Checking Conformance of Applications against GUI Policies

Zhen Zhang  
University of Washington  
Seattle, USA  
zgzhen@cs.washington.edu

Yu Feng  
University of California, Santa Barbara  
Santa Barbara, USA  
yufeng@cs.ucsb.edu

Michael D. Ernst  
University of Washington  
Seattle, USA  
mernst@cs.washington.edu

Sebastian Porst  
Google  
Mountain View, USA  
sporst@google.com

Isil Dillig  
University of Texas Austin  
Austin, USA  
isil@cs.utexas.edu

ABSTRACT

A good graphical user interface (GUI) is crucial for an application’s usability, so vendors and regulatory agencies increasingly place restrictions on how GUI elements should appear to and interact with users. Motivated by this concern, this paper presents a new technique (based on static analysis) for checking conformance between (Android) applications and GUI policies expressed in a formal specification language. In particular, this paper (1) describes a specification language for formalizing GUI policies, (2) proposes a new program abstraction called an event-driven layout forest, and (3) describes a static analysis for constructing this abstraction and checking it against a GUI policy. We have implemented the proposed approach in a tool called Venus, and we evaluate it on 2361 Android applications and 17 policies. Our evaluation shows that Venus can uncover malicious applications that perform ad fraud and identify violations of GUI design guidelines and GDPR laws.

CCS CONCEPTS

• Software and its engineering → Software verification and validation;  • Security and privacy → Human and societal aspects of security and privacy;  • Software security engineering.

KEYWORDS

ad fraud, static analysis, user interface, Android, mobile app

ACM Reference Format:


Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

1 INTRODUCTION

Good graphical user interfaces (GUIs) are essential for the success and popularity of mobile applications. A bad user interface can significantly degrade the user’s overall experience, causing the app to become unpopular even if it provides otherwise useful functionality. Beyond leading to poor user experience, bad GUI designs can indicate malicious intent — for example, many ad fraud applications provide a misleading user interface to trick their users into clicking on unwanted links. Such behavior violates one of the advertisement policies published by mobile platforms [1, 14, 19], and according to a recent report [13], click fraud (a major type of ad policy violation) accounts for more than 50% of all potentially harmful applications. Furthermore, several companies and governmental agencies have other types of policies concerning the user interface of mobile apps. For instance, both Google and Apple publish UI design guidelines [2, 15], and the European Union’s General Data Protection Regulation (GDPR) laws [10, 20] impose restrictions on how mobile apps may interact with users via their user interfaces.

Despite the increasing importance of ensuring compliance between GUI policies and mobile applications, there are no existing techniques that can be used to check whether an app conforms to such GUI policies. This work aims to address this problem by proposing a new technique, and its implementation in a tool called Venus (figure 1), for checking conformance between a mobile application written in the Android framework and a GUI policy. We envision such a tool being utilized in two different ways: First, Venus can be used by developers to ensure that their user interface is consistent with existing policies, thereby improving overall user experience and ensuring compliance with applicable laws. Second, Venus can be used by security analysts to detect ad fraud applications that trick users through a misleading user interface.
In practice, checking conformance between an app and a GUI policy turns out to be challenging for two key reasons. First, Android applications consists of several interacting activities, all of which provide a different and dynamically changing interface. Thus, checking adherence to a GUI policy requires exploring the (possibly infinite) ways that a user can interact with the app. Second, by studying existing GUI policies, we found that many of them concern not only the static appearance of the app, but also how the interface needs to dynamically evolve as users interact with it. Thus, verifying an app against a GUI policy requires reasoning about the dynamic behavior of the app in relation to the GUI elements it provides.

In this paper, we address these challenges through an end-to-end solution that statically reasons about an app’s GUI-related behaviors. Our solution consists of three ingredients that make it possible to specify and check such properties:

1. **Policy language:** We present a formal policy language called Vesper for expressing realistic GUI design guidelines. Vesper allows specifying both spatial relations between GUI elements as well as their behavioral properties, such as how a button should react to a click event.

2. **ELF abstraction:** We propose a new program abstraction called Event-driven Layout Forest (Elf) that summarizes spatial and behavioral properties of GUI elements. While Elf bears resemblance to other Android abstractions like window transition graph [46] and ICCG [11], it differs from them in that nodes correspond to individual GUI elements (rather than activities) and node labels (computed using numeric abstract domains and pointer analysis) track GUI-related properties.

3. **Conformance checking:** To check whether an Android app corresponds to a Vesper specification, Venus needs to decide whether a given Elf abstraction is a model of the input Vesper specification. Venus achieves this task by encoding both the Elf abstraction and the Vesper policy as logical formulas and reduces conformance checking to a satisfiability query.

To evaluate the effectiveness of our proposed approach, we performed an extensive experimental evaluation on 2361 Android applications. Specifically, we formalized existing GUI policies as Vesper specifications and then used Venus to check each Android application against these policies. Our evaluation shows that Venus is able to accurately pinpoint violations of GDPR policies with a low false positive rate (around 6.9%). Furthermore, Venus can identify previously unknown ad fraud instances and detect violations of a subset of GDPR (General Data Protection Regulation) regulations.

In short, this paper makes the following key contributions:

- We propose a policy language called Vesper for describing GUI policies (Section 4).
- We introduce a new program abstraction called event-driven layout forest that is suitable for checking such GUI policies and present a static analysis technique for automatically constructing the proposed Elf abstraction (Section 5)
- We implement Venus, the first tool for statically checking conformance between Android applications and GUI specifications, and we extensively evaluate Venus by checking conformance between 2361 Android applications and several existing GUI policies (Section 6).

### Background on Android GUI

In Android, the basis of an app’s user interface is an activity, which always has a window associated with it. Activities can start other activities by a message-passing system known as inter-component communication (ICC). An Android ICC message is an Intent, which can be thought of as a description of what the launched component should do. An Intent object specifies both the action to perform (e.g., view, edit, etc.) and provides the relevant data.

The Android framework provides two types of basic GUI elements, namely Views and ViewGroup. A View element is a widget, such as a button or progress bar, that the user can see and interact with. A ViewGroup is an invisible container that stitches
together Views and ViewGroup objects. Android provides different types of ViewGroup objects, such as a LinearLayout for arranging GUI elements horizontally or vertically. The user interface of a GUI activity corresponds to a tree data structure (figure 2(a)), where internal nodes are ViewGroup elements and all leaves are View objects. Each Android GUI element also has a set of attributes that define its properties, including height, width, alignment, position etc.

**Declaring and manipulating GUI elements.** In Android, there are two ways to declare GUI elements. The first option is to specify the layout through an XML file. In addition to defining the hierarchical user interface of an activity, the XML file can also specify the attribute values of each GUI element, such as the text attribute “CONTINUE” of a button on line 5 of listing 1. During compilation, the XML file is translated into a so-called layout resource that can be loaded in the application’s source code by calling setContentView (R.layout.layout_name) (e.g., line 4 of listing 2). An alternative way to create a layout is to do so programmatically by calling methods provided by the Android framework. For instance, rather than statically declaring the text attribute in the XML file, a program can do this at run time by calling the setText method.

In practice, programmers often combine XML-based declaration of GUI elements with programmatic modifications at run time. For example, line 4 of listing 2 loads the layout declared in the XML file, but the two subsequent lines modify the title of the nested TextView element to “Venus Demo” from its original name (“Default title”) declared in line 3 of listing 1. Hence, understanding an application’s user interface requires analysis of both XML files and source code.

**Interacting with GUI elements.** To facilitate interaction with users, GUI elements register callbacks that get invoked upon specific types of user events (e.g., click, hover, etc.). In particular, Android GUI elements can respond to events of type X types of user events. In particular, Android GUI elements can respond to events of type X types of user events. GUI elements can respond to events of type X types of user events.

### 3 OVERVIEW

This section gives an overview of the Venus framework through a simple but realistic motivating example.

#### 3.1 Example GUI Policy for Ad Fraud Detection

Fig. 3 shows the screenshot of an *ad fraud* application called “Super Cleaner” that was recently submitted to the Google Play Store. This app does not conform to a Google AdMob policy [17], which states that transparent backgrounds should not display ads upon a click event. However, as shown in parts (b) and (c) of figure 3, the Super Cleaner application blatantly violates this policy.

In order to use Venus to check conformance between this app and the AdMob policies, the user first needs to formalize the policy in Venus’s specification language. In particular, figure 4 shows a formalization for the policy “transparent backgrounds should not display ads upon click events” in our policy language called Vesper.

The first line declares a View element called bg. Next, the assume statement stipulates that bg is the background of some other View element. Then, on line 3, the let binding defines a custom predicate called popAd(v), which evaluates to true if clicking on v shows a new window v’ that corresponds to an adView GUI element. Finally, the assertion specifies the desired property. Section 4 will present more about Vesper.

Given this Vesper policy and the source code of the Super Cleaner application (shown in listing 3 and listing 4), we next explain how Venus automatically identifies this policy violation.

```java
public class MainActivity extends Activity implements OnClickListener {
    @Override
    public void onClick(View view) {
        // suspicious URL
        loadURL("http://funtest.afatwallet.com");
    }
}
```

**Listing 3: Main activity**

```java
public class BatterySaver extends Activity {
    public void onCreate(Bundle bundle) {
        FrameLayout frameLayout = (FrameLayout)
            findViewById(R.id.bg_view);
        View a = new NativeAdViewBuild().f();
        frameLayout.addView(a);
    }
}
```

**Listing 4: Battery saver activity**

![Image of Venus screenshot](image-url)

**Figure 3:** Clicking on the white space will (surprisingly) trigger the display of an untrusted website.
1. View bg
2. assume (\(\exists v. (\text{View}(v) \land \text{background(bg, v)})\))
3. let popAd(v) = \(\exists v'. (\text{showWindow}(v, \text{click}, v') \land \text{AdView}(v'))\)
4. assert (transparent(bg) \(\rightarrow \neg\text{popAd(bg)}\))

Figure 4: Vesper specification for the policy “Transparent backgrounds should not be clickable”.

3.2 Elf Generation via Static Analysis
As mentioned earlier, Vesper uses static analysis to construct an even-driven layout forest (Elf) abstraction of the application. At a high-level, this abstraction captures all relevant behavior of the app with respect to the Vesper policy language. For example, figure 5 shows the Elf abstraction for the Super Cleaner application. Here, each node corresponds to a GUI element; node labels (e.g., for bgView) indicate attribute values (e.g., opacity, width); and there are two types of edges: (1) a spatial (solid) edge from node \(n\) to \(n'\) indicates that GUI element \(n'\) is nested inside \(n\), and (2) a behavioral (dashed) edge from \(n\) to \(n'\) labeled with \(e\) indicates that GUI element \(n'\) launches another GUI element \(n'\) upon event \(e\). For example, in figure 5, there is a spatial (solid) edge from MainActWindow to saverBtn since the latter is spatially nested within the window of the main activity (see figure 3). On the other hand, there is a behavioral (dashed) edge from saverBtn to BatterySaverWindow labeled with showWindow(click) because clicking on the saverBtn results in opening the window of the BatterySaver activity (see code 3).

In practice, constructing a sufficiently precise Elf abstraction of the application requires non-trivial static analysis. For example, the construction of behavioral edges between GUI elements requires reasoning about heap objects and callbacks as well as analysis of inter-component communication (ICC). On the other hand, reasoning about GUI element attributes (e.g., height, width) requires reasoning about numeric values.

3.3 Checking Conformance
Our method uses the computed Elf abstraction to check conformance against any Vesper policy. At a high-level, we can think of the Elf abstraction as defining a conjunction of ground predicates in Vesper. Thus, checking conformance between the app and policy boils down to determining whether the formula defined by the Elf abstraction implies the specification. Going back to our example, we can determine that Super Cleaner violates the Vesper policy from figure 4 using the following chain of inferences:

- First, since bgView is nested inside nativeAdView and has the same width/height of its parent (figure 5), we determine that bgView satisfies the assumption from line (2) of Figure 4.
- Next, because the opacity attribute of bgView is 0 (see figure 5), transparent evaluates to true for bgView.
- In addition, bgView satisfies the popAd predicate because figure 5 contains a behavioral edge from bgView to adViewWindow labeled with click.
- Finally, because bgView satisfies both the assumption at line (2) as well as the transparent and popAd predicates, the assertion at line (4) of Figure 4 is violated.

Therefore, Vesper reports that the Super Cleaner app does not conform to the Vesper policy from figure 4.

4 Vesper Specifications
As shown in figure 6, a specification in Vesper starts with a set of declarations, is followed by a sequence of statements (i.e., definitions and assumptions), and ends in a set of assertions. While Vesper provides built-in predicates relevant to the spatial and behavioral properties of GUI elements (figure 7), the user can also define custom predicates through let bindings. For instance, in figure 4, showWindow is an example of a built-in predicate, whereas popAd is a custom predicate defined by the user. Vesper also provides a way to define a set of GUI elements through the set comprehension syntax \(\{v | \Phi\}\).

Expressions. In Vesper, the most basic expressions are variables \(v\), integer constants \(c\), and pre-defined Android events \(e\) such as click or touch. Vesper allows performing arithmetic operations over integers as well as aggregation over sets. For instance, the expression \(\text{count}(v)\) returns the number of elements in set \(v\).

Built-in predicates. Vesper provides a core set of built-in predicates that constrain spatial and behavioral properties of GUI elements. Figure 7 shows examples of these predicates, which are classified into three categories:

- **Element type predicates** describe the type of a GUI element (e.g., button, dialog). Note that, unlike the actual Android API, Vesper does not differentiate between views and view groups, and every GUI element is considered to be a view. Thus, views can contain nested views under Vesper’s semantics.
- **Spatial predicates** refer to visual properties of GUI elements (e.g., height, width) as well as spatial relationships between different GUI elements (e.g., containment).
- **Behavioral predicates** constrain how GUI elements react to user events (e.g., what methods they can invoke, which other GUI elements they can display, etc.).

Example 1. Consider the following Vesper specification:

```vesper
View w;
let LView(v) = \exists x, y. \{(\text{width}(v, x) \land x > 100) \land \text{height}(v, y) \land y > 100\}
let LAds = \{v | \text{AdView}(v) \land \text{contains}(w, v) \land LView(v)\}
assert \text{count}(LAds) \leq 1
```

This specification requires that every window contains at most one “large” ad, meaning that the width and height of the ad is above a certain threshold. Here, the combination of set comprehension syntax and the count function allows constraining the number of GUI elements with a certain property.

We present the formal semantics of Vesper policies in Appendix A. At a high level, the semantics of Vesper policies are defined over execution traces, and we consider a predicate \(p(\bar{s})\) to be true in an execution \(\omega\) if it holds on objects \(\bar{s}\) at any time during \(\omega\). For example, the predicate \(\text{startBrowser}(e, v)\) evaluates to true in execution \(\omega\) if \(v\) starts the browser at some point during \(\omega\). Given
Appendix A). Finally, we say that an app’s policy under ELF is a tuple \( \langle N, N_0, E, L \rangle \) where:

- \( N \) is a set of nodes where each node is a pair \((a, r)\) representing an abstract heap object \( a \) of GUI element type \( r \).
- \( N_0 \subseteq N \) are initial nodes that may correspond to the main window of the application.
- \( E = E_S \uplus E_B \) encode relationships between GUI elements. We refer to edges \((n, n') \in E_S\) as spatial edges and \((n, e, n') \in E_B\) as behavioral edges.
- Labeling function \( L : N \times \langle \text{Attrib} \rangle \rightarrow \text{AbstractVal} \) maps attributes of GUI elements to their abstract values.

As mentioned in Section 3, a spatial edge \((n, n')\) encodes that GUI element \( n \) is nested within \( n' \), whereas a behavioral edge \((n, e, n')\) indicates that user/system event \( e \) on GUI element \( n \) directly results in the display of element \( n' \). The labeling function \( L \) can refer to both spatial and behavioral properties of GUI elements. For example, the height attribute refers to a spatial property of the node, whereas click is a behavioral property that identifies which methods may be invoked upon a click event. In general, since \( \text{VENUS} \) cannot exactly determine the values of node attributes using static analysis, the labeling function \( L \) maps these attributes to abstract rather than concrete values; however, the choice of abstract domain depends on the type of attribute (see Section 6).

5.2 XML Analysis for Layout Schema

As mentioned in section 2, GUI elements in Android are typically declared via an XML file and then loaded by the application code at runtime. Thus, to facilitate static analysis, \( \text{VENUS} \) encodes the GUI-related information declared in the XML file as a so-called layout schema. As shown in figure 8, a layout schema \( \Psi \) maps each layout name to its structure, represented as a multi-map from attributes to their type \( T \) and default value \( c \). Given a layout name \( N \) and its definition \( \Psi(N) = \{a_1 \mapsto \langle T_1, c_1 \rangle, a_2 \mapsto \langle T_2, c_2 \rangle, \ldots, a_n \mapsto \langle T_n, c_n \rangle\} \), we write \( \text{DefaultVal}(N) \) to indicate an object with fields \( a_1, \ldots, a_n \) where each field \( a_i \) initialized to \( c_i \).

Example 2. Consider the following layout XML:

```
<LinearLayout id="lin" orientation="vertical">
    <TextView id="txt1" width=100 height=200 />
```

the truth value of built-in predicates under \( \omega \), evaluation of the full policy under \( \omega \) largely follows the standard semantics of first-order logic, with some modifications to handle set comprehension (see Appendix A). Finally, we say that an app \( A \) conforms to a Vesper policy \( P \) if \( P \) evaluates to true in all executions of \( A \).
5.3 Static Analysis

In this section, we describe our static analysis for computing the ELF abstraction using Datalog-style inference rules. Note that the event-driven layout forest is a global abstraction of the entire application; however, our static analysis for computing is both flow- and context-sensitive. Our analysis leverages the layout schema extracted from the XML file (Section 5.2) as well as the results of standard techniques like pointer analysis.

We formalize our static analysis using three different types of predicates (summarized in Table 1):

- **Source code predicates** refer to statements in the source code. For instance, `addView(l, m, v1, v2)` indicates that there is an API call of the form `v1.addView(v2)` at location `l` of method `m`.

- **Pre-analysis predicates** refer to program facts computed by off-the-shelf static analyzers. For example, `pointsTo(c, l, v, o)` indicates that variable `v` points to heap object `o` at program location `l` in calling context `c`. Similarly, `val(c, l, v, a)` indicates that variable `v` has abstract value `a` at location `l` in calling context `c`.

- **Output predicates** collectively define our ELF abstraction. For example, the predicate `sAttrib(o, a, val)` indicates that the abstract value for spatial attribute `a` of object `o` is `val`, and `bEdge(o, e, o')` indicates that there is a behavior edge between `o` and `o'` labeled `e`.

As mentioned earlier, we present our static analysis (see figure 9) using Datalog-style rules of the form:

\[ H(x_1, \ldots, x_n) \iff B_1(\ldots), \ldots, B_k(\ldots) \]

The meaning of such a rule is that the predicate `H(x_1, \ldots, x_n)` is true if all the of the predicates `B_1(\ldots), \ldots, B_k(\ldots)` in the rule body are satisfied. We refer to `H` as the *head predicate* and the `B_i`'s as *body predicates*. In our case, the head predicates are either auxiliary or output predicates computed by our analysis, whereas body predicates also involve source code and pre-analysis predicates. If an argument to a predicate does not matter, we use the symbol "_" to indicate that it matches anything.

We now explain the rules from figure 9 in more detail.

**Nodes.** According to the first rule in figure 9, any (abstract) heap object that corresponds to a GUI element (i.e., is subtype of `View`) is a node in the event-driven layout forest abstraction.

**Root view.** The second rule computes a predicate `rootView(o, o')` indicating that Activity `o` sets its main window to be GUI element `o'`. Since root views are set via an API call `setContentView(v')`, this rule looks up the heap objects pointed to by variables `v, v'` at the program location `l` (in method `m`) where the API call occurs. Note that our analysis is context-sensitive in that we look up the points to sets of `v, v'` in feasible calling contexts of `m`.

**Entry view.** The next rule marks the initial nodes of the ELF abstraction. To determine the initial nodes, we first identify all heap objects `o` that are of instance of type `A`, where `A` is the main activity of the application. We then mark all root views of `o` as initial nodes using the auxiliary `rootView` predicate from rule (2).

**Behavioral attributes.** The next rule, (4), describes how we compute behavioral attributes of each node. In particular, behavioral attributes map each GUI event to a set of methods that can be used to handle that event. Since event handlers are registered via the `setListener` methods, this rule uses the `setListener(l, m, v, m')` source code predicate, which indicates that method `m'` is registered as the listener for event `X` for variable `v`, and if `m` corresponds to the program location and method where the registration occurs respectively. If `v` points to a heap object `o` that is a node in the ELF abstraction, behavioral attribute `X` is mapped to method `m'`.

**Spatial attributes.** The next three rules, (5)–(7), describe the computation of spatial attributes. Unlike behavioral attributes that have a finite domain (i.e., a set of methods), spatial attributes like height have an infinite domain (i.e., all integers). Thus, our method uses abstract interpretation to reason about such attributes. In particular, rule (5) initializes all spatial attributes to `⊥`, as standard.

The next two rules essentially describe a fixed point computation where we take the join of existing values with a new value.
### Table 2: GUI policies that we formalized as Vesper specifications.

<table>
<thead>
<tr>
<th>Category</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraudulent</td>
<td>8</td>
</tr>
<tr>
<td>Unwanted</td>
<td>3</td>
</tr>
<tr>
<td>Appearance</td>
<td>4</td>
</tr>
<tr>
<td>GDPR Consent</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\text{node}(o, \tau) \iff \text{pointsTo}_o(\tau, v, o), \quad \text{hasType}(o, \tau, \tau <: \text{View}).
\]

(1)

\[
\text{rootView}(o, o') \iff \text{setContentView}(l, m, v, v'), \text{inCtx}(m, c), \text{pointsTo}(c, l, v, o), \text{pointsTo}(c, l, v', o').
\]

(2)

\[
\text{entryView}(o) \iff \text{mainAct}(A), \text{instanceOf}(o, A), \quad \text{rootView}(o', o).
\]

(3)

\[
\text{bAttrib}(o, X, m) \iff \text{node}(o, \text{setXListener}(l, v, m), \text{inCtx}(m, c), \text{pointsTo}(c, l, v, o)).
\]

(4)

\[
\text{sAttrib}(o, a, \bot) \iff \text{node}(o, \text{View}), a \in \text{Attribs}(\Psi).
\]

(5)

\[
\text{sAttrib}(o, a, \text{val}'f) \iff \text{loadView}(l, m, v, N), \text{inCtx}(m, c), \text{pointsTo}(c, l, v, o), a \in \text{Dom}((\Psi(N)), a \neq \text{subview}, \Psi(N)(a) = (7, \text{val}_a) \quad \text{sAttrib}(o, a, \text{val}_a), \text{val}'f = \text{val} \cup \alpha(\text{val}_a).
\]

(6)

\[
\text{sAttrib}(o, a, \text{val}'f') \iff \text{setAttrib}(l, m, v, a, v', \text{inCtx}(m, c), \text{pointsTo}(c, l, v, o), \text{val}(c, l, v', \text{val}'f), \text{sAttrib}(o, a, \text{val}_a), \text{val}'f' = \text{val} \cup \alpha(\text{val}_a').
\]

(7)

\[
\text{sEdge}(o, o') \iff \text{loadView}(l, m, v, N), \text{inCtx}(m, c), \text{pointsTo}(c, l, v, o'), o' \in \text{Dom}((\Psi(N)) \text{subview}).
\]

(8)

\[
\text{sEdge}(o_1, o_2) \iff \text{addView}(l, m, v_1, v_2), \text{inCtx}(m, c), \text{pointsTo}(c, l, v_1, o_1), \text{pointsTo}(c, l, v_2, o_2).
\]

(9)

\[
\text{bEdge}(o_1, X, o_2) \iff \text{bAttrib}(o_1, X, m), \text{inCtx}(m, c), \text{call}('c', m, m'), \text{inCtx}(m', c'), \text{showWindow}(l, m', v), \text{inCtx}(m', c'), \text{call}('c', m, m'), \text{inCtx}(m', c'), \text{icent}(l, m', l), \text{pointsTo}(c, l, i, o).
\]

(10)

\[
\text{bEdge}(o_1, X, o_2) \iff \text{bAttrib}(o_1, X, m), \text{inCtx}(m, c), \text{call}('c', m, m'), \text{inCtx}(m', c'), \text{rootView}(o', o, \text{get'}, \text{rootView}(o', o_2)).
\]

(11)

**Figure 9:** Datalog-style inference rules describing Elf construction. Here, \( \cup \) is an abstraction function for the underlying abstract domain, and \( \text{corresponding join operator. } \Psi \text{ refers to the layout schema from section 5.2.} \)

Specifically, in rule (6), we deal with API calls that load a view from the XML file. In particular, suppose that we have determined that attribute \( a \) of layout name \( N \) can have default value \( c \) according to the analysis from Section 5.2. Now, if we encounter an API call that loads layout \( N \) into variable \( v \), we first look-up the points-to target \( o \) of \( v \) and add \( c \) to the set of possible values of \( o.a \) by taking the join with the old abstract value with \( c \).

Next, rule (7) deals with spatial attributes that are set programmatically via an API call. We represent such API calls using the source code predicate \( \text{setAttrib}(l, m, v, a, v') \) indicating that attribute \( a \) of variable \( v \) is set to variable \( v' \) at program location \( l \) inside method \( m \). To update the Elf abstraction, we first look up the abstract value \( a \) of variable \( v' \) at program location \( l \) in some calling context \( c \) of method \( m \). If \( v \) points to heap object \( o \) at the same program location \( l \) and calling context \( c \), we then update \( o.a \) to be the join of \( a \) and \( o.a \)’s old abstract value. Our implementation uses the interval abstract domain for numeric attributes and the so-called bounded set abstraction for strings. [7, 31]

**Spatial edges.** The next two rules, (8) and (9), describe the introduction of spatial edges due to loading views from the XML file and programmatically adding sub-views respectively. Since these rules are very similar, we only focus on (9). Consider an API call for adding view \( v_2 \) as a sub-view of \( v_1 \) at program point \( l \) in method \( m \). If \( v, v' \) point to heap objects \( o, o' \) at program location \( l \) in the same calling context \( c \) of method \( m \), we introduce a spatial edge from \( o \) to \( o' \) in the Elf abstraction. In general, \( v, v' \) can have multiple points-to targets; thus, this rule can end up introducing multiple spatial edges for the same source code statement.

**Behavioral edges.** The last two rules, (10) and (11), deal with the introduction of behavioral edges. Recall that a behavioral edge indicates that GUI element \( o \) launches GUI element \( o' \) upon event \( X \). In general, \( o \) can launch \( o' \) in one of two ways: The handler of event \( X \) (transitively) calls a method that \( (a) \) either directly displays \( o' \) by calling an API (e.g., \text{showWindow}) or \( (b) \) indirectly displays \( o' \) by performing inter-component communication via an intent object whose target has root view \( o' \). In figure 9, rule (10) deals with case (a), and rule (11) deals with case (b). Since both of these rules rely on knowing the handler method for event \( X \), the body of the rule matches the \text{bAttrib} predicate computed by the other rules.

5.4 Checking Conformance

Once Venus generates the Elf abstraction, it translates attributes and edges in the Elf abstraction to ground built-in predicates in the Vesper specification language in the expected way. For instance, the spatial edge \((o, o')\) in the Elf corresponds to the predicate \text{contains}(o, o') in the Vesper DSL. Similarly, a behavioral edge \((o, o')\) corresponds to the Vesper predicate \text{showWindow}(o, e, o') if \( o' \) is another window and, for instance, to \text{startBrowser}(e, o) if \( o' \) is the browser. Thus, Venus can directly convert the Elf abstraction to a formula \( \mathcal{F} \) that is a conjunction of ground predicates.

Next, to decide whether the input program \( \mathcal{P} \) entails specification \( \Psi \), Venus checks whether \( \mathcal{F} \) implies \( \Psi \). To do so, Venus first
converts $\psi$ to a logical formula $\phi$ using the $\downarrow$ function defined in Appendix A and then checks the satisfiability of the formula $F \land \neg \phi$ using a Datalog solver. If this formula is satisfiable, the specification is violated under the computed Elf abstraction, and Venus produces a model of $F \land \neg \phi$ as a potential counterexample. On the other hand, the unsatisfiability of $F \land \neg \phi$ constitutes a proof of conformance since $F$ over-approximates the app’s relevant behavior with respect to the Vesper specification language.

6 IMPLEMENTATION AND EVALUATION

We implemented our core static analysis on top of the Soot framework [42] and the IC3 tool for Android [34]. We use the SPARK framework [27] provided by Soot to perform pointer analysis and construct a call-graph. Our implementation uses the interval abstract domain for reasoning about numeric attributes and the bounded-set abstraction for strings. Venus also leverages the Soufflé [41] Datalog solver for checking conformance between the Elf abstraction and the Vesper specification. As described in Section 5, our analysis is context-sensitive and uses the call site representation proposed in [35]. Venus is openly available on Github. 1

Experimental set-up. All of our experiments are conducted on a shared 48-core server with Intel Xeon E7-8890 CPU and 500G memory, running the CentOS 7.6 operating system.

6.1 Benchmarks

To evaluate Venus, we collected 2361 Android applications from three different sources:

• Google Play: We collected 1488 popular applications that were available on the Google Play Store in Jan 2019.

• GPP benchmarks: The Google Play Protect (GPP) team provided us with a labeled data set consisting of 773 Android apps and their label (benign or type of malware). All of these applications were flagged as potential malware by Google’s internal tools and manually audited by Google security analysts.

• AdFraudBench: We also evaluate our approach on a dataset taken for detecting ad fraud [8]. This dataset includes 57 ad fraud samples and 43 benign applications.

6.2 Properties

To evaluate Venus, we collected a total of 49 representative GUI policies from Google Play Ads Policy [19], AdMob Help [14], Material Design [18], and EU General Data Protection Regulation [9]. Among those 49 policies, 25 are too vague to formalize (e.g., “Ensure that none of the ad attributes look like navigation features within the app.”). Among the remaining 24, seven of them cannot be expressed in Vesper (e.g., require temporal logic). This leaves us with a total of 17 policies that we formalized in Vesper. To give the reader some intuition, table 2 shows a categorization of these policies and provides some examples of the types of policies we formalized. (See the Appendix for their Vesper formalizations.)

6.3 Results on Google Play Dataset

We evaluated Venus on the 1488 Google Play apps by checking conformance against all 17 policies summarized in table 2. As shown in the first row of table 3, Venus reports a total of 1645 violations across 711 apps, with an average running time of 465.3 seconds per app. Among the 1645 reports, 1258 reports pertain to violations of ad-related policies, 127 reports concern GDPR regulations, and the remaining 260 reports pertain to Material design guidelines.

Manual inspection. Since there is no ground truth label for the apps in the Google Play dataset, we manually inspected 50 of the 711 apps for which Venus reports at least one violation. For these 50 apps, Venus reports a total of 195 warnings. We now report on the findings from our manual inspection.

• GDPR violations: Among the 50 apps we inspected, Venus reports a total of 18 GDPR violations, and we manually confirmed that 16 of them indeed access private user information without ever displaying a user consent form.

• Ad-fraud: Across the 50 manually inspected apps, Venus reports 40 of them to violate an ad-related property. In particular, 37 of these are true positives, and 11 are previously unknown ad fraud instances (confirmed by Google security auditors).

• Design guidelines: Venus reports 24 of the 50 apps to violate a Material design guideline-related property, and 18 of these indeed violate the design guidelines we encoded.

False positive analysis. Among all 50 sampled apps, Venus reported 195 violations, of which 174 are true positives. Based on our manual inspection, most of the false positives are due to imprecision in the pointer analysis. Using the estimation of proportion method [24], we conclude that it is 95% likely that the false positive rate for the whole dataset is between 4% and 18%.

Result #1: Among the 50 apps we manually inspected, Venus identified 11 previously unknown ad fraud instances (confirmed) and 16 Google Play apps that violate GDPR regulations. Furthermore, Venus’s false positive rate for the inspected apps is around 10%.

6.4 Results on GPP Dataset

The GPP dataset consists of 773 apps where each app is either labeled as benign or malicious. If the app is malicious, the label also indicates the type of malware (e.g., ad fraud, spyware). For this dataset, we used Venus to detect ad fraud instances by checking
We can see from this bar chart, we compare Venus's results with those of VirusTotal [43], which is a widely-used service for detecting several types of malware. VirusTotal uses more than sixty state-of-the-art malware detection engines to analyze an app and shows the aggregate results.

Since VirusTotal does not report a single result and covers a broader class of malware than just ad fraud, there is no "obviously right" way to compare against it for the purposes of our evaluation. Thus, we consider two different, but equally plausible, ways of interpreting VirusTotal results:

- **VirusTotal-a**: As in prior work on ad fraud detection [8], we consider VirusTotal to classify an app as ad fraud if at least two of its underlying malware detection engines label it as ad fraud.
- **VirusTotal-b**: Since the security community typically uses VirusTotal as a binary classifier [5], we consider an app to be ad fraud if at least two of the underlying malware detectors label the app as not benign.\(^2\)

The results of our comparison are shown in figure 10. Here, blue bars (with "\" pattern) show recall, whereas dark magenta bars (with "\" pattern) indicate precision. As we can see from this bar chart, both variants of VirusTotal yield much lower recall and precision compared to Venus.

### Analysis of false positives and negatives

We manually inspected the apps that are incorrectly classified by Venus to better understand the root causes of false positives and false negatives. Most of the false positives are caused by imprecision in the pointer analysis (e.g., additional spurious methods are identified as event handlers). On the other hand, false negatives are mainly caused by the security community's criteria.

### Comparison against VirusTotal

To put these results in context, we also compare Venus's results against those of VirusTotal as well as FraudDroid, which is a dynamic analysis tool specifically for detecting ad fraud [8]. The results of this comparison are shown in table 4, which shows that Venus outperforms VirusTotal and FraudDroid both in terms of precision and recall.\(^3\)

<table>
<thead>
<tr>
<th>Dataset</th>
<th># apps</th>
<th># violating apps</th>
<th># violations</th>
<th>recall</th>
<th>precision</th>
<th>avg. time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Play</td>
<td>1488</td>
<td>711</td>
<td>1645</td>
<td>N/A</td>
<td>89.2%</td>
<td>465.3</td>
</tr>
<tr>
<td>GPP</td>
<td>773</td>
<td>243</td>
<td>391</td>
<td>86.8%</td>
<td>95.7%</td>
<td>464.7</td>
</tr>
<tr>
<td>AdFraudBench</td>
<td>100</td>
<td>54</td>
<td>90</td>
<td>91.2%</td>
<td>96.3%</td>
<td>302.1</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>2361</td>
<td>1008</td>
<td>2126</td>
<td>N/A</td>
<td>91.3%</td>
<td>458.2</td>
</tr>
</tbody>
</table>

### Results on AdFraudBench Dataset

In our next experiment, we evaluate Venus on the AdFraudBench dataset used in prior work [8]. Since this data set is specifically targeted for ad fraud detection, we check these apps against the eight ad-fraud-related policies formalized in Vesper. As shown in table 3, Venus has a precision of 96.3% and recall 91.2% on this dataset, and its average running time per app is 302.1 seconds.

To put these results in context, we also compare Venus's results against those of VirusTotal as well as FraudDroid, which is a dynamic analysis tool specifically for detecting ad fraud [8]. The results of this comparison are shown in table 4, which shows that Venus outperforms VirusTotal and FraudDroid both in terms of precision and recall.\(^4\)

### Evaluation of the ELF Abstraction

In our final experiment, we evaluate the benefits of our proposed ELF abstraction by performing ablation studies and comparing it against the windown transition graph (WTG) abstraction proposed in prior work [46].

WTG abstraction. As mentioned earlier, the WTG abstraction from the GATOR tool [39] is somewhat similar to ELF in that it is a graph abstraction of Android applications where nodes are windows, and edges (annotated with events) represent communication between them. However, WTG differs from our proposed ELF abstraction in two important ways: First, nodes in a WTG correspond to main windows of activities, so it does not contain nodes for any nested GUI elements. Second, a WTG does not contain any information about spatial attributes of windows. To use the WTG abstraction to check Vesper specifications, we use the following

---

\(^2\)Recall that all applications in these datasets are either benign or ad fraud.

\(^3\)Dong et al. ([8]) report 92% recall on 100 apps from the AdFraudBench instead of the 12000 apps dataset. After we manually inspected the ground truth for those 100 apps, we noticed that FraudDroid actually mislabeled 7 malicious apps as benign. To resolve this discrepancy, we further confirmed our results by uploading those 7 apps to VirusTotal, which also marked those apps as malware. We further contacted the co-authors of FraudDroid and they also agreed that those 7 apps should all be ad fraud. That is why the actual recall is around 80%.

---

<table>
<thead>
<tr>
<th>Dataset</th>
<th># apps</th>
<th># violating apps</th>
<th># violations</th>
<th>recall</th>
<th>precision</th>
<th>avg. time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FraudDroid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VirusTotal-a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VirusTotal-b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3: Summary of Venus results across all three datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th># apps</th>
<th># violating apps</th>
<th># violations</th>
<th>recall</th>
<th>precision</th>
<th>avg. time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Play</td>
<td>1488</td>
<td>711</td>
<td>1645</td>
<td>N/A</td>
<td>89.2%</td>
<td>465.3</td>
</tr>
<tr>
<td>GPP</td>
<td>773</td>
<td>243</td>
<td>391</td>
<td>86.8%</td>
<td>95.7%</td>
<td>464.7</td>
</tr>
<tr>
<td>AdFraudBench</td>
<td>100</td>
<td>54</td>
<td>90</td>
<td>91.2%</td>
<td>96.3%</td>
<td>302.1</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>2361</td>
<td>1008</td>
<td>2126</td>
<td>N/A</td>
<td>91.3%</td>
<td>458.2</td>
</tr>
</tbody>
</table>

### Table 4: Results on AdFraudBench

<table>
<thead>
<tr>
<th>Dataset</th>
<th>precision</th>
<th>recall</th>
<th>(^4)</th>
<th>(^5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venus</td>
<td>96.3%</td>
<td>91.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FraudDroid</td>
<td>91.8%</td>
<td>79.6%</td>
<td>75.0%</td>
<td>89.5%</td>
</tr>
<tr>
<td>VirusTotal-a</td>
<td>79.6%</td>
<td>75.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VirusTotal-b</td>
<td>75.0%</td>
<td>75.4%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

\(^2\)Recall that all applications in these datasets are either benign or ad fraud.

\(^3\)Dong et al. ([8]) report 92% recall on 100 apps from the AdFraudBench instead of the 12000 apps dataset. After we manually inspected the ground truth for those 100 apps, we noticed that FraudDroid actually mislabeled 7 malicious apps as benign. To resolve this discrepancy, we further confirmed our results by uploading those 7 apps to VirusTotal, which also marked those apps as malware. We further contacted the co-authors of FraudDroid and they also agreed that those 7 apps should all be ad fraud. That is why the actual recall is around 80%.

---

\(^4\)Venus outperforms FraudDroid (a dynamic analysis tool for ad fraud detection) significantly in terms of recall, while also attaining better precision.

---

\(^5\)FraudDroid is not available, so we cannot evaluate it on GPP apps.
methodology: First, since WTG only contains main windows of activities, we consider any GUI element mentioned in the VESPER specification but not in the WTG as being non-existent in the app. Clearly, this may result in GATOR reporting false negatives. Second, since a WTG does not contain any information about spatial attributes, we consider the abstract value of any spatial attribute to be $\top$, which can result in false positives. Thus, in principle, using GATOR to check for VESPER specifications can suffer from both false positives as well as false negatives.

**Ablations of ELF.** In this evaluation, we also compare our proposed ELF abstraction against two of its own ablations. Since one of our claims is that many GUI policies require reasoning about both spatial and behavioral properties in practice, we consider the following two ablations of ELF:

- **VESPER**$^{-S}$: This is a variant of VESPER that does not contain spatial attributes. In other words, we do not perform abstract interpretation to reason about values of spatial attributes such as height, size etc., and simply map all of them to $\top$.
- **VESPER**$^{-B}$: This is a variant of VESPER that does not contain any behavioral edges or attributes. In particular, we do not reason about event handlers of GUI elements (i.e., behavioral attributes), and we also do not reason about communication between different GUI elements (i.e., behavioral edges).

At first glance, it might seem that VESPER$^{-S}$ should have only false positives whereas VESPER$^{-B}$ would suffer from only false negatives. However, since VESPER predicates may appear negated in the specification, in principle, VESPER$^{-S}$ and VESPER$^{-B}$ can have both false negatives and false positives.

Table 5 presents the results of our evaluation of the ELF abstraction by comparing it against WTG, VESPER$^{-S}$, and VESPER$^{-B}$ on both the GPP and AdFraudBench datasets for which we know the ground truth. Our first observation is that GATOR has high precision but very poor recall. While the poor recall is perhaps expected, the high precision is surprising since we treat spatial attributes as $\top$ when using the WTG abstraction to check VESPER policies. However, the reason for this is that GATOR reports a grand total of 3 violations (among the 258 actual violations) in the GPP dataset, and all of these three reports turn out to be real violations. However, the recall is extremely poor, resulting in F1-scores of 0.024 and 0.387 on the GPP and AdFraudBench datasets compared to that of 0.906 and 0.937 of VESPER.

Next, we compare VESPER against its two ablations. While the recall of both ablations is significantly higher than the WTG abstraction, the overall F1-scores of substantially worse than VESPER.

Table 5: Evaluation of our abstraction. Prec. is "Precision"

<table>
<thead>
<tr>
<th>Tool</th>
<th>GPP</th>
<th>AdFraudBench</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec.</td>
<td>Recall</td>
</tr>
<tr>
<td>GATOR</td>
<td>100.0%</td>
<td>1.2%</td>
</tr>
<tr>
<td>VESPER$^{-S}$</td>
<td>53.8%</td>
<td>85.2%</td>
</tr>
<tr>
<td>VESPER$^{-B}$</td>
<td>69.0%</td>
<td>80.9%</td>
</tr>
<tr>
<td>VESPER</td>
<td>94.7%</td>
<td>86.8%</td>
</tr>
</tbody>
</table>

These results indicate that our proposed ELF abstraction is highly beneficial for checking apps against GUI policies.

**Result #4:** Our proposed ELF abstraction significantly outperforms the WTG abstraction in terms of recall, and it also outperforms its own ablations in terms of F1-score.

### 7 RELATED WORK

#### 7.1 Program Analysis for User Interfaces

**GUI analysis for mobile apps.** In the space of GUI analysis tools of mobile apps, the most related one is GATOR [39], which statically analyzes Android applications to build models of their GUI-related behavior. These models include so-called constraint graphs [40] and (more related to this work) window transition graphs [46]. However, as shown in Section 6.6, the models produced by GATOR do not provide sufficient information to check an app against VESPER specifications. Another static analyzer that is related to this work is the BackStage tool [25] for identifying which sensitive API functions can be invoked through which UI elements. BackStage checks for specific unintended behaviors of GUI elements, such as leaking a user’s location when she clicks the “upload picture” button. In contrast to BackStage, VESPER supports a general class of policies expressed in the VESPER policy language and also reasons about spatial properties of GUI elements as well as communication patterns between them.

There are also some GUI-related analysis tools based on dynamic techniques. For instance, Corridroid [29] tests an application against a set of UI constraints given by the user. As another example, GVT [32] dynamically checks whether the user interface of a mobile app is implemented according to its design mock-up by monitoring its visual appearance. Similarly, REMAUI [33] can automatically identify certain types of UI elements (e.g., images and text) using optical character recognition (OCR) and computer vision techniques. Compared to these dynamic techniques, static techniques like VESPER provide complementary advantages such as higher coverage for behaviors that are hard to trigger at run-time.

**GUI analysis for web applications.** Beyond mobile applications, GUI analysis has also attracted some interest in the context of web applications. For example, Cilla [30] finds unused CSS selectors by dynamically monitoring the relationship between CSS rules and webpage elements selected by those rules. Another related work in this space is the Cassius framework [36, 37] for building semantics-aware CSS tools. Specifically, Cassius formalizes the semantics of CSS in first-order logic and can be used to check spatial properties of GUI elements displayed on a webpage. However, since the user interface of web applications is rendered exclusively based on declarative HTML and CSS code, Cassius does not need to analyze JavaScript programs. In contrast, checking an Android application against a VESPER specification requires both precise reasoning about Java code as well as the declarative layout definitions provided in XML files. Besides Cassius, there are other tools specifically built for addressing accessibility problems in web pages [38, 44]. Compared to these tools that are typically based on dynamic testing, VESPER has the potential to cover code that is hard to reach by dynamic analysis.
analysis. Furthermore, accessibility tools can only check spatial properties of GUI elements while VENUS reasons about both spatial and behavioral properties.

7.2 Static Analysis of Android Applications

Due to the popularity and security-critical nature of Android applications, there is a rich literature of program analysis techniques for the Android framework [4, 6, 11, 22, 23, 28, 34, 47]. A key challenge in statically analyzing Android applications is reasoning about dependencies between different components, such as activities and services. Thus, several papers focus on inter-component communication (ICC) analysis for Android [11, 34]. In this work, we leverage the ICC analysis techniques proposed in prior research.

Among techniques for analyzing Android applications, a particularly relevant work is the Apposcopy system for malware detection [11]. Similar to VENUS, Apposcopy provides a specification language for describing semantic behaviors of Android apps and allows statically checking an app against such a specification. However, the specification language of Apposcopy is tailored for spyware detection and does not allow referring to GUI elements. Thus, beyond ICC analysis, the underlying static analyses performed by Apposcopy and VENUS are quite different.

7.3 Android Malware Detection

Since one of the use cases for VENUS is to detect ad fraud, VENUS is also related to a long line of work on Android malware detection [3, 8, 11, 12, 26, 45]. Most malware detection tools in this space focus on information leakage [11, 28], rather than GUI-related behavior and are therefore not suitable for accurately detecting ad fraud applications. As mentioned earlier, the most relevant work in this space is the FraudDroid tool [8] for detecting malware in the ad fraud category. However, unlike VENUS, FraudDroid is based on dynamic analysis, and, as demonstrated in section 6.5, it has significantly worse recall compared to VENUS.

8 CONCLUSION

We introduced a new framework called VENUS for checking conformance between Android apps and GUI policies expressed in a policy language called VESPER. We manually studied GUI policies from multiple different sources, and, among English policies that are precise enough to be formalized, we showed that around 70% are expressible in the VESPER policy language. We used VENUS to check conformance between these policies and over 2000 Android applications and showed that VENUS can uncover previously unknown ad fraud instances as well as violations of GDPR regulations. Our comparison against VirusTotal and FraudDroid indicates that VENUS advances the state-of-the-art in ad fraud detection in terms of both precision and recall. Finally, our comparison against GATOR as well as the two ablation studies highlight the benefits of our proposed ELF abstraction.

ACKNOWLEDGEMENTS

This work was partially supported by NSF Grants #1908494, #1908304, CCF-#2005889, and CNS-#1822251 as well as a Google Faculty Research Award.

REFERENCES
