 2 ** x < x ** 2);
```

The most interesting data type in SETLX is the *set* type. You can create a *set* by writing

```
{1, 2, 3};
```

which is exactly the same as writing

```
{2, 1, 3};
```

as order doesn't matter in *sets*. SETLX also provides other convenient ways of creating sets, for example writing

```
{1..15};
```

will create a *set* containing all elements counting from 1 till 15 with a step of 1. While writing

```
{a,b..c};
```

will create a *set* containing all elements counting from a till c with a step of $b - a$. The same also applies for descending orders.

There are some basic operators on sets which are

1. "+" to compute the union of two sets.
2. "*" to compute the intersection of two sets.
3. "-" to compute the difference of two sets.
4. "><" to compute the cartesian product of two sets.
5. "** 2" to compute the cartesian product of a set with itself.
6. "2 **" to compute the power set of a set.
7. "%" to compute the symmetric difference of two sets.

Another interesting thing about *sets* in SETLX is *set comprehension* which is used to build sets which has the general formula of

```
{ expr : x1 in s1, ..., xn in sn | cond }.
```

where *expr* is an expression that contains x_1, \dots, x_n in it which are bound to the values inside s_1, \dots, s_n while *cond* is an extra optional condition if needed. For example

```
{ a * b : a in { 1 .. 3 }, b in { 1 .. 3 } };
```

computes the set

```
{1, 2, 3, 4, 6, 9}.
```

This is actually very powerful, as by very simple code like

```
s := {2..100};
s - { p * q : p in s, q in s };
```

yields to an output containing all prime numbers between 1 and 100. SETLX also supports

some set functions which make things easier like

`first(s)`

which returns the *first* element of the set s , and

`last(s)`

which returns the *last* element of the set s . Note that sets in SETLX are internally ordered, so `first(s)` is the smallest element with respect to this order, while `last(s)` yields the largest element.

SETLX also supports lists which have 2 main differences with sets which are:

1. A syntactical difference: the curly braces “{” and “}” of sets are substituted with the square brackets “[” and “]” for lists.
2. A logical difference: A list is an ordered collection of elements which can contain an element more than once unlike sets.

Apart from these two points, lists are almost the same as sets and they support everything mentioned above concerning sets. Note that for a list l , `first(l)` is indeed the first element of the list, while `last(l)` is the last element of l which is different than the case of sets.

Alike SETL the old version of SETLX, SETLX pairs are supported. A pair is represented in SETLX as a list of length two. A binary relation can be represented as a set of pairs. So if we can consider r as a binary relation, then we have a domain and a range represented in the formulas

$$\text{domain}(r) = \{ x : [x,y] \text{ in } r \} \quad \text{and} \quad \text{range}(r) = \{ y : [x,y] \text{ in } r \}.$$

Furthermore, binary relations can be used as maps. In that case if r is the binary relation, we have a definition for $r[x]$

$$r[x] := \begin{cases} y & \text{if the set } \{y \mid [x,y] \in r\} \text{ contains exactly one element } y; \\ \Omega & \text{otherwise.} \end{cases}$$

The symbol *om* in SETLX appears when there is no solution or result to a certain expression.

Binary relations seem very helpful that they actually can be used as functions, but that won't be smart as that would consume a lot of memory and will compute even cases that aren't needed. Thus in SETLX *procedures* are used for that. Figure 2.1 for example defines a procedure that computes prime numbers starting at 2 up to a given number n .

```

1  primes := procedure(n) {
2      s := { 2..n };
3      return s - { p*q : p in s, q in s };
4  };

```

Figure 2.1: A procedure to compute the prime numbers.

In Figure 2.1, the block starts with assigning “`procedure(n) {`” to a variable `primes`

which is the name of the procedure where n in this case is the parameter given to this *procedure*, then the procedure's block ends with a closing bracket “}” followed by a semicolon ‘;’. The procedure's block itself contains the logic that the procedure has to perform.

In SETLX procedures can be

1. assigned to a variable,
2. used as an argument to another procedure,
3. be returned from other procedures,

more about procedures will be discussed later on.

Strings in SETLX are a sequence of characters enclosed by double quotes. Here is a list of functions that can be applied to *Strings* in SETLX :

1. `s1 + s2`: + to concatenate strings $s1$ and $s2$.
2. `s * n` : * to concatenate multiple instances of the string s n times where n is a natural number.
3. `s[i]` : to get the i_{th} character of string s .

SETLX also provides *string interpolation* where any string containing substring of it enclosed between two “\$” signs is evaluated and the result is substituted into the string. For example, if n has the value of 6

```
print("$n$! = $n!$");
```

will actually print

```
6! = 720.
```

If you wish to turn off such kinds of processing, you just have to use single quotes instead of double quotes. For example

```
print('$n$! = $n!$');
```

will actually print

```
$n$! = $n!$.
```

SETLX provides also *first order terms* similar to the one provided by the programming language *Prolog*. Terms consists of *functors* and *arguments*. The difference between a *functor* and a *function* is that a *functor* starts with a capital letter and doesn't evaluate anything. An example for using *Terms* is implementing *ordered binary trees*.

1. An empty tree is represented as

```
Nil()
```

2. A non-empty tree has three components

- (a) The *root* node.
- (b) A left subtree.
- (c) A right subtree.

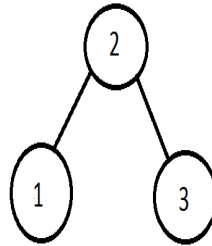
represented as

`Node(k, l, r),`

where *k* is the element stored at the root, *l* is the left subtree and *r* is the right subtree. For example the term

`Node(2,Node(1,Nil(),Nil()), Node(3,Nil(), Nil()))`

is actually representing the tree that looks as follows



SETLX supports 3 main functions for terms.

1. `fct(t)` which returns the functor of a term *t*.
2. `args(t)` which returns the arguments of a term *t*.
3. `makeTerm(f,l)` which creates a term of functor *f* and arguments *l*.

So executing

`fct(Node(3,Nil(),Nil()))`

yields a result of

`"Node",`

and executing

`args(Node(3,Nil(),Nil()))`

yields a result of

`[3, Nil(), Nil()],`

while executing

`makeTerm("Node",[makeTerm("Nil",[]), makeTerm("Nil",[])])`

constructs the term

`Node(3, Nil(), Nil()).`

Figure 2.2 shows how terms can be used in SETLX to implement ordered binary trees.

```

1  insert := procedure(m, k1) {
2      switch {
3          case fct(m) == "Nil" :
4              return Node(k1, Nil(), Nil());
5          case fct(m) == "Node":
6              [ k2, l, r ] := args(m);
7              if (k1 == k2) {
8                  return Node(k1, l, r);
9              } else if (compare(k1, k2) < 0) {
10                 return Node(k2, insert(l, k1), r);
11             } else {
12                 return Node(k2, l, insert(r, k1));
13             }
14         }
15     };

```

Figure 2.2: Inserting an element into a binary tree.

In an ordered binary tree any node will have nodes with smaller values inserted on left of it and any nodes with bigger values inserted on right of it, and in case the inserted node has the same value as an existing node then it won't be inserted into the tree. This is exactly what happens in the example shown in Figure 2.2 where m represents the tree and $k1$ represents the value of the node to be inserted. The tree's parent node is first checked if it's empty by checking if the functor of this node is `Nil` which will match the first case of the switch statement. If that condition is true then the returned tree will just be the newly inserted node, otherwise if it's a non-empty `Node` it will then match the second statement, in this case at line '6' $k2$ will have the value of this node while l and r will hold the left and right sub-trees of this node respectively. After that three nested *if-conditions* are executed. If the inserted node's value has the same value of the current node being checked then it just returns the original tree m , otherwise if it has a smaller value then a recursion happens only this time checking the inserted node with the left sub-tree, the same goes for the case if the inserted node is bigger only this time with the right sub-tree.

2.2 Statements in setlX

In this section we are going to talk about different kinds of statements in SETLX. The most basic statements are the assignment statements. In SETLX, you can have a single assignment like

```
x := 2;
```

which assigns the value of 2 to the variable x . Chained assignments can also be used as

```
a := b := 2;
```

which assigns the value of 2 to both variables a and b , you can also have simultaneous assignments like

```
[x, y, z] := [1, 2, 3];
```

which assigns variable x to a value of 1, variable y to a value of 2 and variable z to a value of 3. Simultaneous assignment can also be useful when swapping values. For example, the statement

```
[x, y] := [y, x];
```

actually swaps the values of x and y .

Functions and their structure have been discussed before but let's see some more examples. Looking at Figure 2.3, the first function 'factors' is a function that computes the factors of a given number while the second one 'primes' make use of the function 'factors' to compute all prime numbers upto a given number p . We can give a grammar rule for defining a function as

```
fctDef -> VAR " := " "procedure" "(" paramList ")" "{" block "}" ";"
```

where the symbols in this grammar means the following

1. *VAR* is a variable which will be used as the name of the function further on.
2. *paramList* is the list of parameters used in the function separated by commas. Parameters can just be a variable name or a variable name preceded by "rw" which means that this variable is a "read-write" variable and after the procedure is executed, that variable's value might actually be changed.
3. *block* is the code holding the logic of the function itself.

```

1  factors := procedure(p) {
2      return { f in { 1 .. p } | p % f == 0 };
3  };
4  primes := procedure(n) {
5      return { p in { 2 .. n } | factors(p) == { 1, p } };
6  };
7  print(primes(100));

```

Figure 2.3: A naive program to compute primes.

There is actually a simpler way to define a procedure in SETLX if the procedure's logic is a single expression, which is called the lambda definition and has the grammar rule

```
fctDef -> ID " := " "lambdaParams" "|->" exp ";"
```

where *lambdaParams* is either a single parameter or a set of parameters enclosed by square brackets. An example of using the lambda definition is

```
double := x |-> x*2;
```

SETLX supports *if-then-else statements*, *switch statements* and *match statements*. An example of the use of the *if-then-else statements* is shown in Figure 2.4 while an example for the use of the *switch statements* is shown in Figure 2.5.

```

1  toBin := procedure(n) {
2      if (n < 2) {
3          return str(n);
4      } else {
5          r := n % 2;
6          n := floor(n / 2);
7          return toBin(n) + toBin(r);
8      }
9  };

```

Figure 2.4: A function to compute the binary representation of a natural number.

```

1  sort3 := procedure(l) {
2      [ x, y, z ] := l;
3      switch {
4          case x <= y && y <= z: return [ x, y, z ];
5          case x <= z && z <= y: return [ x, z, y ];
6          case y <= x && x <= z: return [ y, x, z ];
7          case y <= z && z <= x: return [ y, z, x ];
8          case z <= x && x <= y: return [ z, x, y ];
9          case z <= y && y <= x: return [ z, y, x ];
10         default: print("Impossible error occurred!");
11     }
12 };

```

Figure 2.5: Sorting a list of 3 elements using a `switch` statement.

As for match statements, there are different types of matching statements. One kind is called *string-matching*. One example for it is shown in Figure 2.6. Calling `reverse("abc");` will yield an output of "cba". There are other kinds of matching like *list-matching*, *set-matching* and *term-matching*. Term-Matching is the most elaborate form of matching, it's similar to the matching provided in programming languages *Prolog* [2] and *ML* [5]. An example for *term-matching* is shown in Figure 2.7 which does the same work as the code shown in Figure 2.2 discussed previously.

A more complex example is shown in Figure 2.8 which computes the derivative of t relative to x . In order to understand this example better, we have to discuss some predefined functions which convert strings to terms which are the `canonical` and the `parse` functions. These functions are going to be discussed in the last section entitled *predefined functions*.

```

1  reverse := procedure(s) {
2      match (s) {
3          case [] : return s;
4          case [c|r]: return reverse(r) + c;
5          default : abort("type error in reverse($s$)");
6      }
7  };

```

Figure 2.6: A function that reverses a string.

```

1  insert := procedure(m, k1) {
2      match (m) {
3          case Nil() :
4              return Node(k1, Nil(), Nil());
5          case Node(k2, l, r):
6              if (k1 == k2) {
7                  return Node(k1, l, r);
8              } else if (compare(k1, k2) < 0) {
9                  return Node(k2, insert(l, k1), r);
10             } else {
11                 return Node(k2, l, insert(r, k1));
12             }
13         default: abort("Error in insert($m$, $k1$, $v1$)");
14     }
15 };

```

Figure 2.7: Inserting an element into a binary tree using matching.

SETLX offers two kinds of loops which are the **while** loops and the **for** loops. The while loops follows the grammar rule

$$\text{statement} \rightarrow \text{"while" "(" boolExpr ")" "{" block "}" } .$$

Figure 2.9 is an example for testing *Collatz conjecture*[3] which is represented as the recursive function f with definition

1. $f(n) := 1$ if $n \leq 1$,
2. $f(n) := \begin{cases} f(n/2) & \text{if } n \% 2 = 0; \\ f(3 \cdot n + 1) & \text{otherwise.} \end{cases}$

using the **while** loop.

On the other hand **for**-loops follow the grammar rule

$$\text{statement} \rightarrow \text{"for" "(" iterator("," iterator)* ")" "{" block "}" } .$$

A simple iterator would be x in s where s is a set or a list or a string. An example for the

```

1  diff := procedure(t, x) {
2      match (t) {
3          case t1 + t2 :
4              return diff(t1, x) + diff(t2, x);
5          case t1 - t2 :
6              return diff(t1, x) - diff(t2, x);
7          case t1 * t2 :
8              return diff(t1, x) * t2 + t1 * diff(t2, x);
9          case t1 / t2 :
10             return ( diff(t1, x) * t2 - t1 * diff(t2, x) ) / t2 * t2;
11         case f ** g :
12             return diff( @exp(g * @ln(f)), x);
13         case ln(a) :
14             return diff(a, x) / a;
15         case exp(a) :
16             return diff(a, x) * @exp(a);
17         case ^variable(x) : // x is defined above as second argument
18             return 1;
19         case ^variable(y) : // y is undefined, matches any other variable
20             return 0;
21         case n | isNumber(n):
22             return 0;
23     }
24 };

```

Figure 2.8: A function to perform symbolic differentiation.

```

1  f := procedure(n) {
2      if (n == 0) {
3          return 1;
4      }
5      while (n != 1) {
6          if (n % 2 == 0) {
7              n /= 2;
8          } else {
9              n := 3 * n + 1;
10             }
11     }
12     return n;
13 };

```

Figure 2.9: A program to test the Collatz conjecture.

usage of *for-loops* is represented in Figure 2.10 and has an output represented in Figure 2.11.

```

1  rightAdjust := procedure(n) {
2      switch {
3          case n < 10 : return "  " + n;
4          case n < 100: return " " + n;
5          default:     return  " " + n;
6      }
7  };
8  for (i in [1 .. 10]) {
9      for (j in [1 .. 10]) {
10         nPrint(rightAdjust(i * j));
11     }
12     print();
13 }

```

Figure 2.10: A simple program to generate a multiplication table.

1	2	3	4	5	6	7	8	9	10
2	4	6	8	10	12	14	16	18	20
3	6	9	12	15	18	21	24	27	30
4	8	12	16	20	24	28	32	36	40
5	10	15	20	25	30	35	40	45	50
6	12	18	24	30	36	42	48	54	60
7	14	21	28	35	42	49	56	63	70
8	16	24	32	40	48	56	64	72	80
9	18	27	36	45	54	63	72	81	90
10	20	30	40	50	60	70	80	90	100

Figure 2.11: Output of the program in Figure 2.10.

2.3 Regular Expressions

SETLX like most modern programming languages supports *regular expressions*^[1] which is a very powerful tool to process strings. Regular Expressions can be used in match statements. One example for that is shown in Figure 2.12 which is a procedure that takes a string and recognizes if it's a word, integer or just a white space.

```

1  classify := procedure(s) {
2      match (s) {
3          regex '0|[1-9][0-9]*': print("found an integer");
4          regex '[a-zA-Z]+'      : print("found a word");
5          regex '\s+'           : // skip white space
6          default               : print("unkown: $$");
7      }
8  };

```

Figure 2.12: A simple function to recognize numbers and words.

Regular expressions also can be used for extracting substrings. Consider the example given in Figure 2.13 which extracts parts of a phone-code in the format of +49-711-6673-4504 where 49 is the country code, 711 is the area code, 6673 is the company code and 4504 is the extension. In this example if an expression is matched, variable *c* will have the country code, variable *a* will have the area code, variable *co* will have the company code, variable *x* will have the extension, finally variable *e* will hold the string that matched the whole regular expression.

```

1  extractCountryArea := procedure(phone) {
2      match (phone) {
3          regex '\+([0-9]+)-([0-9]+)-([0-9]+)-([0-9]+)' as [e, c, a, co, x]:
4              return [c, a];
5          default: abort("The string $phone$ is not a phone number!");
6      }
7  };

```

Figure 2.13: A function to extract the country and area code of a phone number.

Regular expressions can also be used with *scan-statements* as shown in Figure 2.14. *Scan statements* have the general form of

```

scan (s) {
    regex  $r_1$  as  $l : b_1$ 
    :
    regex  $r_n$  as  $l : b_n$ 
}

```

```

1  printComments := procedure(file) {
2      s := join(readFile(file), "\n");
3      scan (s) {
4          regex '//[^\n]*'                as c: print(c[1]);
5          regex '/\*([^\*]|\*+[/])*\*/' as c: print(c[1]);
6          regex '.*\n'                    : // skip every thing else
7      }
8  };

```

Figure 2.14: Extracting comments using the scan statement.

Given bellow are some examples for predefined functions for regular expressions.

1. `matches('(a*)(a+)b', "aaab", true);` with three arguments with the last one defined as a boolean *true* has the exact same effect as the `testRegexp` and has the same return value of ["aaab", "aa", "a"].
2. `matches('(a*)(a+)b', "aaab");` with two arguments returns true if the string given matches the regular expression or false otherwise. In this case given in the example the result would be true.
3. `replace(s, r, t)` replaces every substring of *s* that matches the regular expression *r* by the string *t*.

2.4 Functional Programming

It has been stated previously that SETLX is a full-fledged functional language. Functions can be used as arguments to other functions and as return values. There is no difference between the type of a function and any other type that can be assigned to a variable.

Using functions as arguments to other functions is not that commonly seen in conventional programming languages. Figure 2.15 shows an example for that. In this example the function *reduce* takes two arguments which are *l* as a list and *f* as a function. What this function does is that it reduces the list to one single value at the end by combining all its elements by the function *f*, so if we could say that *f* was an add function then the function *reduce* computes the sum of all its elements. In case the list was empty then the function returns the value $\text{om}(\Omega)$.

```

1  reduce := procedure(l, f) {
2      match (l) {
3          case []      : return;
4          case [x]      : return x;
5          case [x,y|r]: return reduce([f(x,y) | r], f);
6      }
7  };
8  add      := procedure(a, b) { return a + b; };
9  multiply := procedure(a, b) { return a * b; };
10
11 l := [1 .. 36];
12 x := reduce(l, add      );
13 y := reduce(l, multiply);

```

Figure 2.15: Implementing a second order function.

```

1  sort := procedure(l, cmp) {
2      if (#l < 2) { return l; }
3      m := #l \ 2;
4      [l1, l2] := [l[.. m], l[m+1 ..]];
5      [s1, s2] := [sort(l1, cmp), sort(l2, cmp)];
6      return merge(s1, s2, cmp);
7  };
8  merge := procedure(l1, l2, cmp) {
9      match ([l1, l2]) {
10         case [[], _] : return l2;
11         case [_, []] : return l1;
12         case [[x1|r1], [x2|r2]] :
13             if (cmp(x1, x2)) {
14                 return [x1 | merge(r1, l2, cmp)];
15             } else {
16                 return [x2 | merge(l1, r2, cmp)];
17             }
18     }
19 };
20 less      := procedure(x, y) { return x < y; };
21 greater   := procedure(x, y) { return y < x; };
22 l := [1,3,5,4,2];
23 s1 := sort(l, less      );
24 s2 := sort(l, greater);

```

Figure 2.16: A generic sort function.

An example for second order functions is shown in Figure 2.16. In this example the function *sort* takes two arguments. The first argument *l* is a list and the second argument *cmp* is a function that compares two list elements and returns either **true** or **false**. What this function does is that it orders the list according to the function 'f' given as the last argument. If for example the function **less** defined in line 20 was given as an argument, then the resulting list would be ordered ascendingly. If instead the function **greater** was the argument given, then the resulting list would be ordered descendingly.

The previous two examples showed how functions can be used as arguments to other functions. As mentioned before functions can also be used as return values from other functions. Shown in Figure 2.17 is an example that computes the *discrete-derivative* of a certain function *f*. This example shows how functions can actually be used as a return value from other functions.

```
1  delta := procedure(f) {
2      return n |-> f(n+1) - f(n);
3  };
4  g := n |-> n;
5  h := n |-> 2 ** n;
6  deltaG := delta(g);
7  deltaH := delta(h);
8
9  print([ deltaG(n) : n in [1 .. 10]]);
10 print([ deltaH(n) : n in [1 .. 10]]);
```

Figure 2.17: Computing the discrete derivative of a given function.

2.5 Exceptions and Backtracking

SETLX supports means of exception handling and backtracking. But since they are irrelevant to the project itself it won't be discussed here. To know more about them kindly check the official tutorial referenced at the beginning of this chapter.

2.6 Objects and Classes

SETLX supports some basic means of *object-oriented programming* most importantly classes. A *class* is a collection of *member-variables* and *methods* where *member-variables* are some defined variables and the *methods* are some defined functions. Figure 2.18 shows an example for creating a class *point* which represents a point in a plane.

```

1  class point(x, y) {
2      mX := x;
3      mY := y;
4
5      getX := procedure() { return mX;          };
6      getY := procedure() { return mY;          };
7      setX := procedure(x) { this.mX := x;       };
8      setY := procedure(y) { this.mY := y;       };
9      toStr := procedure() { return "<mX$, mY$>"; };
10
11     distance := procedure(p) {
12         return sqrt((mX - p.getX()) ** 2 + (mY - p.getY()) ** 2);
13     };
14 }

```

Figure 2.18: The class *point*.

The first line of this example defines the class along with the constructor of this class where the constructor takes two arguments x and y . The general form of a class definition is as follows

```

class name( $x_1, \dots, x_n$ ) {
    member-and-method-definitions
}

```

where *name* is the name of the class, x_1, \dots, x_n are the formal parameters of the constructor and *member-and-method-definitions* are the definitions of member variables and methods. In this example *mX* and *mY* are the defined member variables while *getX*, *getY*, *setX*, *setY*, *toStr* and *distance* are the class methods.

We can define an object *origin* of this class *point* by writing

```
origin := point(0, 0);
```

then if we would print the object itself by writing

```
print(origin);
```

that would yield the output shown in Figure 2.19. This output has been formatted this way to make it readable. Note that the constructor of the class `point` is part of the object `origin`.

```

1  object<
2      { distance := procedure(p) {
3          return sqrt((mX-p.getX()) ** 2 + (mY-p.getY()) ** 2);
4      };
5      toStr := procedure() { return "<mX$, $mY$>"; };
6      setX  := procedure(x) { this.mX := x; };
7      getY  := procedure() { return mY; };
8      mX    := 0;
9      setY  := procedure(y) { this.mY := y; };
10     getX  := procedure() { return mX; };
11     mY    := 0;
12     class (x, y) {
13         mX := x;
14         mY := y;
15         distance := procedure(p) {
16             return sqrt((mX - p.getX()) ** 2 + (mY - p.getY()) ** 2);
17         };
18         getX := procedure() { return mX; };
19         getY := procedure() { return mY; };
20         setX := procedure(x) { this.mX := x; };
21         setY := procedure(y) { this.mY := y; };
22         toStr := procedure() { return "<mX$, $mY$>"; };
23     }
24 }
25 >

```

Figure 2.19: The output of the command “`print(origin);`”.

The fact that all methods are stored as part of the object may seem redundant but in fact it’s useful in case we needed to disable a method of the class from being used for a certain object. For example to disable the setter methods for the object `origin` we simply write:

```
origin.setX := om;
origin.setY := om;
```

If we don’t want to change methods on a per-object basis, we just need to declare these methods to be static as shown in Figure 2.20.

Now in this case the command

```
print(origin);
```

```

1  class point(x, y) {
2      mX := x;
3      mY := y;
4
5      static {
6          getX := procedure() { return mX;          };
7          getY := procedure() { return mY;          };
8          setX := procedure(x) { this.mX := x;       };
9          setY := procedure(y) { this.mY := y;       };
10         toStr := procedure() { return "<$mX$, $mY$>"; };
11
12         distance := procedure(p) {
13             return sqrt((mX - p.getX()) ** 2 + (mY - p.getY()) ** 2);
14         };
15     }
16 }

```

Figure 2.20: The class `point` implemented using static methods.

will yield the output shown in Figure 2.21.

```

1  object<
2      { mX := 0;
3        mY := 0;
4        class (x, y) {
5            mX := x;
6            mY := y;
7            static {
8                distance := procedure(p) {
9                    return sqrt((mX-p.getX()) ** 2 + (mY-p.getY()) ** 2);
10                };
11                getX := procedure() { return mX;          };
12                getY := procedure() { return mY;          };
13                setX := procedure(x) { this.mX := x;       };
14                setY := procedure(y) { this.mY := y;       };
15                toStr := procedure() { return "<$mX$, $mY$>"; };
16            }
17        }
18    }
19  >

```

Figure 2.21: Output of “`print(origin);`”.

2.7 Predefined Functions

SETLX provides a lot of predefined functions for

1. dealing with sets and lists,
2. dealing with strings,
3. working with terms,
4. working with mathematical functions,
5. dealing with objects,
6. supporting interactive debugging,
7. dealing with I/O.

Some of the predefined functions have been mentioned earlier. Since there are too many predefined functions, we are only going to discuss the most important of them that are used in the implementation of the static analysis.

1. **+**: computes the union of its arguments which are either sets or lists.
2. **first**: The function **first**(*s*) picks the first element from the sequence *s*. The argument *s* can either be a set, a list, or a string.
3. **fromB**: The function **fromB**(*s*) picks the first element from the sequence *s*. The argument *s* can either be a set, a list, or a string. This element is removed from *s* and returned. This function returns the same element as the function **first** discussed previously.
4. **fromE**: The function **fromE**(*s*) picks the last element from the sequence *s*. The argument *s* can either be a set, a string, or a list. This element is removed from *s* and returned. This function returns the same element as the function **last** discussed previously.

5. **domain**: If *r* is a binary relation, then the equality

$$\text{domain}(r) = \{ x : [x, y] \text{ in } R \}$$

holds. For example, we have

$$\text{domain}(\{ [1, 2], [1, 3], [5, 7] \}) = \{ 1, 5 \}.$$

6. **execute**: The function **execute** is called as

execute(*s*).

Here, *s* has to be a string that can be parsed as a SETLX statement. This statement is then executed in the current variable context and the result of this evaluation is returned. Note that *s* can describe a SETLX expression of arbitrary complexity. For example, the statement

```
execute("f := procedure(x) { return x * x; };");
```

has the same effect as the following statement:

```
f := procedure(x) { return x * x; };
```

7. **split**: The function **split** is called as

```
split(s, t).
```

Here, s and t have to be strings. t can either be a single character or a regular expression. The call **split**(s, t) splits the string s at all occurrences of t . The resulting parts of s are collected into a list. If t is the empty string, the string s is split into all of its characters. For example, the expression

```
split("abc", "");
```

returns the list ["a", "b", "c"]. As another example,

```
split("abc xy z", " +");
```

yields the list

```
["abc", "xy", "z"].
```

Note that we have used the regular expression “+” to specify one or more blank characters.

Certain *magic* characters, i.e. all those characters that serve as operator symbols in regular expressions have to be escaped if they are intended as split characters. Escaping is done by prefixing two backslash symbols to the respective character as in the following example:

```
split("abc|xyz", "\\|");
```

The function **split** is very handy when processing comma separated values from CVS files.

8. **str**: The function **str** is called as

```
str(a)
```

where the argument a can be anything. This function computes the string representation of a . For example, after defining the function **f** as

```
f := procedure(n) { return n * n; };
```

the expression **str**(**f**) evaluates to the string

```
"procedure(n) { return n * n; }".
```

9. **canonical**: Given a term t , the expression **canonical**(t) returns a string that is the canonical representation of the term t . The point is, that all operators in t are replaced by functors that denote these operators internally. For example, the expression

```
canonical(parse("x+2*y"));
```

yields the string

```
^sum(^variable("x"), ^product(2, ^variable("y"))).
```

This shows that, internally, variables are represented using the functor `^variable` and that the operator “+” is represented by the functor `^sum`.

10. `parseStatements`: Given a string *s*, the expression

```
parseStatements(s)
```

tries to parse the string *s* as a sequence of SETLX statements. In order to visualize the structure of this term, the function `canonical` discussed above can be used. For example, the expression

```
canonical(parseStatements("x := 1; y := 2; z := x + y;"));
```

yields the following term (which has been formatted for easier readability):

```
^block([ ^assignment(^variable("x"), 1),
         ^assignment(^variable("y"), 2),
         ^assignment(^variable("z"), ^sum(^variable("x"), ^variable("y")))
       ])
```

Chapter 3

Implementation

In this chapter we are going to discuss the implementation of the analyzer itself. Since the code is too long, we are going to give an overview and state the purpose of every function in it, but we are only going to explain the inner workings of some of them.

First of all, in order to run the analyzer you have to type

```
setlx analyzeCode.stlx --params fileName.stlx
```

in the command line prompt for windows users or in the terminal for Mac or Unix users after setting up SETLX on your computer, where `fileName` is either a file or a list of files separated by commas or blanks to be analyzed.

In the upcoming sections we will discuss each part of the code separately. Each section will discuss a certain function.

3.1 main Function

The main function of the code is shown in Figure 3.1.

1. The main function takes one argument '`params`' which is the list of files to be checked.
2. The main function consists of a *for-loop* that loops through the files to be checked printing the start and the end of the checking of each file in the list.
3. After reading the file into the variable `listOfLines`, in *line 4* a new line "`\n`" is then added to the end of each line in the code in order to separate them. The reason for that is to avoid the commenting sign "`/"`" at any line from commenting all code lines followed by it when merging all code lines into the string `codeSeparated`.
4. After that in *line 6* the code is then parsed and saved in the variable `statements`.
5. In lines 8 to 11, four empty sets are created to be used later for checking the code. Sets `definedVars` and `allUsedVars` are used to save the defined and used variables in the code, while `allProceduresIdentities` and `allClassesIdentities` are used to save all existing procedures and classes in the code in a certain format that is going to be discussed later.

```

1  main := procedure(params){
2    for(file in params){
3      print("NOW CHECKING $file$");
4      listOfLines := readFile(file);
5      codeSeparated := +/ [ g + "\n" : g in listOfLines ];
6      statements := parseStatements(codeSeparated);
7
8      definedVars := {};
9      allUsedVars := {};
10     allProceduresIdentities := {};
11     allClassesIdentities := {};
12
13     identifyAllProcedures(statements, allProceduresIdentities);
14     identifyAllClasses(statements, allClassesIdentities);
15     prefix := "";
16     if("/") in str(file)){
17       splitting := split(str(file),"/");
18       fromE(splitting);
19       for(x in splitting){
20         prefix := prefix + "$x$"/";
21       }
22     }
23     analyzeAllLoadedFiles(statements, definedVars, allUsedVars,
24       allProceduresIdentities, allClassesIdentities);
25     checkCode(statements, definedVars, allUsedVars, false, "", {},
26       {}, allProceduresIdentities, false, {}, {}, {}, false, "",
27       allClassesIdentities);
28     unusedVars := definedVars-allUsedVars;
29     if(#unusedVars != 0)
30     {
31       ...
32     }
33     print("DONE CHECKING $file$");
34   }
35 };

```

Figure 3.1: The main function of the analyzer's code.

6. In *line 13* and *line 14* all existing procedures and classes in the code being checked are saved by a certain format in the variables called *allProceduresIdentities* and *allClassesIdentities*.
7. From *line 15* upto *line 22* any prefix in the name of the file being checked is extracted into the variable *prefix*. For example the prefix in the file name */User/example.stlx* is */User/*. The prefix of a file name is actually the directory of its location which is

very important when checking the loaded files in it.

8. After that in *line 23* the analyzer checks all files being loaded by the analyzed file using the *analyzeAllLoadedFiles* function so that any important information in them would be available further on when checking the file being analyzed.
9. The most important part then comes at *line 25* where the function *checkCode* tries to find all problem in the code that should be reported to the user.
10. Then starting *line 28* upto *line 32* the analyzer checks if there were any variables defined throughout the whole code that were never used and in such case it prints to the user a notification concerning these variables.
11. This process is then repeated in case there were more files to be checked.

3.2 *checkCode* Function

The main job of the *checkCode* function is to decompose the code into simple statements to extract the used variables and defined variables in each of them in order. The *checkcode* function takes 15 arguments which are:

1. *l* : which holds the code to be analyzed after being parsed in the main part.
2. *rw definedVars* : which is a set containing all defined variables at the point.
3. *rw allUsedVars* : which is set containing all used variables at the point.
4. *isProcedure* : which is a boolean value that indicates if the current code being checked belongs to a procedure or not.
5. *procName* : which holds the name of the procedure currently being checked. If the current code being checked didn't belong to a procedure then it's just an empty string.
6. *procParameters* : which is the set of parameters of the current procedure being checked. If the current code being checked didn't belong to a procedure then this argument is just an empty set.
7. *rw procEmbeddedProcedures* : which is a set that holds all procedures defined inside the procedure being checked in case they existed, if not then it is just the empty set.
8. *rw allProceduresIdentities* : which is a set containing the identities of all existing procedures in the code in a certain format which is discussed in the section entitled *identifyAllProcedures Function*.
9. *forLoopCase* : which is a boolean value that indicates if the current code being checked belonged to a for loop or not.
10. *rw forLoopTmpDefVars* : which is a set containing the temporary defined variables in the for-loop case which are the iterators of the loop itself.

11. **exceptionTmp** : which is a set containing the temporary defined variables for an exception in case an exception block was being checked. Otherwise, it's just an empty set.
12. **matchTmp** : which is a set containing the temporary defined variables in a match case branch in case a match statement was being checked at the moment. Otherwise, it's just the empty set.
13. **isClass** : which is a boolean value that indicates if the current code being checked belonged to a class or not.
14. **className** : which is the name of the class in case the current code being checked belonged to a class. Otherwise it's just an empty string.
15. **rw allClassesIdentities** : which is a set containing the identities of all existing classes in the code in a certain format which is discussed in the section entitled "*identifyAllClasses Function*".

The *checkCode* function is composed of a big match statement. Its functionality is actually to decompose any block of code to simple statements in order to extract from every statement the defined and used variables in it. An example for that is the if-condition block shown below.

```

1  case ^ifThenBranch(cond, then):
2                                checkCode(cond, ...);
3                                checkCode(then, ...);
4                                return;

```

where **cond** is the condition of the if-statement and the **then** is the block of code executed in case the condition was satisfied. The *checkCode* function does the same with all other constructors that contain blocks such as the *while-loops*, *for-loops*, *try-catch*, *check* and *switch* statements. However the *match statements* and the *scan statements* are held kind of differently by the *checkCode* function where they are only decomposed at the beginning as shown below :

```

1  case ^scan(x, y, z):
2                                checkCode(x, ...);
3                                checkCode(y, ...);
4                                for(w in z){
5                                    checkCode(w, ...);
6                                }

```

where *x* is the variable being scanned, *y* is the optional variable being used and *z* is the list of cases.

```

1  case ^match(x, y):
2      checkCode(x, ...);
3      for(z in y){
4          checkCode(z, ...);
5      }

```

In the previous case x is the variable being matched and y is the list of cases.

Then checking each case is shown in Figure 3.2 where "v" is the case itself, "cond" is

```

1  case ^matchCaseBranch([v],cond,do) :
2      getMatchedVariables(v, matchTmp);
3      usedVarsInMatch := {};
4      checkCode(cond, definedVars, usedVarsInMatch, ...);
5      checkCode(do, definedVars, usedVarsInMatch, ...);
6      allVariablesAreDefined := true;
7      undefinedVariables := {};
8      for(v in usedVarsInMatch){
9          v := str(v);
10         if((v in definedVars) &&
11             (!(v in forLoopTmpDefVars) || (v in exceptionTmp)
12             || (v in matchTmp)))){
13             allUsedVars := allUsedVars + {v};
14         }
15         if(!(v in definedVars || v in forLoopTmpDefVars
16             || v in exceptionTmp || v in matchTmp)){
17             allVariablesAreDefined := false;
18             undefinedVariables := undefinedVariables + {v};
19         }
20     }
21     if(allVariablesAreDefined == false){
22         ...
23     }

```

Figure 3.2: Match case branch

the case condition and "do" is what happens when this case is true. For example in the case:

```
case sum(x, y) | x != y : z := x + y;
```

"v" is `sum(variable(x),variable(y))`,

"cond" is `notEqual(variable(x), variable(y))` and

"do" is `assignment(variable(z), sum(variable(x), variable(y)))`.

1. In *line 2* of figure 3.2, all defined variables for a certain case are saved in the set

`matchTmp` using the function `getMatchedVariables`. The defined variables for the case

```
case sum(variable(x), variable(y)) : ...
```

are the variables x and y .

2. In *line 3*, variable `usedVarsInMatch` is assigned to an empty set and then all used variables in the condition '`cond`' and in the do part '`do`' are added to this set in *lines 4 and 5*.
3. After that in *line 6 and 7* the boolean variable `allVariablesAreDefined` and the set `undefinedVariables` are defined.
4. Then starting *line 8* upto *line 19* all the used variables are checked if they were defined before or not, if they were not defined before, the boolean variable `allVariablesAreDefined` is set to false and the undefined variable is added to the set `undefinedVariables`.
5. From *line 21* to *line 23* the undefined variables are then printed to the user indicating their location.

The case of procedure assignment is not decomposed as the other cases but is held in a different way as shown in figure 3.3 where " x " is the name of the procedure, " y " is the list of parameters of this procedure and " z " is the code of the procedure itself.

1. The case of procedure assignment shown in figure 3.3 is composed of a big *if-else statement*. The *then* part is for the normal procedure assignment while the *else* part is for the procedures created inside other procedures. That is known through the boolean value of the variable `isProcedure`.
2. The *then* part then checks if the procedure being checked belonged to a class or not. In both cases the function `checkProcedure` is called to check the procedure, the only difference is that the class procedures are allowed to use class variables beside the parameters while the normal procedures are allowed to use only the parameters given, so the `checkProcedure` function is given the class variables `allClassesIdentities[className][2]` as a parameter in case of class procedures while in case of other procedures the empty set is given as a parameter instead.
3. In the *else* part the name of the procedure represented in the variable x is added to the set `procEmbeddedProcedures` in *line 16*.
4. In *line 17* the procedure's name is also added to the set `definedVars` as it might be used as a return value. It is also added to the set `allUsedVars` in *line 18* so the user won't be notified that it wasn't used in case it wasn't used as a return.
5. In *line 19* the procedure's parameters are added to the set `definedVars` using the function `getProcedureParameters`.
6. The code of the procedure's code is then checked normally in *line 20* by recursively calling the `checkCode` function.

```

1  case ^assignment(^variable(x), ^procedure(y, z)):
2      if(!isProcedure){
3          if(!isClass){
4              checkProcedure(x, ^procedure(y,z),
5                  allProceduresidentities, allClassesidentities,
6                  {}, isClass, className);
7          }
8          else{
9              checkProcedure(x, ^procedure(y,z),
10                 allProceduresidentities, allClassesidentities,
11                 allClassesidentities[className][2] + {"this"},
12                 isClass, className);
13             }
14         }
15     else{
16         procEmbeddedProcedures += {x};
17         definedVars := definedVars + {x};
18         allUsedVars += {x};
19         getProcedureParameters(y, definedVars, {});
20         checkCode(z, definedVars, allUsedVars, isProcedure,
21             procName, procParameters, procEmbeddedProcedures,
22             allProceduresidentities, forLoopCase,
23             forLoopTmpDefVars, exceptionTmp, matchTmp,
24             isClass, className, allClassesidentities);
25     }
26     return;

```

Figure 3.3: Procedure Assignment case

The cached procedures and the lambda procedures are also held the same way. Object's procedure assignment on the other hand has a total different logic as creating a procedure for a certain object makes this procedure only available for this exact object and not for all objects of the same class. The way this is handled internally is adding an imaginary class in the set `allClassesIdenteties` exclusive only for the object itself having the same procedures and class variables as its class adding to it the newly created procedure. In case the object's class was unknown because it wasn't declared directly in the code, it is assumed that it has all other classes' properties. After that the procedure is checked normally.

Now anything that reaches the default branch would actually be a single statement. The default branch's first part is shown in Figure 3.4.

1. In *lines 2 and 3* in figure 3.4 the sets `usedVariables` and `tempDefinedVars` are defined to be used to store the defined and used variables for the statement being

```

1  default:
2      usedVariables := {};
3      tempDefinedVars := {};
4      iteratorCase := false;
5      getUsedVariables(l, usedVariables, allUsedVars, tempDefinedVars,
6                      definedVars, isProcedure, procName, procParameters,
7                      procEmbeddedProcedures, iteratorCase, allProceduresidentities,
8                      forLoopCase, forLoopTmpDefVars, exceptionTmp, matchTmp,
9                      allClassesidentities, isClass, className);
10     allVariablesAreDefined := true;
11     undefinedVariables := {};
12     for(v in usedVariables){
13         v := str(v);
14         if("(" in v && ")" in v){
15             pos := 1;
16             while(true){
17                 if(v[pos] == "{"){ break;}
18                 pos += 1;
19             }
20             v := v[..(pos-1)];
21         }
22         if(v in definedVars){
23             allUsedVars := allUsedVars + {v};
24         }
25         if(!(v in definedVars || v in tempDefinedVars ||
26             v in forLoopTmpDefVars || v in exceptionTmp ||
27             v in matchTmp)){
28             allPredefinedProcedures := { ... };
29             if(!((v in domain(allProceduresidentities)) ||
30                 (v in allPredefinedProcedures))){
31                 allVariablesAreDefined := false;
32                 undefinedVariables := undefinedVariables + {v};
33             }else{
34                 print("ATTENTION! Procedure \"$v$\" is being used
35                     not as a procedure maybe a parameter to be called,
36                     please check that this is correct.");
37             }
38         }
39     }

```

Figure 3.4: Default case branch's first part

checked.

2. In *line 5* all used variables in the statement are stored in the set `usedVariables`

using the function `getUsedVariables`.

3. After that in *lines 10 and 11*, the variable `allVariablesAreDefined` is assigned to the boolean value of *true* initially and the variable `undefinedVariables` is assigned to the empty set. The set and the boolean variable are to be used further on detecting and printing the unused variables for the user.
4. Then starting *line 12* upto *line 39* a *for-loop* loops over the used variables to check if they were defined previously or not.
5. What happens in short from *line 14* till *line 21* that if one of the used variables was in the form of a collection access "`a[n]`" where `n` is a natural number, the part of the access "`[n]`" is removed to check if the variable '`a`' was defined previously or not.
6. From *line 22* till *line 38*, the definedness of the used variable '`v`' is checked, if it was defined in the set `definedVariables` only then it will be added to the set `allUsedVars`. If it wasn't defined at all, we check if it is defined as a procedure, if it wasn't defined as a procedure then the variable `allVariablesAreDefined` is set to *false* and the variable '`v`' is added to the set `undefinedVariables`.
7. In the rest of the default branch that isn't shown in the figure all the undefined variables are printed out to the user with their location.
8. Then the defined variables of the statement are extracted using the `addDefinedVariables` function as shown:

```
addDefinedVariables(l, definedVars, isClass, className,
                    allClassesidentities, isProcedure, false);
```

3.3 addDefinedVariables Function

The main job of this function is to extract all defined variables from a certain statement and add them to the set `definedVars`.

The `addDefinedVariables` function takes seven arguments which are :

1. `l` : which is the statement that the function is analyzing.
2. `rw definedVars` : which is the set that the function saves the defined variables of a statement in.
3. `isClass` : which is a boolean value that indicates if the statement being analyzed belonged to a class or not.
4. `className` : which is the name of the class that the statement being analyzed belongs to. If the statement didn't belong to a class then this argument is just an empty string.
5. `rw allClassesIdentities` : which is the list of all classes in the code along with their properties in a certain format.

6. **isProcedure** : which is a boolean value that indicates if the statement being analyzed belonged to a procedure or not.
7. **classCheck** : which is a boolean value that indicates if the analyzing was part of a class check or not.

The **addDefinedVariables** function explores the statement given to it using a match statement until it reaches an assignment, and when it finds one it saves the defined variable in it in the **definedVars** set. The assignment case would be as follows:

```

1  case ^assignment(^variable(x), y):
2      definedVars := definedVars + {x};
3      addDefinedVariables(y, definedVars, isClass,
4                          className, allClassesidentities, isProcedure,
5                          classCheck);
6      if(isClass && (x in allClassesidentities[className][2]) &&
7          isProcedure){
8          print("Warning: A new variable $x$ is being
9              defined instead of assigning the member variable
10             of class $className$, if you wish to assign
11             the member variable add a prefix this. to it.");
12      }
13      return;

```

where in *line 2* the defined variable '*x*' is added to the set **definedVars**, and in *line 3* the function **addDefinedVariables** is recursively called on whatever the variable *x* was assigned to. After that starting *line 6* the function checks if the defined variable was inside a class and had the same name as a member variable of the class, in this case it prints out to the user a warning that a new variable is created instead of assigning a new value to the member variable, the same goes concerning this warning for assignments in the form `+=`, `-=`, `*=` and `/=`.

In case of mutual assignments, the function recursively calls each assignment by its own as shown below :

```

1  case ^assignment(^assignListConstructor([a|b]), y):
2      for(z in [a|b]){
3          addDefinedVariables(^assignment(z,y), definedVars,
4                              isClass, className, allClassesidentities,
5                              isProcedure, classCheck);
6      }
7      return;

```

where `[a|b]` is the set of variables being assigned to the same value *y*.

3.4 getUsedVariables Function

The main job of the function `getUsedVariables` is to get all used variables in a statement and add them to the set `usedVariables`. However there is so much more logic in this function for handling certain cases which will be discussed in this section.

The function `getUsedVariables` takes 18 arguments which are :

1. `l` : which is the statement being checked by the function.
2. `rw usedVariables` : which is the set the function uses to store the used variables in the statement being checked.
3. `allUsedVars` : which is the set containing all used variables throughout the whole code so far.
4. `rw tempDefinedVars` : which is the set containing the temporary defined variables.
5. `rw definedVars` : which is the set containing all defined variables so far in the code.
6. `isProcedure` : which is a boolean value indicating if the statement `l` belonged to a procedure or not.
7. `procName` : If the statement being checked `l` belonged to a procedure, then this argument is the name of that procedure.
8. `procParameters` : If the statement being checked `l` belonged to a procedure, then this argument is a set containing the parameters of this procedure.
9. `rw procEmbeddedProcedures` : If the statement being checked `l` belonged to a procedure, then this argument is a set containing any embedded procedure if existed.
10. `iteratorCase` : which is a boolean value indicating if an iterator is being checked at the time.
11. `rw allProceduresidentities` : which is a list containing all procedures in the code and their properties in a certain format.
12. `forLoopCase` : which is a boolean value indicating if a for-loop is being checked at the time.
13. `rw forLoopTmpDefVars` : which is a set containing all temporary defined variables for a *for-loop*.
14. `exceptionTmp` : which is a set containing temporary defined variables for *try-catch* statements.
15. `matchTmp` : which is a set containing all temporary defined variables for a match statement case.
16. `rw allClassesidentities` : which is a list containing all classes in the code and their properties in a certain format.

17. **isClass** : which is a boolean value indicating if the statement being checked belonged to a class or not.
18. **className** : If the statement being checked *l* belonged to a class, then this argument is the class name.

The **getUsedVariables** function consists of a big match statement, it decomposes the statement given to it until it reaches a variable and then add it to the set **usedVariables** as shown below :

```

1  case ^variable(x):
2      usedVariables := usedVariables + {x};
3      return;

```

Of course the assigned variables are not counted as used variables and that's why they are skipped in the decomposition as shown :

```

1  case ^assignment(^variable(_), y):
2      getUsedVariables(y, usedVariables, allUsedVars, tempDefinedVars,
3          definedVars, isProcedure, procName, procParameters,
4          procEmbeddedProcedures, iteratorCase, allProceduresidentities,
5          forLoopCase, forLoopTmpDefVars, exceptionTmp, matchTmp,
6          allClassesidentities, isClass, className);
7      return;

```

The case of a call '**call(x, y)**' can happen under so many different circumstances, we will go through all the possible cases and state what happens.

1. The case of calling the '**load**' function: This case should do nothing unless it was inside a procedure as the analyzer has already extracted all the needed information from all loaded files that are outside procedures in the main function before checking the code. If the **load** function was inside a procedure then the file being loaded is checked for any problems and all needed information for the analyzing is extracted from it. An example for that is shown below:

```

1  loadInformationFile := procedure(){
2      load(informationFile.stlx);
3  }

```

In this case the file **informationFile.stlx** is checked.

2. In case the function called was a class procedure being called through an object of that class, the analyzer checks that this procedure already exists in the class, in case it didn't exist then a message is printed out to the user that the procedure doesn't exist. After that all variables given as parameters to that function are added to the set `usedVariables` and the case branch ends. For example given the code below:

```

1  class complex(real, imag) {
2      mReal := real;
3      mImag := imag;
4      static {
5          sum :=
6              that |-> complex(mReal + that.mReal, mImag + that.mImag);
7      }
8  }
9  p1 := complex(0,0);
10 p2 := complex(1,1);
11 p3 := p1.sum(p2);
12 p4 := p1.difference(p2);

```

The sum procedure called in *line 11* wouldn't be a problem while the difference called in *line 12* would actually be a problem as it doesn't exist in the class `complex` and a message concerning that will be printed to the user.

3. If the call wasn't any of the previous two cases then the function checks if the function was a predefined SETLX function. In case it was then all the variables used as parameters are added to the set `usedVariables` and the case branch ends.
4. After that the the analyzer checks if the function was defined in the code or not by looking in the set `allProceduresIdentities`, if it was indeed defined in the code the analyzer checks that it is called by the correct number of parameters, if not a message is printed to the user to notify him/her that the function is called with the wrong number of parameters. Then all parameters used are added to the set `usedVariables` and to the set `definedVars` as well in case they were rw parameters and that could be known by looking into the set `allProceduresIdentities`. An example to demonstrate that is shown below:

```

1  sum := procedure(x, y){
2      return x+y;
3  };
4  x1 := 1;
5  x2 := 2;
6  a := sum(x1,x2);
7  b := sum(x1,x2,x1);
8  c := add(x1,x2);

```

In this piece of code the `sum` procedure called at *line 6* goes fine as the procedure is actually defined with the correct number of parameters 2 while the procedure calling at *line 7* is reported to the user as it is given three parameters instead of two. Lastly, the procedure calling at *line 8* is reported as the procedure `add` is undefined at all.

5. The analyzer after that checks if the call existed in a class and that the procedure was defined inside it, if that was true then all parameters used are added to the set `usedVariables` and the case branch ends.
6. Next the analyzer checks if the call existed inside a procedure and if the function was one of the procedure's arguments or existed in the set `procEmbeddedProcedures` and if one of these cases was satisfied then all the parameters given are added to the set `usedVariables` and the case branch ends. An example for that is shown below:

```

1  dumbSum := procedure(x,y){
2      sum := [x,y] |-> x+y;
3      return sum(x,y);
4  };

```

Note that in this example the defined procedure `sum` is only defined inside the procedure `dumbSum`.

7. The analyzer then checks if the call was actually for a class and if it actually was, then all parameters given to the class are added to the set `usedVariables` and the case branch ends.
8. The last thing the analyzer checks is if the function called was in the set `definedVars` as that would be a possible case if the procedure was assigned to the variable as a return from another procedure for example. If it did exist in the set then all the parameters given are added to the set `usedVariables`.
9. If non of the above cases were satisfied then the function called is definitely undefined and a message is printed out to the user that the procedure is undefined.

When the analyzer is dealing with an iteration, it has to check it in a different way than other statements. Shown below a simple demonstration about how iterators are treated:

$$\begin{aligned} \text{getUsedVariables}(\{e: x_1 \text{ in } s_1, \dots, x_n \text{ in } t_n \mid c\}) := \\ \text{getUsedVariables}(e) \cup \text{getUsedVariables}(s_1) \cup \dots \cup \\ \text{getUsedVariables}(s_n) \cup \text{getUsedVariables}(c) \setminus \{x_1, \dots, x_n\}. \end{aligned}$$

The previous equation needs some discussion. To understand it, consider the following set definition:

$$\{ x * y : x \text{ in } s, y \text{ in } s \}.$$

This set definition reads only the variable `s` when it is evaluated. The reason is that both

x and y are local variables that are created for the construction of the set. Therefore, these variables need not be defined when the expression given above is evaluated.

Actually, the equation given above is still a simplification. The reason is that the expression s_1 must not use the variable x_1 , while the expression s_2 might use this variable.

The `getUsedVariables` function makes sure that any variable used inside the iteration is defined either inside the iteration or outside. In case a variable is used inside an iteration without being defined a message is printed to the user concerning this variable.

When a member access '`memberAccess(x, y)`' is reached in the `getUsedVariables` function, the function checks if the accessed variable y existed in a class or not, if it didn't then a message is printed to the user concerning it.

3.5 identifyAllClasses Function

The function `identifyAllClasses` with the help of the function `extractClassProperties` adds all existing classes in the code to the set `allClassesIdentities` along with their properties in a certain format. Every element of the set `allClassesIdentities` would have the format

```
[Name, [{parameters},{class variables},{class proc.s},{class objects}]]
```

where

1. **Name** : is the name of the class.
2. **parameters** : are the parameters given for creating an object of that class.
3. **class variables** : are the class member variables.
4. **class proc.s** : are the member procedures available for objects of that class.
5. **objects** : are the objects created from that class. This set is actually given empty when identifying a class and is filled later on while analyzing the code.

After this function fills the set `allClassesIdentities`, the set is used further on a lot in the analysis for accessing information concerning any class.

3.6 identifyAllProcedures Function

The function `identifyAllProcedures` with the help of the function `getProcedureIdentity` adds all existing procedures in the code to the set `allProceduresIdentities` in a certain format. Every element of the set `allProceduresIdentities` would have the format

```
[Name, [argument-type]]
```

where

1. **Name** : is the name of the procedure defined.

2. **argument-type** : is a list of 1s and 0s indicating the types of the parameters given to the procedure in order where '0' indicates a normal parameter while '1' indicates a rw parameter.

After this function fills the set **allProceduresIdenteties**, the set is used further on a lot in the analysis for accessing information concerning any procedure.

3.7 checkProcedure Function

The function **checkProcedure**'s main function is to check on the code of a procedure and if there was anything wrong in it, messages should be printed out to the user. The function **checkProcedure** does more or less the normal analysis done to the whole code but it is separate as procedures have their own space apart from the code and the only defined variables for their code are parameters and class variable in case the procedure existed in a class.

3.8 getProcedureParameters Function

The function **getProcedureParameters** takes three arguments which are :

1. **y** : which is the piece of code that has the parameters of a certain procedure.
2. **rw allProcedureDefinedVariables** : which is a set used to store all procedure parameters.
3. **rw allProcedureRWParameters** : which is a set used to store all procedure rw parameters.

getProcedureParameters simply extracts the parameters given to a certain function and adds them to the set argument **allProcedureDefinedVariables** and also adds the rw parameters to the set argument **allProcedureRWParameters**.

3.9 analyzeAllLoadedFiles Function

The function **analyzeAllLoadedFiles** checks the code for any loaded files, in case it found any, it checks the loaded file first for any problems in it, and extracts all useful information as defined procedures or classes from it to be used further on in the analysis.

This function only detects loads that are not inside any procedures, other loads that are indeed inside procedures are not targeted by this function and are tackled later.

3.10 getMatchedVariables Function

The function **getMatchedVariables** extracts all defined variables within a match case and adds them to the set argument **list**. The reason that a whole separate function is needed to do this job instead of the function **getUsedVariables** is that the names of the matched functors can't be predicted as they can be named anything as long as they start with a

capital letter

Ex: `Node(_,_,_), Parent(_,_)` ...

so this function deals with it in a different way other than a big match statement with cases in it. This function can't be used in analyzing the code but it works perfectly for extracting defined variables in a match case branch.

3.11 Helper Functions

There are two helper function that haven't been discussed in the previous sections which are :

1. `extractClassProperties` : which is used only by `identifyAllClasses` to help filling the set `allClassesIdentities` with class properties.
2. `getProcedureIdentity` : which helps mainly the `identifyAllProcedures` and `identifyAllClasses` in filling the list that indicates if the parameters given to a function were read/write parameters or not.

Chapter 4

Regression Tests

Regression tests were created in order to sustain the analyzer's functionalities in further implementations. In order to run the regression tests just navigate to the directory where the file and the tests are placed. Then run the command

```
setlx regressionTests.stlx
```

and all tests will run printing out a message in case any of them failed.

The `regressionTests.stlx` file has one main function shown in Figure 4.1.

```
1  test := procedure(fileName, expectedOutput, rw allPassed){
2      print("Now testing $fileName$");
3      output := run("setlx analyzeCode.stlx --params $fileName$");
4      if(output[1] != expectedOutput){
5          //print(output[1]);
6          print("Testing $fileName$ has failed");
7          allPassed := false;
8      }
9  };
```

Figure 4.1: Test procedure in the regression tests' file

1. The function 'test' shown in Figure 4.1 takes three arguments which are
 - (a) `fileName` : which is the name of the test file being checked.
 - (b) `expectedOutput` : which is the expected output from running the the analyzer on the test file.
 - (c) `rw allPassed` : which is a boolean variable to indicate if any test has failed.
2. In *line 2* the function prints the start of testing a certain file so the user would know which test file is being checked at the moment.
3. In *line 3* the output of the test file is saved in the variable `output`.

4. After that the function checks if the output of the test file matches the expected output or not. If they didn't match it prints in *line 6* that this test has failed.

The rest of the `regressionTests.stlx` is composed of three parts which are :

1. Creating a boolean variable `allPassed` and setting it initially to `true`.
2. Calling the function `test` for each and every test file. An example for that is shown below :

```
1  fileName := "basicTest1.stlx";
2  expectedOutput := ["NOW CHECKING basicTest1.stlx",
3      "Variables {\\"y\\"} is
4      called to be used in procedure called f without being defined !",
5      "Variable {\\"x\\"} is
6      defined but never used in the procedure called f!!",
7      "DONE CHECKING basicTest1.stlx"];
8  test(fileName, expectedOutput, allPassed);
```

3. At the end of the file the boolean variable '`allPassed`' is checked, and if its value was `true` then the file prints that all tests have passed successfully.

Chapter 5

Conclusion

5.1 Sample Run

In short, SETLX is an untyped language and type errors in it are only caught at run time. However, after the implementation of this analyzer, errors can actually be caught before run time. To be specific, the analyzer successfully covers the following checks:

1. Checking whether a variable is defined before it is used, so the code shown below:

```
1  x := 2;
2  z := x + y;
3  print("z = $$");
```

would have the following output printed to the user:

```
1  NOW CHECKING new.stlx
2  Variable {"y"} is called to be used without being defined !
3  DONE CHECKING new.stlx
```

assuming that `new.stlx` is the name of the file being checked.

2. Checking whether a variable is ever used after being defined, so the code shown below:

```
1  w := 1;
2  x := 2;
3  y := 3;
4  z := x + y;
5  print("z = $$");
```

would have the following output printed to the user:

```

1  NOW CHECKING new.stlx
2  Variable {"w"} is defined but never used in the whole code!
3    unless it was assigned to a returned procedure
4  DONE CHECKING new.stlx

```

assuming that `new.stlx` is the name of the file being checked.

3. Checking whether a used procedure is defined, so the code shown below :

```

1  x := add(1, 2);
2  print(x);

```

would have the following output printed to the user :

```

1  NOW CHECKING new.stlx
2  PROCEDURE add is undefined!
3  DONE CHECKING new.stlx

```

assuming that `new.stlx` is the name of the file being checked.

4. Checking whether any procedure or class are called with the correct number of parameters, so the code shown below:

```

1  add := procedure(x, y){
2    return x+y;
3  };
4
5  class complex(real, imag) {
6    mReal := real;
7    mImag := imag;
8  }
9
10 x := add(1);
11 y := complex(1, 2, 3);
12

```

would have the following output printed to the user:

```

1  NOW CHECKING new.stlx
2  ATTENTION! Procedure add is being called with 1 parameter(s)
3      instead of 2 parameter(s)
4  Object y is called to be created of class complex with 3 parameters
5      instead of 2
6  Variables {"x", "y"} are defined but never used in the whole code!
7      unless they were assigned to a returned procedure
8  DONE CHECKING new.stlx

```

assuming that `new.stlx` is the name of the file being checked.

5. Checking if all member variables and procedures called for objects actually do exist or not, so the code shown below:

```

1  class complex(real, imag) {
2      mReal := real;
3      mImag := imag;
4      sum := that |-> complex(mReal + that.mReal, mImag + that.mImag);
5  }
6  x:= complex(1, 2);
7  print(x.real);
8  print(x.add(x));

```

would have the following output printed to the user :

```

1  NOW CHECKING new.stlx
2  Usage of variable real for object x is uncorrect
3  Attention! calling class procedure add for object x is uncorrect!
4  DONE CHECKING new.stlx

```

assuming that `new.stlx` is the name of the file being checked.

6. Giving a warning in classes in case the user defined a new variable with the same name of a class variable due to not adding a prefix `this.` before the variable's name. So the code shown below:

```
1  class complex(real, imag) {
2      mReal := real;
3      mImag := imag;
4      setmReal := procedure(x){
5          mReal := x;
6      };
7  }
```

would have the following output printed to the user :

```
1  NOW CHECKING new.stlx
2  Warning: A new variable mReal is being defined instead of
3      assigning the member variable of class complex, if you
4      wish to assign the member variable add a prefix "this." to it.
5  DONE CHECKING new.stlx
```

assuming that `new.stlx` is the name of the file being checked.

5.2 Future Work

For future work on this analyzer, more checks can be implemented. An example for these checks is *Type checking*. *Type checking* is a program analysis that verifies something about the types that are used in the program. In the strictest sense, the sense found in programming language analysis conferences, the analysis verifies (when it is successful) that the analyzed program will not have type errors when it is executed. Given the following example:

```
1  x := "string";
2  y := 2;
3  z := x/y;
```

After *Type checking* is implemented, the analyzer would report a message to the user that dividing the string `x` by 2 in *line 3* is actually wrong and will cause a type error so that the user could fix it before running the program.

Bibliography

- [1] Alfred V Aho. *Algorithms for finding patterns in strings*. The MIT Press., 1990.
- [2] Ivan Bratko. *PROLOG Programming for Artificial Intelligence*. Addison-Wesley, 2nd edition, 1990.
- [3] Richard K. Guy. *Unsolved problems in number theory (3rd ed.)*. Springer-Verlag, 2004.
- [4] G. Steele G. Bracha J. Gosling, B. Joy. *The Java Language Specification, 3rd Edition*. 2008.
- [5] Robin Milner, Mads Tofte, and Robert Harper. *The Definition of Standard ML*. MIT Press, 1990.
- [6] Guido van Rossum. Python tutorial. Technical report, Centrum Wiskunde & Informatica, Amsterdam, 1995.
- [7] Whaley. *Datalog with Binary Decision Diagrams for Program Analysis*. 2004.