PyCG: Practical Call Graph Generation in Python

Vitalis Salis,§† Thodoris Sotiropoulos,§ Panos Louridas,§ Diomidis Spinellis§ and Dimitris Mitropoulos§†
§ Athens University of Economics and Business
† National Technical University of Athens
‡ National Infrastructures for Research and Technology - GRNET

vitalsis@gmail.com, {theosotr, louridas, dds, dimitro}@aueb.gr

Abstract—Call graphs play an important role in different contexts, such as profiling and vulnerability propagation analysis. Generating call graphs in an efficient manner can be a challenging task when it comes to high-level languages that are modular and incorporate dynamic features and higher-order functions. Despite the language's popularity, there have been very few tools aiming to generate call graphs for Python programs. Worse, these tools suffer from several effectiveness issues that limit their practicality in realistic programs. We propose a pragmatic, static approach for call graph generation in Python. We compute all assignment relations between program identifiers of functions, variables, classes, and modules through an inter-procedural analysis. Based on these assignment relations, we produce the resulting call graph by resolving all calls to potentially invoked functions. Notably, the underlying analysis is designed to be efficient and scalable, handling several Python features, such as modules, generators, function closures, and multiple inheritance.

We have evaluated our prototype implementation, which we call PyCG, using two benchmarks: a micro-benchmark suite containing small Python programs and a set of macro-benchmarks with several popular real-world Python packages. Our results indicate that PyCG can efficiently handle thousands of lines of code in less than a second (0.38 seconds for 1k LoC on average). Further, it outperforms the state-of-the-art approach for call graph generation in Python. We compute all assignment relations between program identifiers of functions, variables, classes, and modules through an inter-procedural analysis. Based on these assignment relations, we produce the resulting call graph by resolving all calls to potentially invoked functions. Notably, the underlying analysis is designed to be efficient and scalable, handling several Python features, such as modules, generators, function closures, and multiple inheritance.

Despite Python's popularity [17], there have been surprisingly few tools aiming to generate call graphs for programs written in the language. Pyan [18] parses the program's Abstract Syntax Tree (AST) to extract its call graph. Nevertheless, it has drawbacks in the way it handles the inter-procedural flow of values and module imports. code2graph [19], [20] visualizes Pyan-constructed call graphs, so it has the same limitations. Depends [21] infers syntactical relations among source code entities to generate call graphs. However, functions assigned to variables or passed to other functions are not handled by Depends, thus it does not perform well in the context of a language supporting higher-order programming. We will expand on the shortcomings of the existing tools in the remainder of this work. That said, developing an effective and efficient call graph generator for a dynamically typed language like Python is no minor task.

We introduce a practical approach for generating call graphs for Python programs and implement a corresponding prototype that we call PyCG. Our approach works in two steps. In the first step we compute the assignment graph, a structure that shows the assignment relations among program identifiers. To do so, we design a context-insensitive inter-procedural analysis operating on a simple intermediate representation targeted for Python. Contrary to the existing static analyzers, our analysis is capable of handling intricate Python features, such as higher-order functions, modules, function closures, and multiple inheritance. In the next step, we build the call graph of the original program using the assignment graph. Specifically, we utilize the graph to resolve all functions that can be potentially pointed to by callee variables. Such a programming pattern is particularly common in higher-order programming.
Similar to previous work [14], our analysis follows a conservative approach, meaning that the analysis does not reason about loops and conditionals. To make our analysis more precise, especially when dealing with features like inheritance, modules or programming patterns such as duck typing [22], we distinguish attribute accesses (i.e., e.x) based on the namespace where the attribute (x) is defined. Prior work uses a field-based approach that correlates attributes of the same name with a single global location without taking into account their namespace [14]. This leads to false positives. Our design choices make our approach achieve high rates of precision, while remaining efficient and applicable to large-scale Python programs.

We evaluate the effectiveness of our method through a micro- and a macro-benchmarking suite. Also, we compare it against Pyan and Depends. Our results indicate that our method achieves high levels of precision (≈99.2%) and adequate recall (≈69.9%) on average, while the other analyzers demonstrate lower rates in both measures. Our method is able to handle medium-sized projects in less than one second (0.38 seconds for 1k LoC on average). Finally, we show how our method can accommodate the fine-grained tracking of vulnerable dependencies through a real-world case study.

**Contributions.** Our work makes the following contributions.

- We propose a static approach for pragmatic call graph generation in Python. Our method performs inter-procedural analysis on an intermediate language that records the assignment relations between program identifiers, i.e., functions, variables, classes and modules. Then it examines the documented associations to extract the call graph (Section II).
- We develop a micro-benchmark suite that can be used as a standard to evaluate call graph generation methods in Python. Our suite is modular, easily extendable, and covers a large fraction of Python’s functionality related to classes, generators, dictionaries, and more (Section V-A1).
- We evaluate the effectiveness of our approach through our micro-benchmark and a set of macro-benchmarks including several medium-sized Python projects. In all cases our method achieves high rates of precision and recall, outperforming the other available analyzers (Sections V-B, V-C).
- We demonstrate how our approach can aid dependency impact analysis through a potential enhancement of GitHub’s “security advisory” notification service (Section V-E).

**Availability.** PyCG is available as open-source software under the Apache 2.0 Licence at [https://github.com/vitsalis/pycg](https://github.com/vitsalis/pycg). The research artifact is available at [https://doi.org/10.5281/zenodo.4456583](https://doi.org/10.5281/zenodo.4456583).

II. BACKGROUND

Generating precise call graphs for Python programs involves several challenges. Existing static approaches fail to address these challenges leaving opportunities for improvement.

```
import cryptops
class Crypto:
    def __init__(self, key):
        self.key = key
    def apply(self, msg, func):
        return func(self.key, msg)
crp = Crypto("secretkey")
encrypted = crp.apply("hello world",
                   → cryptops.encrypt)
decrypted = crp.apply(encrypted,
                     → cryptops.decrypt)
```

Fig. 1: The crypto module. Existing tools fail to generate a corresponding call graph effectively.

A. Challenges

- **Higher-order Functions:** In a high-level language such as Python, functions can be assigned to variables, passed as parameters to other functions, or serve as return values.
- **Nested Definitions:** Function definitions can be nested, meaning that a function can be defined and invoked within the context of another function.
- **Classes:** As an object-oriented language, Python allows for the creation of classes that inherit attributes and methods from other classes. The resolution of inherited methods from parent classes requires the computation of the Method Resolution Order (MRO) of each class.
- **Modules:** Python is highly extensible, allowing applications to import different modules. Keeping track of the different modules that are imported in an application, as well as the resolution order of those imports, can be a challenging task.
- **Duck Typing:** Duck typing is a programming pattern that is particularly common in dynamic languages such as Python [22]. Through duck typing, the suitability of an object is determined by the presence of specific methods and properties, rather than the type of the object itself. In this context, given a method defined by two (or more) classes, it is not trivial to identify its origins when it is invoked.

B. Limitations of Existing Static Approaches

We focus on two open-source static analyzers: Pyan [18] and Depends [21]. We do not examine code2graph [19], [20] separately, as it is based on Pyan to generate call graphs. We discuss the limitations of the two existing analyzers in terms of efficiency and practicality. To do so, we introduce a small Python module named crypto (see Figure 1), which is used to encrypt and decrypt a “hello world” message. First, it imports an external Python module named cryptops, which defines two functions, namely: encrypt(key, msg) and decrypt(key, msg). Then, the Crypto class is defined. To use it, we instantiate it with an encryption key and we can encrypt or decrypt messages by calling apply(self,
Our approach for generating call graphs employs a context-insensitive inter-procedural analysis operating on an intermediate representation of the input Python program. The analysis uses a fixed-point iteration algorithm, and gradually builds the assignment graph, which is a structure that shows the assignment relations between program identifiers (Section III-A). In a language supporting higher-order programming, the assignment graph is an essential component that we use for resolving functions pointed to by variables. Function resolution takes place at the final step where we build the 1

1In Python, __init__() is the name of a special function called during object construction.
The second one lives in the namespace \( \{ \text{main, mod} \} \), while the first resides in the namespace \( \{ \text{main, mod}, (\text{cls}, \text{els}) \} \).

The analysis distinguishes the objects created at lines 1 and 3, as the first one resides in the namespace \( \{ \text{main, mod} \} \), while the second one lives in the namespace \( \{ \text{main, mod}, (\text{cls}, \text{els}) \} \).

Our approach treats every object as the value given from the evaluation of the expressions supported by the language. In particular, our representation contains expressions that capture the inter-procedural flow, assignment statements, class and function definitions, module imports, and iterators / generators (see the \( \text{Expr} \) rule). Note that the language is able to abstract different features, including lambda expressions, keyword arguments, constructors, multiple inheritance, and more.

As with prior work focusing on JavaScript \cite{15, 16, 24}, we use evaluation contexts \cite{23} that describe the order in which sub-expressions are evaluated. For example, in an attribute assignment \( E.x := e \), the \( E \) symbol denotes that we are currently evaluating the receiver of the attribute \( x \), while \( o.x := E \) indicates that the receiver has been already evaluated to an object \( o \in \text{Obj} \) (recall that evaluating expressions results in objects), and the evaluation now proceeds to the right-hand side of the assignment.

**Remarks.** When calling Python functions that produce a generator (i.e., they contain a \texttt{yield} statement instead of \texttt{return}), these calls take place only when the generator is actually used. To model this effect, when encountering such lazy calls (e.g., \texttt{gen = lazy\_call(x)}), we create a thunk (e.g., \texttt{gen = lambda: lazy\_call(x)}) that is evaluated only when we iterate the generator (through the \texttt{iter} construct). Furthermore, dictionaries and lists are treated as regular objects. For example, we model a dictionary lookup \( x[\text{"key"}] \), as an attribute access \( x.key \).

2) State: After converting the original Python program to our intermediate representation, our analysis starts evaluating each expression, and gradually constructs the assignment graph. To do so, the analysis maintains a state consisting of four domains as shown in Figure 4, namely, **scope**, **class hierarchy**, **assignment graph**, and **current namespace**.

A scope is a map of definitions to a set of definitions. Conceptually, a scope is a tree where each node corresponds to a definition (e.g., a function), and each edge shows the parent/child relations between definitions, i.e., the target node is defined inside the definition of the source node. The domain of scopes is useful for correctly resolving the definitions that are visible inside a specific namespace. Figure 5a illustrates the scope tree of the program depicted in Figure 1 and shows all program definitions and their inter-relations. Orange nodes correspond to module definitions, red nodes are class definitions, black nodes indicate functions, while blue nodes denote variables. Based on this scope tree, we infer that the function \texttt{apply} is defined inside the class \texttt{Crypto}, which is in turn defined inside the module \texttt{crypto}, i.e., notice the path \texttt{crypto -> Crypto -> apply}. This domain enables us to properly deal with Python features such as function closures and nested definitions.

A class hierarchy is a tree representing the inheritance relations among classes. An edge from node \( u \) to node \( v \) indicates that the class \( u \) is a child of the class \( v \). The analysis uses this domain for resolving class attributes (either methods or fields) defined in the base classes of the receiver object. Through this domain we are able to handle the object-oriented nature of Python, addressing features such as multiple inheritance, and the method resolution order.

The assignment graph is defined as a map of objects to an element of the power set of objects \( \mathcal{P}(\text{Obj}) \). This graph holds the assignment relations between objects, capturing the assignments and the inter-procedural flow of the program. Figure 5b illustrates the assignment graph corresponding to the program of Figure 1. Each node in the graph (e.g., \{\texttt{crypto.Crypto.apply, func}\}) represents an object. The first component of the node label (e.g., crypto.Crypto.apply) indicates the namespace where each identifier (e.g., \texttt{func}) is defined. Colors reveal the type of the identifier as explained in a previous paragraph (e.g., the blue color implies variable definitions). An edge shows the possible values that a variable may hold. For example, the variable \texttt{func} defined in the crypto.Crypto.apply namespace may point to the functions \texttt{decrypt} and \texttt{encrypt}, both defined in the cryptops namespace. As another example, notice the edge originating from the node \{\texttt{crypto.Crypto.apply, msg}\} and leading to \{\texttt{crypto, encrypted}\}. This edge shows that the parameter \texttt{msg} of the function crypto.Crypto.apply points to the variable \texttt{encrypted} when the function is invoked on line 12. The assignment graph domain enables us to address the challenge regarding higher-order programming in Python.

Finally, we use the current namespace to track the location where new variables, classes, modules, and functions are defined. This domain is important for establishing a more precise analysis than field-based analysis employed by prior work. Through namespaces, objects and attribute accesses are distinguished based on their namespace, addressing challenges...
In the following, we describe each rule in detail.

3) Analysis Rules: The analysis examines every expression found in the intermediate representation of the initial program, and transitions the analysis state according to the semantics of each expression. The algorithm repeats this procedure until the state converges, and the assignment graph is given by the final state of the analysis.

Figure 5 demonstrates the state transition rules of our analysis. The rules follow the form:

$$\langle \pi, s, n, h, E[e] \rangle \rightarrow \langle \pi', s', n', h', E[e'] \rangle$$

In the following, we describe each rule in detail.

According to the $[\text{E-CTX}]$ rule, when we have an expression $e$ in the evaluation context $E$, an assignment graph $\pi$, a scope $s$, a namespace $n$, a class hierarchy $h$, we can get an expression $e'$ in the evaluation context $E$, if the initial expression $e$ evaluates to $e'$. For what follows, the binary operation $x \cdot y$ stands for appending the element $y$ to the list $x$.

The $[\text{COMPOUND}]$ rule states that when we have a compound expression consisting of two objects $o_1$, $o_2$, we return the last object $o_2$ as the result of the evaluation. Observe that the evaluation of the compound expression requires each subterm to have been evaluated to an object according to the evaluation contexts shown in Figure 3. The rest of the rules also follow this behavior.

The $[\text{IDENT}]$ rule describes the scenario when the initial expression is an identifier $x$. In this case, the analysis retrieves the object $o$ corresponding to the identifier $x$, in the namespace $n$, based on the scope tree $s$. To do so, the analysis uses the function $\text{getObject}(s, n, x)$, which iterates every element $y$ of the namespace $n$ in the reverse order. Then, by examining the scope tree $s$, it checks whether the element node $y$ has any child matching the identifier $x$. In case of a mismatch, the function $\text{getObject}$ proceeds to the next element of the namespace. Notice that the $[\text{IDENT}]$ rule does not have any side-effect on the analysis state.

The $[\text{ASSIGN}]$ rule assigns the object $o$ to the identifier $x$. First, the analysis adds the identifier $x$ in the current namespace $n$ of the scope tree $s$, using the function $\text{addScope}(s, n, x, \tau)$. This function adds an edge from the node accessed by the path $n$ to the target node given by the definition $(x, \tau)$. Second, this rule updates the assignment graph by adding an edge from the object corresponding to the left-hand side of the assignment (i.e., $o'$) to that of the right-hand side (i.e., $o$). This update says that the variable $x$ defined in the namespace $n$ can point to the object $o$.

$[\text{FUNC}]$ updates the scope tree. In particular, it adds the function $x$ to the current namespace $n$, leading to a new scope tree $s'$. Then, it creates a new namespace $n'$ by adding the function definition $(x, \text{func})$ to the top of the current namespace. It adds all function parameters, and a virtual variable named $\text{ret}$—which stands for the variable holding the return value of the function—to the newly-created namespace $n'$. This results in a new scope tree $s''(1)$. Finally, the analysis proceeds to the evaluation of the body of the function $x$ in the fresh namespace $n'$, i.e., observe that the rule evaluates to $E[e]$. The new namespace $n'$ correctly captures that any variable defined in $e$, is actually defined in the body of the function.

$[\text{RETURN}]$ assigns the object $o$ to the virtual variable $\text{ret}$, which is used for storing the return value of a function (recall the $[\text{FUNC}]$ rule). To do so, the analysis updates the assignment graph by adding a new edge from the object $o'$ corresponding to the return variable $\text{ret}$ to the object $o$ which is the operand of $\text{return}$. Finally, this rule evaluates to the object $o'$ related to the return virtual variable $\text{ret}$.

The inter-procedural flow is captured by the $[\text{CALL}]$ rule. Specifically, when we encounter a call expression $o_1(y = o_2...)$, we examine the callee object $o_1$ associated with a function $f$ defined in a namespace $n'$. Then, the rule connects every parameter of $f$ with the appropriate argument passed during function invocation (e.g., the counterpart object of the parameter $y$ at call-site is $o_2$), leading to a new assignment graph $\pi'$. As an example, consider again the graph of Figure 5b. The outgoing edges of the $\text{crypto.Crypto.apply, func}$ node are created by this rule. These edges imply that the parameter $\text{func}$ of the $\text{crypto.Crypto.apply}$ function may hold the functions $\text{encrypt}$ and $\text{decrypt}$ passed when calling $\text{crypto.Crypto.apply}$ (Figure 1).
After this, the analysis works on the body of the class \( c \) in a new namespace \( n' \). The new namespace contains the class definition to the top of the current namespace (i.e., \( n \cdot (x, \text{cls}) \)). Then, the analysis starts examining the body of the class using the new namespace.

The \([\text{ATTR}]\) rule is similar to \([\text{IDENT}]\). However, this time, in order to correctly retrieve the object corresponding to the attribute \( x \) of the receiver object \( o \), the analysis examines the hierarchy of classes \( h \) through the function \( \text{getClassAttrObject}(o, x, h) \). This is the point where our analysis is able to distinguish attributes according to the location (i.e., \( o \)) where they are defined.

To deal with multiple inheritance, the function \( \text{getClassAttrObject}() \) respects the method resolution order implemented in Python. For example, consider the following code snippet.

```python
1 class A:
2     def func1():
3         pass
4
class B:
5     def func1():
6         pass

8 class C(B, A):
9     pass
12 c = C()
13 c.func1()
```

In the example above, the method resolution order is \( C \rightarrow B \rightarrow A \), because the class \( B \) is the first parent class of \( C \), while \( A \) is the second one. As a result, \( c.func1() \) leads to the invocation of function \( \text{func1} \) defined in class \( B \), as it is the first matching function whose name is \( \text{func1} \) in the method resolution order. Correctly resolving class members explains why the domain of the class hierarchy maps every object to a sequence of objects rather than a set—we need to track the order in which the parents of a class are registered.

For object initialization, we introduce the \([\text{NEW}]\) rule. This rule gets the object \( o_2 \) associated with the definition of the class \( x \). Using the \( \text{getClassAttrObject}() \) function, the rule inspects the method resolution order of the object \( o_3 \) to find the first object \( o_2 \) matching the function \( \text{__init__} \). Recall that this function is called whenever a new object is created. Observe how the \( \text{new} \) evaluates; it reduces to \( o_2(y = o_1 \ldots ; o_3) \). That is, we first call the constructor of the class with the same arguments passed as in the initial expression (i.e., \( o_2(y = o_1) \)), and then we return the object \( o_3 \) corresponding to the class definition, which is eventually the result of the \( \text{new} \) expression.

The rule for attribute assignment \( o_1.x := o_2 \) describes the case when the attribute \( x \) is defined somewhere in the class hierarchy of the receiver object \( o_1 \). In this case, \( \text{getClassAttrObject}() \) function, the rule inspects the method resolution order of the object \( o_3 \) to find the first object \( o_2 \) matching the function \( \text{__init__} \). The attribute assignment is similar to \([\text{ASSIGN}]\), i.e., we first add
Algorithm 1: Call Graph Construction

```
Input : p ∈ Program
        σ ∈ State
Output: cg ∈ CallGraph
1 foreach e in Program do
    while e ∉ Obj do
        (σ, E[e]) → (σ′, E[e′])
        if e′ = o1(y = o2...) then // Call Expression
            (π, s, n, f, h) ← σ′
            c ← getReachableFuns(π, o1)
            σ ← getScope(s, n, f)
        cg ← cg[o3 → cg(o3) ∪ e] // Add Call Edges
        end
        e ← e′
2 end
3 return cg
```

The graph contains only objects associated with functions. An element \( o ∈ Obj \) mapped to a set of objects \( t ∈ P(Obj) \) means that the function \( o \) may call any function included in \( t \).

The algorithm inspects every expression \( e \) found in the program (line 1), and it evaluates \( e \) based on the state transition rules described in Figure 6. The algorithm repeats the state transition rules, until \( e \) eventually reduces to an object (lines 2, 3). Every time when \( e \) reduces to a call expression of the form \( o1(y = o2...) \) (line 4), the algorithm gets the namespace where this invocation happens and retrieves the top element of that namespace (see \( n • f \), line 5). After that, the algorithm gets all functions that the callee object \( o1 \) may point to. To do so, it consults the assignment graph through the function \( getReachableFuns(π, o1) \), which implements a simple Depth-First Search (DFS) algorithm and gets the set of functions \( c \) that are reachable from the source node \( o1 \). In turn, the algorithm updates the call graph \( cg \) by adding all edges from the top element of the current namespace to the set of the callee functions \( c \) (lines 7, 8). In other words, the object \( o3 \) (line 7) representing the top element of the namespace, where the call occurs, is actually the caller of the functions pointed to by the object \( o1 \).

C. Discussion & Limitations

One of our major design decisions is to ignore conditionals and loops. For instance, when we come across an `if` statement, our analysis over-approximates the program’s behavior and considers both branches. This design choice enables efficiency without highly compromising the analysis precision (as we will discuss in Section V). Other static analyzers [9–11] choose to follow a more heavyweight approach and reason about conditionals. These static analyzers, though, do not only focus on call-graph construction, but rather they attempt to compute the set of all reachable states based on an initial one. However, for call-graph generation, providing such an initial state that exercises all feasible paths (which is required in order to compute a complete call graph), especially when analyzing libraries, is not straightforward.

In Python where object-oriented features, duck typing [22], and modules are extensively used, it is important to separate attribute accesses based on the namespace where each attribute is defined. This design choice both—contrary to prior work [14]—the precision of our analysis without sacrificing its scalability.

Our analysis does not fully support all of Python’s features. First, we ignore code generation schemes, such as calls to the `eval` built-ins. In general, such dynamic constructs hinder the effectiveness of any static analysis, and dynamic approaches are often employed as a countermeasure [25, 26]. Second, our approach does not store information about variables’ built-in types, and does not reason about the effects of built-in functions. Therefore, attribute calls that depend on a specific built-in type (e.g., `list.append()`) are not resolved, while the effects of functions such as `getattr` and `setattr` are ignored. Third, we can only analyze modules for which their source code has been provided. When a function—for which

the attribute \( x \) to the current scope through \( addScope() \), and then update the graph. This case is omitted for brevity.

When we encounter an `import` \( x \) from \( m \) as \( y \) expression, we retrieve the object \( o2 \) corresponding to the imported identifier \( x \), which is defined in the module \( m \). Then, we create an alias \( y \) for \( x \). To do so, we add \( y \) to the scope tree of the current namespace, and update the assignment graph by adding an edge from the object of \( y \) to that of \( x \). Through this rule, we are able to deal with Python’s module system.

Consuming iterables and generators is supported through the `iter \( x \)` expression. When the identifier \( x \) points to an iterable, (i.e., the object pointed to by \( x \) has an attribute named `__next__`), we get the object \( o' \) related to `__next__`. Then, `iter` evaluates to a call of `\( o'() \)` (see the `ITER-ITERABLE` rule). If this is not the case, we treat \( x \) as a generator (`ITER-GENERATOR`). In this case, `iter` reduces to a call of `\( x() \)`. Recall from Section III-A1 that we model generators as thunks, therefore this scenario describes the evaluation of these thunks (generators) when they are actually used (iterated).

Remark about analysis termination. The analysis traverses expressions, and transitions the analysis state based on the rules of Figure 5 until the state converges. The analysis is guaranteed to terminate, because the domains are finite. Even in the presence of the domain of class hierarchy \( h ∈ ClassHier \) (Figure 4), which is theoretically infinite, the analysis eventually terminates, because a Python program cannot have an unbounded number of classes.

B. Call Graph Construction

After the termination of the analysis, we build the call graph by performing a final pass on the intermediate representation of the given Python program. Algorithm I describes the details of this pass. The algorithm takes two elements as input: (1) a program \( p ∈ Program \) of the model language whose syntax is shown in Figure 3 and (2) the final state \( σ ∈ State \) stemming from the analysis step. The algorithm produces a call graph:

\[
cg ∈ CallGraph = Obj → P(Obj)
\]
its code definition is not available—is called, our method will add an edge to the function, but no edges stemming from that function will ever be added, and its return value will be ignored.

IV. IMPLEMENTATION

We have developed PyCG, a prototype of our approach in Python 3. For each input module, our tool creates its scope tree and its intermediate representation by employing the `ast` [28] and `symtable` [27] modules respectively.

Our prototype discovers the file locations of the different imported modules to further analyze them by using Python’s `importlib` module. This is the module that Python uses internally to resolve import statements. We perform two steps. First, the file location of the imported module is identified, and then a `loader` is used to import the module’s code. In Python one can define custom loaders for import statements, which allowed us to use a loader that logs the file locations discovered and then exit without loading the code. Then, in the second step, our tool takes over and uses the discovered file’s contents to iterate its intermediate representation in a recursive manner. This allows us to resolve imports in an efficient way. Currently, we only analyze discovered modules that are contained in the package’s namespace.

V. EVALUATION

We evaluate our approach based on three research questions:

**RQ1** Is the proposed approach effective in constructing call graphs for Python programs? (Sections V-B and V-C)

**RQ2** How does the proposed approach stand in comparison with existing open-source, static-based approaches for Python? (Sections V-B and V-C)

**RQ3** What is the performance of our approach? (Section V-D)

Further, we show a potential application through the enhancement of GitHub’s “security advisory” notification service.

A. Experimental Setup

We use two distinct benchmarks: (1) a micro-benchmark suite containing 112 minimal Python programs, and (2) a macro-benchmark suite of five popular real-world Python packages. We ran our experiments on a Debian 9 host with 16 CPUs and 16 GBs of RAM.

1) Micro-benchmark Suite: We propose a test suite for benchmarking call graph generation in Python. Based on this suite, researchers can evaluate and compare their approaches against a common standard. Reif et al. [29] have provided a similar suite for Java, containing unique call graph test cases, grouped into different categories.

Our suite consists of 112 unique and minimal micro-benchmarks that cover a wide range of the language’s features. We organize our micro-benchmarks into 16 distinct categories, ranging from simple function calls to more complex features such as twisted inheritance schemes. Each category contains a number of tests. Every test includes (1) the source code, (2) the corresponding call graph (in JSON format), and (3) a short description. Categorizing and adding a new test is relatively easy. The source code of each test implements only a single execution path (i.e., no conditionals and loops) so there is a straightforward correspondence to its call graph. Table I lists the categories along with the number of benchmarks they incorporate and a corresponding description.

### Table I: Micro-benchmark suite categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>#tests</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>parameters</td>
<td>6</td>
<td>Positional arguments that are functions</td>
</tr>
<tr>
<td>assignments</td>
<td>4</td>
<td>Assignment of variables to variables</td>
</tr>
<tr>
<td>built-ins</td>
<td>3</td>
<td>Calls to built in functions and data types</td>
</tr>
<tr>
<td>classes</td>
<td>22</td>
<td>Class construction, attributes, methods</td>
</tr>
<tr>
<td>decorators</td>
<td>7</td>
<td>Function decorators</td>
</tr>
<tr>
<td>dicts</td>
<td>12</td>
<td>Hashmap with values that are functions</td>
</tr>
<tr>
<td>direct calls</td>
<td>4</td>
<td>Direct call of a returned function (func(()))</td>
</tr>
<tr>
<td>exceptions</td>
<td>3</td>
<td>Exceptions</td>
</tr>
<tr>
<td>functions</td>
<td>4</td>
<td>Vanilla function calls</td>
</tr>
<tr>
<td>generators</td>
<td>6</td>
<td>Generators</td>
</tr>
<tr>
<td>imports</td>
<td>14</td>
<td>Imported modules, functions classes</td>
</tr>
<tr>
<td>kwargs</td>
<td>3</td>
<td>Keyword arguments that are functions</td>
</tr>
<tr>
<td>lambdas</td>
<td>5</td>
<td>Lambdas</td>
</tr>
<tr>
<td>lists</td>
<td>8</td>
<td>Lists with values that are functions</td>
</tr>
<tr>
<td>mro</td>
<td>7</td>
<td>Method Resolution Order (MRO)</td>
</tr>
<tr>
<td>returns</td>
<td>4</td>
<td>Returns that are functions</td>
</tr>
</tbody>
</table>

### Addressing Validity Threats: The internal validity of the micro-benchmark suite depends on the range of Python features that it covers. To address this threat, we presented the suite to two researchers, who have professionally worked as Python developers (other researchers have applied similar methods to verify their work [30]). Then, we asked them to rank the suite (from 1 to 10) based on the following criteria:

1) Completeness: Does it cover all Python features?
2) Code Quality: Are the tests unique and minimal?
3) Description Quality: Does the description adequately describe the given test case?

The first reviewer provided a 9.7 ranking in all cases. The second indicated an excellent (10) code and description quality but ranked lower (6) the completeness of the benchmarks.

Both reviewers provided corresponding feedback. In their comments, they suggested some code cleanups and asked for more comprehensive descriptions on some complex benchmarks. Regarding the completeness of the suite, they pointed out missing tests for some common features such as built-in functions and generators. We applied the reviewers’ suggestions by refactoring the affected benchmarks and improving their descriptions. Furthermore, we implemented more tests for some of the missing functionality.

2) Macro-benchmarks: We have manually generated call graphs for five popular real-world packages. The packages were chosen as follows. First, we queried the GitHub API for Python repositories sorted by their number of stars. Then, we downloaded each repository and counted the number of lines of Python code. If the repository contained less than 3.5k lines of Python code, we kept it. Table II presents the GitHub repositories we chose along with their lines of code, GitHub stars and forks, together with a short description.

Currently, there is no acceptable implementation generating Python call graphs in an effective manner, so the first author manually inspected the projects and generated their call graphs in JSON format, spending on average 10 hours for each project.
TABLE II: Macro-benchmark suite project details.

<table>
<thead>
<tr>
<th>Project</th>
<th>LoC</th>
<th>Stars</th>
<th>Forks</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fabric</td>
<td>1,225</td>
<td>12.1k</td>
<td>1.8k</td>
<td>Remote execution &amp; deployment</td>
</tr>
<tr>
<td>autojump</td>
<td>2,662</td>
<td>10.8k</td>
<td>5.8k</td>
<td>Directory navigation tool</td>
</tr>
<tr>
<td>asciinema</td>
<td>1,409</td>
<td>7.9k</td>
<td>6.7k</td>
<td>Terminal session recorder</td>
</tr>
<tr>
<td>face_classification</td>
<td>1,455</td>
<td>4.7k</td>
<td>1.4k</td>
<td>Face detection &amp; classification</td>
</tr>
<tr>
<td>Sublist3r</td>
<td>1,269</td>
<td>4.4k</td>
<td>1.1k</td>
<td>Subdomains enumeration tool</td>
</tr>
</tbody>
</table>

TABLE III: Micro-benchmark results for PyCG and Pyan. Depends is unsound in all cases and complete in 110/112 cases and is omitted.

<table>
<thead>
<tr>
<th>Category</th>
<th>PyCG</th>
<th>Pyan</th>
<th>Depends</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>assignments</td>
<td>4/4</td>
<td>3/4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>built-ins</td>
<td>3/3</td>
<td>1/3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>classes</td>
<td>22/22</td>
<td>22/22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decorators</td>
<td>6/7</td>
<td>5/7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dicts</td>
<td>12/12</td>
<td>11/12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>direct calls</td>
<td>4/4</td>
<td>4/4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exceptions</td>
<td>3/3</td>
<td>3/3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>functions</td>
<td>4/4</td>
<td>4/4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>generators</td>
<td>6/6</td>
<td>6/6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>imports</td>
<td>14/14</td>
<td>14/14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kwargs</td>
<td>3/3</td>
<td>3/3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lambdas</td>
<td>5/5</td>
<td>5/5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lists</td>
<td>8/8</td>
<td>7/8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mro</td>
<td>7/7</td>
<td>5/7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parameters</td>
<td>6/6</td>
<td>6/6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>returns</td>
<td>4/4</td>
<td>4/4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>111/112</td>
<td>103/112</td>
<td>43/112</td>
<td>36/112</td>
<td></td>
</tr>
</tbody>
</table>

We opted for medium sized projects (less than 3.5k LoC), so that we could minimize human errors. To further verify the validity of the generated call graphs, we examined the output of PyCG Pyan, and Depends and identified 90 missing edges from a total of 2506.

B. Micro-benchmark suite results

The benchmarks included in the micro-test suite have a limited scope and are designed to cover specific functionalities (such as decorators and lambdas). Table [II] lists the results of our evaluation. For each benchmark belonging to a specific category, we show if our prototype and Pyan generated complete or sound call graphs. Note that a call graph is complete when it does not contain any call edges that do not actually exist (no false positives), and sound when it contains every call edge that is realized (no false negatives).

PyCG produces a complete call graph in almost all cases (111/112). In addition, it produces sound call graphs for 103 out of 112 benchmarks. The lack of soundness is attributed to not fully covered functionalities, i.e., Python’s starred assignments.

Pyan produces either complete or sound call graphs at a much lower rate. However, for assignments, Pyan turns out as a more sound method because it supports them in a better manner. We performed a qualitative analysis on the call graphs generated by Pyan to check the reasons behind its performance. We observed that Pyan produces incomplete call graphs because it creates call edges to class names as well as their __init__ methods (see also Section II-B). Also it generates imprecise results because it does not support all of Python’s functionality, (0/6 generators and 0/3 exceptions), ignores the inter-procedural flow of functions (0/6 parameters and 0/4 returns), misses calls to imported ones (4/14), and fails to support classes (10/22).

The evaluation of Depends shows both its fundamental strengths and limitations. Recall that each benchmark implements a single execution path and includes a call coming from the module’s namespace. Our results indicate that Depends does not identify calls from module namespaces, and therefore soundness is never achieved (0/112). In terms of completeness, Depends achieves an almost perfect score (110/112) due to its conservative nature—i.e., it adds an edge when it has high confidence that it will be realized.

C. Macro-benchmark results

By using our macro-benchmark, we have examined the three tools in terms of precision and recall. Precision measures the percentage of valid generated calls over the total number of generated calls. Recall measures the percentage of valid generated calls over the total number of calls. To do so, we manually generated the call graphs of the examined packages.

Table [IV] presents our results. The missing entries for Pyan indicate that the tool crashed during the execution. Our findings show that PyCG generates high precision call graphs. On all cases, more than 98% of the generated call edges are true positives, while on one case none of the generated call edges are false positives. Recall results show that on average, 69.9% of all call edges are successfully retrieved. The missing call edges are attributed to the approach’s limitations (recall Section III-C), and missing support for some functionalities.

Pyan shows average precision and low recall. Pyan’s average precision appears because the tool adds call edges to class names instead of just their __init__ methods. Also, it does not track the inter-procedural flow of functions, which is the reason why it has low recall. For instance, the implementation of the face_classification package mostly depends on functions declared in external packages. Pyan ignores such calls which in turn leads to a 7.6% recall.

Finally, Depends shows high precision (98.7%) and low recall. The high precision of Depends can be attributed to its conservative nature. Furthermore, Depends does not track higher order functions and does not include calls coming from module namespaces. This in turn, leads to its low recall.

D. Time and Memory Performance

We use the macro-benchmark suite as a base for our time and memory evaluation. Table [V] presents the time and memory performance metrics of the three tools. The execution time was calculated using the UNIX time command, while the memory usage was calculated using the `ps` command.

<table>
<thead>
<tr>
<th>Project</th>
<th>PyCG</th>
<th>Pyan</th>
<th>Depends</th>
<th>Pyan</th>
<th>Depends</th>
<th>Pyan</th>
<th>Depends</th>
</tr>
</thead>
<tbody>
<tr>
<td>autojump</td>
<td>99.5</td>
<td>60.6</td>
<td>99.2</td>
<td>68.2</td>
<td>28.3</td>
<td>22.3</td>
<td></td>
</tr>
<tr>
<td>fabric</td>
<td>98.3</td>
<td>7.9k</td>
<td>100</td>
<td>61.9</td>
<td>.6</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>asciinema</td>
<td>100</td>
<td>98.1</td>
<td>68.8</td>
<td>69.8</td>
<td>15.5</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td>face_classification</td>
<td>99.5</td>
<td>86.8</td>
<td>96.2</td>
<td>89.7</td>
<td>7.6</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td>Sublist3r</td>
<td>98.8</td>
<td>69.8</td>
<td>100</td>
<td>61.6</td>
<td>25.1</td>
<td>21.9</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>99.2</td>
<td>74.4</td>
<td>98.7</td>
<td>69.9</td>
<td>20.6</td>
<td>14.4</td>
<td></td>
</tr>
</tbody>
</table>

TABLE IV: Macro-benchmark results and tool comparison.
TABLE V: Time and memory comparison.

<table>
<thead>
<tr>
<th>Project</th>
<th>Time (sec)</th>
<th>Memory (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PyCG</td>
<td>Pyan</td>
</tr>
<tr>
<td>audeojump</td>
<td>0.76</td>
<td>0.32</td>
</tr>
<tr>
<td>fabric</td>
<td>0.77</td>
<td>-</td>
</tr>
<tr>
<td>ascinoma</td>
<td>0.87</td>
<td>-</td>
</tr>
<tr>
<td>face_classification</td>
<td>0.92</td>
<td>0.38</td>
</tr>
<tr>
<td>Sublime3r</td>
<td>0.51</td>
<td>0.33</td>
</tr>
<tr>
<td>Average</td>
<td>0.77</td>
<td>0.38</td>
</tr>
</tbody>
</table>

consumption was measured using the UNIX `pmap` command. The metrics presented are the average out of 20 runs.

The results show that `Pyan` is more time efficient, and that `Depends` is more memory efficient. `PyCG` and `Pyan` generate a call graph for the programs in the benchmark ($\leq 3.5k$ LoC) in under a second, while `Depends` requires more than two seconds on average. Furthermore, all tools use a reasonable amount of memory, with `PyCG`, `Pyan` and `Depends` using on average $\sim61.2$, $\sim36.3$ and $\sim22$MBs of memory respectively. Overall, `PyCG` is on average 2 times slower than `Pyan`, and uses 2.8 times the amount of memory that `Depends` uses. We attribute the differences in execution time between `Pyan` and `PyCG` to the fact that `Pyan` performs two passes of the AST in comparison to `PyCG` performing a fixpoint iteration (Section [III]). `Depends` is overall slower, because it spends most of its execution time parsing the source files. In terms of memory, `Pyan` and `Depends` store less information about the state of the analysis leading to better memory performance.

E. Case Study: A Fine-grained Tracking of Vulnerable Dependencies

GitHub sends a notification to the contributors of a repository when it identifies a dependency to a vulnerable library. However, this notification does not indicate if the project invokes the function containing the defect. We show that `PyCG` can be employed to enhance the service with method-level information that may further warn the contributors.

To highlight the usefulness of our method in this context, we performed the following steps. First we accessed GitHub's “Advisory Database” [31]. Then, we searched for vulnerable Python packages sorted by the severity of the defect. In many occasions the accompanying CVE (Common Vulnerabilities and Exposures) entries did not include further details about the defects. We disregarded such instances and focused on the first two cases that provided information about the vulnerabilities that contained the vulnerability: (1) `PYYAML` [32] (versions before 5.1), a YAML parser affected by CVE-2017-18342 [33], and (2) Paramiko [34] (multiple versions before 2.4.1), an implementation of the SSHv2 protocol affected by CVE-2018-7750 [35]. Both packages were imported by thousands of projects, 9226 for `PYYAML` and 1097 for Paramiko. We could not clone all dependent repositories because some were private and others did not exist any more: we managed to download 570 `PYYAML` and 322 Paramiko dependent projects. Then, we ran our tool on each project and generated corresponding call graphs for 106 out of the 570 `PYYAML` dependent projects and 76 out of the 322 Paramiko dependent projects—the projects that `PyCG` failed to generate call graphs were written in Python 2. Finally, we queried the generated call graphs to check if the vulnerable functions were included. We found that the vulnerable function in `PYYAML` (i.e., `load`) was invoked by 42/106 projects. In Paramiko we found that the problem method (`start_server`) was not utilized at all by any of the 76 projects. We also observed that 12 projects did not invoke any library coming from Paramiko. Paramiko was needlessly included in the requirement files of the dependents. That was not a false negative from our part: we manually checked that `pyCG` did not miss any invocation.

VI. RELATED WORK

Call Graph Generation. Methods that generate call graphs can be either dynamic [36], or static [37]. Dynamic approaches usually produce fewer false positives, but suffer from performance issues. Also, they are able to analyze a single execution path, and their effectiveness relies on the program’s input. Static approaches are more time efficient and can typically cover a wider range of execution paths, trying to capture all possible program’s behaviors. Several approaches [38]–[40], try to combine the two so they can get improved results.

There are plenty of methods and tools targeting call graph generation for statically-typed programming languages such as Java. DOOP [41] and WALA [42] follow a context-sensitive, points-to analysis method. PADDLE [43], a similar approach, employs Binary Decision Diagrams (BDDs) [44]. Finally, OPAL [45] is a lattice-based approach written in Scala. Ali et al. [46], implement CGC, a partial call graph generator for Java, with the main focus being efficiency. They ignore calls coming from externally imported libraries, and only analyze the source code of a given package. We are currently following a similar approach, but we aim to efficiently analyze external dependencies in the future.

Moving to dynamic languages, Ali et al. [47] convert Python source code into JVM bytecode, and use the existing implementations for Java [42], [48], [49] to generate its call graph. However, they argue that generating precise call graphs using this method is infeasible, and sometimes the output has more than 96% of false positives. `pycallgraph` [50] generates Python call graphs by dynamically analyzing one execution path. Thus the analysis is not practical and one should pair it with another method (e.g., fuzzing) to retrieve meaningful results. On the JavaScript front, Feldthaus et al. [14] implement a flow-based approach for the generation of call graphs. They evaluate against call graphs generated by a dynamic approach paired with instrumentation, achieving $\geq 66\%$ precision and $\geq 85\%$ recall. Other JavaScript call graph generators include, IBM WALA [42], NPM call graph [51], Google closure compiler [52]. Approximate Call Graph (ACG) [14], and Type Analyzer for JavaScript (TAJS) [9]. TAJS implements a lattice-based flow-sensitive approach using abstract interpretation. Although, such an approach yields more promising results, it comes with a performance cost.

Call Graph Benchmarking and Comparison. Reif et al. present Judge [29], a toolchain for analyzing call graph generators for Java. At its core, the toolchain contains a test
suite with benchmarks for a range of Java features. The authors then proceed to compare Java call graph generators, namely Soot [48], WALA [42], DOOP [41] and OPAL [45]. Sui et al. [53] also present a test suite of Java benchmarks, and they use it to evaluate and compare Soot [48], WALA [42], and DOOP [41]. The above benchmark suites are very similar, leading to judge consolidating them into one benchmark suite. Recall our very similar implementation of a micro-benchmark suite from Section V-A.


Other JavaScript approaches target different aspects of its functionality. Madsen et al. implement RADAR [54] a tool that identifies bugs in event-driven JavaScript programs. Sotirooulos et al. [15] propose an analysis targeting asynchronous functions. Bae et al. [55], implement SAFE-WAPP a tool aimed at identifying possible API misuses. Park et al. [56] propose SAFE-WAPP, a static analyzer for client-side JavaScript.

Fromherz et al. [57] implement a prototype that soundly identifies run-time errors by evaluating the data types of Python variables through abstract interpretation. In comparison, our approach does not infer the data types of variables and focuses on the generation of call graphs.

VII. CONCLUSION

We have introduced a practical static approach for generating Python call graphs. Our method performs a context-insensitive inter-procedural analysis that identifies the flow of values through the construction of a graph that stores all assignment relationships among program identifiers. We used two benchmarks to evaluate our method, namely a micro- and a macro-benchmark suite. Our prototype showed high rates of both precision and recall. Also, our micro-benchmark suite can serve as a standard for the evaluation of future methods. Finally, we applied our approach in a real-world case scenario, to highlight how it can aid dependency impact analysis.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments and constructive feedback. This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 825328.

REFERENCES


S. Fink and J. Dobby, “WALA—the T.J. Watson libraries for analysis,” 2012.


