Modelgen: Mining Explicit Information Flow Specifications from Concrete Executions

Lazaro Clapp*, Saswat Anand*, Alex Aiken*
*Stanford University, USA

***Abstract*—We present a technique to mine explicit information flow specifications from concrete executions. These specifications can be consumed by a static taint analysis, enabling static analysis to work even when method definitions are missing or portions of the program are too difficult to analyze statically (e.g., due to dynamic features such as reflection). We present an implementation of our technique for Java and the Android platform. When compared to a set of manually written specifications for 309 methods across 51 classes, our technique is able to recover 96.36% of these manual specifications and produces many more correct annotations that our manual models missed (97.12% vs 79.12% precision). We incorporate the specifications generated by our technique into an existing static taint analysis system, and show that they enable it to find additional true flows. Although our implementation is Android-specific, our approach is applicable to other application frameworks.**

I. Introduction

Scaling a precise and sound static analysis to real-world software is challenging, especially for software written in modern object-oriented languages such as Java. Typically such software builds upon large and complex frameworks (e.g., Android, Apache Struts, and Spring). For soundness and precision, any analysis of such software entails analysis of the framework. However, there are at least four problems that make the analysis of framework code challenging. First, a very precise analysis of a framework may not scale because most frameworks are very large. Second, framework code may use dynamic language features, such as reflection in Java, which are difficult to analyze statically. Third, frameworks typically use non-code artifacts (e.g., configuration files) that have special semantics that must be modeled for accurate results. Fourth, frameworks themselves usually build on abstractions written in lower-level languages for which a comprehensive static analysis may not be available (e.g., Java's native methods). Such foreign functions appear simply as missing code to the static analysis of the higher-level language.

One approach to address these problems is to use specifications (also called *models*) for framework classes and methods. From a high-level, a specification reflects those effects of the framework code on the program state that are relevant to the analysis. The analysis can then use these specifications instead of analyzing the framework. Use of specifications can improve the scalability of an analysis dramatically because specifications are usually much smaller than the code they specify. In addition to scalability, use of specifications can also improve the precision of the analysis because specifications are also simpler (e.g., no dynamic language features or non-code artifacts) than the corresponding code.

Although use of specifications can improve both scalability and precision of an analysis, obtaining specifications is a challenging problem in itself. If specifications are computed by static analysis of the framework code, the aforementioned problems arise. An alternative approach is to manually write specifications. This approach is not impractical because once the specifications for a framework are written, those specifications can be used to analyze any piece of software that uses that framework. However, writing and maintaining specifications manually for a large framework is still laborious and susceptible to human error. Dynamic analysis, which observes concrete executions of a program and generalizes to produce specifications, represents an attractive third alternative. Mining specifications from execution traces is not a novel idea. For example, some techniques produce control-flow specifications (e.g., [2, 46, 32, 19, 34]), while others discover general pre- and post-conditions on methods (e.g., Daikon [14]). However, we are interested in inferring information flow specifications from program executions, a problem that, to the best of our knowledge, has not been previously explored.

We mine explicit information flow specifications by executing each method for which we wish to construct a model and recording a trace of all operations performed by the method. Using this trace, we reconstruct the view the method has of the structures in the heap reachable from the method's arguments. We apply a specialized from of dynamic taint tracking to capture the information flows between locations inside those structures. We then lift these dynamic information flows to a static signature summarizing the flows between a method's arguments or between an argument and the return value. Finally, we combine the flows mined from different executions of a method to produce its overall specification.

We evaluate our generated specifications in three ways. First, we compare them to a set of specifications which were manually written over a period of two years. Our technique independently discovers 96.36% of the manual models and finds many additional correct specifications missed by human model writers. Second, we give our specifications as models for a static taint analysis. The specifications allow the analysis to discover over 31% additional flows, many of which we found to be true positives, while preserving a 98.12% recall compared to the same tool using only manual models. Third, we show that we are able to mine useful specifications from only a few executions of a method (1.38 in average) that are as good as those mined from large sets of traces.

```
// Set-up objects
SocketChannel socket = ...;
CharBuffer buffer = ...;
CharsetEncoder encoder =
    Charset.forName("UTF-8").newEncoder();
TelephonyManager tMgr = ...;

// Leak phone number:
String mPhoneNumber = tMgr.getLine1Number();
CharBuffer b1 = buffer.put(mPhoneNumber,0,10);
ByteBuffer bytebuffer = encoder.encode(b1);
socket.write(bytebuffer);
```

Fig. 1. Leak phone number to Internet

TelephonyManager.getLine1Number()	$PHONE_NUM → return
CharBuffer.put(String,int,int)	arg#1 → this this → return arg#1 → return
CharsetEncoder.encode(CharBuffer)	arg#1 → return
SocketChannel.write(ByteBuffer)	arg#1 → !INTERNET

TABLE I
SPECIFICATIONS FOR PLATFORM METHODS

We begin by giving a motivating example for the value of our technique (Section II) and describe the overall architecture of our implementation (Section III). We then present our specification mining technique in detail (Section IV). Next, we describe our empirical evaluation and present our results (Section V). Finally, we summarize related work (Section VI) and conclude (Section VII).

II. MOTIVATION

As part of a long term research project to improve malware detection techniques for mobile platforms, our group has developed STAMP. STAMP is a hybrid static/dynamic program analysis tool for Android applications: The core analysis performed by STAMP is a static taint analysis that aims to detect privacy leaks. Given the code fragment in Figure 1, STAMP should infer that the device's phone number (retrieved by `getLine1Number()`) is sent to the Internet (using `socket`) and flag it as a potential leak.

STAMP performs whole-program analysis of the Android application code and any libraries bundled into its installer (.apk file). However, because of the challenges discussed in Section I, STAMP does not directly analyze the Android platform's libraries. In Figure 1, STAMP's static analysis component has no way of inspecting the behavior of `tMgr.getLine1Number()`, `buffer.put()`, `encoder.encode()` or `socket.write()`.

The simplest solution to this problem is to manually write a specification of the information flow properties of each platform method. These specifications can then be loaded by the static analysis and assumed to be an accurate representation of the corresponding methods. This is the approach we adopted for early versions of STAMP. Table I shows the specifications for the methods in Figure 1. The notation is as follows:

$a \to b$ indicates that there is a possible flow from a to b. Whatever information was accessible from a before the call is now potentially accessible from b after the call. If a is a reference, the information accessible from a includes all objects transitively reachable through other object references in fields.

this is the instance object for the modeled method.

return is the return value of the method.

arg#i is the i-th positional argument of the method. For a static method, argument indices begin at 0. For instance methods, *arg#0* is an alias for *this* and positional arguments begin with *arg#1*.

$SOURCE is a source of information flow and represents a resource, external to the program, from which the API method reads some sensitive information (e.g. `$CONTACTS`, `$LOCATION`, `$FILE`).

!SINK is an information sink and represents a location outside of the program to which the information flows (e.g. `!INTERNET`, `!FILE`).

Given the specifications in Table I, STAMP can track the flow of sensitive information from `$PHONE_NUM`—through parameters and return values—to `!INTERNET`, via static analysis of the code in Figure 1.

Over a period of two years, we produced a large set of manually-written models. Generating these models was a non-trivial task, as it required running STAMP on various Android applications, discovering that it failed to find some flows, figuring out the platform methods involved in breaking the static flow path and reading the Android documentation before finally writing a model for each missing method.

In the rest of this paper we describe a technique for automatically mining explicit information flow specifications between the parameters (*this*, *arg#i*) and return values of arbitrary platform methods.

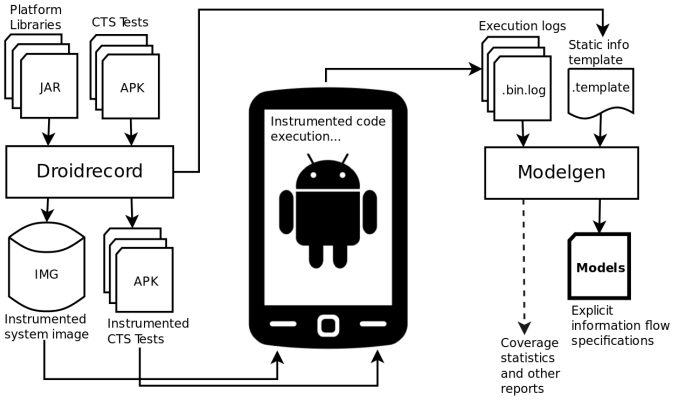

Fig. 2. Architecture of Droidrecord/Modelgen

III. ARCHITECTURAL OVERVIEW

An architecture diagram of our system is given in Figure 2. The first part of our system, Droidrecord, takes binary libraries (.jar) and application archives (.apk) in their compiled form as Android's DEX bytecode. Droidrecord inserts a special logger class into every executable. Using the Soot Java Optimization Framework [45] with its Dexpler [6] DEX frontend,

Droidrecord then modifies each method to use this logger to record the results of every bytecode operation performed. We call each such operation an *event* and the sequence of all events in the execution of a program is a *trace*.

Once instrumented, the modified Android libraries are put together into a full Android system image that can be loaded into any standard Android emulator. For specification mining, we capture the traces generated by running the test suite for the platform methods we wish to model. In particular, we make use of the Android Compatibility Test Suite (CTS) [21].

Running the instrumented tests over the instrumented system image produces a collection of traces. Modelgen is the component of our system that analyzes these traces off-line and generates explicit information flow specifications.

Droidrecord Instrumentation. Since events are recorded to the file system during the instrumented code's execution, it is important that their representation be compact. A compact representation reduces both the disk space taken inside the emulator and the slowdown resulting from performing many additional disk writes as the instrumented code executes.

Droidrecord generates a template file (.template) containing all the information for each event that can be determined statically. The instrumented code stores only a number identifying the event in the template file and those values of the event that are only known at runtime. As an example, consider a single method call operation, shown below in Soot's internal bytecode format (slightly edited for brevity):

```
$r5 = v_invoke $r4.<StringBuilder.append(int)>(i0);
```

When encountering this instruction, Droidrecord outputs the following event template into its .template file:

```
17533:[MethodCallRecord{Thread: _,
       Name: <java.lang.StringBuilder.append(int)>,
       At: [...],
       Parameters: [obj:StringBuilder:_,int:_],
       ParameterLocals: [$r4,i0],
       Height: int:_}]
```

The bytecode is then instrumented to record the runtime values of the method's parameters:

```
staticinvoke <TraceRecorder.recordEvent(long)>(17533L);
staticinvoke <TraceRecorder.writeThreadId()>();
staticinvoke <TraceRecorder.writeObjectId(Object)>(r4);
staticinvoke <TraceRecorder.write(int)>(i0);
$r5 = v_invoke $r4.<StringBuilder.append(int)>(i0);
```

When reading the trace, these values are plugged into the placeholder positions ('_' above) of the event template. For some events (simple assignments, arithmetic operations, etc) all the values can be inferred statically as a simple function of the values of previous events. These events generate event templates but incur no dynamic recording overhead.

Modelgen Trace Extraction. After tests are run and traces extracted from the emulator, they are first pre-processed and combined with the static information in the .template file. The result is a sequential stream of events for each method invocation; we write $(m : i)$ for the ith invocation of method m. Calls made by $(m : i)$ to other methods are included in this stream, together with all the events and corresponding calls within those other method invocations. Spawning a new thread is an exception: events happening in a different thread are absent from the stream for $(m : i)$, but appear in the streams for enclosing method invocations in the new thread. This separation may break flows that involve operations of multiple threads and is a limitation of our implementation. We did not find any cases where a more precise tracking of explicit information flow across threads would have made a difference in our experimental results.

$$
\begin{array}{rll}
T ::= & e^* & \text{(trace)} \\
e \in event ::= & x = pv & \text{(literal load)} \\
& \mid x = newObj & \text{(new object)} \\
& \mid x = y & \text{(variable copy)} \\
& \mid x = y \diamond z & \text{(binary op)} \\
& \mid x = y.f & \text{(load)} \\
& \mid x.f = y & \text{(store)} \\
& \mid x = m(\overline{y}) & \text{(call)} \\
& \mid return\ x & \text{(return)} \\
& \mid throw\ x & \text{(throw exception)} \\
& \mid catch\ x = a & \text{(caught exception)}
\end{array}
$$

$$
\begin{array}{ll}
pv \in Primitive\ Value & \diamond \in BinOp \\
a \in Address & r \in Rec = \{\overline{f : v}\} \\
x, y, z \in Var & \rho \in Env : Var \rightharpoonup Value \\
f \in Field & h \in Heap : Address \rightharpoonup Rec \\
m \in Method &
\end{array}
$$

Fig. 3. Structure of a trace

IV. Specification Mining

To explain Modelgen's core model generation algorithm, we describe its behavior on a single *invocation subtrace* $T_{(m:i)}$, which is the sequence of events in the trace corresponding to method invocation $(m : i)$. Recall $T_{(m:i)}$ includes the invocation subtraces for all method invocations called from m during invocation $(m : i)$, including any recursive calls to m. We now describe a simplified representation of $T_{(m:i)}$ (Section IV-A) and give its natural semantics (Section IV-B), that is, the meaning of each event in the subtrace with respect to the original program execution. Modelgen analyzes an invocation subtrace by processing each event in order and updating its own bookkeeping structures. We represent this process with a non-standard semantics: the modeling semantics of the subtrace (Section IV-C). After Modelgen finishes scanning $T_{(m:i)}$, interpreting it under the modeling semantics, it saves the resulting specification which can then be combined with the specifications for other invocations of m (Section IV-D).

A. Structure of a Trace

Figure 3 gives a grammar for the structure of traces, consisting of a sequence of events. Events refer to constant primitive values, field or method labels, and variables. The symbol $\diamond$ stands for binary operations between primitive values. Objects are represented as records mapping field names to values, which might be either addresses or primitive values. This grammar is similar to that of a 3-address bytecode representing Java operations. However, it represents not static program structure, but the sequence of operations occurring during a concrete program run, leading to the following characteristics:

$$\frac{}{\langle h, \rho, x = pv\rangle \downarrow \langle h, \rho[x \to pv]\rangle} \quad \text{(LIT)}$$

$$\frac{a \notin dom(h)}{\langle h, \rho, x = newObj\rangle \downarrow \langle h[a \to \{\}], \rho[x \to a]\rangle} \quad \text{(NEW)}$$

$$\frac{\rho(y) = v}{\langle h, \rho, x = y\rangle \downarrow \langle h, \rho[x \to v]\rangle} \quad \text{(ASSIGN)}$$

$$\frac{\rho(y) = pv_1 \qquad \rho(z) = pv_2 \qquad pv_1 \diamond pv_2 = pv_3}{\langle h, \rho, x = y \diamond z\rangle \downarrow \langle h, \rho[x \to pv_3]\rangle} \quad \text{(BINOP)}$$

$$\frac{\rho(y) = a \qquad h(a) = r \qquad r(f) = v}{\langle h, \rho, x = y.f\rangle \downarrow \langle h, \rho[x \to v]\rangle} \quad \text{(LOAD)}$$

$$\frac{\rho(x) = a \qquad h(a) = r \qquad \rho(y) = v \qquad r' = r[f \to v]}{\langle h, \rho, x.f = y\rangle \downarrow \langle h[a \to r'], \rho\rangle} \quad \text{(STORE)}$$

$$\frac{\begin{array}{c} m = \mathit{fun}(z_1, ..., z_n)\{\mathit{var}\ \overline{x'}; \overline{e}; \mathit{return}\ y'\} \\ \forall i\ \rho(y_i) = v_i \qquad \rho'(y') = v' \\ \langle h, [z_1 \to v_1, ..., z_n \to v_n, \overline{x'} \to \mathit{undef}\,], \overline{e}\rangle \overline{\downarrow} \langle h', \rho'\rangle \end{array}}{\langle h, \rho, x = m(y_1, ..., y_n)\rangle \downarrow \langle h', \rho[x \to v']\rangle} \quad \text{(INV)}$$

$$\frac{\langle h_i, \rho_i, e_i\rangle \downarrow \langle h_{i+1}, \rho_{i+1}\rangle}{\langle h_0, \rho_0, e_0; \ldots; e_{n-1}\rangle \overline{\downarrow} \langle h_n, \rho_n\rangle} \quad \text{(SEQ)}$$

Fig. 4. Natural semantics

1) Conditional (`if`, `switch`) and loop (`for`, `while`) operations are omitted and unnecessary; the events in T represent a single path through the program. The predicates inside conditionals are still evaluated, usually as binary operations.
2) The values of array indices in recorded array accesses are concrete, which allows us to treat array accesses as we would object field loads and stores (e.g., $a[i]$ becomes $a.i$, and note i is a concrete value).
3) For each method call event $x = m_1(\overline{y})$ in $T_{(m:i)}$ there is a unique invocation subtrace of the form $T_{(m_1:j)} = fun(\overline{z})\{var\ \ \overline{x}; \overline{e}; e_f\}$ where e_f is a return or throw event and $\overline{x}$ is a list of all variable names used locally within the invocation. Again, since we cover only one path through m for each invocation, invocation subtraces may have at most one return event and must end with a return or throw event.

We avoid modeling static fields explicitly by representing them as fields of a singleton object associated with each class.

B. Natural Semantics of a Subtrace

Figure 4 gives a natural semantics for executing the program path represented by an invocation subtrace. Understanding these standard semantics makes it easier to understand the custom semantics used by Modelgen to mine specifications, which extend the natural semantics. The natural semantics of a subtrace are similar but not quite identical to the semantics of Java bytecode. The differences arise from the fact that subtrace semantics represent a single execution path.

During subtrace evaluation, an environment ρ maps variable names to values. A heap h maps memory addresses to object records. Given a tuple $\langle h, \rho, e\rangle$ representing event e under heap h and environment ρ, the operator $\downarrow$ represents the evaluation of e in the given context and produces a new tuple $\langle h', \rho'\rangle$ containing a new heap and a new environment. The operator $\overline{\downarrow}$ represents the evaluation of a sequence of events which consists of evaluating each event ($\downarrow$) under the heap and environment resulting from the evaluation of the previous event. The rules in Figure 4 describe the behavior of $\downarrow$ and $\overline{\downarrow}$ for different events and their necessary pre-conditions. We omit the rules for handling exceptions since they do not add significant new ideas with respect to our specification mining technique and exception propagation complicates both the natural and modeling semantics.

We now consider how the natural semantics represent the evaluation of the following example subtrace fragment which increments a counter at $x.f$:

$$t ::= y = x.f; z = 1; w = y + z; x.f = w$$

Assuming x contains the address a (i.e., $\rho(x) = a$) of heap record $r = \{f : 0\}$ (i.e., $h(a) = r$), LOAD gives us:

$$\langle h, \rho, y = x.f\rangle \downarrow \langle h, \rho[y \to 0]\rangle$$

Applying LIT, BINOP and STORE, respectively, we get:

$$\langle h, \rho[y \to 0], z = 1\rangle \downarrow \langle h, \rho[y \to 0; z \to 1]\rangle$$

$$\langle h, \rho[y \to 0; z \to 1], w = y{+}z\rangle \downarrow \langle h, \rho[y \to 0; z \to 1; w \to 1]\rangle$$

$$\langle h, \rho[...; w \to 1], x.f = w\rangle \downarrow \langle h[a \to \{f : 1\}], \rho[...; w \to 1]\rangle$$

Using those evaluations for each expression, SEQ gives the full evaluation of the fragment as

$$\langle h, \rho, t\rangle \overline{\downarrow} \langle h[a \to \{f : 1\}], \rho[y \to 0; z \to 1; w \to 1]\rangle$$

where, in addition to some changes to the environment, field f of record r in the heap has been incremented by one.

C. Modeling Semantics of a Subtrace

The modeling semantics augment the natural semantics by associating *colors* with every heap location and primitive value. For subtrace $T_{(m:i)}$, each argument to m is initially assigned a single unique color. The execution of $T_{(m:i)}$ under the modeling semantics preserves the following invariants:

Invariant I: Computed values have all the colors of the argument values used to compute them.

Invariant II: At each point in the trace, if a heap location l is accessed from an argument a using a chain of dereferences that exists at method entry, then l has the color of a.

Invariant III: At each point in the trace, every argument and the return value have all the colors of heap locations reachable from that argument or return value.

These invariants are easily motivated. Invariant I is the standard notion of taint flow: the result of an operation has the taint of the operands. Invariant II captures the granularity of our specifications on entry to a method: all the locations reachable from an argument are part of the taint class associated with that argument (recall the semantics of our specifications described in Section II). Similarly, Invariant III captures reachability on method exit. For example, if part of the structure of *arg#1*

$$\frac{l = new\_loc() \qquad c = new\_color()}{\langle h, \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}, \mathbb{D}, x = pv\rangle \downarrow \; \langle h, \rho[x \to pv], \mathscr{L}[x \to l], \mathbb{C}[l \to \{c\}], \mathbb{G}, \mathbb{D}\rangle} \quad (\textsc{mLit})$$

$$\frac{a \notin dom(h) \qquad l = new\_loc() \qquad c = new\_color()}{\langle h, \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}, \mathbb{D}, x = newObj\rangle \downarrow \; \langle h[a \to \{\}], \rho[x \to a], \mathscr{L}[x \to l], \mathbb{C}[l \to \{c\}], \mathbb{G}, \mathbb{D}\rangle} \quad (\textsc{mNew})$$

$$\frac{\rho(y) = v \qquad \mathscr{L}(y) = l}{\langle h, \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}, \mathbb{D}, x = y\rangle \downarrow \; \langle h, \rho[x \to v], \mathscr{L}[x \to l], \mathbb{C}, \mathbb{G}, \mathbb{D}\rangle} \quad (\textsc{mAssign})$$

$$\frac{\begin{array}{c}\rho(y) = pv_1 \qquad \rho(z) = pv_2 \qquad pv_1 \diamond pv_2 = pv_3 \\ \mathscr{L}(y) = l_1 \qquad \mathscr{L}(z) = l_2 \qquad l_3 = new\_loc() \\ C = \mathbb{C}(l_1) \cup \mathbb{C}(l_2)\end{array}}{\langle h, \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}, \mathbb{D}, x = y \diamond z\rangle \downarrow \; \langle h, \rho[x \to pv_3], \mathscr{L}[x \to l_3], \mathbb{C}[l_3 \to C], \mathbb{G}, \mathbb{D}\rangle} \quad (\textsc{mBinop})$$

$$\frac{\begin{array}{c}\rho(y) = a \qquad h(a) = r \qquad r(f) = v \\ \mathscr{L}(a) = l_1 \qquad \mathscr{L}(y, f) = l_2 \\ C = \mathbb{D}(a, f)?\mathbb{C}(l_2) : \mathbb{C}(l_1) \cup \mathbb{C}(l_2)\end{array}}{\langle h, \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}, \mathbb{D}, x = y.f\rangle \downarrow \; \langle h, \rho[x \to v], \mathscr{L}[x \to l_2], \mathbb{C}[l_2 \to C], \mathbb{G}, \mathbb{D}\rangle} \quad (\textsc{mLoad})$$

$$\frac{\begin{array}{c}\rho(x) = a \qquad h(a) = r \qquad \rho(y) = v \qquad r' = r[f \to v] \\ \mathscr{L}(y) = l_1 \qquad \mathscr{L}(a) = l_2 \\ \mathbb{G}' = \mathbb{G} + \{c_1 \to c_2 | \forall c_1 \in \mathbb{C}(l_1), c_2 \in \mathbb{C}(l_2)\}\end{array}}{\langle h, \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}, \mathbb{D}, x.f = y\rangle \downarrow \; \langle h[a \to r'], \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}', \mathbb{D}[(a, f) \to \mathit{True}]\rangle} \quad (\textsc{mStore})$$

$$\frac{\begin{array}{c}m = \mathit{fun}(z_1, ..., z_n)\{\mathit{var}\ \overline{x'}; \overline{e}; \mathit{return}\ y'\} \\ \rho_m = [z_1 \to v_1, ..., z_n \to v_n, \overline{x'} \to \mathit{undef}\,] \\ \mathscr{L}_m = \mathscr{L}[z_1 \to \mathscr{L}(y_1), ..., z_n \to \mathscr{L}(y_n), \overline{x'} \to new\_loc()] \\ \langle h, \rho_m, \mathscr{L}_m, \mathbb{C}, \mathbb{G}, \mathbb{D}, \overline{e}\rangle \overline{\downarrow} \langle h', \rho', \mathscr{L}', \mathbb{C}', \mathbb{G}', \mathbb{D}', t\rangle \\ \mathscr{L}'' = \mathscr{L}'[z_1 \to \mathscr{L}(z_1), ..., z_n \to \mathscr{L}(z_n), \overline{x'} \to \mathscr{L}(\overline{x'})] \\ \forall i\ \rho(y_i) = v_i \qquad \rho'(y') = v' \qquad \mathscr{L}(y') = l\end{array}}{\langle h, \rho, \mathscr{L}, \mathbb{C}, \mathbb{G}, \mathbb{D}, x = m(y_1, ..., y_n)\rangle \downarrow \; \langle h', \rho[x \to v'], \mathscr{L}''[x \to l], \mathbb{C}', \mathbb{G}', \mathbb{D}'\rangle} \quad (\textsc{mInv})$$

$$\frac{\begin{array}{c}\langle h_i, \rho_i, \mathscr{L}_i, \mathbb{C}_i, \mathbb{G}_i, \mathbb{D}_i, e_i\rangle \downarrow \\ \langle h_{i+1}, \rho_{i+1}, \mathscr{L}_{i+1}, \mathbb{C}_{i+1}, \mathbb{G}_{i+1}, \mathbb{D}_{i+1}\rangle\end{array}}{\begin{array}{c}\langle h_0, \rho_0, \mathscr{L}_0, \mathbb{C}_0, \mathbb{G}_0, \mathbb{D}_0, e_0; \ldots; e_{n-1}\rangle \overline{\downarrow} \\ \langle h_n, \rho_n, \mathscr{L}_n, \mathbb{C}_n, \mathbb{G}_n, \mathbb{D}_n\rangle\end{array}} \quad (\textsc{mSeq})$$

Fig. 5. Modeling semantics

is inserted into the structure reachable from *arg#2* by the execution of the trace, then *arg#2* will have the color of *arg#1* on exit. At every step of the modeling semantics these invariants are preserved for every computed value and heap location seen so far; the invariants need not hold for heap locations and values that have not yet been referenced by any event in the examined portion of the subtrace. In addition, reachability in Invariants II and III applies only to the paths through the heap actually accessed during subtrace execution.

The natural semantics differentiate between primitive values or addresses stored in variables of ρ and objects stored in the heap h. Although this distinction is useful in representing the subtrace's execution, for specification mining we want to associate colors with both heap and primitive values. For uniformity, we introduce a mapping $\mathscr{L}$ which assigns a "virtual location" (*VLoc*) to every variable, object and field based on origin (i.e., where the value was first created) rather than the kind of value. Because virtual locations may be tainted with more than one color (recall Invariant I), we introduce a map $\mathbb{C} : \mathit{VLoc} \rightharpoonup 2^{\mathit{Color}}$ from virtual locations to sets of colors. The modeling semantics also use $\mathbb{G} : \{(\mathit{Color}, \mathit{Color})\}$, which is a relation on colors or, equivalently, a directed graph in which nodes are colors, and $\mathbb{D} : (\mathit{Address}, \mathit{Field}) \rightharpoonup \mathit{Boolean}$, which stands for "destructively updated" and maps object fields to a boolean value indicating that the field of that location has been written in the currently executed subtrace. We explain the use of $\mathbb{G}$ and $\mathbb{D}$ below.

Figure 5 lists the modeling semantics corresponding to the natural semantics in Figure 4. We now explain how the first 4 rules preserve Invariant I, as well as how MLOAD and MSTORE preserve Invariants II and III, respectively.

Rule MLIT models the assignment of literals to variables. A new literal value is essentially a new information source within the subtrace and is assigned a new location with a new color. The location is associated with the variable now holding the value, preserving Invariant I. Rule MNEW, which models new object creation, is similar. Rule MASSIGN models an assignment $x = y$ where x and y are both variables in ρ and does not create a new location, but instead updates $\mathscr{L}(x)$ to be the location of y, indicating that they are the same value, again preserving Invariant I.

Rule MBINOP gives the modeling semantics for binary operations. Assuming locations l_1 and l_2 for the operands, the rule adds a new location l_3 to represent the result. Because of Invariant I, l_3 must be assigned all the colors of l_1 and all the colors of l_2, thus $\mathbb{C}(l_3)$ becomes $\mathbb{C}(l_1) \cup \mathbb{C}(l_2)$.

Rules MLOAD and MSTORE deal with field locations. The virtual location of field $a.f$ (denoted $\mathscr{L}(a, f)$) is defined as either the location of the object stored at $a.f$, if the field is of reference type, or as an identifier which depends on $\mathscr{L}(a)$ and the name of f, if f is of primitive type.

Rule MLOAD models load events of the form $x = y.f$ by assigning the location $l_2 = \mathscr{L}(y, f)$ to x and computing the color set for this location (which will be the colors for both x and $y.f$). There are three cases to consider:

- If this is the first time the location $\mathscr{L}(y, f)$ has been referenced within the subtrace $T_{(m:i)}$, then $y.f$ has no color (all heap locations except the arguments start with the empty set of colors in $\mathbb{C}$). Furthermore, since this is the first access, $y.f$ has not been previously written in the subtrace, so $\mathbb{D}(\rho(y), f) = \mathit{False}$. Therefore, l_2 is assigned the colors $\mathbb{C}(l_1) \cup \mathbb{C}(l_2)$ where $l_1 = \mathscr{L}(y)$. Since $\mathbb{C}(l_2) = \emptyset$ before the load event, we end up with $\mathbb{C}(l_2) = \mathbb{C}(l_1)$. If $y.f$ is reachable from a method argument through y, this establishes Invariant II for $y.f$ on its first access.
- If l_2 has been loaded previously in the trace but not previously overwritten, then $\mathbb{C}(l_2) = \mathbb{C}(l_1) \cup \mathbb{C}(l_2)$, indicating that l_2 now has the colors of all of its previous accesses plus a possibly new set of colors $\mathbb{C}(l_1)$. This handles the case where a location is reachable from multiple method arguments and preserves Invariant II.
- If $y.f$ has been written previously then $\mathbb{D}(\rho(y), f) = \mathit{True}$. In this case it is no longer true that $\mathscr{L}(y, f)$ was reachable from $\mathscr{L}(y)$ on method entry and so it is not necessary to propagate the color of $\mathscr{L}(y)$ to $\mathscr{L}(y, f)$ to preserve Invariant II and we omit it. Also, note that if $y.f$ has been written, that implies the value stored in $y.f$

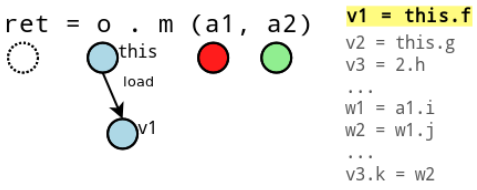

Fig. 6. Processing a load event

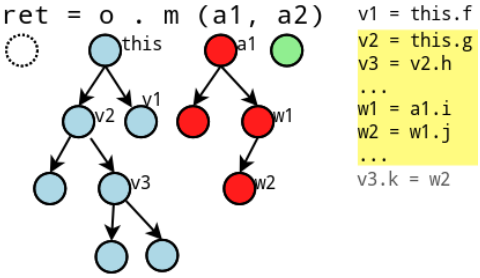

Fig. 7. Loads reconstruct the heap structure reachable from each argument

was loaded before the write and so $y.f$ will already have at least one color.

Figure 6 shows the effect of a single load operation from an argument to m, while Figure 7 depicts the coloring of a set of the heap locations after multiple load events.

Rule MSTORE models store events of the form $x.f = y$. The rule updates $\mathbb{D}(\rho(x), f) = \mathit{True}$ since it writes to $x.f$. We

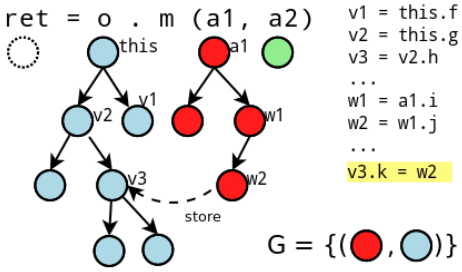

Fig. 8. Processing a store event

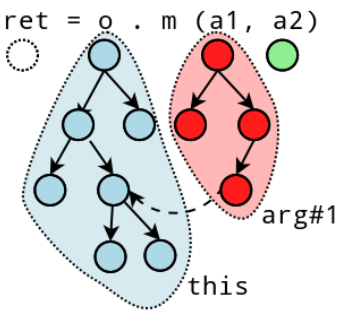

Fig. 9. Stores induce connections between colored argument structures

could satisfy Invariant III by implementing MSTORE in a way that traverses the heap backwards from x to every argument of m that might reach x and associates every color of y with those arguments (and possibly intermediate heap locations). As an optimization, we instead use $\mathbb{G}$ to record an edge from each color c_1 of $\mathscr{L}(y)$ to each color c_2 of $\mathscr{L}(x.f)$ with the following meaning: $c_1 \to c_2 \in \mathbb{G}$ means every virtual location with color c_2 has color c_1 as well. Figure 8 depicts the results of a store operation, while figure 9 depicts how $\mathbb{G}$ serves to associate two colored heap subgraphs.

Rule MINV implements standard method call semantics, mapping the virtual locations of arguments and the return value between caller and callee. Rule MSEQ is the same as SEQ in the natural semantics but adds $\mathscr{L}$, $\mathbb{C}$, $\mathbb{G}$ and $\mathbb{D}$ where needed.

As a consequence of Invariants I and II, the modeling semantics associate the color of each argument to every value and heap location that depends on the argument values on entry to m. Then, because of Invariant III, when the execution reaches the end of subtrace $T_{(m:i)}$ every argument and the return value have all the colors of heap locations reachable from that argument or return value (as represented by $\mathbb{G}$). We construct our specifications by examining the colors of each argument a_j and the return value r after executing the subtrace: for every color of r (or a_j) that corresponds to the initial color of a different argument a_k, we add $a_k \to r$ ($a_k \to a_j$) to our model.

D. Combining Specifications

For each invocation subtrace $T_{(m:i)}$, the process just outlined produces an underapproximation of the specification for m, based on a single execution $(m : i)$. We combine the results from different invocations of m by taking an upper bound on the set of argument-to-argument and argument-to-return flows discovered for every execution, which is simply the union of the results of $(m : i)$ for every i.

For example, consider the method `max(a,b)` designed to return the larger of two numbers, disregarding the smaller one. Suppose that we have two subtraces for this method: one for invocation `max(5,7)`, which returns 7 and produces the model $M_1 = \{\mathit{arg\#2} \to \mathit{return}\}$ and one for invocation `max(9,2)`, which returns 9 and produces the model $M_2 = \{\mathit{arg\#1} \to \mathit{return}\}$. Clearly the correct specification reflecting the potential explicit information flow of method `max(a,b)` is $M_1 \cup M_2 = \{\mathit{arg\#1} \to \mathit{return},\ \mathit{arg\#2} \to \mathit{return}\}$.

We should note that combining specifications in this way inherently introduces some imprecision with respect to the possible flows on a given execution of the method. The effects of this imprecision in our overall system depend on the characteristics of the static analysis that consumes the specifications. For example, the above specification for `max(a,b)` would be strictly less precise than analyzing the corresponding code (assuming the natural implementation) with an ideal path-sensitive analysis, since it merges two different control paths within the max function: one in which the first argument is greater and one in which the second argument is greater. For context-sensitive but path-insensitive analysis such as

STAMP (see Section V-B), loss of precision due to combining specifications is less common, but still possible in theory. Consider a method `do(a,b) { a.doImpl(b) }` and two invocations of this method in which `a` has different types and each type has its own implementation of `a.doImpl(b)`. A context-sensitive analysis can tell which version of `doImpl` is executed, but Modelgen will simply merge the flows observed for every version of `doImpl` seen in any trace of `do(a,b)`.

E. Calls to Uninstrumented Code

Our approach to specification mining is based on instrumenting and executing as much of the platform code as we can. Unfortunately recording the execution of every method in the Android platform is challenging. In particular, any technique based on Java bytecode instrumentation cannot capture the behavior of native methods and system calls. Since our inserted recorder class is itself written in Java, we must also exclude from instrumentation some Java classes it depends upon to avoid introducing an infinite recursion.

Thus, traces are not always full traces but represent only a part of a program's execution. We need to deal with two separate problems during event interpretation: (1) How should Modelgen interpret calls to uninstrumented methods? (2) How can we detect that a trace has called uninstrumented code?

For the first problem, Modelgen offers two separate solutions. The user can provide manually written models for some methods in this smaller uninstrumented subset (as we do, for example, for `System.arraycopy` and `String.concat`). If a user-supplied model is missing for a method, Modelgen assumes a worst-case model in which information flows from every argument of the method to every other argument and to its return value. In many cases, this worst-case model, although imprecise, is good enough to allow us to synthesize precise specifications for its callers.

The problem of detecting uninstrumented method calls inside traces is surprisingly subtle. Droidrecord writes an event at the beginning of each method and before and after each method call. In the simplest case we would observe these before-call and after-call markers adjacent to each other, allowing us to conclude that we called an uninstrumented method. However, because uninstrumented methods often call other methods which are instrumented, this simple approach is not enough. A call inside instrumented code could be followed by the start of another instrumented method, distinct from the one that is directly called. Dynamic dispatch and complex class hierarchies further complicate determining if the method we see start after a call instruction is the instruction's callee.

Our solution for detecting holes in the trace due to invoking uninstrumented code is to record the height of the call stack at the beginning of every method and before and after each call operation. Since the stack grows for every method call, whether instrumented or not, we use the stack height to determine when we have called into uninstrumented code.

V. Evaluation

We perform three studies to evaluate the specifications generated by Modelgen. First, we compare them directly against our existing manually-written models (Section V-A). Second, we contrast the results of running the full STAMP static information-flow analysis system using these specifications as input, against the results of the same system using the manual models (Section V-B). Third, we study the effect of testsuite quality on the generated specifications (Section V-C).

A. Comparison Against Manual Models

To evaluate Modelgen's ability to replace the manual effort involved in writing models for STAMP (see Section II), we compare the specifications mined by Modelgen against existing manual models for 309 Android platform methods.

We conducted all of our evaluations on the Android 4.0.3 platform, which has a total of 46,559 public and protected methods. STAMP includes manual models for 1,116 of those methods, of which 335 are inside the `java.lang.*` package and therefore are uninstrumented in DroidRecord, and 321 have only source or sink annotations, leaving 460 methods for which Modelgen could infer comparable specifications.

For our evaluation, we obtained traces by running tests from the Android Compatibility Test Suite (CTS) [21]. The Android CTS is designed to ensure compatibility between multiple implementations and variations of the Android platform and our positive results are due at least in part to the fact that CTS is a high quality set of tests. The CTS contains static references to 14,435 android platform methods, but might exercise an even larger portion of the platform due to dynamic dispatch and some platform methods being called from other methods under test. For our experiments, we restricted ourselves to a subset of the CTS purporting to test those classes in the `java.*` and `android.*` packages, but outside of `java.lang.*`, for which we have manual models (not counting simple source or sink annotations). This smaller subset of the CTS contains static references to 712 platform methods and produced Modelgen specifications for 660 platform methods. Note that for some packages for which we have manual models, such as `com.google.*`, the CTS contains no tests.

Table II summarizes our findings, organized by Java package. For each package we list the number of classes and methods for which we have manual specifications, as well as the total number of correct individual flow annotations (e.g. *arg#X* $\rightarrow$ *return*) included either in our manual specifications or generated by Modelgen. We then list separately the flows discovered by Modelgen and those in our manual specifications. We consider only those flows in methods for which we have manual specifications and only those classes for which we ran any CTS tests, which gives us 309 methods to compare.

We evaluate Modelgen under two metrics: precision and recall. Precision is a measure of soundness: the percentage of all possible flows through the methods that are discovered by Modelgen. Given that neither Modelgen nor our manual models are guaranteed to be precise, we approximate precision by comparing the flows discovered by each approach to the union of the flows discovered by both approaches. Formally, let *FModelgen* be the set of flows discovered by Modelgen

Package	Classes	Methods	Missing trace info.	Total correct flows	Modelgen correct flows	Manual correct flows	Modelgen false positives	Manual errors	Modelgen precision	Manual precision	Modelgen recall
java.nio.*	2	26	4	50	50	42	0	0	100.00%	84.00%	100.00%
java.io.*	28	146	23	280	275	234	2	0	98.21%	83.57%	97.86%
java.net.*	7	37	4	104	100	65	0	1	96.15%	62.50%	93.85%
java.util.*	4	28	0	36	36	31	0	1	100.00%	86.11%	100.00%
android.text.*	3	5	2	3	3	3	0	0	100.00%	100.00%	100.00%
android.util.*	2	8	1	11	4	7	0	0	36.36%	63.64%	0.00%
android.location.*	3	13	3	12	12	9	0	0	100.00%	75.00%	100.00%
android.os.*	2	46	3	60	60	49	0	0	100.00%	81.67%	100.00%
Total	51	309	40	556	540	440	2	2	97.12%	79.14%	96.36%

TABLE II
COMPARING MODELGEN SPECIFICATIONS AND MANUAL MODELS

and *FManual* the set of flows in our manual models. Modelgen's precision is:

$$PModelgen = \frac{|FModelgen|}{|FModelgen \cup FManual|}$$

And the precision of the manual models is:

$$PManual = \frac{|FManual|}{|FModelgen \cup FManual|}$$

Table II lists the precision of each approach for each package. Overall, Modelgen's precision is above 97%, whereas the manual models are about 79% precise by comparison.

Recall measures how many of our manual models are also discovered by Modelgen, and is calculated as:

$$Recall = 1 - \frac{|FModelgen \cup FManual| - |FModelgen|}{|FManual|}$$

As we can see from Table II, Modelgen finds about 96% of our manual specifications. The specifications Modelgen misses were written to capture implicit flows, which is not surprising since Modelgen is designed to detect only explicit flows. The most visible example of this limitation is the row corresponding to `android.util`, in which 7 of the 8 analyzed methods are part of the `android.util.Base64` class, which performs base64 encoding and decoding of byte buffers via table lookups, inducing implicit flows. The last remaining method in this package is a native method and the four new correct flows identified by Modelgen have to do with flags being stored inside the `Base64` class.

When collecting our results and contrasting the manual models to Modelgen's specifications, we found two false positives in Modelgen, both in the same method and due to an unexpected hole in the trace resulting from a bug in how our current implementation handles inner classes. Modelgen detected the hole in the trace and processed it under worst-case assumptions, resulting in two spurious flows. Fixing the bug will remove these flows. Notably, we found two errors in the manual models: one was a typo (*arg#2* $\rightarrow$ *arg#2* instead of *arg#2* $\rightarrow$ *return*) and the other was a reversed annotation (*arg#1* $\rightarrow$ *this* instead of *this* $\rightarrow$ *arg#1*).

Our current implementation of Modelgen failed to produce traces for a few methods that have manual annotations, listed under the column "Missing trace info." of Table II. Reasons for missing traces include: the method for which we tried to generate a trace is a native method, the Android CTS lacks tests for the given method, or an error occurred while instrumenting the class under test or while running the tests. This last case often took the form of triggering internal responsiveness timers inside the Android OS, known as ANR (Application Not Responding) [22]—because our instrumentation results in a significant slowdown (about 20x), these timers are triggered more often than they would be in uninstrumented runs. Since capturing the traces is a one-time activity, this high overhead is otherwise acceptable.

Given these results, we are confident that Modelgen can be used to replace most manual information flow models as it managed to reproduce almost all our manual flow annotations (96.38% recall) and produced many more correct annotations that our manual models missed (97.12% vs 79.14% precision), while significantly reducing manual effort. Although our evaluation focuses on Java and Android, the results should generalize to any platform for which good test suites exist.

B. Whole-System Evaluation of STAMP and Modelgen

The STAMP static analysis component is a bounded context-sensitive, flow- and object-insensitive information flow analysis. A complete description of this system can be found in Section 4 of [17]. STAMP never analyzes platform code and treats platform methods for which it has no explicit model under best-case assumptions. That is, platform methods without models are assumed to induce no flows between their arguments or their return values[1].

To evaluate the usefulness of our specifications in a full static analysis, we ran STAMP under two configurations: base and augmented. In the base configuration, we used only the existing manually-written models. In the augmented configuration, we included (1) all source and sink annotations from the manual models (annotating sources and sinks is outside of the scope of Modelgen), (2) the Modelgen specifications generated in the experiment of Section V-A, and (3) the existing manual models for those methods for which Modelgen did not construct any specifications (e.g. the `java.lang.*` classes). The base and augmented configurations included 1215 and 2274 flow annotations, respectively.

[1]The alternative, analyzing under worst-case assumptions, produces an overwhelming number of false positives.

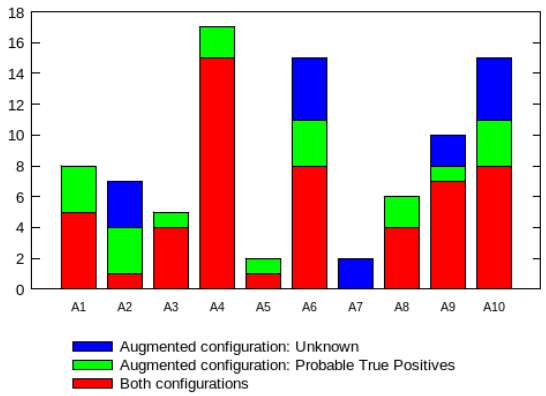

Fig. 10. Number of flows produced by STAMP with and without specifications generated by Modelgen

We compared the results of both configurations on 242 apps from the Google Play Store. These apps were randomly selected among those for which STAMP was able to run with a budget of 8GB of RAM and 1 hour time limit in both configurations. The average running time per app is around 7 minutes in either configuration.

STAMP finds a total of 746 (average 3.08 per app) and 986 (average 4.07) flows in the base and augmented configuration, respectively. The union of the flows discovered in both configurations is exactly 1000. In other words, STAMP finds 31% (254) new flows in the augmented configuration. Like most static analysis systems, STAMP can produce false positives, even when given sound models. Additionally, Modelgen may produce unsound models for some methods (recall the discussions in sections IV-D and IV-E). Given this, we would like to know what proportion of these new flows are true positives. To get a sense of what the true positive rate of the new flows might be, we took 10 random apps from the subset of our sample (109 of 242 apps) for which the augmented configuration finds any flows not discovered by the base configuration. We manually inspected these apps and marked those flows for which we could find a feasible source-sink path, and for which control flow could reach such path, as "probable true positives". Although this sort of inspection is always susceptible to human error, we tried to be conservative in declaring flows to be "probable true positives". In most cases, the flows are contained in advertisement libraries and would trigger as soon as the app started or a particular view within the app was displayed to the user.

Figure 10 shows the results of our manual inspection. The flows labeled as "Augmented configuration: Unknown" are those for which we could not find a source-sink path, but are not necessarily false positives. The flows labeled "Augmented configuration: Probable True Positives" represent a lower bound on the number of new true positives that STAMP finds under the augmented configuration. The lower portion of the bar corresponds to those flows found in both configurations, without attempting to distinguish whether they are false or true positives. For the 10 apps, the augmented configuration produces 64% more flows than the base configuration; of these new flows, at least 55% are true positives.

The recall of the augmented configuration, which is the percentage of all flows found under the base configuration that were also found when running the augmented configuration, is 98.12%. A flow found in the base configuration could be missed in the augmented configuration if Modelgen infers a different specification for a method, which is relevant for the flow, than the manually-written model. In other words, the delta in the recall for the flows STAMP reports follows from the delta in the recall of Modelgen specifications compared to the manual models (Table II).

C. Controlling for Testsuite Quality

Specification mining based on concrete execution traces depends on having a representative set of test cases for each method for which we want to infer a specification. One threat to the validity of our experiment is that it could be that our results are good only because the standard Android compatibility tests are unusually thorough. In this section we attempt to control for the quality of the test suite.

We measure how strongly our specification mining technique depends on the available tests by the number of method executions it needs to observe before it converges to the final specification. Intuitively, if few executions of a method are needed to converge to a suitable specification of the method's behavior, then our specification mining technique is more robust that if it requires many executions, and therefore many test cases. Additionally, if a random small subset of the observed executions is enough for our technique to discover the same specification as the full set of executions, we can gain some confidence that observing additional executions won't dramatically alter the results of our specification mining.

We take all methods from Table II for which we are able to record traces and Modelgen produces non-empty specifications, which are 264 methods in total. We restrict ourselves to those methods, as opposed to the full set for which we have mined specifications, since we have studied their quality during the comparison of Section V-A and found them close to the ideal models a manual writer would find. For each such method m, we consider the final specification produced by Modelgen (S_m) as well as the set $\mathbb{S}$ of specifications for each invocation subtrace of m. Starting with the empty specification we repeatedly add a random specification chosen from $\mathbb{S}$ until the model matches S_m, recording how many such subtrace specifications are used to recover S_m.

Figure 11 shows a log scale plot of the number of methods (y axis) that required n traces (x axis) to recover the full specification over each of 20 trials. That is, we sampled the executions of each method to recover its specification and then counted the number of methods that needed one execution, the number that needed two, and so on, and then repeated this process 19 more times. The multiple points plotted for each number of executions give an idea of the variance due to the random choices of method executions to include in the specification.

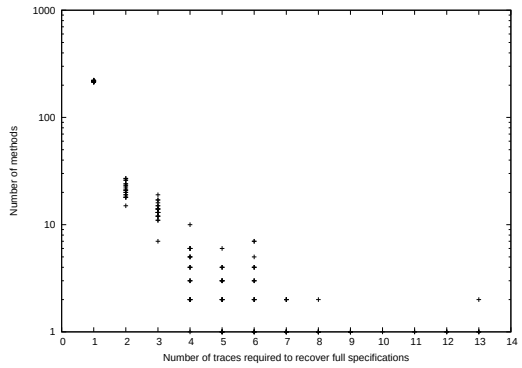

Fig. 11. Specification Convergence

It is also useful to consider aggregate statistics over all method specification inferences. In our experiment, 83.7% of the methods needed just one subtrace specification to recover the specification and no method required more than an average of 9 random subtrace specifications. The maximum number of subtraces needed to converge to a method specification (when taking the worst out of 20 iterations of the algorithm) was 13 for `java.util.Vector.setElementAt(Object, int)`. The average number of subtraces required to converge to a specification is 1.38. For comparison, the specifications evaluated in Section V-A were inferred using a median of 4 traces (the average, 207, is dominated by a few large outliers). We conclude that explicit information flow typically requires few observations to produce useful specifications.

VI. Related Work

Static taint analysis. A number of static techniques and tools [12, 24, 36, 31, 44] have been developed for whole-application taint analysis. See [41] for a survey of work in this field. For applications that run inside complex application frameworks these analyses often must include some knowledge of the framework itself. F4F [43] is a scheme for encoding framework-specific knowledge in a way that can be processed by a general static analysis. In F4F, any models for framework methods must be written manually. Flowdroid [3] is a context-, flow- and object-sensitive static taint analysis system for Android applications, which can analyze Android platform code directly. By default, it uses models or 'shortcuts' for a few platform methods as a performance optimization and to deal with hard-to-analyze code. Flowdroid's shortcuts are also information-flow specifications of a slightly more restrictive form than that used by Modelgen. Thus, it seems likely the FlowDroid shortcuts could also be mined successfully from tests.

There has also been some previous work on identifying sources and sinks in the Android platform based on the information implicitly provided by permission checks inside API code [16, 4, 7] or by applying machine learning to some of the method's static features [40]. This work could be combined with our method for inferring explicit information flow specifications to enable fully automatic explicit information flow analysis (i.e., with no manual annotations).

Dynamic taint tracking. Dynamic taint tracking uses instrumentation and run-time monitoring to observe or confine the information flow of an application. Many schemes have been proposed for dynamic taint tracking [23, 9, 13, 5]. An exploration of the design space for such schemes appears in [42]. Dytan [9] is a generic framework capable of expressing various types of dynamic taint analyses.

Our technique for modeling API methods is similar to dynamic taint tracking, and could in principle be reformulated to target Dytan or some similar general dynamic taint tracking framework. However, heap-reachability and all of our analysis would have to be performed online, as the program runs, which might exacerbate timing dependent issues with the Android platform (recall the discussion in Section V-A).

Dynamic techniques for creating API specifications. Many schemes have been proposed for extracting different kinds of specifications of API methods or classes from traces of concrete executions. However, unlike our information flow specifications, most such specifications focus on describing control-flow related properties of the code being modeled.

A large body of work (e.g. [8, 2, 46, 32, 48, 47, 10, 19, 34, 33, 30]) constructs Finite State Automata encoding transitions between abstract program states. Other approaches focus on inferring program invariants from dynamic executions, such as method pre- and post-conditions (Daikon [38, 14, 15]), array invariants [37] and algebraic "axioms" [25]. Another relevant work infers static types for Ruby programs based on the observed run-time types over multiple execution traces [26]. Finally, program synthesis techniques have been used to construct simplified versions of API methods that agree with a set of given traces on their input and output pairs [39].

Tools for Tracing Dynamic Executions. Tools that allow tracing and analyzing program executions are plentiful. Query languages such as PTQL [20] and PQL [35] can be used to formulate questions about program executions in a high-level DSL, while tools like JavaMaC [28], Tracematches [1], Hawk [11] and JavaMOP [27] permit using automata and formal logics for the same purpose. Frameworks such as RoadRunner [18] and Sofya [29] allow analyses to subscribe to a stream of events representing the program execution as it runs.

VII. Conclusions

We have described an effective technique for generating explicit information flow specifications for platform methods that outperforms manual flow annotations in practice. We presented Modelgen, an implementation of this technique for Java and the Android platform. Finally, we have show that Modelgen specifications are highly precise, have high recall with respect to our existing manual models, allow our static analysis to find true flows it misses despite years of manual model construction effort and can be inferred from a relatively small set of execution traces.

References

[1] Chris Allan, Pavel Avgustinov, Aske Simon Christensen, Laurie J. Hendren, Sascha Kuzins, Ondrej Lhoták, Oege de Moor, Damien Sereni, Ganesh Sittampalam, and Julian Tibble. Adding trace matching with free variables to AspectJ. In *OOPSLA*, pages 345–364, 2005.

[2] Glenn Ammons, Rastislav Bodík, and James R. Larus. Mining specifications. In *Conference Record of POPL 2002: The 29th SIGPLAN-SIGACT Symposium on Principles of Programming Languages, Portland, OR, USA, January 16-18, 2002*, pages 4–16, 2002.

[3] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Octeau, and Patrick McDaniel. Flowdroid: precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for Android apps. In *PLDI*, page 29, 2014.

[4] Kathy Wain Yee Au, Yi Fan Zhou, Zhen Huang, and David Lie. PScout: analyzing the Android permission specification. In *ACM Conference on Computer and Communications Security*, pages 217–228, 2012.

[5] Thomas H. Austin and Cormac Flanagan. Multiple facets for dynamic information flow. In *POPL*, pages 165–178, 2012.

[6] Alexandre Bartel, Jacques Klein, Martin Monperrus, and Yves Le Traon. Dexpler: Converting Android dalvik bytecode to jimple for static analysis with Soot. *CoRR*, abs/1205.3576, 2012.

[7] Alexandre Bartel, Jacques Klein, Martin Monperrus, and Yves Le Traon. Static analysis for extracting permission checks of a large scale framework: The challenges and solutions for analyzing Android. *IEEE Trans. Software Eng.*, 40(6):617–632, 2014.

[8] Alan W Biermann and Jerome A Feldman. On the synthesis of finite-state machines from samples of their behavior. *Computers, IEEE Transactions on*, 100(6):592–597, 1972.

[9] James A. Clause, Wanchun Li, and Alessandro Orso. Dytan: a generic dynamic taint analysis framework. In *ISSTA*, pages 196–206, 2007.

[10] Valentin Dallmeier, Christian Lindig, Andrzej Wasylkowski, and Andreas Zeller. Mining object behavior with ADABU. In *Proceedings of the 2006 International Workshop on Dynamic Systems Analysis*, WODA '06, pages 17–24, New York, NY, USA, 2006. ACM.

[11] Marcelo d'Amorim and Klaus Havelund. Event-based runtime verification of java programs. *ACM SIGSOFT Software Engineering Notes*, 30(4):1–7, 2005.

[12] Dorothy E. Denning and Peter J. Denning. Certification of programs for secure information flow. *Commun. ACM*, 20(7):504–513, 1977.

[13] William Enck, Peter Gilbert, Byung gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol Sheth. Taintdroid: An information-flow tracking system for real-time privacy monitoring on smartphones. In *OSDI*, pages 393–407, 2010.

[14] Michael D. Ernst, Jeff H. Perkins, Philip J. Guo, Stephen McCamant, Carlos Pacheco, Matthew S. Tschantz, and Chen Xiao. The Daikon system for dynamic detection of likely invariants. *Sci. Comput. Program.*, 69(1-3):35–45, 2007.

[15] Viktoria Felmetsger, Ludovico Cavedon, Christopher Kruegel, and Giovanni Vigna. Toward automated detection of logic vulnerabilities in web applications. In *USENIX Security Symposium*, pages 143–160, 2010.

[16] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android permissions demystified. In *ACM Conference on Computer and Communications Security*, pages 627–638, 2011.

[17] Yu Feng, Saswat Anand, Isil Dillig, and Alex Aiken. Apposcopy: Semantics-based detection of Android malware through static analysis. In *FSE*, 2014.

[18] Cormac Flanagan and Stephen N. Freund. The RoadRunner dynamic analysis framework for concurrent programs. In *PASTE*, pages 1–8, 2010.

[19] Mark Gabel and Zhendong Su. Javert: fully automatic mining of general temporal properties from dynamic traces. In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering, 2008, Atlanta, Georgia, USA, November 9-14, 2008*, pages 339–349, 2008.

[20] Simon Goldsmith, Robert O'Callahan, and Alexander Aiken. Relational queries over program traces. In *OOPSLA*, pages 385–402, 2005.

[21] Google. Compatibility test suite (Android) - https://source.Android.com/compatibility/cts-intro.html.

[22] Google. Keeping your app responsive - https://developer.Android.com/training/articles/perf-anr.html.

[23] Vivek Haldar, Deepak Chandra, and Michael Franz. Dynamic taint propagation for Java. In *ACSAC*, pages 303–311, 2005.

[24] Nevin Heintze and Jon G. Riecke. The SLam calculus: Programming with secrecy and integrity. In *POPL*, pages 365–377, 1998.

[25] Johannes Henkel, Christoph Reichenbach, and Amer Diwan. Discovering documentation for Java container classes. *IEEE Trans. Software Eng.*, 33(8):526–543, 2007.

[26] Jong hoon (David) An, Avik Chaudhuri, Jeffrey S. Foster, and Michael Hicks. Dynamic inference of static types for Ruby. In *POPL*, pages 459–472, 2011.

[27] Dongyun Jin, Patrick O'Neil Meredith, Choonghwan Lee, and Grigore Roşu. JavaMOP: Efficient parametric runtime monitoring framework. In *Proceeding of the 34th International Conference on Software Engineering (ICSE'12)*, pages 1427–1430. IEEE, 2012.

[28] Moonjoo Kim, Sampath Kannan, Insup Lee, Oleg Sokolsky, and Mahesh Viswanathan. Java-MaC: a run-time assurance tool for Java programs. *Electronic Notes in Theoretical Computer Science*, 55(2):218–235, 2001.

[29] Alex Kinneer, Matthew B Dwyer, and Gregg Rothermel. Sofya: A flexible framework for development of dynamic program analyses for Java software. *CSE Technical reports*, 2006.

[30] Ivo Krka, Yuriy Brun, Daniel Popescu, Joshua Garcia, and Nenad Medvidovic. Using dynamic execution traces and program invariants to enhance behavioral model inference. In *ICSE (2)*, pages 179–182, 2010.

[31] V. Benjamin Livshits and Monica S. Lam. Finding security vulnerabilities in Java applications with static analysis. In *Proceedings of the 14th conference on USENIX Security Symposium*, volume 14, pages 18–18, 2005.

[32] David Lo and Siau-Cheng Khoo. SMArTIC: towards building an accurate, robust and scalable specification miner. In *Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering, FSE 2006, Portland, Oregon, USA, November 5-11, 2006*, pages 265–275, 2006.

[33] David Lo, Leonardo Mariani, and Mauro Pezzè. Automatic steering of behavioral model inference. In *ESEC/SIGSOFT FSE*, pages 345–354, 2009.

[34] Davide Lorenzoli, Leonardo Mariani, and Mauro Pezzè. Automatic generation of software behavioral models. In *30th International Conference on Software Engineering (ICSE 2008), Leipzig, Germany, May 10-18, 2008*, pages 501–510, 2008.

[35] Michael C. Martin, V. Benjamin Livshits, and Monica S. Lam. Finding application errors and security flaws using PQL: a program query language. In *OOPSLA*, pages 365–383, 2005.

[36] Andrew C. Myers. Jflow: Practical mostly-static information flow control. In *POPL*, pages 228–241, 1999.

[37] ThanhVu Nguyen, Deepak Kapur, Westley Weimer, and Stephanie Forrest. Using dynamic analysis to discover polynomial and array invariants. In *ICSE*, pages 683–693, 2012.

[38] Jeremy W. Nimmer and Michael D. Ernst. Automatic generation of program specifications. In *ISSTA*, pages 229–239, 2002.

[39] Dawei Qi, William N. Sumner, Feng Qin, Mai Zheng, Xiangyu Zhang, and Abhik Roychoudhury. Modeling software execution environment. In *WCRE*, pages 415–424, 2012.

[40] Siegfried Rasthofer, Steven Arzt, and Eric Bodden. A machine-learning approach for classifying and categorizing Android sources and sinks. In *NDSS*, 2014.

[41] Andrei Sabelfeld and Andrew C. Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003.

[42] Edward J. Schwartz, Thanassis Avgerinos, and David Brumley. All you ever wanted to know about dynamic taint analysis and forward symbolic execution (but might have been afraid to ask). In *IEEE Symposium on Security and Privacy*, pages 317–331, 2010.

[43] Manu Sridharan, Shay Artzi, Marco Pistoia, Salvatore Guarnieri, Omer Tripp, and Ryan Berg. F4F: taint analysis of framework-based web applications. In *OOPSLA*, pages 1053–1068, 2011.

[44] Omer Tripp, Marco Pistoia, Stephen J. Fink, Manu Sridharan, and Omri Weisman. Taj: effective taint analysis of web applications. In *PLDI*, pages 87–97, 2009.

[45] Raja Vallée-Rai, Phong Co, Etienne Gagnon, Laurie J. Hendren, Patrick Lam, and Vijay Sundaresan. Soot - a Java bytecode optimization framework. In *CASCON*, page 13, 1999.

[46] Westley Weimer and George C. Necula. Mining temporal specifications for error detection. In *Tools and Algorithms for the Construction and Analysis of Systems, 11th International Conference, TACAS 2005, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2005, Edinburgh, UK, April 4-8, 2005, Proceedings*, pages 461–476, 2005.

[47] Tao Xie, Evan Martin, and Hai Yuan. Automatic extraction of abstract-object-state machines from unit-test executions. In *28th International Conference on Software Engineering (ICSE 2006), Shanghai, China, May 20-28, 2006*, pages 835–838, 2006.

[48] Jinlin Yang, David Evans, Deepali Bhardwaj, Thirumalesh Bhat, and Manuvir Das. Perracotta: mining temporal API rules from imperfect traces. In *28th International Conference on Software Engineering (ICSE 2006), Shanghai, China, May 20-28, 2006*, pages 282–291, 2006.