Specification Inference Using Context-Free Language Reachability

Osbert Bastani
Stanford University
obastani@cs.stanford.edu

Saswat Anand
Stanford University
saswat@cs.stanford.edu

Alex Aiken
Stanford University
aiken@cs.stanford.edu

Abstract
We present a framework for computing context-free language reachability properties when parts of the program are missing. Our framework infers candidate specifications for missing program pieces that are needed for verifying a property of interest, and presents these specifications to a human auditor for validation. We have implemented this framework for a taint analysis of Android apps that relies on specifications for Android library methods. In an extensive experimental study on 179 apps, our tool performs verification with only a small number of queries to a human auditor.

Categories and Subject Descriptors F.3.2 [Semantics of Programming Languages]: Program analysis; F.3.1 [Specifying and Verifying Reasoning about Programs]: Mechanical verification

Keywords program analysis; verification; specification inference

1. Introduction
Interprocedural analyses are the building blocks for many applications of program analysis, including program slicing [23], bug finding [1, 6, 18], and taint analysis [32]. One limitation of most interprocedural analyses is that they assume the entire program’s source code can be analyzed. In practice, many programs call methods in libraries that may be written in lower-level languages, use dynamic features such as Java reflection, or do not have source code available for analysis. In fact, in our experience, for large systems it is unusual if any of these situations does not occur many times. Handling missing or hard-to-analyze code in a fully automatic way generally results in using either very pessimistic and imprecise assumptions or very optimistic and unsound assumptions.

One pragmatic approach to address these problems is to manually write specifications encoding the relevant behavior of the library (or hard-to-analyze) methods so that the library source code does not need to be analyzed. For this approach to work, it is critical that (i) the manual effort of writing the specifications is small, and (ii) the analysis produces sound results even if some specifications are missing. Prior work has addressed these issues to some extent: [3, 4, 7, 15, 16, 21, 22, 25, 33] have applied data mining techniques to infer specifications, which helps address (i), and [34] has developed techniques for inferring minimal sets of specifications needed for program verification, which addresses both (i) and (ii).

In this paper we consider the problem of inferring missing specifications in the context of solving a general context-free language (CFL) reachability problem. Many important interprocedural analyses can be formulated as CFL reachability problems, and our framework makes these algorithms more practical and easier to apply to large programs where parts of the code are either missing or too difficult to analyze statically.

We instantiate our framework for taint analysis of Android apps, and perform an extensive experimental study on 179 apps. Many of these apps have hundreds of thousands of bytecode instructions and thousands of calls to Android library methods. Instead of analyzing the Android library methods, we infer two kinds of specifications that help capture their taint behavior:

- missing taint flows between parameters and return values of library methods, and
- missing alias relations introduced by library methods.

Our work has three main contributions. First, we develop a general framework for describing potentially missing specifications for CFL reachability analyses (Section 4). In particular, the framework can be instantiated to compute the two kinds of specifications above, among others. Second, we present a specification inference algorithm based on this framework (Section 5) that produces sound results, infers accurate specifications, scales to large applications, and allows a human auditor to refine the results. Third, we evaluate an implementation (Section 6) of our specification inference framework with the aforementioned experimental study in which our tool infers a large collection of library specifications and verifies whether or not apps have taint flows (Section 7).

2. Motivating Example
We begin with an example that motivates the need for specification inference to verify reachability properties of partial programs. The technical development begins in Section 3.

The popularity of the Android framework (which is installed on over 1 billion phones) has led to a proliferation of Android malware. These malicious apps exhibit behaviors including theft of contact information, sending SMS text messages to premium phone numbers, and unauthorized location tracking. Many of the malicious behaviors exhibited by current Android malware can be described as the flow of sensitive data to untrusted recipients, such as location data flowing to an untrusted web server, or an untrusted phone number flowing to an SMS send request. In principle, standard static taint analysis can identify such malware.

For example, suppose we want to look for malware that leaks location data via SMS messages. Consider the Java-like code in...
Figure 1. Naïvely, we may expect that flow specifications are sufficient to capture all information flows. For example, we may consider using taint specifications that handle field accesses, and replace the alias specification for List.add with the flow specification @Flow(arg, this.val). However, this specification is unsound—it fails to capture the flow from LOC to SMS in Figure 3:

1. Double lat = getLatitude();
2. List list = new List();
3. list.add(lat);
4. Double latAlias = list.get(0);
5. String latStr = latAlias.toString();
6. sendSMS(latStr);

Figure 3. An information flow not captured by taint specifications.

boxAlias.f is tainted by LOC, so List.val.f is tainted (since boxAlias and list.val are aliased), but the proposed specification for List.add does not transfer this taint to box.f (even though taint flows to box.f). In general, we need alias specifications to precisely and soundly capture flows due to aliasing.

Manually writing library specifications is expensive: the Android library contains over 5000 classes, many exhibiting complex data return and native methods, making it very difficult to construct a precise call graph or to perform a sound and precise context sensitive points-to analysis. Examples of such problems in practice include:

- System.arraycopy is a native method.
- Bundle.putLong indirectly calls the native Android library method Parcel.nativeWriteLong (bundles are used to pass data to and receive data from the Android system).
- GeoPoint.getLatitude is part of a closed-source Google library, so the source code is unavailable.

While implementations of the Android library methods may be missing, the net taint flows through these methods are generally simple. For example, we can summarize the net taint flows through List.add, List.get, and Double.toString as follows: (i) arg may be aliased with ret.get (i.e., they may point to the same object), and (ii) if taint flows to this.toString, then taint also flows to toString. Here, ret is the return value of fn, this is the receiver of fn, and argName is the parameter passed to fn named argName. The library specifications in Figure 2 capture these properties. Note that the code in Figure 2 is only a specification for the List and Double classes (i.e., it soundly captures the information flows and aliasing introduced by their methods), and is not a valid implementation of the classes. There are two kinds of specifications:

1. An alias specification summarizes the potential aliases caused by a library method. These are written as short methods that do nothing except introduce the desired aliasing. For example, the specification for List.add means that calling List.add may cause arg.add and this.add.val to be aliased. Also, the specification for List.get says that calling List.get may cause this.get.val and ret.get to be aliased. Note that if this.arg and this.get are aliased, then these specifications cause arg.add and ret.get to be aliased.

2. A taint specification describes how taint flows through a library method. These specifications are expressed as @Flow annotations, where the taint on the first argument (either a parameter arg, a return value ret, the receiver this, or a special label such as LOC representing a taint source) is transferred to the second argument (similar to the first argument, except in this case a special label would represent a sink, e.g. SMS). For example, the specification of Double.toString means that if this.toString is tainted, then toString is also tainted.

The specifications in Figure 2 enable the analysis to find the flow from the source label LOC to the sink label SMS in Figure 1. First, the return value ret.setLatitude is tainted with the source label LOC. Second, this tainted is passed to lat, which is stored in list.val. Third, the value is retrieved from list and stored in latAlias, before being converted into a string and passed as the text argument of the Android library method sendSMS(String text). Finally, our specification says that the text argument of the method sendSMS is sent to the sink labeled SMS, so the code exhibits a flow from LOC to SMS.

Naïvely, we may expect that flow specifications are sufficient to capture all information flows. For example, we may consider using taint specifications that handle field accesses, and replace the alias specification for List.add with the flow specification @Flow(arg, this.val). However, this specification is unsound—it fails to capture the flow from LOC to SMS in Figure 3:

1. class Box: String f;
2. List list = new List();
3. Box box = new Box();
4. list.add(box);
5. Box boxAlias = list.get(0);
6. boxAlias.f = getLatitude();
7. sendSMS(box.f);

Figure 2. Specifications for Android library classes.

boxAlias.f is tainted by LOC, so List.val.f is tainted (since boxAlias and list.val are aliased), but the proposed specification for List.add does not transfer this taint to box.f (even though taint flows to box.f). In general, we need alias specifications to precisely and soundly capture flows due to aliasing.

While implementations of the Android library methods may be missing, the net taint flows through these methods are generally simple. For example, we can summarize the net taint flows through List.add, List.get, and Double.toString as follows: (i) arg may be aliased with ret.get (i.e., they may point to the same object), and (ii) if taint flows to this.toString, then taint also flows to toString. Here, ret is the return value of fn, this is the receiver of fn, and argName is the parameter passed to fn named argName. The library specifications in Figure 2 capture these properties. Note that the code in Figure 2 is only a specification for the List and Double classes (i.e., it soundly captures the information flows and aliasing introduced by their methods), and is not a valid implementation of the classes. There are two kinds of specifications:

1. An alias specification summarizes the potential aliases caused by a library method. These are written as short methods that do nothing except introduce the desired aliasing. For example, the specification for List.add means that calling List.add may cause arg.add and this.add.val to be aliased. Also, the specification for List.get says that calling List.get may cause this.get.val and ret.get to be aliased. Note that if this.arg and this.get are aliased, then these specifications cause arg.add and ret.get to be aliased.

2. A taint specification describes how taint flows through a library method. These specifications are expressed as @Flow annotations, where the taint on the first argument (either a parameter arg, a return value ret, the receiver this, or a special label such as LOC representing a taint source) is transferred to the second argument (similar to the first argument, except in this case a special label would represent a sink, e.g. SMS). For example, the specification of Double.toString means that if this.toString is tainted, then toString is also tainted.

The specifications in Figure 2 enable the analysis to find the flow from the source label LOC to the sink label SMS in Figure 1. First, the return value ret.setLatitude is tainted with the source label LOC. Second, this tainted is passed to lat, which is stored in list.val. Third, the value is retrieved from list and stored in latAlias, before being converted into a string and passed as the text argument of the Android library method sendSMS(String text). Finally, our specification says that the text argument of the method sendSMS is sent to the sink labeled SMS, so the code exhibits a flow from LOC to SMS.

Naïvely, we may expect that flow specifications are sufficient to capture all information flows. For example, we may consider using taint specifications that handle field accesses, and replace the alias specification for List.add with the flow specification @Flow(arg, this.val). However, this specification is unsound—it fails to capture the flow from LOC to SMS in Figure 3: Figure 1. A flow through the List and Double classes.

```java
1. class List:
2. Object val;
3. void add(Object arg) { val = arg; }
4. Object get(Integer index) { return val; }
5. class Double:
6. @Flow(this, return)
7. String toString() {}
8. class SMS:
9. @Flow(LOC, return)
10. static String getLatitude() {}  
11. class LocationManager:
12. @Flow(text, SMS)
13. static void sendSMS(String text) {}
```
to $\Sigma$. However, $\arg_{\text{add}}$ and $\text{this}_{\text{add}}.\text{val}$ are no longer aliased, so the flow is broken. Only one additional assumption (i.e., $\arg_{\text{add}}$ is aliased with $\text{this}_{\text{add}}.\text{val}$) is needed to complete the flow.

This example motivates an approach to specification inference that searches for specifications that complete broken flows. Our approach proceeds in two steps. First, our tool finds potential flows by making worst-case assumptions about the missing specifications. Second, for each potential flow, our tool keeps track of which assumptions are sufficient to prove that the potential flow is a true flow, which we call sufficient assumptions for the potential flow. Finally, the tool proposes that these sufficient assumptions are true. These assumptions correspond to specifications that are the inferred specifications produced by the tool.

Continuing our example, our tool would find a potential flow from $\text{L}O\text{C}$ to $\Sigma$ by making worst-case assumptions about the specification for $\text{List}.\text{add}$, which includes introducing aliasing between $\arg_{\text{add}}$ and $\text{this}_{\text{add}}.\text{val}$. This specification for $\text{List}.\text{add}$ is a sufficient assumption for the potential flow from $\text{L}O\text{C}$ to $\Sigma$, so it is inferred by our tool. Upon seeing this inferred specification, the human auditor can confirm that it is correct and thereby determine that the potential flow is a true flow.

In practice, there may be multiple sufficient assumptions for each potential flow. Our tool keeps track of a minimal set of sufficient assumptions—i.e., it looks for flows that are broken in the fewest places possible. The guiding principle is that potential flows that require fewer assumptions are more likely to be real flows than potential flows that require more assumptions. By extension, potential flows that produce the fewest inferred specifications are most likely to be correct and should be checked first by an auditor.

We use (interprocedural) taint analysis to motivate and validate our general framework. Taint analysis is a prototypical example of a program analysis that can be expressed as a CFL reachability problem and has many practical applications. Furthermore, it is built on top of points-to analysis, which is a key part of many interprocedural program analyses. One of our goals is to infer aliasing relationships potentially introduced by library methods that are missing specifications.

### 3. CFL Reachability for Explicit Taint Flows

In this section we present a static analysis for finding explicit taint flows and illustrate it using the example in Section 2. Our analysis performs CFL reachability on the portion of the code that is available, and uses specifications for the portion of the code that is unavailable. The specifications are usually manually generated, for example by a human auditor. Our understanding is that many practical systems have this design, and we do not claim that it is novel. However, as far as we know this approach is not well-documented in the literature, so we describe it in some detail. In Section 4, we generalize this analysis to one where some specifications (in addition to the code) are also unavailable (for example, because the auditor has only produced partial specifications), and the task is to infer the possible missing specifications.

Let $C = (U, \Sigma, P, T)$ be a context-free grammar (CFG), where $U$ is the set of non-terminals, $\Sigma$ is the set of terminals, $P$ is the set of productions, and $T$ is the start symbol. We assume $C$ is normalized so that every production has at most two symbols on the right-hand side [19]. We write $A \Rightarrow^* \alpha$, where $A$ is a non-terminal and $\alpha$ is a string of terminals and non-terminals, if $\alpha$ can be derived from $A$.

Let $G$ be a directed graph such that the edges $v \xrightarrow{\ell_1} w_1 \xrightarrow{\ell_2} \ldots \xrightarrow{\ell_k} v'$ in $G$ are labeled with terminal symbols $\sigma \in \Sigma$. A path $v \xrightarrow{\ell^*} v' \in G$ is a sequence of edges $v \xrightarrow{\ell_1} w_1 \xrightarrow{\ell_2} \ldots \xrightarrow{\ell_k} v'$ such that $\alpha = \sigma_1 \ldots \sigma_k$.

The transitive closure of $G$ under $C$ is the graph $G^C$ such that $v \xrightarrow{\ell^*} v' \in G^C$ if and only if there exists $v \xrightarrow{\ell} v' \in G$ such that $A \Rightarrow^* \alpha$. If $v \xrightarrow{\ell^*} v'$, we say $v'$ is $C$-reachable from $v$.

#### Definition 3.1

Given a CFG $C$, a graph $G = (V, E)$, and subsets of vertices $V_{\text{src}}, V_{\text{sink}} \subseteq V$, the CFL reachability problem is to determine whether there exist $v \in V_{\text{src}}$ and $v' \in V_{\text{sink}}$ such that $v'$ is $C$-reachable from $v$.

Typically, formulating a program analysis as a CFL reachability problem involves converting the input program into a graph $G = (V, E)$ and a CFG $C$, and then solving the CFL reachability problem for $C$ and $G$ for some given sets of sources and sinks $V_{\text{src}}, V_{\text{sink}} \subseteq V$. This problem can be solved by finding the closure of $G$ under $C$ using a dynamic programming algorithm [19].

#### 3.1 Explicit Taint Flows

We now describe a standard flow- and path-insensitive explicit taint analysis. Explicit taint analysis only tracks taint through data flows—i.e., it follows tainted data as it is copied in the course of program execution. We defer discussion of implicit flows, where the information transferred depends on control flow decisions, to Section 3.2. As a part of the explicit taint analysis, we perform points-to analysis to find taint flows due to aliasing. For Java, points-to analysis can be expressed as a CFL reachability problem [26]. We extend the points-to analysis to a taint analysis by including source and sink edges that pass taint into and out of the system.

A set of taint specifications $S$ consists of three relations: the source specifications $\text{Src}(v, \ell)$, the sink specifications $\text{Sink}(v, \ell)$, and the flow specifications $\text{Flow}(v, v')$. Here, $v, v' \in G$ are formal method parameters or return values, and $\ell \in \mathcal{L} = \mathcal{L}_{\text{src}} \cup \mathcal{L}_{\text{sink}}$ is a label representing a taint source ($\ell \in \mathcal{L}_{\text{src}}$) or a sink ($\ell \in \mathcal{L}_{\text{sink}}$). The relations have the following semantics:

- $\text{Src}(v, \ell)$ says $v$ is tainted with source $\ell \in \mathcal{L}_{\text{src}}$.
- $\text{Sink}(v, \ell)$ says $v$ is passed to sink $\ell \in \mathcal{L}_{\text{sink}}$.
- $\text{Flow}(v, v')$ says that any taint on $v$ flows to $v'$.

Given a program $P$, the taint graph for $P$ is $G = (V, E)$, where $V = \mathcal{O} \cup \mathcal{U} \cup \mathcal{L}$. Here, $\mathcal{O}$ is the set of abstract objects in the program (each of which uniquely corresponds to an object allocation site), $\mathcal{U}$ denotes the program’s set of reference variables, and $\mathcal{L}$ is the set of labels introduced above. Figure 4 gives rules for generating an initial set of edges for the graph. Rules 1-4 handle primitive forms of statements. Rule 1 says that the contents of the abstract object $o$ flow to the reference $v$ on the left-hand side of the assignment (more formally, $v$ may point to $o$). Rule 2 similarly encodes the flow when a reference variable $v$ is assigned to another reference variable $u$. Rules 3 and 4 record the flows induced by field writes (or puts) and field reads (or gets) respectively; note that there is a distinct put/get operation for each field $f$.

Rules 5-7 formalize the description of the taint specifications $S$ given above. Rule 5 (symbol $\text{SrcRef}$) says that a source taints a reference variable, Rule 6 (symbol $\text{RefSink}$) says that the contents of a reference variable flow to a particular sink, and Rule 7 (symbol $\text{RefRef}$) says that the contents of one reference variable flow to another reference variable. Rules 8 and 9 are technical devices. Intuitively, Rule 8 ensures at least one abstract object flows to the target reference variable of any flow specification, which eliminates the need to include alias specifications for every method allocating a non-primitive return value (Rule 8 is discussed further in Section 3.2). Finally, Rule 9 allows us to express paths with “backwards” edges by introducing a label $\tau$ to represent the reversal of an edge labeled $\sigma$.

We handle interprocedural taint flow as follows: arguments passed by the caller are assigned to formal parameters, which are
assigned to the corresponding references in the callee. Values returned by the callee are assigned to a formal return value, which is assigned to the corresponding reference in the caller. Vertices representing formal parameters and formal return values are added to $G$ by the analysis, for example $\text{lat} \xrightarrow{\text{Put}} \text{this}\_\text{add} \xrightarrow{\text{Assign}} \text{list} \xrightarrow{\text{Assign}} \text{this}\_\text{get} \xrightarrow{\text{Get}} \text{ret}\_\text{get}$. This direction ensures that method boundaries are clear.

Figure 6 shows the taint graph generated from the code in Figure 1 and the specifications in Figure 2 using the rules in Figure 4. For clarity, we have not included formal parameters and formal return values in the graph, but have assigned caller arguments directly to the corresponding callee references and callee return values directly to corresponding caller return values. The graph describes the explicit flows in the program. For example, the edges $\text{arg}\_\text{add} \xrightarrow{\text{Put}} \text{this}\_\text{add} \xrightarrow{\text{Assign}} \text{list} \xrightarrow{\text{Assign}} \text{this}\_\text{get} \xrightarrow{\text{Get}} \text{ret}\_\text{get}$ capture the fact that any value stored in the list through the List.add method can potentially be the result of the List.get method. More precisely, the middle two edges show that the reference, the receiver of List.add, and the receiver of List.get all potentially point to the same abstract object.

The next step is to identify the paths through the graph that correspond to explicit taint flows, which we specify using the CFG $C_{\text{taint}}$ defined as follows:

$$\Sigma_{\text{taint}} = \{\text{New, Assign, SrcRef, RefSink, RefRet}\}$$

$$U_{\text{taint}} = \{\text{FlowTo, SrcObj, SrcSink}\} \cup \{\text{FlowTo}[f]|f \in F\}$$

where $F$ is the set of fields used in the program. The start symbol of the grammar is $T_{\text{taint}} = \text{SrcSink}$. The source vertices we consider are the taint source labels (i.e. $V_{\text{source}} = L_{\text{source}}$), and the sink vertices we consider are the sink labels (i.e. $V_{\text{sink}} = L_{\text{sink}}$).

The productions are shown in Figure 5. Rules 10-13 build the graph. Instead of keeping track of taint on the reference variables, it is simpler to keep track of taint on the objects. In Figure 6, Rule 14 adds the edge $L_{\text{oc}} \xrightarrow{\text{SrcObj}} \text{lat}_{\text{oc}}$, which says that $\text{lat}_{\text{oc}}$ is tainted. For this taint to flow to the sink, the analysis must be able to pass the taint to $\text{ret}\_\text{toString}$ (the object allocated to $\text{ret}\_\text{toString}$).

But here we encounter a problem: there is no explicit path from the object to the output of the method. Instead, there is an implicit flow through a sequence of look-ups converting digits in the double value to characters in the string. That is, information still flows from the input to the output of the method, but through control flow decisions rather than through explicit data flow.

Implicit flows are much more difficult to analyze precisely than explicit flows, because it is much harder to avoid tainted far too much of the program when analyzing implicit flows. Another way to handle implicit flows is to use explicit taint analysis but include appropriate specifications for methods that have implicit flows. We consider the set of specifications that pass taint from one argument to another argument or to the return value. In our specification $S$, this behavior is described by the relation $\text{Flow}(v, v')$, where $v \in V$ is a library method parameter and $v' \in V$ is a library method parameter or return value. Such specifications require Rule 8 from Figure 4, which adds phantom objects to the graph.

Recall that we attach taint to objects, not references. Thus, no taint can flow to parts of the code that have no objects associated with them. If a library method does not have an alias specification, then the method will have no object associated with its return value. Rather than manually add alias specifications for every method with a non-primitive return object, Rule 8 in Figure 4 automatically adds a (fresh) abstract object that the return value points to. For example, Rule 8 adds $\text{oc}_{\text{toString}} \xrightarrow{\text{ret}\_\text{toString}}$, causing Rule 10 to add
4. Problem Statement

In this section, we formulate the problem of performing a sound and precise CFL reachability analysis when some specifications are missing, along with the problem of inferring the missing specifications. Our formulation extends the framework defined in Section 3.

4.1 Missing Specifications CFL Reachability

Suppose we want to perform a CFL reachability analysis on a program $P$. Assume $G^* = (V^*, E^*)$ is the graph constructed from $P$ with complete specifications. If specifications are missing, then the constructed graph $\tilde{G} = (\hat{V}, \hat{E})$ may be missing vertices and edges, i.e., $\hat{V} \subseteq V^*$ and $\hat{E} \subseteq E^*$. The goal is to perform a sound and precise worst-case analysis given information about the missing vertices and edges. We encode the possible missing data as a family of graphs $G$, where we only know that $G^* \in G$.

DEFINITION 4.1. Suppose we are given $\tilde{G} = (\hat{V}, \hat{E})$, a set of sources $V_{source} \subseteq \hat{V}$, a set of sinks $V_{sink} \subseteq \hat{V}$, along with some family $G$ of graphs such that for each $G \in G$, $G$ is a subgraph of $G^*$, i.e., $G \subseteq G$. We call $G$ a completion of $\tilde{G}$. Let $A : V_{source} \times V_{sink} \rightarrow \{\text{true}, \text{false}\}$ be the result of a static analysis, where $A(v, v') = \text{true}$ indicates that taint flows from $v$ to $v'$.

1. class List:
   2.   void add(Object arg) {
        arg.f = arg;
   4.   this.val = arg; }

![Figure 7. An alternative (and incorrect) specification for List.add.](image)

- $A$ is sound if for every $v \in V_{source}$ and $v' \in V_{sink}$, $A(v, v') = \text{false}$ if and only if there does not exist any $G \in G$ such that $v \xrightarrow{T} v' \in G^C$.
- $A$ is precise if for every $v \in V_{source}$ and $v' \in V_{sink}$, $A(v, v') = \text{true}$ if and only if there exists $G \in G$ such that $v \xrightarrow{T} v' \in G^C$.

The idea behind this definition is that an analysis is sound if it does not miss any taint flow present in at least one of the possible completions of $\tilde{G}$, and the algorithm is precise if it does not report any flows that do not occur in any completion of $\tilde{G}$. An analysis $A$ solves the missing specifications CFL reachability problem for a family $G$ if it is both sound and precise.

While Definition 4.1 captures the notion of performing a worst-case analysis that is sound and precise, we are also interested in keeping track of the assumptions that the worst-case analysis makes about missing specifications. In practice, many assumptions may produce the same results. Therefore we are interested in producing a minimal set of assumptions. Suppose we have a partial order $(G, \leq)$, where $G_1 \leq G_2$ should mean that the graph $G_1$ makes at most as many assumptions as $G_2$. The definition of $\leq$ depends on the family $G$. In addition to producing sound and precise results $A$, we would like to produce a minimal $G \in G$ (with respect to $\leq$) such that performing the CFL reachability analysis on $G$ yields $A$.

DEFINITION 4.2. Suppose we are given the inputs as in Definition 4.1, along with a partial order $(G, \leq)$. We use the notation $e \in G$; this expression evaluates to true if $e \in G$ and false otherwise. The CFL reachability specification inference problem is to produce sound and precise results $A$, along with sufficient assumptions, encoded as a graph $G \in G$ satisfying $A(v, v') = (v \xrightarrow{T} v' \in G^C)$ for every $v \in V_{source}$ and $v' \in V_{sink}$. Furthermore, we require that $G$ is minimal, i.e., there does not exist sufficient assumptions $G' \in G$ such that $G' < G$.

In the remainder of this section, we describe how we apply this framework to inferring alias and flow specifications.
In general, there may be infinitely many possible paths \( \text{arg}_{\text{add}} \to \text{this}_{\text{add}} \), because the sequence of operations \( \alpha \) can be arbitrarily long. Thus, we need some compact way to represent an infinite language of strings; the regular languages are a natural choice. This discussion motivates our definition of the family \( G_{\text{add}}^R \):

**Definition 4.3.** Let \( W \subseteq \hat{V} \times \hat{V} \) and let \( R \) be a regular language over \( \Sigma \). The family of graphs \( G_{\text{add}}^R \) contains the graph \( G \) if

- \( \hat{G} \subseteq G \)
- If \( (w, w') \in W \), then the nondeterministic finite automaton (NFA) \( N \) with the transition matrix given by the subgraph \( G_{w,w'} \) satisfies \( L(N) \subseteq L(R) \).

Here, \( G_{w,w'} \) is the subgraph connecting \( w \) to \( w' \) (not including \( w \) or \( w' \)). This definition exploits the insight that we can minimize the subgraph \( G_{w,w'} \) as the transition graph of an NFA \( N \) with start state \( w \) and final state \( w' \). Any path \( w \to^\alpha w' \) through \( G_{w,w'} \) satisfies \( \alpha \in L(N) \). Conversely, for any \( \alpha \in L(N) \), there exists a path \( w \to^\alpha w' \in G_{w,w'} \).

In general, choosing \( R = \Sigma^* \) will produce sound results, since this choice imposes no constraints on the allowed paths connecting \( w \) and \( w' \). In practice, a more restrictive language may be chosen either to incorporate known constraints on potential specifications, or to trade some soundness for improved scalability.

### 4.3 Inferring Specifications for \( G_{\text{add}}^R \)

We now formulate the corresponding missing specifications problem. Our goal is to infer specifications of the following form: there exists a path (or set of paths) connecting \( w \) and \( w' \). As discussed above, both alias and flow specifications can be described in this manner. We need to define a partial order \( \leq \) on \( G_{\text{add}}^R \) that captures the notion of making minimal assumptions about missing specifications. Because we are searching for source-sink paths, it is natural to define \( \leq \) in terms of source-sink paths through \( G \). To simplify notation, we assume there is a single source and a single sink, i.e.,

\[ V_{\text{source}} = \{ v_{\text{source}} \} \quad \text{and} \quad V_{\text{sink}} = \{ v_{\text{sink}} \}. \]

Let \( \mathcal{P}(G) = V_{\text{source}} \to \).

\[ G_{\text{add}} \to \text{this}_{\text{add}} \text{ in Figure 6 is missing, giving us the graph } G. \text{ Without access to static fields, the only way to complete a flow through } G \text{ is if there is a path connecting } G_{\text{add}} \text{ to } \text{this}_{\text{add}}. \]

In general, for a method \( m \), the only possible taint flows through \( m \) are from a parameter of \( m \) to \( m \)'s return value, or from one parameter to another parameter. We use \( V_m \subseteq \hat{V} \) to denote the vertices of \( G \) corresponding to the formal parameters and return value of \( m \).

To be sound, we must assume that the specification of List.add could execute any sequence of operations. In other words, \( G \) consists of \( \hat{G} \) with some additional subgraph \( G_{\text{arg}_{\text{add}}} \text{this}_{\text{add}} \) connecting \( \text{arg}_{\text{add}} \to \text{this}_{\text{add}} \). Note that for any subgraph \( G_{\text{arg}_{\text{add}}} \text{this}_{\text{add}} \) corresponding to a possible specification of List.add’s behavior, the only information relevant to the CFL reachability problem is the possible sequences of terminals \( \alpha \in \Sigma^* \) that can occur along paths \( \text{arg}_{\text{add}} \to \text{this}_{\text{add}} \). Generalizing from this example, for a missing method \( m \) it suffices to consider the family of graphs \( \hat{G} \) consisting of all graphs containing \( \hat{G} \) as a subgraph with additional paths \( w \to w' \), where \( w, w' \in V_m \).

For example, one completion of \( \hat{G} \) is the taint graph \( G \) in Figure 6, which is just \( \hat{G} \) with the edge \( \text{arg}_{\text{add}} \to \text{this}_{\text{add}} \) added back in. But there are other ways to complete \( \hat{G} \), even for this simple example. Consider the graph \( G' \) obtained when the specification for List.add is given in Figure 7. Then \( G' \) is \( \hat{G} \) with the additional path

\[ \text{arg}_{\text{add}} \to \text{this}_{\text{add}} \]

In general, there may be infinitely many possible paths \( \text{arg}_{\text{add}} \to \text{this}_{\text{add}} \), because the sequence of operations \( \alpha \) can be arbitrarily long. Thus, we need some compact way to represent an infinite language of strings; the regular languages are a natural choice. This discussion motivates our definition of the family \( G_{\text{add}}^R \):

**Definition 4.3.** Let \( W \subseteq \hat{V} \times \hat{V} \) and let \( R \) be a regular language over \( \Sigma \). The family of graphs \( G_{\text{add}}^R \) contains the graph \( G \) if

- \( \hat{G} \subseteq G \)
- If \( (w, w') \in W \), then the nondeterministic finite automaton (NFA) \( N \) with the transition matrix given by the subgraph \( G_{w,w'} \) satisfies \( L(N) \subseteq L(R) \).

Here, \( G_{w,w'} \) is the subgraph connecting \( w \) to \( w' \) (not including \( w \) or \( w' \)). This definition exploits the insight that we can minimize the subgraph \( G_{w,w'} \) as the transition graph of an NFA \( N \) with start state \( w \) and final state \( w' \). Any path \( w \to^\alpha w' \in G_{w,w'} \) satisfies \( \alpha \in L(N) \). Conversely, for any \( \alpha \in L(N) \), there exists a path \( w \to^\alpha w' \in G_{w,w'} \).

In general, choosing \( R = \Sigma^* \) will produce sound results, since this choice imposes no constraints on the allowed paths connecting \( w \) and \( w' \). In practice, a more restrictive language may be chosen either to incorporate known constraints on potential specifications, or to trade some soundness for improved scalability.

### 5. Algorithms for Specification Inference

In this section we present an algorithm that solves the missing specifications problem stated in Definition 4.1 for \( G_{\text{add}}^R \). Next, we discuss an optimization that enables our algorithm to scale to large programs. Finally, we describe how to extend the algorithm to solve the specification inference problem stated in Definition 4.2 by using a shortest-path extension of the CFL reachability algorithm. This allows us to construct an algorithm that interacts with a human auditor to produce results that are sound and precise with respect to \( G^* \). An overview of our system is given in Figure 8.

#### 5.1 Algorithms for \( G_{\text{add}}^R \)

Consider \( (w, w') \in W \). Recall that every potential path \( w \to^\alpha w' \) satisfies \( \alpha \in L(R) \). To be sound and precise with respect to \( G_{\text{add}}^R \), it suffices to construct a subgraph connecting \( w \) and \( w' \) such that there is a path \( w \to^\alpha w' \) through this subgraph if and only if \( \alpha \in L(R) \).

The subgraphs that satisfy this property are the transition graphs for nondeterministic finite automata (NFAs) that accept \( L(R) \). For every \( (w, w') \in W \), Algorithm 1 constructs the transition graph \( N_{w \to w'} \) for one such NFA, and then adds this transition graph to \( G \) to connect \( w \) to \( w' \), resulting in graph \( G^* \). Finally, we compute
The transitive closure \((G')^C\) of \(G'\) with respect to the context-free language \(C\). The following correctness result follows from the correspondence between \(L(R)\) and \(N(s \xrightarrow{R} f)\) described above.

**Theorem 5.1.** Algorithm 1 is sound and precise for \(G^R_W\).

### 5.2 Optimizations

Consider the subgraph \(G_i = N(s_i \xrightarrow{R} f_i)\) constructed by Algorithm 1 for a pair \((w_i, w'_i) \in W\) (where \(i\) is an index over pairs \((w_i, w'_i)\)). One issue scaling Algorithm 1 is that the CFL reachability algorithm may add a large number of edges that are only among the vertices within \(G_i\). Any such internal edge \(n_i \xrightarrow{A} m_i\), where \(n_i\) and \(m_i\) are vertices in \(G_i\), is added whenever there is a path \(n_i \xrightarrow{\alpha} m_i \in G_i\), such that \(A \Gamma\alpha\). The problem is that the subgraphs \(G_i\) are all isomorphic—thus, the same edges are recomputed many times by the standard CFL reachability algorithm. This observation suggests that we can benefit from precomputing the internal edges.

One convenient way to implement this optimization is to preprocess the grammar \(C\) instead of adding graphs to \(G\). That is, we embed an (optimized) version of \(G_i\) in \(C\). Since the graphs \(G_i\) are all isomorphic to one another, this embedding only needs to be performed once. Intuitively, such an transformation is possible because \(G_i\) encodes a regular language and \(C\) is context-free.

The essential idea is that for every \((w_i, w'_i) \in W\), we replace \(G_i\) by a single vertex. Because we are only interested in summarizing the transitive closure of \(G_i\) with respect to \(C\), we only need one vertex to represent the net effect of \(G_i\), though this vertex may have many incident edges. More specifically, we modify \(G\) in the following ways to define a new graph \(\tilde{G}\):

- We add a new vertex \(v_i\) to \(\tilde{G}\); here the single vertex \(v_i\) will represent \(G_i\).
- We add two new, distinct terminal symbols \(b\) and \(e\) to \(\Sigma\), standing for “beginning \(G_i\) at \(w_i\)” and “exiting \(G_i\) at \(w'_i\),” respectively. These terminals are needed to mark in the modified grammar where we enter and exit \(G_i\).
- We add the edges \(w_i \xrightarrow{b} v_i \xrightarrow{e} w'_i\) to \(\tilde{G}\).

We now turn to incorporating the transitions of each \(G_i\) into \(C\), defining a new grammar \(\bar{C}\). Let \(n_i\), \(m_i\), and \(v_i\) denote vertices in \(G_i\) (corresponding to NFA states \(n_i\), \(m_i\), and \(v_i\), respectively). Let \(v \in V\) (recall that \(V\) is the set of vertices of \(\bar{G}\)—i.e., all the vertices not in any \(G_i\)). Finally, \(G'\) is the graph constructed by Algorithm 1. We want \(C\) and \(\bar{C}\) to correspond in the following way:

1. If there is an edge \(v \xrightarrow{A} n_i \in (G')^C\), then there should be an edge \(v \xrightarrow{A} s_i \xrightarrow{f_i} v'\) in \(\bar{G}\). Intuitively, the non-terminal \(A\) records that \(A\) was matched starting at the vertex in \(G_i\) corresponding to state \(n_i\).
2. If there is an edge \(n_i \xrightarrow{A} v \in (G')^C\), then there should be an edge \(v_i \xrightarrow{A} s_i \xrightarrow{f_i} v\) in \(\bar{G}\). Intuitively, the non-terminal \(A\) records that \(A\) was matched starting at the vertex in \(G_i\) corresponding to state \(n_i\).
3. If there is an edge \(n_i \xrightarrow{A} m_j \in (G')^C\) (that is not an internal edge), then there should be an edge \(v_i \xrightarrow{A} v_j \in \bar{G}\). Intuitively, the non-terminal \(A\) records that \(A\) was matched starting at the vertex in \(G_i\) corresponding to state \(n_i\), and ending at the vertex in \(G_j\) corresponding to state \(m_j\).

Finally, we need to define the productions for each of the additional non-terminals so that conditions 1-3 above are satisfied. We generate these additional productions of \(\bar{G}\) from the productions in \(C\) and the \(C\)-closure of \(G_i\) using the rules in Figure 10. In the figure, we refer to vertices \(s_i\) and \(f_i\) in \(G_i\) introduced by Algorithm 1, which correspond to the start state \(s\) and end state \(e\) of the NFA, respectively.

We briefly explain Rules 1a-f (the rules for non-terminals of the form \(A_{m_i}\)) for producing productions of \(\bar{G}\) in Figure 10; Rules 2a-f and 3a-h are similar.

(a) Suppose we have edge \(v \xrightarrow{\alpha} n_i \in (G')^C\). Because we have edge \(w_i \xrightarrow{\gamma} s_i \in G_i\), we produce \(v \xrightarrow{\alpha} s_i \in (G')^C\). In \(\bar{G}\), we
1. Productions for $A_n$:

(a) $A_n \rightarrow Ab \in C$

(b) $s_i \xrightarrow{A_n} n_i \in G_C$

(c) $n_i \xrightarrow{A_n} m_i \in G_C$  \[A_m \rightarrow A_n \in \bar{C}\]

2. Productions for $A^n$:

(a) $A^f \rightarrow e A \in \bar{C}$

(b) $n_i \xrightarrow{A^n} f_i \in G_C$

(c) $m_i \xrightarrow{A^n} n_i \in G_C$  \[A_m \rightarrow A^n \in \bar{C}\]

3. Productions for $A^n_m$:

(a) $A^n_m \rightarrow A^n b \in \bar{C}$

(b) $A^n_m \rightarrow e A_n \in \bar{C}$

(c) $n_i \xrightarrow{A^n_m} m_i \in G_C$  \[A_m \rightarrow A^n \in \bar{C}\]

(d) $m_i \xrightarrow{A^n_m} n_i \in G_C$  \[A^n_m \rightarrow A^n \in \bar{C}\]

4. Stitching productions:

(a) $A_n \rightarrow A_{f_{c_{v}} \in \bar{C}}$

(b) $A_n \rightarrow b A_{f_{c_{v}} \in \bar{C}}$

(c) $A_n \rightarrow B_{D_{f_{c_{v}}} \in \bar{C}}$

(d) $A_n \rightarrow B_{D_{f_{c_{v}}} \in \bar{C}}$

(e) $A_n \rightarrow B_{D_{f_{c_{v}}} \in \bar{C}}$

(f) $A_n \rightarrow B_{D_{f_{c_{v}}} \in \bar{C}}$

Algorithm 2 Optimized algorithm for $G^{R_W}_W$. Here, $T(G, C)$ applies the rules in Figure 10 to $C$ for the given graph $G$. Also, $s$ and $f$ are fresh vertices.

function PREPROCESS($C, \bar{G}, W, v, v'$)

|\[G_i \leftarrow N(s \xrightarrow{f}; \bar{G}, C)|\]

return $v \xrightarrow{f} v' \in \text{PREPROCESSHELP}(\bar{G}, W)$

end function

function PREPROCESSHELP($\bar{G}, W$)

|\[\bar{G} \leftarrow \bar{G}|\]

for all $(w, u') \in W$ do

if $w$ is a fresh vertex

|\[\bar{G} \leftarrow \bar{G} \cup \{v \xrightarrow{w} v, v \xrightarrow{w} w', v \xrightarrow{w} v, v \xrightarrow{w} \}|\]

end for

end function

Figure 10. Productions for $\bar{C}$.

Figure 11. Examples of production rules added by Figure 10, along with the rules that generate them.

Figure 12. Algorithm 2 adds the dashed edges to Figure 6 if the specification for $L_{\bar{e}}$ add is missing. We only show edges relevant to the production of the edge labeled (4).

have edges $v \xrightarrow{A_n} w_i \xrightarrow{b} v_i$, and we need to produce $v \xrightarrow{A} \bar{v}_i$. This is achieved by the production $A_n \rightarrow Ab \in \bar{C}$.

(b) Suppose we have internal edge $s_i \xrightarrow{A} n_i \in G_C$. Because we have edge $w_i \xrightarrow{b} s_i \in G'$, we produce $w_i \xrightarrow{A} n_i \in (G')^C$. In $G_{\bar{e}}$, we have edge $w_i \xrightarrow{b} v_i$, and we need to produce $w_i \xrightarrow{A_n} v_i$. This is achieved by the production $A_n \rightarrow b \in \bar{C}$.

(c) Suppose we have internal edge $n_i \xrightarrow{m_i} n_i \in G_C$ and edge $v \xrightarrow{A} n_i \in (G')^C$. Then we produce $v \xrightarrow{A_n} m_i \in (G')^C$. In $G_{\bar{e}}$, we have edge $v \xrightarrow{A_n} v_i$, and we need to produce $v \xrightarrow{A_m} v_i$. This is achieved by the production $A_m \rightarrow A_n \in \bar{C}$.

(d) Suppose we have edge $v \xrightarrow{A_n} n_i \in (G')^C$ and production $A \rightarrow B \in C$. Then we produce $v \xrightarrow{A_n} n_i \in (G')^C$. In $G_{\bar{e}}$, we have edge $v \xrightarrow{A_n} v_i$, and we need to produce $v \xrightarrow{A_n} v_i$. This is achieved by the production $A_n \rightarrow B_{D_{f_{c_{v}}} \in \bar{C}}$.

(e) Suppose we have edges $v \xrightarrow{A_n} v' \xrightarrow{D_{f_{c_{v}}} \in \bar{C}} n_i \in (G')^C$ and production $A \rightarrow B \in C$. Then we produce $v \xrightarrow{A_n} v_i \xrightarrow{A_m} v_i$. This is achieved by the production $A_m \rightarrow B_{D_{f_{c_{v}}} \in \bar{C}}$.

(f) Suppose we have edge $v \xrightarrow{A_n} n_i \in (G')^C$, internal edge $n_i \xrightarrow{m_i} n_i \in G_C$, and production $A \rightarrow B \in C$. Then we produce $v \xrightarrow{A_n} m_i \in (G')^C$. In $G_{\bar{e}}$, we have edge $v \xrightarrow{B_{D_{f_{c_{v}}} \in \bar{C}}} v_i$, and we need to produce $v \xrightarrow{A_n} v_i$. This is achieved by the production $A_m \rightarrow B_{D_{f_{c_{v}}} \in \bar{C}}$.

Note that cases (e) and (f) correspond to two possibilities for binary productions. (c) handling the case where one edge is fully outside of $G_i$, and (f) handling the case where one edge is fully inside $G_i$. In the case where only the middle vertex is in $G_i$ and both endpoints are outside, then we need the “stitching production” (c) in Figure 10. We describe stitching productions (a), (b), and (c):

(a) Suppose we have edge $v \xrightarrow{A} f_{i} \in (G')^C$. Because we have edge $f_{i} \xrightarrow{w'_i} G' \in G'$, we produce $v \xrightarrow{A} w'_i \in (G')^C$. In $G_{\bar{e}}$, we have edges $v \xrightarrow{A} v_i \xrightarrow{w'_i} v'$, and we need to produce $v \xrightarrow{A} w'_i$. This is achieved by the production $A \rightarrow A_{f_{C}} \in C$.

(b) Suppose we have edge $s_i \xrightarrow{v} v \in (G')^C$. Because we have edge $w_i \xrightarrow{b} s_i \in G'$, we produce $w_i \xrightarrow{A} v \in (G')^C$. In $G_{\bar{e}}$, we have edges $w_i \xrightarrow{b} v_i \xrightarrow{A} v$, and we need to produce $w_i \xrightarrow{A} v$. This is achieved by the production $A \rightarrow b A_{C} \in C$. 
(c) Suppose we have edges \( v \xrightarrow{B} n_i \xrightarrow{D} v' \in (G')^C \) and production \( A \rightarrow BD \in C \). Then we produce \( v \xrightarrow{A} v' \in (G')^C \). In \( G^{'P} \), we have edges \( v \xrightarrow{B} v_i \xrightarrow{D^m} v' \), and we need to produce \( v \xrightarrow{A} v' \). This is achieved by the production \( A \rightarrow B_i D^m \).

The stitching productions (d), (e), and (f) are similar to (c).

Finally, we show that the rules given in Figure 10 are complete. As above, we focus on Rules 1a-f first. Note that we need to add an edge \( v \xrightarrow{A} v_i \) whenever there exists a path \( v \xrightarrow{\alpha} w_i \) and there exists \( \beta \in \Sigma^* \) such that \( A \Rightarrow \alpha \beta \) and \( s_j \xrightarrow{\beta} n_i \in G_i \). In other words, \( \alpha \) is the portion of the path in \( G \) and \( \beta \) is the portion of the path in \( G_i \), and the path ends at vertex \( n_i \in G_i \). Consider the production that is the first step in the derivation of \( A \Rightarrow \alpha \beta \):

- Case \( A \rightarrow e \): then \( \alpha = e \), so \( v = w_i \), and we need to add edge \( w_i \xrightarrow{A} v_i \). The fact that \( \alpha = e \) also implies that \( A \Rightarrow \beta \), so \( s_i \xrightarrow{A} n_i \in G^{'P} \). Hence this case is handled by Rule 1b.
- Case \( A \rightarrow BD \): either \( \alpha \) is a prefix of \( BD \) and \( \beta \) is a suffix of \( D \) (handled by Rule 1e), or \( \alpha \) is a prefix of \( B \) and \( \beta \) is a suffix of \( B \) (handled by Rule 1f).
- Case \( A \rightarrow B \): then \( \alpha \) is a prefix of \( B \) and \( \beta \) is a suffix of \( B \), so this case is handled by Rule 1d.

Rule 1a is added to satisfy the semantics of the symbol \( b \in \Sigma \). Finally, we have to consider \( e \) transitions that occur in \( G_i^{'P} \)—i.e., \( n_i \xrightarrow{A} m_i \in G^{'P}_i \) (these transitions are used in conjunction with the implicit productions \( A \rightarrow eA \) and \( A \rightarrow Ae \)). These transitions are handled by Rule 1c. Rules 2a-f and 3a-h follow similarly.

Next, we show that Rules 4a-f are complete. Note that we need to add an edge \( v \xrightarrow{A} v_i \) whenever there exist paths \( v \xrightarrow{\alpha} w_i \) and \( w_i \xrightarrow{A} v' \), and there exists \( \beta \in L(R) \) such that \( A \Rightarrow \alpha \beta \gamma \). Here, \( \alpha \) and \( \gamma \) are the portions of the path in \( G \), and \( \beta \) is the portion of the path in \( G_i \). As before, we can consider production that is the first step in the derivation of \( A \Rightarrow \alpha \beta \gamma \). This time, we only need to handle the case where the production is split at vertex \( v_i \)—i.e., \( A \rightarrow BD \), where \( B \Rightarrow \alpha \beta \) and \( D \Rightarrow \beta_2 \alpha \beta_3 \); this is handled by Rule 4c. Rules 4d-f follow similarly when considering productions for \( A_n, A^*_n \), and \( A^n \). Finally, the semantics of the symbols \( b \) and \( e \) are handled by Rules 4b and 4a, respectively. While we only described the case where the path passes through a single pair \((w_i, w'_i)\), the general case follows because the first step in the derivation can be split only at a single vertex \( v_i \).

We denote the subroutine constructing \( \tilde{G} \) by \( \mathcal{T} \), i.e., \( \tilde{G} = \mathcal{T}(G, C) \). Note that any \( G_i \) can be used, since \( G_i \) (and hence \( G^{'P}_i \)) is the same for every \((w_i, w'_i) \in W \). Algorithm 2 calls \( \mathcal{T} \) to obtain a new grammar \( \tilde{G} \). It then computes the transitive closure \( \tilde{G}^{'P} \). We have the following correctness result:

**Theorem 5.2.** Algorithm 2 is sound and precise for \( G^{'P}_W \).

We briefly discuss the complexity of Algorithm 2. The rules in Figure 10 are not recursive, so the number of productions in \( C \) is a constant multiple of the number of productions in \( C \). Similarly, the graph \( G \) constructed by Algorithm 2 is a constant multiple of the size of \( G \). The complexity of Algorithm 2 is dominated by the complexity of computing the transitive closure \( \tilde{G}^{'P} \). This is \( O(|G|^3|C|^3) \) (where \(|G|\) is the number of vertices in \( G \), and \(|C|\) is the number of terminals and non-terminals in \( C \)) [19].

As an example, consider \( G^{'P}_W \) defined in Section 4. Let \( \Sigma_W = \{ \text{New, Assign}, \{ \text{Get}[f'], \text{Put}[f'] \} \mid f' \in \mathcal{F} \} \).

Recall that \( R_{\text{alias}} = (\text{Assign} + \text{Assign})(\Sigma_W \cup \Sigma_T)'(\text{Assign} + \text{Assign}) \). Then \( N(s, R_{\text{alias}} f) \) produces the transition graph for a single state \( NFA \ N = \{(s, n, f), \delta, s, f\} \), where \( n \in \delta(s, \text{Assign}), n \in \delta(s, \text{Assign}), n \in \delta(\sigma, n) \) for \( \sigma \in \Sigma_W \cup \Sigma_T, f \in \delta(\text{Assign}, n) \), and \( f \in \delta(\text{Assign}, n) \). The productions for \( \text{FlowsTo}_n \) and \( \text{FlowsTo}[f']_n \) are shown in Figure 11.

Consider the code in Figure 1, and suppose the specification for \( \text{Assign} \) is the shortest string in a context-free grammar [14]. This algorithm generalizes to solving the shortest-path CFL reachability problem: we replace the worklist of edges in [19] with a heap of edges, where the priority of an edge is the weight of its shortest path (see Algorithm 4 in Appendix A for details). By using Algorithm 4 to compute the closure \( G \), we find the source-sink path \( p^* \in \mathcal{P}(G) \).
that passes through the fewest possible edges \( w \xrightarrow{b} v \xrightarrow{r} w' \). Then the sufficient assumptions \( G \) constructed from \( p^* \) has minimum weight.

Finally, we design an algorithm for interactively refining the static analysis results with the help of an auditor. Recall that the graph \( G \) is missing some vertices and edges from \( G^* \). Suppose we can query an oracle to obtain information about \( G^* \):

**Definition 5.4.** We say \( O \) is an oracle for \( G \) if for every \((w, w') \in W, O(w, w') = G^*_w,w'\).

We use a human auditor as an oracle \( O \). On input \((w, w')\), the auditor examines the library documentation and return the true specification \( G^*_w,w' \). The problem is to produce static analysis results that are sound and precise with respect to \( G^* \), while making as few queries \( O(w, w') \) as possible.

Algorithm 3 solves this problem. It obtains the shortest path \( p^* \in P(G) \) for the edge \( e' = v_{\text{source}} \overset{r}{\to} v_{\text{sink}} \) by calling \( p^* \leftarrow \text{SHORTESTPATH}(G, G, e') \). Then the algorithm replaces every edge \( w \xrightarrow{b} v \xrightarrow{r} w' \) in \( p^* \) with \( O(w, w') \). Algorithm 3 repeats this process until either weight \((p^*) = 0, \) or until \( P(G^*) = \emptyset \). In the former case, the path \( p^* \) does not contain any symbols \( b \) or \( e \), i.e. \( p^* \) does not pass through any potentially missing specifications. This proves that \( p^* \subseteq G^* \), i.e. \( e' \in G^* \subseteq (G^*)^* \). In the latter case, because Algorithm 2 is sound, it only returns \( p^* = 0 \) if there does not exist any \( G \in G^*_W \) such that \( e' \in G \). Since we have assumed that \( G^* \in G^*_W \), this proves that \( e' \notin (G^*)^* \). Therefore:

**Theorem 5.5.** Algorithm 3 computes \( v_{\text{source}} \overset{r}{\to} v_{\text{sink}} \in (G^*)^* \).

### 6. Implementation

We have implemented the system described in Figure 8 for Java, and specifically for the Android framework. A number of extensions are required beyond the essentials we have described, but these extensions do not introduce any new ideas. For example, we include rules for primitive variables that are essentially rules for reference variables without fields. Our current implementation does not consider specifications involving static fields; as discussed in Section 4, in our experience such specifications lead to many false positives as there are few constraints on the possible flows between static variables, and in practice there are few flows among them.

Prior to this work, we had manually written many specifications over a period of more than one year. These baseline specifications \( S \) cover 176 Android library classes, including a number of taint sources (location, device ID, SIM data, contacts, and calendar data) and a number of sinks (network sockets, SMS messages, and user settings modifications). This baseline also includes specifications for important container objects including ArrayList, LinkedList, and HashMap. Since our focus is on the long tail of missing specifications, we bootstrap our implementation of the specification inference framework with these baseline specifications.

In our implementation, program fact extraction is performed using the Chord platform [20] modified to work with the Jimple intermediate representation provided by Soot [30].

### 7. Experimental Results

We ran our tool on a corpus of 179 Android apps. Our results are for the optimized version of our specification inference algorithm, i.e. Algorithm 2, since preliminary experiments with Algorithm 1 did not scale even to apps of moderate size. The running time for one iteration of Algorithm 3 is plotted in Figure 16(c). The flow specification inference algorithm ran on all 179 apps, running in fewer than 10 seconds per iteration on average for most apps, which is fast enough to allow a human auditor to interactively run the analysis. The alias specification inference algorithm successfully ran on 156 apps. The worst-case assumptions cause a substantial increase in the points-to relation size (see Figure 16(d)), which proved to be too large on the remaining 23 apps. Still, inferring alias specifications runs in under 100 seconds for most apps with up to 100,000 lines of Jimple code.
<table>
<thead>
<tr>
<th>App Name</th>
<th>Jimple LOC</th>
<th># Spec.</th>
<th># Critical Spec.</th>
<th>Total # Spec.</th>
<th>Rounds</th>
<th>Accuracy</th>
<th>Flows</th>
<th>Run Time (s)</th>
<th>Spec. Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>41124</td>
<td>471943</td>
<td>20</td>
<td>9</td>
<td>5553</td>
<td>5</td>
<td>0.5116</td>
<td>42</td>
<td>178.6</td>
<td>Flow</td>
</tr>
<tr>
<td>APG-M</td>
<td>421026</td>
<td>24</td>
<td>16</td>
<td>10361</td>
<td>16</td>
<td>0.2662</td>
<td>48</td>
<td>607.3</td>
<td>Flow</td>
</tr>
<tr>
<td>browser</td>
<td>265875</td>
<td>142</td>
<td>15</td>
<td>4242</td>
<td>18</td>
<td>0.1690</td>
<td>64</td>
<td>12.20</td>
<td>Flow</td>
</tr>
<tr>
<td>OC2B78</td>
<td>263014</td>
<td>38</td>
<td>15</td>
<td>4937</td>
<td>5</td>
<td>0.5789</td>
<td>38</td>
<td>14.24</td>
<td>Flow</td>
</tr>
<tr>
<td>highrail</td>
<td>193518</td>
<td>48</td>
<td>26</td>
<td>4388</td>
<td>5</td>
<td>0.6042</td>
<td>44</td>
<td>10.38</td>
<td>Flow</td>
</tr>
<tr>
<td>iyw</td>
<td>136800</td>
<td>4</td>
<td>4</td>
<td>3454</td>
<td>2</td>
<td>1.0</td>
<td>2</td>
<td>2.357</td>
<td>Flow</td>
</tr>
<tr>
<td>ConnectBot</td>
<td>109286</td>
<td>0</td>
<td>0</td>
<td>5440</td>
<td>1</td>
<td>N/A</td>
<td>0</td>
<td>3.769</td>
<td>Flow</td>
</tr>
<tr>
<td>yaaic</td>
<td>44478</td>
<td>14</td>
<td>5</td>
<td>3029</td>
<td>6</td>
<td>0.3571</td>
<td>2</td>
<td>1.607</td>
<td>Flow</td>
</tr>
<tr>
<td>highrail</td>
<td>265875</td>
<td>2</td>
<td>2</td>
<td>5167</td>
<td>2</td>
<td>1.0</td>
<td>1</td>
<td>9623</td>
<td>Alias</td>
</tr>
<tr>
<td>andmj</td>
<td>239227</td>
<td>0</td>
<td>0</td>
<td>5229</td>
<td>1</td>
<td>N/A</td>
<td>0</td>
<td>1555</td>
<td>Alias</td>
</tr>
<tr>
<td>ce667f</td>
<td>193518</td>
<td>0</td>
<td>0</td>
<td>5509</td>
<td>1</td>
<td>N/A</td>
<td>0</td>
<td>3234</td>
<td>Alias</td>
</tr>
<tr>
<td>ConnectBot</td>
<td>136800</td>
<td>0</td>
<td>0</td>
<td>3293</td>
<td>1</td>
<td>N/A</td>
<td>0</td>
<td>3193</td>
<td>Alias</td>
</tr>
<tr>
<td>05ed82</td>
<td>134230</td>
<td>1</td>
<td>1</td>
<td>12632</td>
<td>2</td>
<td>1.0</td>
<td>3</td>
<td>2254</td>
<td>Alias</td>
</tr>
<tr>
<td>SMSBot</td>
<td>134230</td>
<td>15</td>
<td>3</td>
<td>4693</td>
<td>7</td>
<td>0.467</td>
<td>2</td>
<td>284.3</td>
<td>Alias</td>
</tr>
<tr>
<td>yaaic</td>
<td>109286</td>
<td>0</td>
<td>0</td>
<td>4161</td>
<td>1</td>
<td>N/A</td>
<td>0</td>
<td>816.1</td>
<td>Alias</td>
</tr>
<tr>
<td>ca70f4</td>
<td>81974</td>
<td>3</td>
<td>0</td>
<td>3307</td>
<td>3</td>
<td>0.0</td>
<td>0</td>
<td>45.57</td>
<td>Alias</td>
</tr>
<tr>
<td>tomroid</td>
<td>44478</td>
<td>0</td>
<td>0</td>
<td>3697</td>
<td>1</td>
<td>N/A</td>
<td>0</td>
<td>68.154</td>
<td>Alias</td>
</tr>
<tr>
<td>a1d5b8</td>
<td>41682</td>
<td>2</td>
<td>1</td>
<td>2285</td>
<td>3</td>
<td>0.5</td>
<td>1</td>
<td>7.847</td>
<td>Alias</td>
</tr>
</tbody>
</table>

Figure 14. Specification inference results on large Android apps. Critical specifications are both correct and lie along a true source-sink path. “Total # specifications” is the number of worst-case specifications, “Rounds” is the number of iterations in Algorithm 3, “Accuracy” is the proportion of proposed specifications that are correct, and “Flows” is the number of new flows discovered (accuracy is N/A if no specifications are inferred).

Results for selected apps, including the largest four that ran successfully for each analysis, are shown in Figure 14. We show the number of specifications proposed by our tool (# Spec.), along with the number of worst-case specifications (Total # Spec.). Our tool may propose specifications that are correct (i.e., represent valid paths through library methods), but do not contribute to a correct source-sink path in the program. Therefore, we also show the number of specifications that are both correct and contribute to a true source-sink flow (# Critical Spec.), which is a lower bound for the number of specifications proposed. We call such specifications critical specifications, because they are the specifications that an auditor must examine in order to find all source-sink flows. The number of inferred specifications is plotted as black circles in Figure 16(a) and (b), along with the number of inferred specifications that are correct (plotted as red triangles), and the number of critical specifications (plotted as blue diamonds). For readability, (a) is a log-log plot, and the x-axis of (b) is log-scale. The accuracy of the aggregated specifications are shown in Figure 13. Note that the accuracy is directly correlated with the manual labor required by the oracle: higher accuracy means that the oracle will have to examine fewer incorrect specifications.

### 7.1 Specification Inference Accuracy

Our first experiment demonstrates the accuracy of the specifications inferred. For each app, we ran our inference algorithm with the baseline specifications $S$. The inferred alias specifications are very accurate, in part because of type filters. We show some examples of inferred specifications in Figure 15.

We compare our results to randomly chosen specifications. We randomly choose 50 possible flow specifications in the following way: randomly choose a method, randomly choose a pair of formal parameters $v$ and $v'$ (or a formal parameter $v$ and the formal return value $v'$), and propose the flow specification $v \xrightarrow{t} v'$. The accuracy of a specification randomly chosen in this way is only 0.140, whereas the overall accuracy of the specifications inferred by our tool is 0.433.

The number of specifications proposed, which grows roughly linearly with app size, is very manageable. It is usually a small multiple of the number of critical specifications. For alias specification inference, each app produced fewer than 20 proposals, each of which could be checked in under a minute. Significantly more flow specifications are inferred, but these are even faster to check. All but five of the apps required fewer than 100 proposals. In total, the tool helped discover hundreds of new specifications and flows, a task that we estimate would have taken weeks without the tool.

### 7.2 Specification Aggregation

Our second experiment demonstrates how our tool can be used to quickly build a useful collection of library specifications. Consider analyzing the apps in some arbitrary order and aggregating specifications along the way: that is, the $i$th app is analyzed using all of the correct specifications discovered in analyzing the first $i-1$ apps. Intuitively, the most frequently used methods should have their specifications discovered relatively early in the process and we should subsequently benefit from already having those specifications and not needing to infer them again.

Figure 16(e) shows the proportion of new specifications proposed by Algorithm 3 with aggregation to the number of specifications proposed when Algorithm 3 starts from the baseline. The red triangle series shows, for each app, the percentage of new specification-
Figure 16. For (a) flow and (b) alias: # specifications proposed (black, circle), # correct (red, triangle), and # critical (blue, diamond). (c) Run time of the flow (black, circle) and alias (red, triangle) specification inference algorithms. (d) Ratio of worst-case points-to relation size to known points-to relation size. (e) Ratio of # specifications with aggregation to # specifications from baseline, averaged over 100 random orders (black line), and for two different random orders (red triangle, blue diamond). (f) Proportion of common specifications proposed, for \( c = 2 \) (black, solid), \( c = 3 \) (blue, dashed), and \( c = 4 \) (red, dotted), averaged over 100 random orders.

Figure 16. For (a) flow and (b) alias: # specifications proposed (black, circle), # correct (red, triangle), and # critical (blue, diamond). (c) Run time of the flow (black, circle) and alias (red, triangle) specification inference algorithms. (d) Ratio of worst-case points-to relation size to known points-to relation size. (e) Ratio of # specifications with aggregation to # specifications from baseline, averaged over 100 random orders (black line), and for two different random orders (red triangle, blue diamond). (f) Proportion of common specifications proposed, for \( c = 2 \) (black, solid), \( c = 3 \) (blue, dashed), and \( c = 4 \) (red, dotted), averaged over 100 random orders.

The required work would drop further after processing more apps.

7.3 Verification

We ran Algorithm 3 to termination, i.e., until no new specifications for missing parts of the program could add any more taint flows, which means that the remaining taint flows all occur in the original graph \( \hat{G} \). In the case of alias inference, this also proves that no additional explicit taint flows can occur. Figure 16(b) shows the number of specifications that had to be checked by an auditor to completely verify the absence of explicit taint flows in an app. This number is very reasonable (at most 15), showing that the tool makes verification of large apps practical. In the case of flow specification inference, taint flows due to missing alias specifications can still
occur, and verification requires that the auditor supply all relevant alias specifications. In practice, this analysis still discovers many library methods that need alias specifications, since taint often flows forwards through these methods. Thus our tool finds many taint flows even if alias specifications are missing.

There are currently two primary limitations to our tool. One we have already discussed: inferring both alias and flow specifications simultaneously is too expensive for our tool on some apps. The second is that while we can infer missing flow and alias specifications, we still require a complete list of the possible sources and sinks in the program to be able to find flows at all. While manually annotating sources and sinks is a much easier problem (by orders of magnitude) than finding flows, it would still be useful to consider how to provide automatic assistance in discovering sources and sinks in large apps.

As can be seen from the total number of potential specifications shown in Figure 14, without our tool an auditor would have to examine a huge number of potential specifications. Even if many of these can be easily eliminated, our experience has been that without the aid of our tool, performing verification on moderately sized apps can take hours or even days, and performing verification on large apps is almost impossible.

8. Related Work

Sound analysis with missing specifications. There has been previous work that shares our goal of inferring specifications for verification based on abductive inference [19]. There are several differences in the two approaches. First, we handle the general class of CFL reachability problems, whereas [34] addresses standard graph reachability problems. Second, abductive inference is a very general tool and solving abductive inference problems is NP-hard. Our algorithm, which is tailored to specification inference, runs in polynomial time. As a result, our system appears to scale considerably better, and we are able to conduct a much larger experiment on many more apps than [34]. There has also been prior work on sound call graph generation [2]. Their work constructs a placeholder library that exhibits every possible behavior that can affect the call graph. Our technique, in addition to being more general, actually proposes library specifications and allows an auditor to interactively refine analysis results.

Specification inference. There has been work on using data mining to infer specifications [3, 4, 7, 15, 16, 21, 22, 25, 33]. Several of these techniques use dynamic traces to propose specifications [4, 33], while others apply domain knowledge and statically infer specifications from the source code [7, 15, 16, 22, 25]. Our tool differs in that it applies to CFL reachability problems. Furthermore, our tool produces sound results, which enables interactive verification.

CFL reachability for program analyses. A large number of program analyses have been expressed as CFL reachability problems, for example points-to analysis [26, 28], many set constraint problems [11, 13, 19], various interprocedural analyses [23, 24], and type qualifier inference [10]. Our work makes these techniques more applicable for whole-program analysis by providing a practical and sound approach to dealing with missing or hard-to-analyze portions of the program. Our work makes use of ideas for combining context-free reachability with additional regular language properties such as [12].

Applications of program analysis to security. Work on taint analysis for Android includes SCANDroid [9], which statically tracks taint flows between applications, TaintDroid [8], which is a dynamic system that performs real-time monitoring, and FlowDroid [5], which uses static analysis to find information leaks. Static analysis has also been applied to finding vulnerabilities in web applications [17, 27, 29, 32].

Algorithm 4 Algorithm for computing shortest-path CFL reachability.

```plaintext
function SHORTESTPATH(G, C, e)
    G’ ← GRAPH(); I ← MAP(); H ← HEAP()
    for all v → v’ ∈ G do
        H.UPDATE(v → v’, weight(σ))
    I[v → v’] ← ∅
    end for
    while ¬H.EMPTY() do
        [v ← v’, P_CU] ← H.DELETE_MIN()
        G’ ← G’ ∪ {v → v’} for all D → AB ∈ C do
            for all v’ → B v’’ ∈ G’ do
                P_new ← P_cu + H.PRIORITY(v’ → v’’)
                if P_new < H.PRIORITY(v → v’’ then
                    H.UPDATE(v → v’, P_new)
                    I[v → v’’] ← [v ← v’, v’ → v’’]
                end if
            end for
        end for
        for all D → BA ∈ C do
            for all v’’ → B v ∈ G’ do
                P_new ← H.PRIORITY(v’’ → v) + P_cu
                if P_new < H.PRIORITY(v’ → v’’ then
                    H.UPDATE(v’ → v’’, P_new)
                    I[v’ → v’] ← [v’ → v’, v ← v’’]
                end if
            end for
        end for
        for all B → A ∈ C do
            if P_cu < H.PRIORITY(v → v’’ then
                H.UPDATE(v → v’, P_cu)
                I[v → v’] ← [v ← v’, v ← v’’]
            end if
        end for
    end while
    if e ∈ I then
        return GETPATH(I, e)
    end if
    return ∅
end function

function GETPATH(I, e)
    if I[e] = ∅ then
        return e
    end if
    [e1, e2, ..., e_k] ← I[e]
    return GETPATH(e1) + ... + GETPATH(e_k)
end function
```

9. Conclusions

We have developed a general framework that applies to any program analysis formulated as a CFL reachability problem. Our framework allows us to perform a sound analysis by inferring missing specifications, and furthermore allows an auditor to interactively refine the results. We have demonstrated the quality of the specifications inferred by our tool on a corpus of 179 real-world Android apps. Our results show that our tool can both help build large collections of specifications very efficiently, and make it practical for an auditor to perform verification.

A. Shortest-Path CFL Reachability

The shortest-path algorithm in Algorithm 4 generalizes Knuth’s algorithm for finding shortest strings in CFLs to computing shortest-path CFL reachability. Essentially, Knuth’s algorithm [14] builds
on an algorithm for determining emptiness of a context-free grammar: it adds a heap that keeps track of the shortest sequence of terminals that can be derived from each non-terminal symbol. Similarly, Algorithm 4 generalizes the algorithm for computing CFL reachability described in [19]. In the pseudocode, arrays are denoted as \([x_1, \ldots, x_k]\), and addition of arrays is defined to be \([x_1, \ldots, x_k] + [y_1, \ldots, y_k] = [x_1 + y_1, \ldots, x_k + y_k]\).

We introduce a heap \(H\) that keeps track of the shortest path for each edge \(v \xrightarrow{\epsilon} v'\). The current priority of the edge \(v \xrightarrow{\epsilon} v'\) is the length of the current shortest path. If a shorter path is found, then the heap is updated with the new path and the new priority. At every iteration of the algorithm, the lowest priority edge \(v \xrightarrow{\epsilon} v'\) (with priority \(P_{cur}\)) is removed from the heap, added to \(G'\), and then processed. Note that once this happens, there can be no way of producing \(v \xrightarrow{\epsilon} v'\) with lower priority; every subsequent edge removed from \(H\) must have priority at least \(P_{cur}\), so any edge produced while processing such an edge must also have priority at least \(P_{cur}\). Since every possible way of producing \(v \xrightarrow{\epsilon} v'\) is considered, the shortest path is correctly identified.

The heap \(H\) supports the following operations: \(\text{UPDATE}(e)\) updates the priority of edge \(e\) (and adds \(e\) to \(H\) if \(e \notin H\)), \(\text{EMPTY}(P_{cur}\) returns true if the heap contains no edges, \(\text{PRIORITY}(e)\) returns the current priority of edge \(e\), and \(\text{DELETEMINS}\) removes the lowest priority element in the heap and returns it (along with its current priority). We assume that \(\text{PRIORITY}\) returns \(\infty\) for edges not yet added to \(H\), and \(0\) for edges already removed from \(H\). The complexity of Algorithm 4 is \(O(|G|^3|C|^3\log |G| + \log |C|))\) because of the additional cost of updating the heap (as before, \(|G|\) is the number of vertices in \(G\), and \(|C|\) is the number of terminals and non-terminals in \(C\)).

The shortest path itself is stored in a map \(I\), which keeps track of the edges \(e_k(1, 2, \ldots, k)\) (where \(k \in \{0, 1, 2\}\)) used to produce \(v \xrightarrow{\epsilon} v'\). The shortest path itself is reconstructed by recursively querying the shortest path for each edge in \(I[v \xrightarrow{\epsilon} v']\).

Algorithm 4 does not handle edges labeled with the empty string \(\epsilon\), or productions \(A \rightarrow \epsilon\). In order to handle the former, our solver introduces a fresh terminal symbol \(\overline{\epsilon}\), replaces every edge \(v \xrightarrow{\epsilon} v'\) with \(v \xrightarrow{\epsilon}\overline{\epsilon} v'\), and adds productions \(A \rightarrow \overline{\epsilon}A\) and \(A \rightarrow \overline{\epsilon}C\) for every non-terminal \(A\) in the input grammar. The latter is handled by adding self loops \(v \xrightarrow{\overline{\epsilon}} v\) for every \(v \in V\) and every \(A \in U\) such that \(A \rightarrow \epsilon \in C\) (see [19]).

Acknowledgments

This material is based on research sponsored by the Air Force Research Laboratory, under agreement number F30602-12-2-0020. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References