Automated, Cost-effective, and Update-driven App Testing

CHANH DUC NGO, SnT Centre, University of Luxembourg, Luxembourg
FABRIZIO PASTORE, SnT Centre, University of Luxembourg, Luxembourg
LIONEL BRIAND, SnT Centre, University of Luxembourg, Luxembourg and School of EECS, University of Ottawa, Canada

Apps' pervasive role in our society led to the definition of test automation approaches to ensure their dependability. However, state-of-the-art approaches tend to generate large numbers of test inputs and are unlikely to achieve more than 50% method coverage.

In this paper, we propose a strategy to achieve significantly higher coverage of the code affected by updates with a much smaller number of test inputs, thus alleviating the test oracle problem.

More specifically, we present ATUA, a model-based approach that synthesizes App models with static analysis, integrates a dynamically-refined state abstraction function and combines complementary testing strategies, including (1) coverage of the model structure, (2) coverage of the App code, (3) random exploration, and (4) coverage of dependencies identified through information retrieval. Its model-based strategy enables ATUA to generate a small set of inputs that exercise only the code affected by the updates. In turn, this makes common test oracle solutions more cost-effective as they tend to involve human effort.

A large empirical evaluation, conducted with 72 App versions belonging to nine popular Android Apps, has shown that ATUA is more effective and less effort intensive than state-of-the-art approaches when testing App updates.

CCS Concepts: • **Software and its engineering** → **Software verification and validation**.

Additional Key Words and Phrases: Android Testing, Regression Testing, Upgrade Testing, Model-based Testing, Information Retrieval

1 INTRODUCTION

The business-critical role played by software applications for mobile devices (Apps) in our society [13] has led to the development of dedicated techniques for their automated testing [31]. Since most of the code in an App concerns the handling of input values and events, test automation approaches automatically generate sequences of events and input values (hereafter, input sequences) that simulate the use of the App under test in its deployed environment. These approaches mainly differ with respect to the strategy used to create input sequences, such as random, evolutionary, and model-based approaches relying either on static or dynamic information [31].

Unfortunately, state-of-the-art automated App testing techniques show limited code coverage capabilities, thus indicating they are unlikely to exercise all the features of the App under test. For example, they typically exercise about half of the methods implemented by commercial apps [56]. As a result, all methods and instructions that are not automatically tested should be exercised by manually implemented test cases, an expensive task that may delay the App release. Also, though existing techniques show a degree of complementarity [56], state-of-the-art approaches do not attempt to integrate them to achieve better results.

Existing testing approaches do not account for the high release frequency of a typical App's lifecycle, which are usually driven by marketing strategies aiming at increasing visibility [9, 17, 34]. As a result, existing work does not include effective means of prioritizing the testing of modified or newly introduced features and are thus not addressing one of the major needs of App developers.

Authors' addresses: Chanh Duc Ngo, SnT Centre, University of Luxembourg, JFK 29, Luxembourg, Luxembourg, chanh-duc.ngo@uni.lu; Fabrizio Pastore, SnT Centre, University of Luxembourg, JFK 29, Luxembourg, Luxembourg, fabrizio.pastore@uni.lu; Lionel Briand, SnT Centre, University of Luxembourg, JFK 29, Luxembourg, Luxembourg, School of EECS, University of Ottawa, Ottawa, Canada, lionel.briand@uni.lu.

However, this is an important requirement for any testing strategy as exercising all the features of an App in each release is enormously wasteful. Existing work on testing App upgrades is limited to the selection of subsets of events that may trigger modified code [49] or the selection of regression test cases [11]. This is, however, not adequate when, to start with, available test cases do not exercise all the new and modified features of the software.

Finally, the current body of work does not address the *oracle problem* [7, 31, 45]. More precisely, testing techniques cannot discover functional failures beyond crashes and the manual verification of the App outputs is difficult due the large number of inputs they exercise [11]. However, in the context of frequent App updates, with a test input generation strategy that effectively exercises updated features, it is conceivable to address the oracle problem by relying on dedicated strategies to minimize test inputs. Failures affecting unchanged features (i.e., regressions) can be automatically detected by comparing the output of different App versions for a same input [19, 27, 40, 48], whereas the output of new and modified features can instead be verified, at reduced costs, by relying on internal or external crowdsourcing [38]. Nevertheless, such solutions are only practical if the number of test inputs is kept down to a reasonable number.

Keeping the number of test inputs to the strict minimum is important to minimize manual intervention since it may be required when executing the same inputs on different software versions, e.g., to adapt input sequences to changes in the GUI [30, 36]. Further, screenshots of the results must be visualized after every input. Unfortunately, state-of-the-art App testing approaches generate large test suites, while test suite reduction approaches require to perform runtime monitoring of the App, which slows down execution and diminishes test automation effectiveness [11].

In summary, to address the limitations above, we aim to achieve the following two objectives: (O1) maximize the number of updated methods and their instructions that are automatically exercised by testing throughout the App lifecycle, within practical test execution time, and (O2) generate a significantly reduced set of inputs, compared to state-of-art approaches, thus decreasing manual effort.

To achieve the two objectives above, it is necessary to integrate multiple analysis strategies. Objective O1 can be effectively achieved by means of static analysis, to determine updated features (e.g., through the identification of updated methods[49]) and identify the inputs that may trigger a specific feature (e.g., the input that leads to a particular Window) [61]. Unfortunately, static analysis alone may not enable the effective testing of Apps; indeed, they typically rely on APIs dedicated to input handling that are hardly processed by static analysis tools, as discussed in related work [45]. Random exploration is thus required to discover, at runtime, inputs that may trigger a potentially large subset of modified methods. Unfortunately, random exploration might be particularly inefficient and conflict with objective O2 (e.g., it may require thousands of inputs to exercise features that depend on specific App states). For this reason it is necessary to determine which inputs bring the App into distinct program states by relying on dynamically-refined state abstraction functions [21] and by identifying dependencies among App features (e.g., to determine that an option in the settings page enables a specific feature).

In this paper we present *ATUA*[1] *(Automated Testing of Updates for Apps)*, the first approach that integrates multiple test strategies to efficiently use the test budget throughout the App lifecycle. More precisely, ATUA implements a model-based approach that integrates a *dynamically-refined state abstraction function* and complementary testing strategies, including (1) coverage of the *model structure*, (2) coverage of the *App code*, (3) *random* exploration, and (4) coverage of *dependencies* among App windows.

[1] Atua is also the name of spirits of the Polynesian peoples, https://maoridictionary.co.nz/word/494

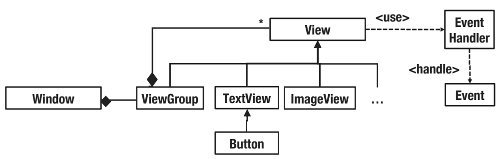


Fig. 1. Overview of the Android Apps's GUI architectural components.

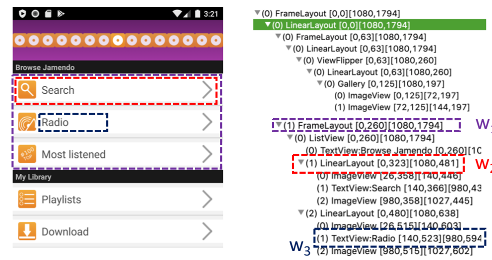


Fig. 2. Example of a GUITree. We use colored dottext boxes to match widgets in the tree to the pixels in the screen.

ATUA generates models of the App under test by combining static and dynamic program analysis. It extends static program analysis approaches [61] to automatically generate extended window transition graphs (EWTG), i.e., finite state machines that capture which inputs trigger window transitions and updated methods. Also, it introduces a state abstraction function that refines the states of the EWTGs to capture differences in the user interface that are not detected by means of static analysis (e.g., the presence of dynamically disabled buttons). The state abstraction function is automatically refined to eliminate or, when not possible, reduce non-determinism while minimizing the number of abstract states.

To automatically exercise Apps, ATUA relies on the generated EWTGs to identify the sequences of inputs that trigger the execution of updated methods. When there are discrepancies between the EWTGs and the observed behavior, random exploration is used to refine the former. Code coverage is used to identify the methods that require additional testing effort. Finally, using information retrieval techniques [53], ATUA identifies dependencies between App windows that may prevent the execution of certain methods.

We assume that, for every software version, engineers are interested in testing the updated methods only. However, the general principles behind ATUA can easily be adopted also for other ways of characterising change, e.g., based on impact analysis [46]. Indeed, other criteria for selecting target methods are straightforward to integrate into ATUA.

An empirical evaluation conducted with nine popular, commercial Apps shows that, compared to state-of-the-art approaches (i.e., DM2 [8], APE [21], and Monkey [4]), ATUA leads to reduced test costs. Indeed, it generates less than 70%, 4%, and 2% of the inputs generated by DM2, APE, and Monkey, respectively. By automatically exercising, on average, 2.6 instructions for every generated test inputs, ATUA is the most cost-effective approach. Further, on average, ATUA, for a same test execution budget (e.g., 1 hour test execution time), can automatically exercise 6% more updated methods and instructions than the second best state-of-the-art approach.

The paper is structured as follows. Section 2 introduces background technologies. Section 3 provides the technical details of the proposed approach. Section 4 reports on the results of our empirical evaluation. Section 5 discusses related work. Section 6 concludes the paper.

2 BACKGROUND

2.1 App Design and Architecture

In this paper, we target Android Apps since Android is the most adopted platform in practice and is widely investigated in research [17]. However, most of the solutions proposed here should be easily tailorable to other platforms.

When an App is running, the end-user interacts with the active *Window* of the App (i.e., the Window being rendered on the screen). Windows consist of a hierarchy tree of widgets (called GUITree). A widget extends the class *View*. Figure 1 shows a portion of the hierarchy tree of the class *View*. Figure 2 shows a portion of the GUITree for a window of Jamendo, a music streaming App [26].

Each widget has a set of properties associated to it. Widgets can be associated to *EventHandlers* that are invoked by the OS when specific *InputEvents* are triggered by the end-user. Typical InputEvents include click, long click, swipe, and keypress.

In Android, the application logic is typically implemented by Activity classes that are instantiated by the framework and act as controllers of the Model-View-Controller design pattern [41]. Inter process communication, instead, is managed by the *Intent* resolution mechanism [2]. More precisely, in Android, a system event (e.g., indicating a battery being low) or a message exchanged between apps (e.g, a URL sent by the browser to a music player App) is referred to as an *Intent*. To handle Intents, an App declares in its XML configuration file an Activity that the OS will instantiate and execute when a specific Intent type should be received by the App.

2.2 App Testing Automation

System-level testing of an App through its GUI (i.e., GUI testing) is performed through sequences of *test inputs* that can be either Events or Intents. Functional GUI testing aims at exercising (i.e., render active) all declared Windows, trigger all event handlers, and cover all the code of the App under test. App testing automation aims to generate input sequences to achieve these objectives at the lowest cost possible.

For a complete overview of App testing automation approaches, the reader is referred to recent surveys [31, 54]. *Model-based* solutions are the most commonly ones reported in the literature [54]. In model-based testing approaches, the model used to drive testing is typically a finite state machine (FSM). It can be formally described as a tuple $(S,A,T,\mathcal{L})$ [21], where

- S is a set of states.
- A is a set of actions.
- T is a set of state transitions. Each transition has a source state and a target state. It is triggered by an action $\alpha \epsilon A$.
- $\mathcal{L}$ is an abstraction function, which might be used to: (1) assign a Window to a state and (2) match an InputEvent or an Intent to an action.

Model-based approaches differ regarding the type of analysis adopted to identify states and transitions (i.e., dynamic [8, 52] or static [61, 62]), the abstraction functions used (i.e., predefined [8, 52] or adaptable [21]), and the model exploration strategies they rely on (i.e., offline [52] or online [8]). Adaptable state abstraction functions have been shown to lead to more effective App testing [21] but they have never been used in testing frameworks that enable the effective combination of static and dynamic analysis. Further, existing model-based approaches do not prioritize the testing of updated methods, which is our objective here, and thus we require dedicated input generation algorithms. To leverage the benefits of static and dynamic program analysis, ATUA integrates (1) Gator, a tool that statically identifies states and transitions, (2) DroidMate2 (DM2), a model-based framework that performs online testing, dynamic identification of states and transitions, and enable the combination of static and dynamic analysis, (3) a dedicated algorithm for test input generation, and (4) a custom and adaptable state abstraction function.

DM2 [8] consists of two main engines for exploration and automation, respectively. The former drives the interaction with the App and derives the model of the App under test. The exploration is driven by a set of user-defined strategies. The latter translates actions into concrete commands on the device. In *DM2*, an App state is univocally identified by the set of UI elements in the user interface and their state-related properties. UI elements are identified by means of their descriptive ID or their image bytes cut from a screenshot. State-related properties are given by the position of the elements and other widget specific characteristics. ATUA relies on the *DM2* automation engine, which provides an API to send inputs to the App under test and retrieve information about the displayed UI elements. In ATUA, App exploration is driven by a custom algorithm (see Section 3.5) that relies on statically derived models and a custom state abstraction function.

Gator [61, 62] is a static analysis tool that creates models of the App under test in the form of window transition graphs (WTGs). WTGs are FSMs representing the possible window sequences and their associated events and callbacks. Though WTGs can be used to enable model-based testing, they have been mostly used to detect resources leaks [59].

3 PROPOSED APPROACH: ATUA

Within an updated App, we can distinguish among unchanged features (i.e, features present in previous versions of the App under test whose functional requirements did not change), modified features (i.e, features present in previous versions of the App under test whose functional requirements did change), repaired features (i.e, features present in previous versions of the App under test whose implementation did not match its functional requirements), and new features (i.e., features introduced in the App under test).

In our work, we aim to automatically exercise *updated features*, that is new, modified, and repaired features. Since App features are typically implemented in the methods of the App code, we focus on features that are implemented either by introducing new methods or by modifying existing methods [2]. In this paper, we use the term *updated methods* to refer to both new and modified methods, which are our test targets.

The testing activity performed by ATUA is driven by an App model that is initially created by static program analysis procedures implemented by ATUA and then refined during testing. The App model is used to drive testing with the objective of exercising a set of test targets (i.e., updated methods). When engineers aim to exercise not only the updated methods but all the methods identified by change impact analysis techniques [1, 29], test targets may be extended to include those methods.

The App model metamodel is shown in Figure 4 and described in Section 3.1. It consists of three parts: (1) an Extended Window Transition Graph (hereafter, EWTG), (2) a Dynamic State Transition Graph (hereafter, DSTG), and (3) a GUI State Transition Graph (hereafter, GSTG). The three graphs are finite state machines capturing how input values trigger changes in the state of the App under test. The *EWTG* models the sequences of windows being visualized after triggering specific inputs (Events or Intents). For every input, the EWTG keeps trace of the name of the handlers associated to the input and the list of test targets that may be invoked during the execution of the input handler. The *GSTG* is a fine-grained model that captures every visual change in the GUI (e.g., the color of a button) that might be triggered by an action performed on the GUI. An action is an instance of an input (e.g., click on a specific Button widget). The visual changes in the GUI caused by an action cannot be determined by means of static analysis. Finally, the *DSTG* models the abstract states of the visualized Windows and the state transitions triggered by events. Abstract states are identified by a state abstraction function to eliminate possible non-determinism. The DSTG plays a critical role to optimize the test budget and identify a reduced set of input events; indeed, it helps

[2]Based on related work, 81% of the updates concern Java files, while only the remaining 19% concerns manifest files (e.g. permissions) or layout declarations in XML files [49].

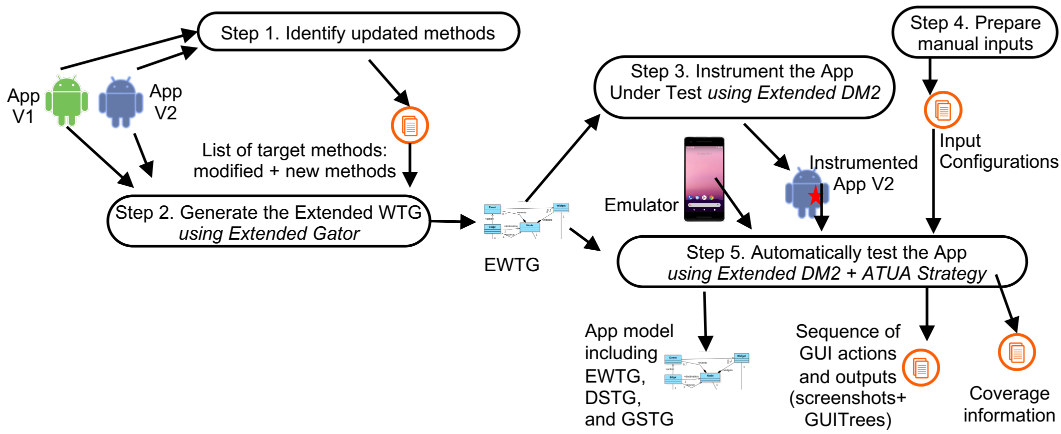


Fig. 3. Overview of the ATUA process to test App updates.

determine a correct and reduced sequence of events necessary to reach a specific Window from another one.

Figure 3 provides an overview of the process implemented by ATUA to test App updates. In Step 1, ATUA compares the previous (App V1 in Figure 3) and the updated (App V2) version of the App under test to identify the updated methods. In Step 2, ATUA relies on Soot [44], a static analysis framework, and an extended version of Gator, to generate the EWTG. In Step 3, ATUA relies on DM2 to generate a version of the App that is instrumented to trace code coverage. In Step 4, engineers manually specify test inputs that are unlikely to be generated automatically (e.g., the login credentials for Apps that require a user to be registered on a remote platform). In Step 5, ATUA relies on an extended version of DM2 that integrates the ATUA test algorithm, to exercise the App under test on the Android emulator. During testing, ATUA refines the App model and relies on it to identify the actions to perform on the GUI. For example, ATUA uses the App model to identify the action that, in the current window, may lead to the execution of a test target.

The main output of Step 5 is the sequence of GUI actions performed during testing and the outputs (i.e., the screenshot of the active Window and the corresponding GUITree) generated by the App under test after every action. This sequence is used by engineers to verify if the behavior of the App is as expected (test oracle). As mentioned earlier, to verify App results, engineers can rely on two complementary state-of-the-art approaches, not addressed by ATUA, that respectively target regression failures in unchanged features and failures in newly implemented, repaired, and modified features. To discover regression failures, engineers can replicate, on a previous App version, the test input sequences generated for the updated App and automatically compare the generated outputs. Differences in the outputs generated by the two versions should indicate the presence of a regression fault. To discover failures in new and repaired features, engineers can visualize the GUITrees or the screenshots of the active Window rendered after each Action. The visual inspection of the App outputs enables an engineer to determine the presence of functional failures, based on expected behavior, whether specifications are implicit or documented. In Section 4, we discuss to what extent ATUA reduces the cost associated to the manual activities entailed by the test oracle strategies above, with respect to other state-of-the-art test automation solutions.

In addition, ATUA provides, as output of Step 5, an App model including the EWTG, the DSTG, and the GSTG. The App model is generated and continuously refined during testing. Further, it reports coverage information, i.e., the sets of updated methods and instructions belonging to updated methods that have been exercised during testing.

In the following, we provide additional details about the App model metamodel (Section 3.1) and describe Steps 1 to 5 (Sections 3.2 to 3.5), except for Step 3 which is already automated by DM2.

3.1 App Model Metamodel

Figure 4 shows the ATUA metamodel as a UML class diagram. It relies on the package notation to group the classes belonging to a specific part, i.e., EWTG, DSTG, and GSTG. Figure 5 shows an example App model built when testing Jamendo.

The EWTG is consistent with the WTG generated by Gator. Each transition is triggered by an *Input*, either an *InputEvent* or an *Intent*. An InputEvent is associated to the Widget that declares its EventHandler. If the EventHandler is not declared by a Widget (e.g., for the event *PressHome*), the InputEvent is not associated to any target Widget. Each Widget belongs to one Window.

In addition to the concepts captured by the WTG, the EWTG generated by ATUA also captures the list of modified methods that can be triggered by the Input (i.e., the attribute *targetMethods* of class *Input*), which are used to drive testing. A Transition triggered by Inputs with associated *targetMethods* is a *target Transition*. Similarly, a Window that is the source for at least one target Transition is a *target Window*. The EWTG also captures the dialogs and menus that can be opened by an Activity (e.g., association *triggeredDialogs*). Finally, it also models the HiddenHandlers of a Window, which are introduced in Section 3.3.

The GSTG captures the same information provided by the models generated by DM2. The state of a Window is captured by its GUITree, which is a composition of Widgets. For each Widget, we record the values associated to its properties and derive the Widget hash (to associate an ID to the current state of the Widget) according to the DM2 strategy. The hash of the GUITree is then derived from the hash of its Widgets. In addition, the GSTG also captures the name of the Activity currently running (see attribute *currentActivity*), which we derive, at runtime, from logcat [3]. The transition between GUITrees is triggered by an Action, which can be handled either by the Widget (WidgetAction) or by the visualized Window. The enumerations WidgetActionType and WindowActionType list the type of actions that can be performed to trigger GUITreeTransitions. Actions might have additional information (i.e., actionData) associated to them; for example, the text provided to the App under test by a TextInput action or the start and end coordinates of a Swipe action.

The DSTG provides abstract states that group together GUITrees in which a same Action triggers a same App behaviour (e.g., leads to a same abstract state). The DSTG enables ATUA to efficiently test the App under test by determining the shortest sequence of Actions that reaches a target Window. Also, abstract states capture the conditions under which a specific Action can trigger a modified method, for a certain Window. Abstract states are thus a mean to minimize the number of Actions generated by the test automation approach.

Each AbstractState consists of a number of AttributeValuationMaps, each one abstracting the state of the widgets belonging to the GUITree that have the same set of attribute valuations. An AttributeValuationMap is a map of pairs ⟨attribute,value⟩. The enumeration Attribute in Figure 4 provides the list of attributes appearing in the AttributeValuationMap. The AttributeValuationMap has a cardinality attribute, which indicates how many widgets have the same attribute valuations. For example, in Figure 5, the AbstractState *as4* includes many LinearLayout widgets (*LL* in the Figure, cardinality *), one for each item in the displayed list.

Since the DSTG is used to drive testing, i.e., to select the Actions to be triggered at runtime, the AbstractState captures only the attributes of widgets that are *interactive*. A widget is *interactive*

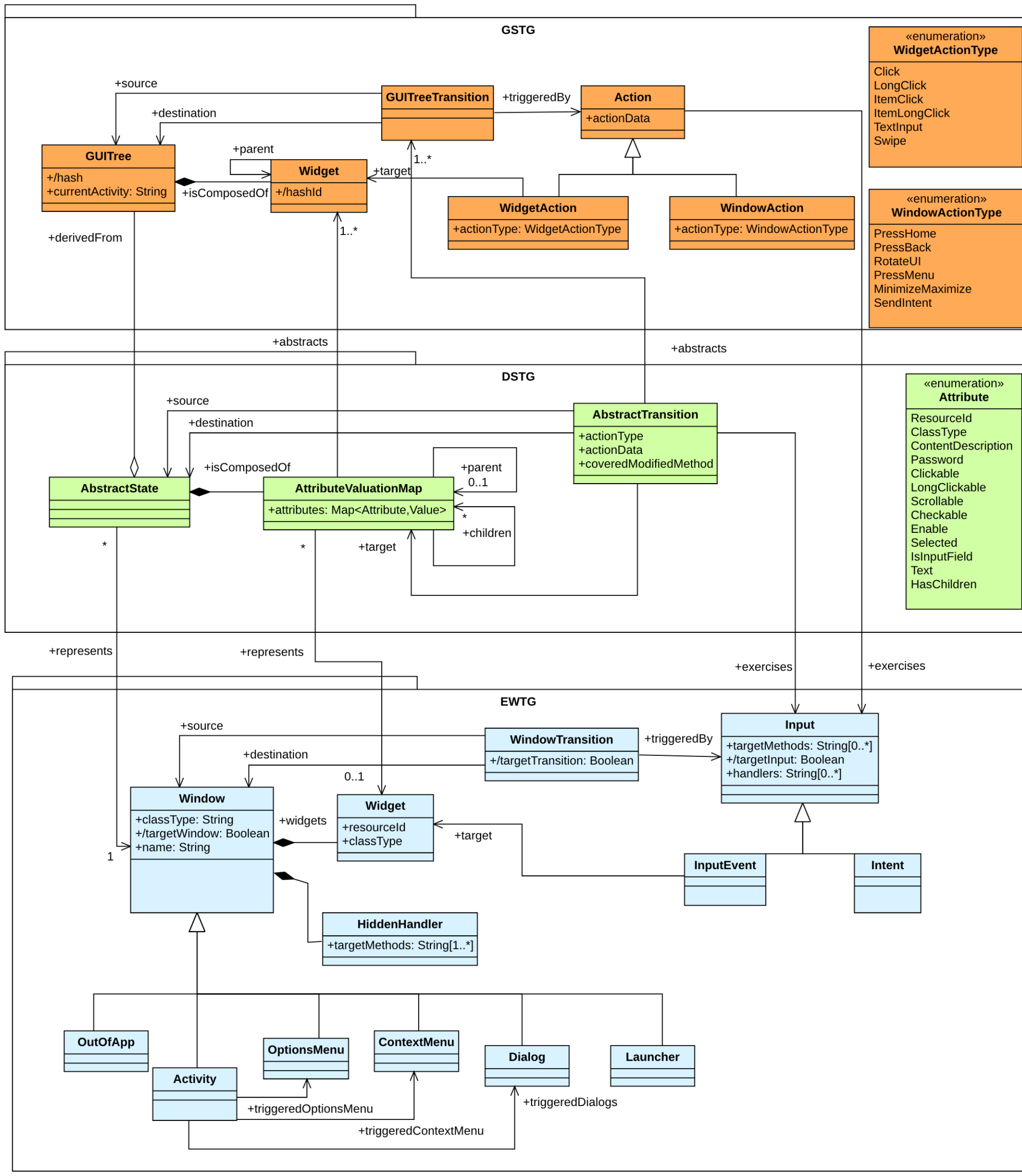
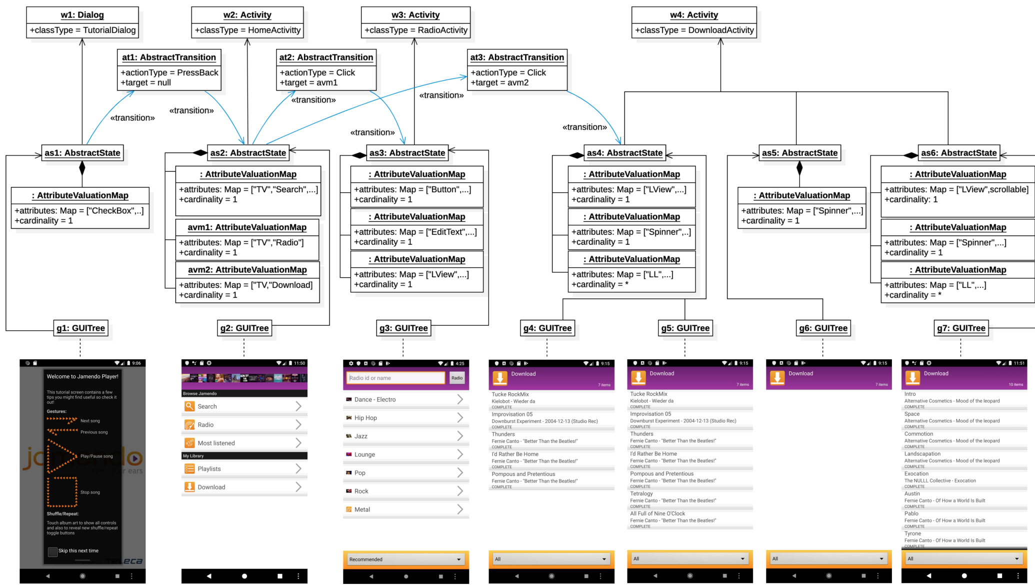


Fig. 4. ATUA Metamodel. Package symbols and colors are used to group classes belonging to a specific metamodel component: EWTG, DSTG, GSTG.

when it is enabled, visible, and is an instance of a class that can be the target of any action of type WidgetActionType (See Figure 4).

At runtime, during testing, ATUA identifies AbstractStates through a dedicated *abstraction function* ($\mathcal{L}$). ATUA automatically defines a distinct $\mathcal{L}$ for each Window of the App under test. $\mathcal{L}$ relies on a predefined set of *reducers*, i.e., functions that extract the value of an abstract property of a widget [21]. Table 1 shows the list of reducers implemented by ATUA. Two AbstractStates differ



Legend: Straight, black arrows capture associations. Curved, blue arrows indicate the source of an AbstractTransition (when the receiving end is an AbstractTransition) and its destination (when the starting end is an AbstractTransition). To avoid cluttering, the target of an AbstractTransition is indicated using the AbstractTransition attribute named *target*. For illustration purposes, we show screenshots instead of GUITree hashes. We use the following abbreviations: LL (LinearLayout), LView (ListView), TV (TextView).

Fig. 5. Example of an App model, represented as a UML object diagram.

when at least one value differs across their respective AttributeValuationMaps, or when they have a different cardinality. For example, in Figure 5, the AbstractStates *as4* and *as5* differ because *as5* does not contain the AttributeValuationMaps for both the ListView containing the downloaded items and the LinearLayouts of each item (the list is empty). The AbstractStates *as4* and *as6* are different because in *as6* the ListView becomes scrollable.

In the DSTG, state transitions are captured by AbstractTransitions. An AbstractTransition is univocally identified by the *actionType*, its *source*, its *target*, its *destination*, and its *actionData*. The *actionType* matches one of the items belonging to the enumerations WidgetActionType or WindowActionType. The *actionData* is captured only for two actionTypes (i.e., Swipe and Intent) that usually lead to distinct AbstractStates depending on their action data. In the case of Swipe, the actionData indicates the direction of the Swipe action (i.e., Up, Down, Left, Right). For Intents, since they are provided as manual inputs by the engineers and, to minimize manual effort, we expect engineers to provide one manual input for each possible Intent type (e.g., one different URL for each of the music file types supported by Jamendo), the actionData matches the Intent input text. We leave to future work the definition of functions that provide an abstract representation for the data associated to other types of actions (e.g., to distinguish between numeric, alphabetic, or non-alphanumeric data provided to TextInputs).

A DSTG may include non-deterministic AbstractTransitions, i.e., transitions with the same actionType, departing from the same AbstractState, but reaching different AbstractStates. To effectively test the App, ATUA minimizes the number of non-deterministic Actions by refining $\mathcal{L}$, which is achieved by accounting for more reducers to distinguish states. Indeed, non-deterministic

	Reducer	Description
1	R_{RID}	Resource ID.
2	R_{CN}	Class name.
3	R_{CD}	Value of *Content description.*
4	R_P	Value of *Password.*
5	R_C	Value of *Clickable.*
6	R_{LC}	Value of *Long Clickable.*
7	R_S	Value of *Scrollable.*
8	R_{Ch}	Value of *Checked.*
9	R_E	Value of *Enabled.*
10	R_S	Value of *Selected.*
11	R_I	True if it is an input field.
12	R_T	Value of *Text.*
13	R_{HC}	True if the widget contains one or more children.

Table 1. ATUA reducers. We indicate the value of the *property* reported by each.

Level	Reducers applied to interactable Widget	Reducers applied to interactable Widget Children
L1	R_{RID},R_{CN},R_{CD},R_{Ch},R_E, R_P,R_S,R_I,R_C,R_{LC},R_S	
L2	R_{RID},R_{CN},R_{CD},R_{Ch},R_E, R_P,R_S,R_I,R_C,R_{LC},R_S, R_T	
L3	R_{RID},R_{CN},R_{CD},R_{Ch},R_E, R_P,R_S,R_I,R_C,R_{LC},R_S,R_T, R_{HC}	
L4	R_{RID},R_{CN},R_{CD},R_{Ch},R_E, R_P,R_S,R_I,R_C,R_{LC},R_S,R_T	R_{RID},R_{CN},R_{CD},R_{Ch},R_E, R_P,R_S,R_I,R_C,R_{LC},R_S
L5	R_{RID},R_{CN},R_{CD},R_{Ch},R_E, R_P,R_S,R_I,R_C,R_{LC},R_S,R_T	R_{RID},R_{CN},R_{CD},R_{Ch},R_E, R_P,R_S,R_I,R_C,R_{LC},R_S, R_T

Table 2. Refinement of ATUA state abstraction function. In blue we show the reducers introduced in finer granularity level.

Actions indicate that the model does not capture all the distinct states of the system. Consequently, this may prevent us from finding the correct sequence of Inputs necessary to reach the states in which the target methods could be triggered.

ATUA detects non-determinism at runtime, during testing, when an Action does not bring the App into the expected AbstractState. When this happens, ATUA refines $\mathcal{L}$ for the AbstractState in which the action had been triggered. It does so according to five levels of granularity, which are captured in Table 2. With level L1, $\mathcal{L}$ distinguishes states based on static information about the widgets (i.e., resource ID and class) and information about how they can be interacted with (i.e., reducers appearing in rows 3 to 12 of Table 1). With level L2, in addition to the information accounted for in L1, $\mathcal{L}$ includes the text associated to the widget, which often affects the behaviour of an app (e.g., invalid characters in a textbox may prevent a state transition). With level L3, $\mathcal{L}$ also reports the number of children of a widget (i.e., R_{HC}). L3 is useful because the interactive widgets captured by $\mathcal{L}$ may include non-interactive children whose state is not captured by $\mathcal{L}$ but may characterize the current state (e.g, through descriptive labels). With levels L4 and L5, $\mathcal{L}$ captures, for every interactive widget, the same information as L2 and, in addition, for every child, the information captured by levels L1 and L2, respectively.

Figure 6 shows the result of the refinement of $\mathcal{L}$ for the abstractState *as2*. By applying $\mathcal{L}$ with level L1, all the clickable elements on the Window, which are LinearLayouts with the same properties, except for the text, resulted into a same AttributeValuationMap with cardinality *. At runtime, ATUA detects non-determinism; indeed, a click on this AttributeValuationMap may lead to two different AbstractStates: *as3* and *as4*. The refinement of $\mathcal{L}$, which leads to level L2, allows ATUA to distinguish all the different clickable elements since the text property is included into the AttributeValuationMap, thus eliminating non-determinism.

Concerning the relationships among the classes of the ATUA metamodel that are used to drive testing (see Section 3.5), we note that a Window of the EWTG can have one or more AbstractStates. Each AbstractState is associated to one or more concrete GUITrees. Multiple Actions performed during testing are associated to a same AbstractTransition. Each AbstractTransition in the DSTG uses one Input. Similarly, each Action in the GSTG uses one Input.

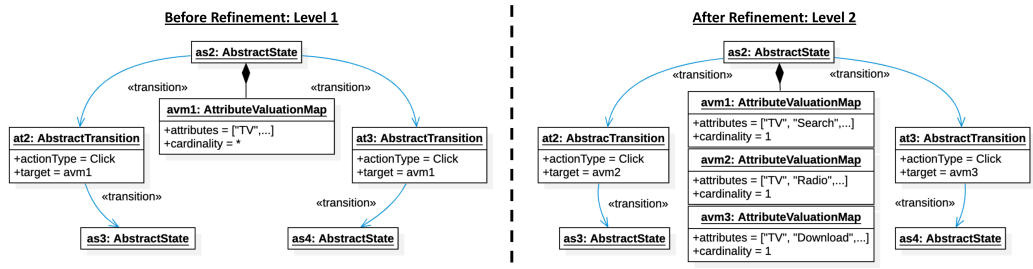


Fig. 6. Example of refinement of $\mathcal{L}$. For the notation used, see the Legend of Figure 5.

3.2 Step 1. Identify Updated Methods

In the first step, ATUA identifies methods that have been modified or introduced by the new version of the App (i.e., V2 in Figure 3). This task might be accomplished through source code comparison across versions [39]. However, to enable experiments with commercial Apps, we developed a toolset (hereafter, Appdiff) that compares compiled Android Apps.

Appdiff is an extension of *LibScout* [14], a light-weight static analysis tool for Android. It first generates a hashtree over the bytecode for each App. The hashtree is a three-layered Merkle tree in which parent hashes are generated from their child nodes. The three layers model the flattened package structure that is preserved in the compiled code, i.e., packages, classes, and methods.

The tree is built bottom up starting with the method hashes. A method hash is computed over the method signature and the opcodes in bytecode instructions. To identify code-level changes across App versions, we additionally store package, class, and method names along with the hashes. To efficiently check for differences, two hash trees are matched top-down starting with the package hashes. Methods that share the same name but have a different hash have been modified. New methods appear only in the most recent version.

3.3 Step 2. Generate the Extended WTG

ATUA generates the Extended WTG by means of static program analysis; more precisely, by performing, on the updated App, the analysis implemented by an extended version of Gator and Soot.

The original version of Gator works by processing Android bytecode and XML layout files. For the analysis of bytecode, Gator relies on Soot. Bytecode analysis is used to identify the types of Window (i.e., Activity, Dialog, OptionsMenu, and ContextMenu) that are programmatically specified in the App. Bytecode analysis is also used to identify the widgets that compose a window and the associated event handlers. Gator identifies widgets that extend the class *android.view.View* and its handlers. Event handlers' code is processed to determine window transitions. XML layout files are processed to identify additional event handlers.

Our extensions to Gator address some known limitations [28]. More precisely, we support the identification of window transitions triggered by Fragments and RecyclerView, which are widget containers that are not identified by Gator as such. Our extensions associate the contained widgets to the window that declares either the Fragment or the RecyclerView.

We rely on Soot to traverse the backward call graph of every updated method. During the traversal, when we encounter a method that has been identified by Gator as an event handler, we update the EWTG to trace the fact that the handler can exercise the updated method. Also, we rely on Soot to extract string literals to be used for testing (see Section 3.5).

```
1 {
2   "BookInsertionAndSearch" : {          //input pattern
3     "Windows" : [   "ACT[bookcatalogue.EditAuthorList]1741", "ACT[bookcatalogue.BookEdit]1802"
        "ACT[bookcatalogue.BookISBNSearch]1843" ],
4     "DataFields" : {
5       "isbn" : {
6         "resourceIdPatterns" : [ "isbn_txt" ]
7       },
8       "title" : {
9         "resourceIdPatterns" : [ "title_txt" ]
10      },
11 ...
12    },
13    "Instances" : [
14      {
15        "isbn" : "0387284540",
16        "title" : "Applied probability and statistics",
17        "publisher" : "Springer",
18        "pages" : "350",
19        "list_prices" : "69",
20        "format" : "Hard Cover",
21        "genre" : "Unfiction",
22        "language" : "English"
23      }
```

Fig. 7. Manual definition of inputs

Finally, we determine if Gator does not identify some of the event handlers of the App, which is a common problem of static analysis tools for Apps. Indeed, these static analysis tools rely on hardcoded procedures for the identification of event handlers (e.g., they look for specific method names [44]); since OS APIs are under continuous evolution, it is unlikely that static analysis tools will ever be able to identify all the event handlers of an App. To address this problem, we introduced into ATUA three solutions, one based on static analysis (described in the next paragraph) and two based on dynamic program analysis (described in Sections 3.5.1 and 3.5.3).

To identify missing event handlers using static analysis, we rely on the observation that if an event handler is not detected by Gator, the backward traversal of the call graph performed by ATUA will not reach any event handler but will terminate in a method that (i) belongs to a Window class and (ii) is not invoked by any other method of the App under test. Such methods are likely event handlers invoked at runtime by the Android APIs. We refer to them as *hidden-handlers*. We keep track of all the hidden-handlers encountered during the analysis along with the list of updated methods reachable from them. We rely on this information during Step 5.

3.4 Step 4. Prepare manual inputs

Certain input values are unlikely to be automatically generated. Examples include login credentials, files of a specific type, and data to be received by the App under test through the Android Intent mechanism. To handle these cases, we enable engineers to specify, in a JSON file, the input values to be used in specific Windows. An example of our input definition format is provided in Figure 7.

According to our format, engineers can specify one or more input insertion patterns (e.g., *BookInsertionAndSearch* in Figure 7, Line 2). For each pattern, they specify the Windows in which the pattern should be used (Line 3), and the widgets which should be used to provide the input data specified (i.e., *DataFields* field in Line 4). Each widget is identified by a name (e.g., *isbn* in Line 5) and a regular expression that enables its selection in the GUITree, based on its name (e.g., *isbn_txt* in Line 6). Finally, multiple input instances (e.g., book names in this case) can be specified (see field *Instances* in line 13).

3.5 Step 5. Automatically test the App

ATUA automatically tests the updated App by triggering, at runtime, the Actions required to exercise target Transitions. When testing starts, the App model consists of an instance of the EWTG for the App under test. GSTG and DSTG are dynamically constructed and extended at runtime by ATUA.

The test execution process includes three distinct phases, each one relying on a different strategy for the generation of Actions. In *Phase 1*, ATUA triggers one Action for every target Input. The goal of Phase 1 is to handle the simplest scenario, i.e., exercise instructions that are executed every time data provided through a target Input is processed. In *Phase 2*, ATUA exercises target Windows with multiple, diverse sets of Inputs. The goal of Phase 2 is to exercise those instructions that are executed only when specific constraints on input values provided in a Window are satisfied. In *Phase 3*, ATUA exercises both target Windows and Windows they depend on. The goal of Phase 3 is to exercise those instructions that can be executed only when certain constraints on the input values provided in related Windows (e.g., preferences Windows) are satisfied.

3.5.1 Detection of the active Window.

A building block of our test automation strategy is the detection of the active Window. This is achieved in two steps. First ATUA looks for a Window in the EWTG whose name matches the name of the current activity in the GSTG (it is provided by DM2). By construction, this name matches the name of either an Activity or a Launcher. It cannot match the other classes of the EWTG Window hierarchy (e.g., Dialogs) because they are always part of an Activity. If the name does not match an Activity or a Launcher in the EWTG for the AUT, then the current Window belongs to another App. In this case, ATUA updates the EWTG to include the new Window (i.e., to remember that an Input may trigger the rendering of a Window of another App).

When the current Window belongs to the App under test, ATUA should determine if a pop-up is open, i.e., if a Dialog, an OptionsMenu, or a ContextMenu is active on top of the active Window. Neither Android nor DM2 provides such information. To determine if a pop-up is open, ATUA relies on the dimensions of the active Window on the screen. Indeed, if an Activity is displayed, its dimensions should match the dimensions of the screen. Otherwise, the currently displayed Window is either a Dialog, an OptionsMenu, or a ContextMenu. To determine which Dialog or Menu is open, ATUA identifies the Window of the EWTG with the highest portion of Widgets visualized on the screen. Indeed, if a Dialog is open, the screen will show more Widgets belonging to the Dialog than Widgets belonging to the underlying Activity. ATUA computes the ratio, R_w, of Widgets belonging to Window w that appear in the displayed GUITree. More precisely, ATUA computes R_w for all the Dialogs and Menus that can be triggered by the current Window. The active Window is the one with the highest values for R_w. To avoid computing R_w every time a same GUITree is visualized, ATUA keeps a mapping between GUITree hashes and the active Windows identified.

If, due to the limitations described in Section 3.3, static analysis does not detect that, for a certain Activity a, there is an event handler that will pop-up a certain Dialog or Menu p, the EWTG will not include an association between a and p. Consequently, at runtime, ATUA may not be able to find a Dialog or Menu to be matched with the displayed GUITree. More precisely, ATUA may observe that (1) the current Window has no Dialogs and Menus associated to it in the EWTG or (2) the score computed for every Dialog and Menu of the current Window is zero. To overcome such a problem, in these scenarios, ATUA updates the EWTG to include a new pop-up Window.

3.5.2 ATUA testing algorithm.

At runtime, after detecting the active Window, ATUA automatically derives the current AbstractState according to the procedure described in Section 3.1. The subsequent activities depend on the current testing phase.

The activities performed in the three phases follow the same algorithm, which is presented in Algorithm 1. What differentiates the three phases are the strategies adopted to perform the specific operations iterated by the algorithm and the test budget allocated to perform each operation. Line 1 in Algorithm 1 shows that the algorithm iterates till the test budget for the current phase is consumed (function *phaseBudgetConsumed*), all the targets for the current phase have been covered (function *coverageTargetsExercised*), or it cannot further improve coverage (function *stagnation*).

The iteration starts by identifying a test target (function *selectTarget*, Line 3). The test target is either a Window or a WindowTransition. A new test target is identified when no target has been selected yet or the current target has already been fully exercised (Line 2). After identifying the test target, ATUA relies on the App model to identify the test target path (Line 4), i.e., a sequence of Actions that makes the App render the target Window or reach the target AbstractState.

The test target path is derived with a breadth-first traversal of the App model. The traversal starts from the current AbstractState. The traversal proceeds through both AbstractTransitions and WindowTransitions. A WindowTransition is taken only if an AbstractTransition is not available. The visit of the model stops when we reach the test target or all the reachable nodes are explored.

So long as a test target is not reached (Line 8), ATUA executes function *reachTargetNode* (Line 9), which triggers the next Action in the test target path. For each Action in the test target path, we know the Window or the AbstractState to expect. After executing an Action, function *reachTargetNode* checks if the App is in the expected Window or AbstractState. If not, the test target path does not reflect the actual behaviour of the AUT and cannot enable ATUA to reach the test target. In this case, function *reachTargetNode* flags the target as not reached, and returns to the main execution loop to look for a different path to reach the test target (Line 5). When a target is reached (Line 10), ATUA exercises the target according to the Action generation strategy for the current phase (function *exerciseTarget*).

Finally, random exploration of the active Window might be triggered by functions *reachTargetNode* and *identifyPathToTarget* to improve the EWTG (Line 14). This is described in Section 3.5.3.

To regulate the allocation of the phase budget (i.e., how many Actions each function invoked by the algorithm is allowed to generate, in each phase), the ATUA algorithm makes use of three budget variables: (1) *reachabilityBudget*, which specifies the maximum number of Actions to be used to reach a target node, (2) *targetBudget*, which specifies the number of Actions to be used to exercise the target node, *randomBudget*, which specifies the number of Actions to be used for random exploration. At runtime, when counting the number of Actions performed, we ignore Actions of type TextInput and Click on checkboxes since they generally do not trigger WindowTransitions. The budget variables are initialized with different values depending on the current test phase. Table 3 provides an overview of the criteria adopted, which are described in details in the following paragraphs. To define budgets, a *scale factor* is used to optimally distribute the test budget across phases and test targets. For example, with a test budget of five hours, we can invest more time in Phase 2 than with a test budget of one hour.

3.5.3 *Random exploration.*

Functions *reachTargetNode* and *identifyPathToTarget* in Algorithm 1 may trigger the random exploration of the App under test to improve the EWTG. This is done when Inputs cannot be exercised and a test target cannot be reached.

Function *reachTargetNode* determines that an Input cannot be exercised when the associated Widget is not visible or enabled in the GUITree. It happens, for example, when the content of a

Algorithm 1 ATUA testing algorithm.

```
while ( NOT stagnation() ) AND ( NOT phaseBudgetConsumed( phaseBudget) ) AND (NOT coverageTargetsExercised() ) do
    if target not selected OR target already exercised OR visitBudget exhausted then
        selectTarget()
        identifyPathToTarget()
    else if target unreachable then
        identifyPathToTarget()
    end if
    if NOT targetReached() then
        reachTargetNode( reachabilityBudget )
    else
        exerciseTarget( targetBudget )
    end if
    if additional random exploration required then
        performRandomExploration( randomExplorationBudget )
    end if
end while
```

Phase	Budget	Strategy
Phase1	Phase	Infinite, i.e., all the target windows are exercised till stagnation is detected or all the targets are covered.
	Reachability	Infinite, i.e., all the paths are traversed in this phase.
	Target	Infinite, i.e., all the target Inputs are tried in this phase.
	Random Exploration	Set to *scaleFactor∗NumberOfActionsForActiveWindow*. In our experiments, we set *scaleFactor* to 1 for an overall test budget of one hour, to 2 for a test budget of five hours. Random exploration is triggered by either *reachTargetNode* or *identifyPathToTarget*.
Phase2	Phase	Set to *scaleFactor∗NumberOfTargetWindows*.
	Reachability	Set to *scaleFactor∗actionsThreshold*. The value is reset every time a new TargetWindow is identified. We set *actionsThreshold* to 25.
	Target	Set to be equal to what remains of the ReachabilityBudget after the target window is reached. In other words, *ReachabilityBudget+TargetBudget=scaleFactor∗actionsThreshold*
	Random Exploration	Set to *scaleFactor∗randomThreshold*. Random exploration is triggered by either *reachTargetNode* or *identifyPathToTarget*. We set *randomThreshold* equal to *actionsThreshold*.
Phase3	Reachability	Not used in this phase.
	Target	Set to *scaleFactor∗actionsThreshold*. It is reset every time a new TargetEvent is identified.
	Random Exploration	Set to *scaleFactor∗actionsThreshold*. Random exploration is triggered (1) to explore the related Window, (2) when the related Window cannot be reached through the identified path, (3) when the target Window cannot be reached through the identified path (see Section 3.5.6). We set *randomThreshold* to 5.

Table 3. Strategies adopted, in different ATUA phases, to define the budget allocated to distinct test activities.

NavigationDrawer varies based on the buttons pressed in the active Window. To make the required Widget visible, function *reachTargetNode* randomly exercises the active Window. This is done by iteratively and randomly selecting one widget among the ones that have been exercised less frequently in the active Window. The selected widget is then exercised according to the strategies listed in Table 4. The exploration of the active Window terminates when the desired widget is found or when the test budget for random exploration is exhausted. In Phase 1, to ensure that all the interactive Widgets are exercised at least once, the budget for random exploration consists of a number of Actions equal to a scale factor multiplied by the number of distinct Actions that can be performed in the active window. The number of distinct Actions that can be performed in a window is based on the interactive information associated to a widget (e.g., we perform a Click Action if the widget is clickable, or four Swipe Actions - one for each swipe direction - if it is scrollable). In Phase 1, ATUA exercises the active window with every possible Action identified.

Widget type	**Input generation procedure**
Any widget	Trigger an InputEvent among the ones for which an event handler has been declared.
Textarea	Randomly apply one of the following: (1) leave it empty, (2) reuse a string already used in the past, (3) reuse a string already used for the same widget, (4) reuse a string literal extracted with static analysis, (5) use a randomly generated alphabetic string [8], (6) use a randomly generated non alphabetic char.
Radio buttons and check boxes	Randomly select one of the possible options (e.g., checked/not checked, for check boxes).
Widgets with manual input	Randomly select one of the available InputInstances, if more than one is available, and then assign the specified value.
Intent	If the current activity declares an Intent, it triggers the Intent specified by the engineer.

Table 4. ATUA Input generation procedures.

In Phase 2 and Phase 3, the budget for random exploration is set by multiplying a scale factor by a constant threshold, as explained in the following paragraphs. If a different Window is reached during random exploration, ATUA returns to the Window to be randomly explored.

Function *identifyPathToTarget* may determine that it is not possible to find a path to a test target. This happens when the EWTG does not include all the WindowTransitions, which is due to the limitations of static analysis tools mentioned in Section 3.3. For example, Gator does not detect the Animation design pattern [23], which leads to a WindowTransition. When a test target path is not found, ATUA performs a random exploration of the active Window and then resumes the execution from the beginning of the main execution loop. ATUA records of all the unreachable targets identified.

3.5.4 Phase 1.

In Phase 1, a test target is any Window with target Inputs that have not been triggered yet. Function *selectTarget* randomly selects a Window with such characteristics (Line 3 in Algorithm 1).

After selecting the target Window, ATUA follows the path to reach the test target. However, to optimize test execution, ATUA exercises any test target accidentally reached before the target Window. In other words, function *targetReached* returns true whenever the active Window is a test target not yet fully exercised.

Function *exerciseTarget*, first produces *user-like inputs* (i.e., input values for text areas, radio buttons, and check boxes), as specified in Table 4. Then it triggers an Action that exercises a randomly selected target Input. Function *exerciseTarget* keeps triggering Actions that exercise target Inputs until all the target Inputs have been exercised or another Window has been visualized. ATUA then resumes the execution of the main loop (i.e., Line 1 in Algorithm 1). When the target Window is associated to HiddenHandlers that can reach target methods, ATUA also performs a random exploration of the Window. If an Action triggers the execution of an HiddenHandler, ATUA introduces a corresponding WindowTransition into the EWTG.

To maximize the chances of exercising every target Input, which is the objective of Phase 1, the reachability and target budgets are infinite. More precisely, we try to reach every TargetWindow by trying every possible path (infinite *reachabilityBudget*); furthermore, we exercise every TargetWindow with all the target Inputs (infinite *targetBudget*), as described above.

In Phase 1, we observe *stagnation* when all the remaining targets either cannot be reached or all their target Inputs cannot be exercised. We did not define a phase budget for Phase 1 (i.e., *phaseBudget* is set to infinite) because we aim to exercise all the target Inputs before proceeding with the following phases.

3.5.5 Phase 2.

In Phase 2, we aim to maximize the coverage of those target methods that have not been fully covered. To increase chances of improving test coverage, the target Window should be the one that can trigger the execution of the highest number of uncovered instructions. Also, since the AbstractState of an App might influence the inputs that reach a target method, we should give higher priority to Windows with target methods that were exercised in fewer AbstractTransitions. Indeed, since AbstractTransitions are associated to distinct pairs of source and destination AbstractStates, methods that were exercised in multiple AbstractTransitions were likely exercised with a diverse set of inputs.

To identify the target Window, we assign a score to each window according to the formula

$$WS_w = \sum_{m \epsilon MT_w} c_w * u_m$$

where MT_w is the set of target methods associated to the target Inputs of Window w. Term u_m is the number of uncovered instructions belonging to method m. A target Window can thus be any Window with $WS_w > 0$. We then randomly select a target Window by assigning to each Window a probability to be selected that is proportional to its score.

In the formula above, c_w is a weight introduced to focus first on those methods that have been covered less. It is computed as the complement of the proportion of AbstractTransitions that exercise the method:

$$c_m = 1 - \frac{AA_m}{AA}$$

where AA_m is the number of AbstractTransitions that covered method m and AA is the number of AbstractTransitions in the App model.

To select the test target path, ATUA first identifies the AbstractState that will more likely increase the coverage of target methods, which is the one with the highest number of uncovered instructions belonging to interactive widgets. Precisely, ATUA selects the AbstractState that maximizes a score computed according to the formula:

$$AS_{as_w} = \sum_{m \epsilon MT_{as_w}} c_w * u_m$$

where as_w is an AbstractState for the Window w, c_w and u_m have already been defined above, and MT_{as_w} is the set of target methods that might be covered through as_w. The set MT_{as_w} consists of all target methods triggered by either (1) an Intent or (2) an InputEvent for an interactive widget in as_w.

Function *exerciseTarget* works in the same way as in Phase 1. However, Phase 2 differs from Phase 1 with respect to the target, phase, and reachability budgets. Indeed, to uniformly distribute the phase budget across the selected target Windows, in Phase 2, the budget for reaching a test target and exercising it is set to "*scaleFactor∗actionsThreshold*", with *actionsThreshold* representing a number of Actions that, based on preliminary experiments, is sufficient to reach a target and exercise it (in our experiments, we set its value to 25). In Phase 2, the *phase budget* is exhausted when ATUA has exercised a number of windows that is equal to "*scaleFactor∗ overall number of TargetWindows*". Also, a same Window can be selected as a target multiple times. By repeatedly generating different sets of Actions for a same Window, ATUA covers different combinations of user-like inputs, which may include combinations that lead to the coverage of different sets of instructions.

In Phase 2, we observe *stagnation* when, after exercising all the available targets, the coverage of target methods has not increased.

3.5.6 Phase 3. In Phase 3, we aim to cover those target method instructions that exhibit data dependencies from state variables defined by Windows different than the one reaching a target method. Examples include instructions that can be executed only after enabling specific options in the preferences Window of the App. For this reason, in Phase 3, the test target is a WindowTransition presenting associated targetMethods that remain to be fully covered. Also, for each WindowTransition to be tested, we need to identify a set of related Windows that should be exercised before executing it.

Function *selectTarget* returns a WindowTransition belonging to a target Window selected according to the same criteria as for Phase 2, i.e., with a probability proportional to its *WS* score. To minimize the effort spent in reaching target Windows, once a target Window has been selected, function *selectTarget* iteratively returns each target WindowTransition belonging to it.

When a test target has been selected, in function *exerciseTarget*, ATUA (1) identifies the related Window that should be exercised first, (2) identifies a path to this related Window, (3) reaches the related Window and randomly exercises it, (4) identifies a path to the closer AbstractState for the target Window in which the target WindowTransition is enabled, and (5) reaches the identified AbstractState and triggers an Action that exercises the Input associated to the WindowTransition. In Phase 3, function *identifyPathToTarget* is not invoked because testing starts from the related Window. Consistent with Phase 2, the *targetBudget* is set to *scaleFactor*∗*actionsThreshold*.

Function *exerciseTarget* relies on random exploration (1) to explore the related Window, (2) when the related Window cannot be reached through the identified path, (3) when the target Window cannot be reached through the identified path. The random exploration budget is set to *scaleFactor*∗*randomThreshold*. Since random exploration has been largely used in previous phases and to limit the time spent in related windows, in Phase 3, we set *randomThreshold* to a value lower than the one used in Phase 2 (e.g., we used 5 in our experiments).

To identify related Windows, we rely on information retrieval techniques, a solution that has been successfully used to perform fault localization [65] and to select string inputs for automated testing of libraries [53]. We chose not to rely on traditional data-flow analysis [5] because data dependencies might be implemented in many different forms (e.g., setting a state variable in a shared object or saving a property in a key-value registry) that are not fully identified by such analysis.

Related Windows are retrieved through the computation of the term frequency (TF) and inverse document frequency (IDF) metrics, which are used by information retrieval approaches to identify similar documents related to a specific topic [32]. In the following, we discuss how we compute these metrics.

Since dependencies between Windows are due to either state variables defined in shared objects or property values in key-value registries, the executable code of Windows presenting such dependencies should share a subset of *class attributes* and *literals*. For this reason, the terms used to identify dependencies are *class attributes* and *literals* appearing in the implementation of the methods of the App (extracted with Soot).

We compute $TF(t,h)$, the frequency of the term t for an Input handler h, as the number of methods in which the term appears, considering the handler itself and any of the methods invoked by the handler. To include only terms that characterize the functionality triggered by the WindowTransition, we consider only methods declared in the same class of the handler or in its inner or outer classes.

The frequency of the term t for a WindowTransition wt is computed as the sum of the term frequency for all the handlers of the Input (HI_{wt}) that triggers the transition,

$$TF(t,wt) = \sum^{h \in HI_{wt}} TF(t,h)$$

The frequency of the term t for a Window w, $TF(t,w)$, instead, is computed as the number of methods in which the term appears, considering the methods that are either declared in the class that implements the Window or in its parent class.

The inverse document frequency of a term is computed as

$$IDF(t) = log\left(\frac{total\ number\ of\ Windows}{number\ of\ Windows\ in\ which\ t\ appears}\right)$$

The related Windows for a WindowTransition wt can be identified by computing a dependency score for every Window w of the App, as follows

$$DS(w,wt) = \sum^{t \in T} NW(t,w) \times NW(t,wt)$$

where T is the set of terms for the App, and $NW(t,d)$ is the normalized term weight, which captures the extent to which a term is representative for either a Window or a WindowTransition. $NW(t,d)$ is computed according to a standard formula [32]:

$$NW(t,d) = \frac{TF(t,d) * IDF(t)}{EL(d)}$$

where EL(d) is the Euclidean Length of the element d (i.e., a Window or a WindowTransition). It is computed as the square root of the sum of the terms' weight squared [32].

ATUA randomly selects related Windows using the dependency score as probability distribution, such that the higher the score, the higher the probability to select a Window.

Phase 3 terminates when the overall test budget is exhausted or all the instructions of the target methods have been covered.

4 EMPIRICAL EVALUATION

The objective of the empirical evaluation is to compare ATUA with state-of-the-art approaches in terms of cost-effectiveness. It is motivated by our need to achieve high test coverage (effectiveness) while enabling the verification of test results within an acceptable budget (cost) in a CI context.

When a new release is ready for testing, a test automation technique is *effective* when it enables engineers to verify updated features in the App under test automatically; more precisely, when it extensively exercises updated methods and their instructions. Measuring the effectiveness of App testing techniques in terms of method and instruction coverage is common practice [12]. Although engineers may aim to exercise all the methods that could be impacted by the changes (e.g., the ones identified by means of change impact analysis as mentioned in Section 3), in our empirical evaluation, we focus on updated methods since exercising them is the minimum requirement of any testing criterion targeting software updates.

What we refer to as *cost* comes in two forms, (1) test execution time and (2) manual effort. In general, *test execution time* does not necessarily need to be minimized but it should be practical. For example, we expect a test budget of one hour to be practical in a continuous integration context, while a budget of five hours might be acceptable when testing overnight.

Manual effort, in our context, is mostly driven from the absence of a solution to automate test oracles, even in the presence of automated test input generation. As described in Section 3, to address the oracle problem in the context of App updates, engineers can rely on two complementary state-of-the-art approaches that respectively address regression failures and failures in newly implemented, repaired, or updated features. To discover regression failures, engineers can replicate, on a previous App version, the test input sequences generated for the updated App. With ATUA, input sequences correspond to the sequences of Actions generated by ATUA to test an App. To discover failures in new, repaired, and updated features engineers should visualize the screenshots of the Windows or the GUITrees generated for the provided inputs. The effort in doing so can potentially be reduced through crowdsourcing. Regardless of the situation and context, manual effort, given the scarcity of qualified human resources, should in general be minimized.

We assume here that manual effort is proportional to the number of inputs generated by test automation. Indeed, to overcome execution errors due to changes in the GUI and be able to replicate test input sequences in a different App version, engineers may need to manually repair the sequence of inputs (e.g., by changing the ID of a widget to be clicked). A large number of inputs may thus lead to numerous manual repair operations and make test oracle automation infeasible[3]. Also, a large number of inputs and, consequently, a large number of outputs generated by the App under test, may lead to prohibitive costs for output verification through visual inspection, even when less costly solutions (e.g., crowdsourcing) are adopted.

Our research questions are organised according to the two cost measures above. They evaluate the extent to which we have achieved the objectives mentioned in the introduction; RQ1 addresses O2, while R2 addresses O1.

- RQ1 *Can ATUA reduce the manual effort required for testing Apps, compared to state-of-the-art approaches?* We aim to determine if the number of inputs generated by ATUA is significantly lower than the number of inputs generated by state-of-the-art approaches, for a same execution time budget. A lower number of inputs makes test automation more widely applicable in practice since it reduces the effort related to the definition of test oracles and repair of input sequences. Also, to determine if the effort of using ATUA is justified by practical benefits, we aim to verify if ATUA provides higher effectiveness per unit of effort than state-of-the-art approaches.
- RQ2 *Can ATUA effectively test Apps within practical time budgets, compared to state-of-the-art approaches?* This research question aims to determine if ATUA performs significantly better than state-of-the-art approaches in terms of coverage of updated methods and their instructions, for a same execution time budget.

A replicability package is made available online [35].

4.1 Study subjects

To perform our experiments, we considered as experimental subjects a number of Apps available on the Android Play Store that are highly popular (i.e., more than 100,000 downloads, on average) and that were used for validation in recent papers reporting on related techniques, i.e., APE [21] and DM2 [8]. We considered only the apps that can be executed on the recent Android simulator version supported by our toolset (i.e., Android API Level above 23). For each App, we considered the latest 10 released versions at the time of running our experiments (hereafter referred to as V0, ..., V9), when available[4]. In Table 5, for each version of each subject, we report the overall number of methods, the number of updated methods, the overall number of bytecode instructions, and the

[3] Related work reports that repairing a single test input takes 15 minutes, on average [22].
[4] We did not consider Jamendo as subject for our study because it was considered for preliminary experiments to setup ATUA. Also, support for older versions of Jamendo had been recently discontinued by the App developers thus preventing experiments involving multiple App versions.

Table 5. Overview of subject systems.

Subject App		Details	
1	Wikipedia	**V**	110,144,146,159,190,198,10239,10263,10264,10269
		AM	3767, 5009, 5646, 6435, 6943, 7477, 8814, 8751, 8759, 8793
		UM	3767, 446, 195, 108, 370, 292, 1430, 535, 13, 94
		AI	32208, 38913, 43753, 48761, 51147, 54759, 68207, 69471, 69533, 69856
		UI	32208, 11000, 4157, 2441, 6606 ,6345, 24536, 12724, 281, 2698
2	Activity Diary	**V**	105, 111, 115, 117*, 118, 122, 125, 130, 131, 134
		AM	260, 333, 333, 333, 333, 450, 479, 540, 540, 659
		UM	260, 18, 3, 7, 12, 117, 39, 28, 1, 49
		AI	3667, 4832, 4831, 4834, 4880, 6613, 7052, 8247, 8251, 10622
		UI	3667, 558, 21, 295, 599, 3393, 2027, 1535, 15, 2459
3	File Manager	**V**	44, 53, 77, 79, 82, 84
		AM	2042, 2132, 3422, 3430, 3648, 3648
		UM	2042, 306, 415, 11, 644, 2
		AI	34389, 34931, 48241, 48294, 51755, 51789
		UI	34389, 14510, 13744, 703, 24960, 143
4	Nuzzel	**V**	302, 303, 318*, 323, 325, 328, 329, 330, 331, 333, 334
		AM	4223, 4220, 4524, 4498, 4527, 4650, 4771, 4832, 4833, 4833, 4834
		UM	4223, 8, 717, 75, 33, 41, 21, 21, 1, 1, 1
		AI	40522, 40449, 43309, 43083, 43403, 44234, 45331, 45908, 45913, 45916, 45940
		UI	40522, 151, 18335, 2952, 1593, 1990, 647, 1378, 35, 69, 45
5	Yahoo weather	**V**	1.16.0, 1.16.1, 1.16.2, 1.17.3, 1.18.1, 1.19.1, 1.20.1, 1.20.3, 1.20.5, 1.20.7
		AM	2932, 2904, 2904, 2630, 3105, 3109, 3178, 3255, 3255, 3303
		UM	2932, 5, 4, 243, 10, 16, 118, 101, 12, 9
		AI	38015, 37867, 37857, 34220, 38219, 38211, 39086, 39439, 39439, 39462
		UI	38015, 417, 272, 10198, 857, 588, 4295, 3842, 689, 961
6	Wikihow	**V**	2.7.3, 2.8.0, 2.8.1, 2.8.3, 2.9.1, 2.9.2, 2.9.3
		AM	333, 333, 333, 333, 325, 322, 319
		UM	333, 111, 1, 1, 65, 4, 18
		AI	3704, 3992, 3941, 3944, 3808, 3761, 3657
		UI	3704, 2279, 39, 42, 1370, 93, 543
7	BBC Mobile	**V**	5.1.0, 5.10.0, 5.11.0, 5.12.0, 5.13.0, 5.4.0, 5.5.0, 5.6.0, 5.8.1, 5.9.0
		AM	10706, 8724, 8792, 8902, 8926, 9945, 10380, 10696, 10200, 8939
		UM	10706, 649, 27, 44, 25, 603, 242, 553, 77, 95
		AI	76649, 61604, 62078, 62937, 63232, 71053, 73082, 73950, 72439, 61618
		UI	76649, 11182, 1557, 2288, 2101, 10637, 6324, 9484, 1638, 3274
8	VLC player	**V**	3.1.4, 3.1.5, 3.1.7, 3.2.12, 3.2.2, 3.2.3, 3.2.6, 3.2.7, 3.2.9
		AM	6796, 6843, 6854, 8681, 8544, 8551, 8621, 8641, 8676
		UM	6796, 672, 26, 3, 149, 13, 51, 33, 42
		AI	86266, 87560, 87886, 117207, 115344, 115412, 116409, 116647, 117071
		UI	86266, 34086, 1522, 150, 9611, 1163, 3527, 1961, 3010
9	City-mapper	**V**	9.1, 9.2, 9.3, 9.4, 9.5, 9.6, 9.7, 9.8, 9.9, 10.0
		AM	9629, 9499, 9599, 9491, 9602, 9761, 9868, 9929, 9884, 10050
		UM	9629, 51, 37, 55, 73, 119, 76, 73, 12, 69
		AI	155117, 154086, 157036, 153200, 155950, 161914, 164267, 165747, 165747, 163303
		UI	155117, 2726, 2286, 2075, 3160, 6262, 2756, 2340, 1372, 3775

Legend: V: Versions. AM: All Methods. UM: Updated Methods. AI: All Instructions. UI: Instructions belonging to updated methods. An asterisk (*) is used to indicate not tested versions.

number of bytecode instructions belonging to the updated methods. In total, we downloaded and processed 83 App versions, 74 being updated versions.

For all subjects, we treat version V0 (i.e., the first one listed in row *V*) as the first released version. The number of updated methods ranges from one (e.g., version V8 of subject App 2) to 1430 (e.g.,

Table 6. Case studies with manual inputs.

Case study	Feature tested	# Windows	# Data fields	# Instances
Wikipedia	Log-in functionality	1	2	1
	Creation of a new account	1	4	14
VLC	Play a video stream using a URL	1	1	2
	Populate the library with all the videos on the device	1	1	1
Nuzzel	Request an e-mail newsletter	1	1	1

version V6 of subject App 1), thus being representative of a wide variety of release scenarios (i.e., from simple bug fixes to major releases). The number of bytecode instructions, ranging from 3667 to 165747, shows that the considered Apps have varied degrees of complexity. Further, the growing number of instructions belonging to the different versions of the Apps (e.g., from 32208 to 69856 in the case of Wikipedia) suggests that they are representative of typical Apps where features are incrementally introduced at every release, thus further motivating the adoption of ATUA.

4.2 Experimental setup

In our experiments, we compare ATUA with three state-of-the art test automation tools, i.e., DM2, Monkey, and APE. These tools do not specifically target updated methods, but simply aim to maximize the coverage of the whole App.

DM2 has been selected because it is the framework on top of which we implemented ATUA. We configured it to use the biased-random testing strategy, which matches the random input selection strategy of ATUA. The comparison with DM2 enables us to determine if the additional analyses performed by ATUA (i.e., static program analysis, adaptable state abstraction function, and inputs generation based on information retrieval) contribute to generating better results than a simpler solution based on dynamic program analysis only.

Monkey is a program that runs on the Android emulator and generates pseudo-random streams of events. It is used as baseline for App testing approaches and surprisingly fares better in many benchmarks [56]. The reason is that the time saved by not processing the App GUITree can be used to further explore the App state space.

APE is a state-of-the-art App testing toolset that overcomes existing approaches thanks to an adaptable state abstraction function (see Section 5).

In our experiments, we considered two possible execution scenarios, with respectively test budgets of one hour (a practical choice in a continuous integration context) and five hours (a reasonable choice for overnight execution).

To use ATUA, for three subjects, we specified a set of manual inputs necessary to exercise the primary features of the Apps (e.g., to login and use the App). Support for manual inputs is a necessary feature of test automation tools because Apps often require domain-specific information that cannot be derived automatically (e.g., login data). Table 6 provides a summary of the manual inputs defined; for each, we provide a description, and the number of windows in which the manual input might be triggered. For Wikipedia, we configured two manual inputs, one with the information for creating a new account, another one with information to log-in. In the case of VLC, we provide the URL of a stream to be reproduced and the indication of a checkbox to be checked in order to populate the library with device data (otherwise, no content can be played and testing is limited). Regarding Nuzzel, we provide the e-mail address to receive a newsletter. The effort required to define manual inputs is limited; indeed, for each input, we have specified a single Window where it is applied, between one and four data fields, and a very limited number of input instances, that is, one for every tested feature except for the creation of a new Wikipedia account and the playback of a video stream with VLC. When creating a new account, it is necessary to specify a larger set of inputs to exercise the feature under test multiple times; indeed, a same e-mail address cannot be shared by distinct Wikipedia accounts. As for VLC, since one of its main features is to play video streams, it makes sense to test it with both a working and a corrupted video stream. This example illustrates that the effort required to specify manual inputs is negligible.

To account for randomness, we executed each tool against each updated version 10 times. We report results for 72 of the 74 versions available since, for two App versions of Nuzzel and Activity Diary (indicated with an asterisk in Table 5), it was not possible to execute all the testing tools. More precisely, for version 318 of Nuzzel, the App starts but gets stuck in the first Activity, while version 117 of ActivityDiary can be tested only with ATUA and DM2, but not with Monkey and APE. In total, we executed 5760 test sessions (4 tools × 72 versions × 10 runs × 2 test budgets) for a total of 17280 test execution hours. To perform our experiments, we relied on the Grid 5000 infrastructure [6, 24] in Luxembourg, which provides access to 800 compute-nodes grouped into homogeneous clusters. We rely on nodes with 16x2.1 GHz and 18x2.2 GHz CPU cores.

In the following sections, we analyze differences in results using a non-parametric Mann Whitney U-test (with $\alpha = 0.05$). We discuss effect size based on Vargha and Delaney's A_{12} statistics [55], a non-parametric effect size measure. The A_{12} statistic, given observations (e.g., code coverage, in our context) obtained with two treatments X and Y (testing tools, in our context), indicates the probability that treatment X leads to higher values than treatment Y. Based on A_{12}, effect size is considered small when $0.56 \leq A_{12} < 0.64$, medium when $0.64 \leq A_{12} < 0.71$, large when $A_{12} \geq 0.71$. Otherwise the two populations are considered equivalent [55]. In contrast, when A_{12} is below 0.50, it is more likely that treatment X leads to lower values than treatment Y. Symmetrically to the case above, effect size is small when $0.36 < A_{12} \leq 0.44$, medium when $0.29 < A_{12} \leq 0.36$, large when $A_{12} \leq 0.29$.

4.3 RQ1: Manual Effort

4.3.1 Experimental setup. To address RQ1, we count the number of inputs generated by each testing tool, for each test execution run. For DM2 and ATUA, we rely on the CSV file generated by the ActionTrace component of DM2, which reports all the inputs triggered during testing. For Monkey and APE, we record the number of test inputs reported by the tool at the end of execution.

For each subject App, we compare distributions of the number of inputs generated across tools. We also analyze the ratio between the number of target instructions (i.e., instructions belonging to updated methods) that are automatically exercised and the number of inputs triggered by the test automation tool. This ratio captures how useful it is for a software engineer, on average, to invest time in repairing a single input of the test sequence or verifying the output produced by an input. For example, a *target instructions/input ratio* of five indicates that, for every input, the test automation approach exercises, on average, five instructions belonging to updated methods.

To answer positively this research question, ATUA, compared to other tools, should generate less test inputs and have the highest *target instructions/input ratio.*

4.3.2 Results. Figures 8 and 9 show boxplots capturing the number of inputs generated by each approach in every run, for every subject App, for the two distinct test budgets considered (i.e., one hour and five hours). Please note that, in all the boxplots presented in this paper: (1) horizontal dashed lines show the average across the data points of the boxplot (i.e., average for the subject App), (2) horizontal dotted lines traversing the whole chart show the average across all the runs, (3) whiskers are used to report min and max values across runs for all versions.

Figures 8 and 9 shows that ATUA generates, on average, the lowest number of inputs: 876.53 for one hour, 4608.57 for five hours. ATUA is thus the most suitable approach to minimize test

Table 7. Statistical significance and effect size for Figure 8 and 9.

	1 hour budget						5 hours budget					
	p-value			A_{12}			**p-value**			A_{12}		
S	D	M	A	D	M	A	D	M	A	D	M	A
1	<0.05	<0.05	<0.05	.018	.000	.000	<0.05	<0.05	<0.05	<u>.630</u>	.000	.000
2	<0.05	<0.05	<0.05	.000	.000	.000	<0.05	<0.05	<0.05	.020	.000	.000
3	<0.05	<0.05	<0.05	.039	.000	.000	<0.05	<0.05	<0.05	.093	.000	.000
4	<0.05	<0.05	<0.05	.099	.000	.000	<0.05	<0.05	<0.05	.278	.000	.000
5	<0.05	<0.05	<0.05	.074	.000	.397	<0.05	<0.05	<u>0.81</u>	.090	.000	<u>.490</u>
6	<0.05	<0.05	<0.05	.360	.000	.000	<u>0.15</u>	<0.05	<0.05	.425	.000	.000
7	<0.05	<0.05	<0.05	.117	.000	.000	<0.05	<0.05	<0.05	.123	.000	.000
8	<0.05	<0.05	<0.05	.405	.000	.000	<0.05	<0.05	<0.05	.075	.000	.000
9	<0.05	<0.05	<0.05	<u>.048</u>	.000	.000	<0.05	<0.05	<0.05	.001	.000	.000

Legend: *S*, subject. *D*, comparison with DM2, *M*, Monkey, *A*, APE. We underline the few cases in which statistics indicate that ATUA shows no significant difference (i.e., *p-value* $\geq$0.05) or no higher chances of generating less instructions (A_{12}>0.44) than state-of-the-art approaches.

automation effort. Monkey generates the largest number of inputs (i.e., 57888.79 for one hour, 291614.69 for five hours) because it does not invest any of the time budget into analyzing execution data but simply generates purely random inputs. APE relies on Monkey to generate inputs; however, APE generates less inputs than Monkey (i.e., 27505.89 for one hour, 134640.71 for five hours) because it spends time refining the state abstraction function (see Section 5). Finally, DM2 generates a number of inputs (i.e., 1325.71 for one hour, 6858.16 for five hours) that is closer to those generated by ATUA. This is mostly due to the fact that both approaches are model-based and share the same dynamic analysis infrastructure; however, on average, ATUA generates less inputs because it invests more of the time budget into the analysis of run-time data.

Figures 10 and 11 present the same boxplots as Figures 8 and 9 but zoom in on ATUA and DM2 data to highlight their differences. Table 7 reports the p-value and A_{12} statistics obtained with the Mann Whitney U-test and the Vargha and Delaney's method, respectively. Recall that we aim to minimize the number of inputs and are thus interested in effect sizes below 0.46, i.e., the probability that ATUA generates a number of inputs higher than another approach should be below 0.46, ideally close to zero.

For a budget of one hour, for all the subject Apps, ATUA generates, on average, less inputs than DM2. Differences are statistically significant and effect size is always in favor of ATUA and large for seven of the nine subjects. Differences with Monkey and APE are always significant and effect size is always large except for subject 5 (YahooWeather), in which APE does not interact properly with one App version, apparently because of a bug in APE.

With a budget of five hours, ATUA also generates, on average, less inputs than DM2 across subjects. But in the case of Wikipedia, effect size is not in favor of ATUA (i.e., it is likely to generate more inputs than DM2). However, this is due to a bug in DM2 rather than a feature; indeed, in the presence of WebViews, the communication between the DM2 device daemon and the DM2 client are delayed, thus reducing the number of inputs being generated. In the case of ATUA, which is built on top of DM2, this problem is less evident because such communication is triggered less frequently. The differences between the number of inputs generated by ATUA and the ones generated by APE and Monkey are always statistically significant with a large effect size in favor of ATUA (except for YahooWeather, as already discussed above).

Figures 12 and 13 show, for each subject, the distribution of the target instructions/inputs ratio. ATUA has the highest ratio: 2.26 for one hour, 0.49 for five hours. For the one-hour budget, ATUA's test automation effort (i.e., manual repair of a GUI input, visual inspection of outputs) is thus more

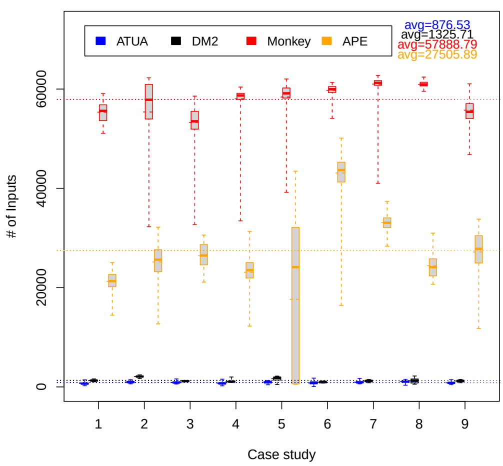


Fig. 8. Number of inputs generated (budget=1 hour).

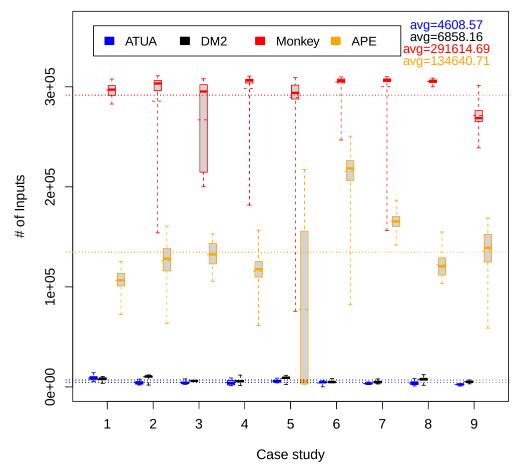


Fig. 9. Number of inputs generated (budget=5 hours).

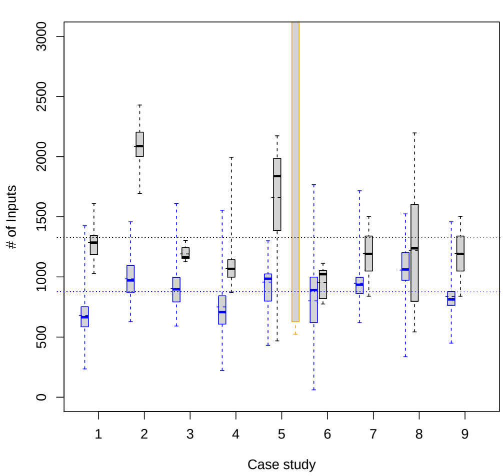


Fig. 10. Number of inputs generated (budget=1 hour). Zoom on ATUA and DM2 data.

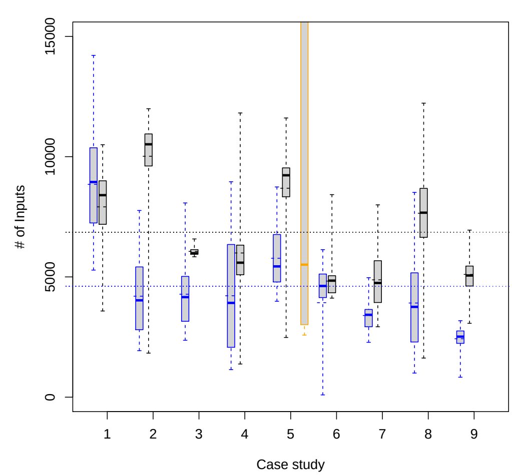


Fig. 11. Number of inputs generated (budget=5 hours). Zoom on ATUA and DM2 data.

beneficial because each input enables the verification of 2.26 additional target instructions. As a comparison, other state-of-the-art approaches yield lower ratios: 1.34 (DM2), 0.03 (Monkey), and 0.10 (APE). These results show that, though Monkey and APE are known for effectively triggering crashes, they are unlikely to be applicable in a testing context where the number of generated inputs should be minimized. For a time budget of five hours, average differences are less pronounced but the same trends hold.

Table 8. Statistical significance and effect size for target instructions/inputs ratios.

	1 hour budget						5 hours budget					
	p-value			A_{12}			p-value			A_{12}		
S	D	M	A	D	M	A	D	M	A	D	M	A
1	0.10	<0.05	<0.05	.728	1.00	.975	0.69	<0.05	<0.05	.555	.963	.901
2	0.17	<0.05	<0.05	.703	.906	.875	0.14	<0.05	<0.05	.719	.938	.875
3	0.29	0.09	0.09	.700	.820	.820	0.25	0.08	0.17	.720	.840	.760
4	0.33	<0.05	0.05	.636	.895	.772	0.38	<0.05	<0.05	.623	.895	.747
5	0.17	<0.05	<0.05	.691	1.00	.901	0.17	<0.05	<0.05	.691	.950	.827
6	0.63	<0.05	<0.05	.583	1.00	1.00	0.52	<0.05	<0.05	.611	1.00	1.00
7	0.27	<0.05	<0.05	.654	1.00	1.00	0.27	<0.05	<0.05	.654	1.00	1.00
8	0.60	<0.05	<0.05	.578	1.00	.953	0.92	<0.05	<0.05	.516	.968	.906
9	0.39	<0.05	<0.05	.642	1.00	1.00	<0.05	<0.05	<0.05	.914	1.00	1.00

Legend: *S*, subject. *D*, comparison with DM2, *M*, Monkey, *A*, APE. We underline the few cases in which statistics indicate that ATUA shows no significant difference (i.e., *p-value* $\geq$0.05) or no higher likelihood of achieving a higher instructions/inputs ratio (i.e., A_{12}<0.56) than state-of-the-art approaches.

Table 8 provides p-values and A_{12} statistics. Since we aim to determine if ATUA is likely to generate a higher target instructions/inputs ratio, we look for A_{12} values above 0.50. For a one-hour budget, effect size is always in favor of ATUA (i.e., is more likely to generate a higher instructions/inputs ratio); effect size is always large with respect to Monkey and APE. Even if in a few cases differences in median are not statistically significant, effect size trends provides a clear picture of the benefits: ATUA is likely to yield a higher instructions/inputs ratio. The same conclusions can be drawn for a five-hour budget, though for two subjects (Wikipedia and VLC) ATUA performs similarly to DM2.

To summarize, regarding the human effort required for practical execution time budgets, ATUA performs better than the other approaches since it saves around 33.8% (1 hour budget) and 32.8% (5 hours budget) of the effort compared with DM2, while it shows huge savings compared to the other two approaches. As for the effectiveness per unit of effort, ATUA provides tangible gains of 68.7% (1h) and 63.3% (5h) compared with DM2, and huge differences with the others. **ATUA therefore significantly decreases the manual effort required for repairing inputs and defining oracles when compared to state-of-the-art approaches.**

4.4 RQ2: Effectiveness within Time Budget

4.4.1 Experiment design. To address RQ2, we focus on code coverage results obtained with the updated versions of our subject Apps, i.e., versions V1 to V9. More precisely, we keep trace of the updated methods (hereafter, *target methods*) and instructions belonging to updated methods (hereafter, *target instructions*) that are exercised by the test automation tools considered in our study. To collect data for ATUA and DM2, we rely on the Soot-based code coverage extension integrated into DM2; for Monkey and APE, we rely on MiniTracing, a toolset developed to measure code coverage with APE [20]. Since all these code coverage tools measure the coverage of the whole App under test, to determine the coverage of target methods and instructions, we filter results based on the list of updated methods generated by AppDiff.

We aim to compare ATUA and state-of-the-art approaches in terms of the percentage of target methods and instructions that are covered for a same execution time budget. Since these approaches require a different degree of manual effort (see RQ1) and manual effort is measured in terms of inputs generated, we also set an identical limit to the number of inputs that might be generated by the test generation techniques. The rationale is that we try to emulate, in our experiments, realistic conditions where testers are limited by both execution time and human resources. This is thus

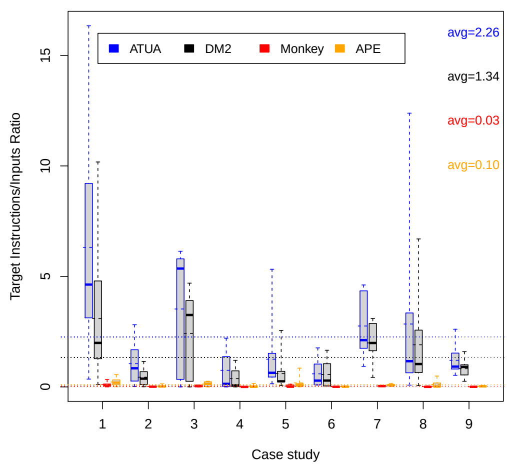


Fig. 12. Target instructions/inputs ratio (budget=1 hour).

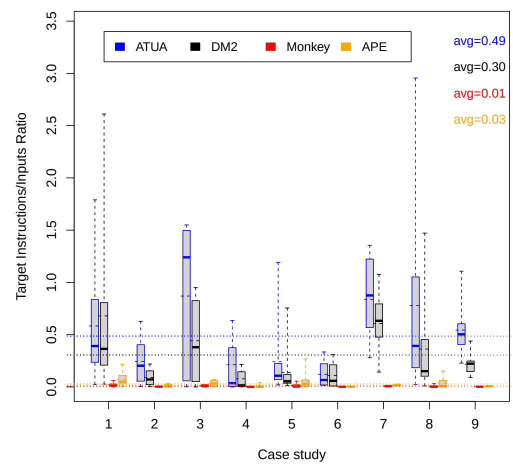


Fig. 13. Target instructions/inputs ratio (budget=5 hours).

expected to yield unbiased comparisons of practical value. It is also consistent with our objective, stated earlier, of minimizing human effort while keeping execution time within acceptable bounds. More precisely, for each software version v, we define an inputs budget equal to the maximum number of inputs generated, over ten runs, by ATUA, which is the approach generating the fewest test inputs for a given time budget, based on RQ1 results. In summary, we thus measure the coverage obtained when testing a subject App for a maximum and practical execution time budget (i.e., one hour and five hours, as discussed in Section 4.2), while not exceeding a maximum input budget determining manual effort.

To positively answer this research question, ATUA should, in statistical terms, exercise more target methods and instructions than the other approaches.

4.4.2 Results. Figures 14 and 16 show the distribution of the percentage of target methods and instructions that have been covered by the selected testing tools for the subject Apps, with a test budget of one hour. Figures 15 and 17 report the same measurements for a budget of five hours.

With a test execution budget of one hour, ATUA is the approach with the highest percentage of target methods and instructions being exercised on average, with 66.34% and 56.14%, respectively. The largest differences are observed when ATUA is compared to Monkey; indeed, on average, ATUA exercises 33.46% and 27.42% more methods and instructions than Monkey, respectively. Since Monkey implements a pure random exploration strategy, our results show that a limit on the number of inputs generated by Monkey highly affects its performance. In contrast, the APE state abstraction function enables a more effective generation of test inputs, thus leading to, on average, higher coverage than Monkey. However, ATUA outperforms APE; indeed, on average, ATUA exercises 21.66% and 18.41% more target methods and instructions than APE, respectively. Though DM2 fares better than Monkey and APE, as it relies on a model-based approach leveraging dynamic analysis, ATUA exercises 7.50% and 6.37% more target methods and instructions. This is explained by the transition-driven exploration based on static analysis (ATUA Phase 1 and 2) and information retrieval (Phase 3), which are not part of DM2.

Table 9. Statistical significance and effect size for RQ2.

	1 hour budget						**5 hours budget**					
	p-value			A_{12}			**p-value**			A_{12}		
C	D	M	A	D	M	A	D	M	A	D	M	A
	Coverage of target methods											
1	<0.05	<0.05	<0.05	.857	.994	.956	<0.05	<0.05	0.14	.703	.935	.564
2	<0.05	<0.05	<0.05	.733	.868	.703	<0.05	<0.05	<0.05	.774	.832	.623
3	0.41	<0.05	<0.05	.548	.795	.684	0.61	<0.05	0.99	.529	.767	.501
4	0.16	<0.05	0.07	.560	.852	.577	0.16	<0.05	0.08	.560	.760	.575
5	<0.05	<0.05	<0.05	.735	.814	.920	<0.05	<0.05	<0.05	.713	.702	.841
6	0.08	<0.05	<0.05	.587	.838	.689	0.2	<0.05	<0.05	.564	.719	.681
7	<0.05	<0.05	<0.05	.614	.836	.762	0.78	<0.05	<0.05	.488	.879	.879
8	0.23	<0.05	<0.05	.554	.997	.972	0.05	<0.05	<0.05	.588	.995	.892
9	<0.05	<0.05	<0.05	.587	.981	.878	<0.05	<0.05	<0.05	.685	.992	.918
	Coverage of target instructions											
1	<0.05	<0.05	<0.05	.853	.991	.936	<0.05	<0.05	0.15	.735	.884	.562
2	<0.05	<0.05	<0.05	.707	.843	.783	<0.05	<0.05	<0.05	.756	.829	.721
3	0.31	<0.05	0.14	.559	.774	.585	0.18	<0.05	0.70	.577	.705	.478
4	<0.05	<0.05	0.07	.600	.861	.579	0.05	<0.05	<0.05	.584	.764	.594
5	<0.05	<0.05	<0.05	.684	.748	.886	<0.05	<0.05	<0.05	.659	.623	.822
6	0.42	<0.05	<0.05	.542	.788	.827	0.47	<0.05	<0.05	.538	.727	.786
7	<0.05	<0.05	<0.05	.598	.735	.695	0.65	<0.05	<0.05	.480	.780	.780
8	0.06	<0.05	<0.05	.586	.999	.984	0.33	<0.05	<0.05	.633	.999	.936
9	0.05	<0.05	<0.05	.584	.980	.822	<0.05	<0.05	<0.05	.671	.988	.896

Legend: *C*, case study. *D*, comparison with DM2, *M*, Monkey, *A*, APE. We underline the few cases in which statistics indicate that ATUA shows no significant difference (i.e., *p-value* ≥ 0.05) or no higher likelihood (i.e., $A_{12} < 0.56$) of covering more targets than state-of-the-art approaches.

When executed with a test budget of one hour, for all the subject Apps, both the median and the average obtained with ATUA are higher than those obtained with other approaches.

To discuss differences across subjects, we report in Table 9 the p-value and A_{12} statistics obtained with the Mann Whitney U-test and Vargha and Delaney's method, respectively. Overall, differences are statistically significant but there are exceptions: when ATUA is compared to APE for Nuzzel (subject 4), and when ATUA is compared to DM2 for Nuzzel, File Manager (subject 3), Wikihow (subject 6), and VLC (subject 8). However, for most of the subjects, ATUA is likely to exercise more target methods and instructions than other approaches; this is shown by the A_{12} statistics being always above 0.56, except for File Manager, Wikihow, and VLC in the case of DM2[5]. Regarding VLC, the effectiveness of ATUA is limited by the need for setup operations that require some manual effort. Indeed, since certain features can be tested only on specific devices (e.g., an Android TV), identifying target methods through static analysis is of limited usefulness and ATUA performs similarly to DM2. However, such limitations could be surmounted after investing some effort to carefully setup ATUA. For example, by configuring ATUA to be executed on an Android TV in addition to a mobile emulator (i.e., what we used in our experiments). Concerning File Manager and Wikihow, ATUA is affected by some limitations of static analysis, which cannot determine that certain WindowTransitions are associated to specific data types provided as input. More precisely, in the case of File Manager, a number of updated features can be exercised only through specific files (e.g., the decompress operation can be executed only with files having ZIP or RAR filename extension). The static analysis currently implemented in ATUA cannot determine that certain

[5]Average A_{12} is 0.77 for target methods and 0.76 for target instructions coverage; median is 0.80 for target methods and 0.77 for target instructions coverage.

features are enabled only in the presence of specific runtime data (e.g., file names) and thus ATUA, similar to DM2, exercises such features only if it accidentally triggers them thanks to random exploration. A similar but more evident problem occurs also in the case of Wikihow, where static analysis does not identify the WindowTransitions triggered by the inputs sent to WebViews. Indeed, the input handlers executed after sending an input to a WebView (e.g., a click on an anchor) depend on the content of the page (e.g., the file type appearing in the URL of the anchor) and thus cannot be identified by static analysis, which does not process the content of the HTML pages displayed at runtime. For this reason, ATUA cannot fully take advantage of static analysis results in the presence of WebViews. In such cases, similar to DM2, ATUA exercises App features thanks to random exploration. However, current ATUA results with Apps using WebViews largely depend on the proportion of features implemented through WebViews. For example, in the case of WikiHow, which mainly relies on WebViews (five out of eight content types are displayed through a WebView), ATUA performs similarly to DM2; instead, in the case of Wikipedia, which implements only one out of 35 Windows using a WebView[6], ATUA outperforms all the other approaches ($A_{12} \geq 0.56$). To overcome the limitations of static analysis and thus improve ATUA results, it might be necessary to develop dedicated strategies relying on dynamic analysis; for example, by extending the state abstraction function of ATUA to use reducers dedicated to HTML anchors or file objects.

With a test budget of five hours, all the approaches achieve better coverage results; however, the ranking observed for a one-hour budget remains unchanged. ATUA is the approach with the highest percentage of exercised target methods and instructions, on average, with 70.37% and 60.18%, respectively. The largest differences are still observed when ATUA is compared to Monkey; indeed, on average, ATUA exercises 27.24% and 22.73% more target methods and instructions than Monkey, respectively. ATUA exercises 18.38% and 15.77% more target methods and instructions than APE. The second best approach remains DM2, as ATUA exercises 6.45% and 6.25% more target methods and instructions than DM2.

A larger time budget enables ATUA to achieve higher coverage. This is set with the *scaleFactor* configuration parameter (see Section 3.5.2), which we increase for a five-hour budget, thus augmenting the time spent to perform random exploration, reach the test target, and exercise targets. We leave to future work the study of the effect of different configuration values for the *scaleFactor* parameter of ATUA.

With a test budget of five hours, the difference between ATUA and other approaches decreases though. Unsurprisingly, with a larger test budget, random-based approaches can more easily reach updated features than with a one-hour budget, for which leveraging static analysis is more important. For example, in subject App 7 (BBC Mobile), in five hours, DM2 achieves the same average coverage as ATUA.

To discuss differences across subjects, we refer to the p-value and A_{12} statistics reported in the rightmost columns of Table 9. Because of the larger test budget benefiting random exploration, differences between ATUA and other approaches are not significant in 7 out of 27 cases (i.e., 3×9, which is the number of pairwise comparisons between ATUA and the other approaches), two cases more than with a one-hour budget. However, ATUA is still likely to exercise more target methods and instructions than other approaches. Indeed, for both method and instruction coverage, the A_{12} statistics is above 0.56 for 24 out of 27 cases. In general, effect size is slightly lower than for a one-hour budget, with an average A_{12} of 0.73 and 0.72 for the coverage of target methods and instructions. In particular, we observe that the larger time budget enables random-driven approaches to achieve the same effectiveness as ATUA when ATUA is negatively affected by static

[6]In Wikipedia, WebViews are used to display Wikipedia pages while other Views are used for other features such as displaying news, image galleries, or editing the content of a page.

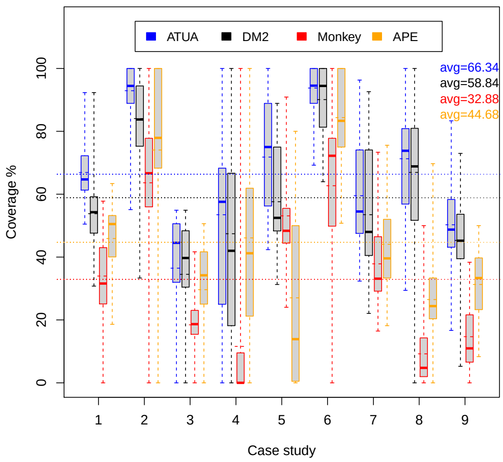


Fig. 14. Percentage of updated methods covered for each version of the case studies (budget=1 hour).

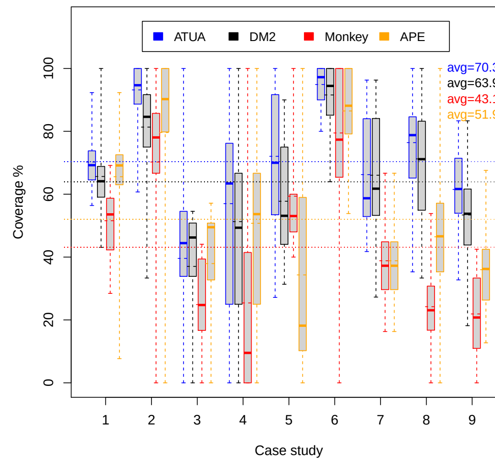


Fig. 15. Percentage of updated methods covered for each version of the case studies (budget=5 hours).

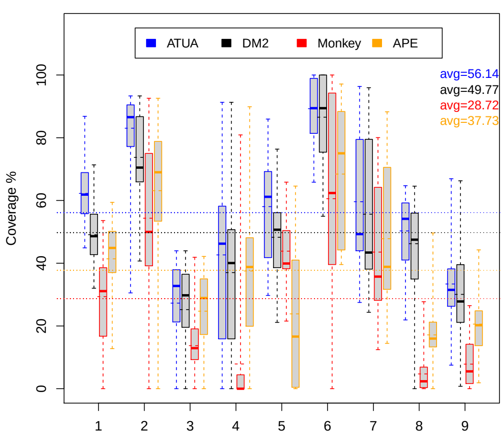


Fig. 16. Percentage of instructions belonging to updated methods that are covered for each version of the case studies (budget=1 hour).

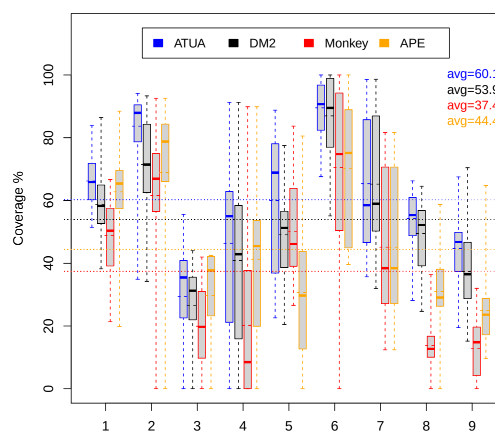


Fig. 17. Percentage of instructions belonging to updated methods that are covered for each version of the case studies (budget=5 hours).

analysis limitations. This happens for File Manager (subject 3), where APE performs similarly to ATUA, Wikihow (subject 6), where DM2 performs similarly to ATUA, and BBC mobile (subject 7), where the additional time budget enables DM2 to exercise the few updated features depending on WebViews (in BBC Mobile, WebViews are used to display BBC Web pages).

To summarize, ATUA is the approach that, on average, most effectively test updated Apps within practical time budgets and human effort. It tends to cover more target methods and instructions than other approaches. The second best approach is DM2. For a one-hour budget, on average, ATUA automatically exercises 7.50% and 6.37% more target methods and instructions than DM2. With a five-hour budget, on average, ATUA automatically exercises 6.45% and 6.25% more target methods and instructions than DM2. For seven out of nine subjects, for both time budgets, ATUA tends to exercise more target methods and instructions than DM2. For the remaining two subjects, DM2 and ATUA are comparable, due mostly to the current limitations of static analysis.

4.5 Discussion

Human effort. RQ1 results have shown that ATUA performs better than the other approaches since it saves around 33.8% (one-hour budget) and 32.8% (five-hour budget) of the effort compared with DM2, the second-best approach. Hereafter, we discuss practical implications concerning testing costs based on related work about the nature of App upgrades [34] and the maintainability of GUI test cases [37].

On average, ATUA generates 450 (one hour) and 2251 (five hours) fewer inputs than DM2 for each App version across all subject Apps. Since related work [34] has shown that roughly 35% of the updates concern the introduction of new features, under the assumption that inputs are uniformly distributed across updated features, we can estimate that ATUA generates, on average for each App version and across all subjects, 158 (one hour) and 788 (five hours) fewer inputs than DM2 for testing new features. Consequently, ATUA generates 292 (one hour) and 1463 (five hours) fewer inputs than DM2 for testing bug fixes and improved features (i.e., changes concerning non-functional requirements).

When testing new features, the output generated by each input should be manually verified; for example, by inspecting the screenshots of the GUI trees visualized after triggering an input (they are automatically captured by ATUA) and determining if they match the expected results. Unfortunately, the software engineering literature lacks studies about the cost of manual verification of GUI trees; assuming, for the sake of illustration, that visual inspection of GUI trees takes a few minutes, say ranging from one minute to five minutes, ATUA may lead to savings within the following intervals of [158-790] and [788-3940] minutes, respectively the for one-hour and five-hour test budgets. In the App development context, where Apps are frequently released (e.g., weekly or bi-weekly) and, additionally, test cases might need to be executed every day following continuous integration practices, such effort savings appear to be particularly beneficial, especially considering that testing should be performed by highly-trained engineers with a deep understanding of the App's features.

When testing updated features, engineers can re-execute the generated test input sequences on previous App versions in order to compare results and, ideally, eliminate oracle costs. However, we have to expect that a number of maintenance operations are required in order to adapt test sequences to a different App version. Pan et al. [37], for example, report that 26.5% of the test inputs need to be repaired. ATUA will thus save engineers from manually repairing 77 and 388 inputs, respectively for the one-hour an five-hour test budget. Under time pressure, which is the case when Apps are frequently released, this is a significant advantage.

Effectiveness. ATUA is the approach that, on average, most effectively tests updated Apps within practical time budgets and human effort. For the one-hour budget, better from competing approaches, it exercises more than 60% of target methods and 50% of target instructions. With a five-hour budget, it exercises more than 70% of target methods and 60% of target instructions. Higher percentages can probably be reached with longer execution budgets, which were not possible in our context given the computational costs of our experiments. Based on these results, we can claim that ATUA can contribute to reducing development costs; indeed, engineers would then be able to focus their manual testing effort on a reduced portion of the developed App.

When comparing with other approaches, we observed that for both one-hour and five-hour budgets, on average, ATUA automatically exercises at least 6% more target methods and instructions than the second-best approach (DM2). The effectiveness of ATUA is comparable to the effectiveness of DM2 and APE only when ATUA cannot fully leverage static analysis to determine the relation between inputs and WindowTransitions, i.e., when Apps integrate input handlers that are selected at runtime based on the nature of input data, which happens, for example, in the presence of WebViews. For the six subjects for which static analysis can effectively be exploited, the percentage of improvement rises above 8%. Among our subject Apps, in the worst case (i.e., five-hour budget), ATUA is comparable to other approaches for one third of the subjects and otherwise fares better; considering that (1) no single competing approach achieves similar coverage as ATUA for these three subject Apps (e.g., DM2 achieves the same results as ATUA for at most two), (2) competing approaches never outperform ATUA but at best reach the same effectiveness, ATUA remains the best choice. To further improve ATUA effectiveness, part of our future work concerns the development of an additional set of reducers that will enable the ATUA state abstraction function to distinguish between widgets containing different types of data.

4.6 Threats to validity

To address threats to external validity we have considered nine popular Apps, downloaded thousands of times worldwide, that have been considered in the empirical evaluation of related work. Also, for each App, we considered up to ten App versions, based on their availability, for a total of 72 App versions tested. The considered Apps greatly vary regarding the overall number of lines of code and updated lines between versions. Because of their diversity, we believe our subjects to be representative of the Apps landscape.

To account for randomness, we tested each App version ten times with every testing tool considered; more than the usual practice of three to five repetitions [21]. Despite the high computational cost (17280 test execution hours, in total), this enabled us to derive solid statistical results for the comparison of different tools.

In our experiments we considered only Android Apps, which is standard practice in most App testing research papers. The prevalence of Android in research papers is mostly due to its worldwide dissemination and the availability of a larger set of tools to test and analyze Android executable bytecode [17]. In our work, the choice of relying on Android Apps enabled the comparison of ATUA with tools working for Android Apps (i.e., Monkey, APE, and DM2). However, since we do not exercise OS-specific features, results should generalize also to Apps running in different execution environments (e.g., HarmonyOS and IOS).

5 RELATED WORK

Automated App testing tools can be grouped according to the strategy adopted to generate test inputs [31, 54]. The most common ones are random, model-based, and evolutionary [54]. Representative approaches of these three categories used in empirical evaluations are Monkey, Stoat [52], and Sapienz [33], respectively. Monkey has been introduced in Section 4. In the following, we discuss state-of-the-art App testing tools and other related work including App regression testing, testing based on information retrieval, and incremental testing.

Stoat performs stochastic model-based testing. It relies on dynamic analysis based on a weighted UI exploration strategy to derive a stochastic finite state machine (FSM) of the App's GUI interactions. Stoat relies on the FSM to generate test suites using an objective function that aims to maximize code coverage, model coverage, and test diversity. The test generation process relies on Gibbs sampling to iteratively mutate and refine the FSM, based on the fitness of the generated test suite.

Sapienz is an evolutionary approach that uses Pareto multi-objective search to automatically explore and optimise test sequences, minimising length, while simultaneously maximising coverage and fault detection. Sapienz combines random fuzzing, systematic exploration and search-based exploration.

Independent empirical evaluations performed by Choudhary et al. [12] and Wang et al. [56] have reported that the different testing strategies are complementary. Further, both studies show that the method and instruction coverage achieved by all test automation approaches are relatively low, that is below 50%. Choudhary et al. [12] note that model-based approaches complement random approaches regarding fault detection, while for code coverage, random approaches fare better. Wang et al. [56] confirm these results. They report that random and evolutionary approaches are complementary regarding method coverage, while both evolutionary and model-based approaches complement random approaches in terms of fault detection. However, the validity of these findings has been weakened by recent advances in model-based approaches. Indeed, more recent results show that model-based approaches that either integrate advanced exploration strategies (i.e., biased random in DM2) or adaptable state abstraction functions (i.e., APE [21]) fare better than random approaches or state-of-the-art model-based approaches. APE, for example, is the most recent technique, and has been reported to perform better than Monkey, Sapienz, and Stoat.

In *APE*, each Window is modeled with sets of *attribute paths* that univocally identify the widgets of the window. Attribute paths resemble the AttributeValuationMaps of ATUA with the difference that ATUA considers a larger set of attributes than APE, which only accounts for type, position with respect to siblings, and appearance (e.g., text). APE and ATUA differ for the strategy used to generate inputs (i.e., ATUA relies on the combination of static and dynamic analysis). They both rely on an adaptable state abstraction function $\mathcal{L}$. However the state abstraction functions integrated in the two approaches present key differences. In APE, a single $\mathcal{L}$ is defined for the whole app, while ATUA specifies one $\mathcal{L}$ for each Window of the App. Also, APE's $\mathcal{L}$ is implemented by means of a decision tree (DT), which enables APE to not rely on a predefined set of reducer functions. The main limitation of APE is that it relies on a global abstraction function for the whole App and thus uses part of the test budget to perform $\mathcal{L}$ refinements that may be avoided by relying on static analysis (i.e., GUITrees belonging to different Windows should be characterized by different abstract states). By relying on a different $\mathcal{L}$ for every Window, ATUA overcomes such limitation. In addition, our empirical evaluation has shown that the combination of static and dynamic program analysis enables ATUA to outperform APE in terms of coverage of updated methods, while minimizing the manual effort required by testing.

Like all the state-of-the-art approaches for App testing [31, 45], ATUA does not address the *oracle problem* [7]. Oracle automation is part of our future work; however, we have presented in Section 4 two solutions to alleviate the oracle problem in the context of App updates. One solution is the automated detection of functional regression faults based on the identification of unexpected changes to outputs across different software versions [19, 27, 40, 48]. The other solution consists in relying on crowdsourcing, a popular solution adopted by industry to reduce the costs of manual GUI testing for Apps [51, 60]. Related work has shown that it is feasible for crowd workers to identify errors after visualizing the inputs and outputs of the functions under test [38]; in the App context, crowd oracles may lower testing costs while test coverage is addressed by test automation. In addition, ATUA could also be integrated with approaches that automatically generate in-program logical assertions [25] or solutions relying on system-independent GUI oracles [64].

Regression testing approaches for Apps concern the selection of events that may trigger modified code [49], the selection of regression test cases [11], and the repair of existing test suites [50]. QADroid is a static analysis toolset that identifies the events (i.e., the Inputs of ATUA EWTG) that may trigger the execution of modified methods. It implements the features of Steps 1 and 2 of ATUA, except that QADroid does not generate EWTGs. QADroid differs from ATUA regarding the underlying static analysis framework used to identify Inputs. It relies on FlowDroid [5], while ATUA relies on Gator because of its capability to generate WTGs. Also, empirical results show that Gator performs better than FlowDroid in identifying valid sequences of callbacks [57]. Different from ATUA, to select the Inputs that may trigger modified methods, QADroid performs a forward traversal of the App control flow graph obtained with Soot (i.e., it starts the traversal from event handler methods) and selects all the Inputs that reach modified methods. Consequently, QADroid cannot determine the presence of *HiddenHandlers* identified by ATUA. In addition, QADroid requires the source code of the App while ATUA works with the bytecode. QADroid has never been applied to automated testing and cannot identify the concrete input values to be used with certain Inputs (e.g., the data to write into a TextArea).

Redroid selects a regression test suite for an updated version of an App [15, 16] from an existing test suite. It relies on state-of-the-art static analysis procedures [43] to perform change impact analysis and uses code coverage information to select any test case that cover modified blocks of code. Redroid does not support the generation of a minimal test suite. *DetReduce* [11], instead, creates a small regression test suite for an App from a test suite generated by a model-based test automation tool. It identifies and removes redundant method call traces and subtraces within the test suite and redundant loops within a test case. Redundancy is identified based on a state abstraction function that considers actionable widgets and their visible attribute values. Widgets are identified by their path in the GUITree. Since executable test suites are often unavailable and state-of-the-art App testing tools can cover only a narrow set of modified methods, the applicability of Redroid and DetReduce remains limited.

Automated repair techniques for the GUI test scripts of mobile Apps are still preliminary [36, 37, 50]; to repair test scripts they include strategies ranging from static program analysis [50], to model-based [30] and computer vision techniques [36, 37]. Existing approaches either leave between 5% [37] and 8% [50] of the test scripts to be manually repaired or do not preserve all the test actions (i.e., the test semantic [30]). Though these results show that automated GUI script repair techniques might be adopted to support the oracle automation approach we suggested for the identification of regression failures, manual intervention would still be required, as indicated in Section 4.5.

In the software testing literature, *information retrieval* techniques applied to the processing of program files have been integrated into fault localization [58, 63], test case prioritization [47], and test input selection approaches [42, 53]. Concerning test input generation, TestMiner relies on information retrieval to select from a large corpus of existing test cases the input values to use in newly generated tests cases, which differs from our purpose [53]. Poster applies a similar approach to derive sequences of Inputs for App testing [42]. Different from ATUA, it relies on existing test suites developed for similar Apps.

Campos et al. [10] have been the first to propose a technique to incrementally test the units in a software project, leading to overall higher code coverage while reducing the time spent on test generation. They apply an evolutionary approach (i.e., EvoSuite [18]) and optimize test generation by providing more test budget to the modified portions of the code and reuse already generated test cases for seeding (i.e., they re-run all the test cases that compile). The main difference with ATUA is that they do not target the GUI testing of Apps but the unit testing of Java libraries. In our context, the reuse of existing test cases is complicated by the presence of a GUI, which is likely updated across versions and may break existing test sequences.

From the above, we can see that existing work lacks a model-based approach combining adaptable state abstraction functions with information retrieved from static program analysis. In ATUA, this is done to maximize the coverage of updated methods while minimizing the number of inputs required for testing. The latter is necessary to make the verification of Apps results feasible. Indeed, fully automated test oracles are currently not an option and manual effort is necessary to identify both regressions and failures in new features. Regression testing approaches, which typically target test case selection and prioritization, are of limited applicability in this context, when test suites covering large portions of the Apps code are not available. ATUA leverages incremental testing to effectively invest the test budget and maximize the coverage of updated methods.

6 CONCLUSION

State-of-the-art App testing techniques are affected by two limitations: limited effectiveness (i.e., low code coverage) and absence of automated oracles. To address the first limitation, given the high release frequency of Apps, we propose a solution (objective O1) to effectively focus the test budget on updated (i.e., modified and new) methods. In other words, within practical test execution time, we aim to maximize the coverage of updated methods and their instructions. To address the second limitation, we aim (objective O2) to generate a significantly reduced set of test inputs, compared to state-of-art approaches, thus proportionally saving the corresponding manual effort required to visualize test outputs or correct test scripts.

To achieve the two objectives above, we developed ATUA, an automated App testing technique that integrates multiple analysis strategies. To achieve O1, it combines static analysis, to determine the inputs that execute updated features, and random exploration, to overcome the limitations of static analysis. To achieve O2, it relies on dynamically-refined state abstraction functions, to determine when distinct inputs lead to a same program state, and relies on information retrieval techniques, to identify dependencies among App features.

We performed an empirical evaluation where we compared ATUA with state-of-the-art approaches implementing testing strategies based on dynamically derived models (DM2), random exploration (Monkey), and dynamic state abstraction (APE). For our experiments, we considered practical execution time budgets of one and five hours, corresponding respectively to approximate time constraints in the context of continuous integration and overnight testing. Concerning manual effort (objective O2), ATUA is the approach that generates the smallest set of inputs with the highest coverage per input. ATUA, on average across subject Apps, saves around 32.6% of the effort, compared to the second-best approach (DM2). Further, it exercises 38.5% more instructions than DM2 per input. Differences with APE and Monkey are much larger. Concerning effectiveness within time budget (objective O1), on average, ATUA automatically exercises up to 70% of updated methods and 60% of instructions belonging to updated methods, 6% more than the second best approach (i.e., DM2). These results show that the analysis strategies integrated in ATUA can drive testing towards an efficient use of the test budget (execution time and effort), thus providing clear benefits when upgrading and testing an App.

ACKNOWLEDGMENTS

This project has received funding from Huawei Technologies Co., Ltd, China, and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 694277). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). Authors would like to thank Erik Derr for the initial implementation of AppDiff.

REFERENCES

[1] Mithun Acharya and Brian Robinson. 2011. Practical Change Impact Analysis Based on Static Program Slicing for Industrial Software Systems. In *Proceedings of the 33rd International Conference on Software Engineering (ICSE '11)*. Association for Computing Machinery, New York, NY, USA, 746–755. https://doi.org/10.1145/1985793.1985898

[2] Android.com. 2020. Intent Resolution. https://developer.android.com/reference/android/content/Intent. Last visited: 11/11/2020.

[3] Android.com. 2020. Logcat command line tool. https://developer.android.com/studio/command-line/logcat.

[4] Android.com. 2020. Monkey - Android ui/application exerciser. http://developer.android.com/tools/help/monkey.html. Last visited: 12/17/2018.

[5] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Octeau, and Patrick McDaniel. 2014. FlowDroid: Precise Context, Flow, Field, Object-Sensitive and Lifecycle-Aware Taint Analysis for Android Apps. *SIGPLAN Not.* 49, 6 (June 2014), 259–269. https://doi.org/10.1145/2666356.2594299

[6] Daniel Balouek, Alexandra Carpen Amarie, Ghislain Charrier, Frédéric Desprez, Emmanuel Jeannot, Emmanuel Jeanvoine, Adrien Lèbre, David Margery, Nicolas Niclausse, Lucas Nussbaum, Olivier Richard, Christian Pérez, Flavien Quesnel, Cyril Rohr, and Luc Sarzyniec. 2013. Adding Virtualization Capabilities to the Grid'5000 Testbed. In *Cloud Computing and Services Science*, Ivan I. Ivanov, Marten van Sinderen, Frank Leymann, and Tony Shan (Eds.). Communications in Computer and Information Science, Vol. 367. Springer International Publishing, 3–20. https://doi.org/10.1007/978-3-319-04519-1_1

[7] Earl T Barr, Mark Harman, Phil McMinn, Muzammil Shahbaz, and Shin Yoo. 2015. The Oracle Problem in Software Testing: A Survey. *IEEE Transactions on Software Engineering* 41, 5 (2015), 507–525.

[8] Nataniel P. Borges Jr., Jenny Hotzkow, and Andreas Zeller. 2018. DroidMate-2: A Platform for Android Test Generation. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE 2018)*. ACM, New York, NY, USA, 916–919. https://doi.org/10.1145/3238147.3240479

[9] Paolo Calciati, Konstantin Kuznetsov, Xue Bai, and Alessandra Gorla. 2018. What Did Really Change with the New Release of the App?. In *Proceedings of the 15th International Conference on Mining Software Repositories (MSR '18)*. Association for Computing Machinery, New York, NY, USA, 142–152. https://doi.org/10.1145/3196398.3196449

[10] José Campos, Andrea Arcuri, Gordon Fraser, and Rui Abreu. 2014. Continuous Test Generation: Enhancing Continuous Integration with Automated Test Generation. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering (ASE '14)*. Association for Computing Machinery, New York, NY, USA, 55–66. https://doi.org/10.1145/2642937.2643002

[11] Wontae Choi, Koushik Sen, George Necula, and Wenyu Wang. 2018. DetReduce: Minimizing Android GUI Test Suites for Regression Testing. In *Proceedings of the 40th International Conference on Software Engineering (ICSE '18)*. Association for Computing Machinery, New York, NY, USA, 445–455. https://doi.org/10.1145/3180155.3180173

[12] Shauvik Roy Choudhary, Alessandra Gorla, and Alessandro Orso. 2015. Automated Test Input Generation for Android: Are We There Yet? (E). In *Proceedings of the 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE) (ASE '15)*. IEEE Computer Society, Washington, DC, USA, 429–440. https://doi.org/10.1109/ASE.2015.89

[13] Deloitte. [n.d.]. 2018 Global Mobile Consumer Survey: US Edition. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/technology-media-telecommunications/us-tmt-global-mobile-consumer-survey-exec-summary-2018.pdf.

[14] Erik Derr, Sven Bugiel, Sascha Fahl, Yasemin Acar, and Michael Backes. 2017. Keep me Updated: An Empirical Study of Third-Party Library Updatability on Android. In *Proceedings of the 24th ACM Conference on Computer and Communication Security (CCS 2017)*. ACM.

[15] Quan Do, Guowei Yang, Meiru Che, Darren Hui, and Jefferson Ridgeway. 2016. Regression Test Selection for Android Applications. In *Proceedings of the International Conference on Mobile Software Engineering and Systems (MOBILESoft '16)*. Association for Computing Machinery, New York, NY, USA, 27–28. https://doi.org/10.1145/2897073.2897127

[16] Quan Chau Dong Do, Guowei Yang, Meiru Che, Darren Hui, and Jefferson Ridgeway. 2016. Redroid: A Regression Test Selection Approach for Android Applications. In *Procedings of the 28th International Conference on Software Engineering and Knowledge Engineering (SEKE 2016)*. 486–491. https://doi.org/10.18293/SEKE2016-223

[17] Daniel Domínguez-Álvarez and Alessandra Gorla. 2019. Release Practices for IOS and Android Apps. In *Proceedings of the 3rd ACM SIGSOFT International Workshop on App Market Analytics (WAMA 2019)*. Association for Computing Machinery, New York, NY, USA, 15–18. https://doi.org/10.1145/3340496.3342762

[18] Gordon Fraser and Andrea Arcuri. 2011. EvoSuite: Automatic Test Suite Generation for Object-Oriented Software. In *Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering (ESEC/FSE '11)*. Association for Computing Machinery, New York, NY, USA, 416–419. https://doi.org/10.1145/2025113.2025179

[19] Z. Gao, C. Fang, and A. M. Memon. 2015. Pushing the limits on automation in GUI regression testing. In *2015 IEEE 26th International Symposium on Software Reliability Engineering (ISSRE)*. 565–575. https://doi.org/10.1109/ISSRE.2015.7381848

[20] Tianxiao Gu. [n.d.]. MiniTracing, APE coverage tool. http://gutianxiao.com/ape/install-mini-tracing. Last visited: 07/07/2020.

[21] Tianxiao Gu, Chengnian Sun, Xiaoxing Ma, Chun Cao, Chang Xu, Yuan Yao, Qirun Zhang, Jian Lu, and Zhendong Su. 2019. Practical GUI Testing of Android Applications via Model Abstraction and Refinement. In *Proceedings of the 41st International Conference on Software Engineering (ICSE '19)*. IEEE Press, 269–280. https://doi.org/10.1109/ICSE.2019.

00042

[22] Mouna Hammoudi, Gregg Rothermel, and Andrea Stocco. 2016. WATERFALL: An Incremental Approach for Repairing Record-Replay Tests of Web Applications. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2016)*. Association for Computing Machinery, New York, NY, USA, 751–762. https://doi.org/10.1145/2950290.2950294

[23] Google Inc. 2020. Animator | Android Developers. https://developer.android.com/reference/kotlin/android/animation/Animator. Last visited: 11/11/2020.

[24] CNRS INRIA. [n.d.]. Grid5000 infrastructure. https://www.grid5000.fr. Last visited: 07/07/2020.

[25] Gunel Jahangirova, David Clark, Mark Harman, and Paolo Tonella. 2016. Test Oracle Assessment and Improvement. In *Proceedings of the 25th International Symposium on Software Testing and Analysis (ISSTA 2016)*. Association for Computing Machinery, New York, NY, USA, 247–258. https://doi.org/10.1145/2931037.2931062

[26] Jamendo. 2020. Music Streaming App. https://www.jamendo.com/.

[27] Bogdan Korel and Ali M. Al-Yami. 1998. Automated Regression Test Generation. In *Proceedings of the 1998 ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '98)*. Association for Computing Machinery, New York, NY, USA, 143–152. https://doi.org/10.1145/271771.271803

[28] K. Kuznetsov, V. Avdiienko, A. Gorla, and A. Zeller. 2018. Analyzing the User Interface of Android Apps. In *2018 IEEE/ACM 5th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. 84–87.

[29] Bixin Li, Xiaobing Sun, Hareton Leung, and Sai Zhang. 2013. A survey of code-based change impact analysis techniques. *Software Testing, Verification and Reliability* 23, 8 (2013), 613–646. https://doi.org/10.1002/stvr.1475 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/stvr.1475

[30] X. Li, N. Chang, Y. Wang, H. Huang, Y. Pei, L. Wang, and X. Li. 2017. ATOM: Automatic Maintenance of GUI Test Scripts for Evolving Mobile Applications. In *2017 IEEE International Conference on Software Testing, Verification and Validation (ICST)*. 161–171.

[31] M. Linares-Vasquez, K. Moran, and D. Poshyvanyk. 2017. Continuous, Evolutionary and Large-Scale: A New Perspective for Automated Mobile App Testing. In *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 399–410. https://doi.org/10.1109/ICSME.2017.27

[32] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.

[33] Ke Mao, Mark Harman, and Yue Jia. 2016. Sapienz: Multi-objective Automated Testing for Android Applications. In *Proceedings of the 25th International Symposium on Software Testing and Analysis (ISSTA 2016)*. ACM, New York, NY, USA, 94–105. https://doi.org/10.1145/2931037.2931054

[34] Stuart Mcilroy, Nasir Ali, and Ahmed E. Hassan. 2016. Fresh Apps: An Empirical Study of Frequently-Updated Mobile Apps in the Google Play Store. *Empirical Softw. Engg.* 21, 3 (June 2016), 1346–1370. https://doi.org/10.1007/s10664-015-9388-2

[35] Chanh Duc Ngo, Fabrizio Pastore, and Lionel Briand. 2020. ATUA replicability package. hidden for now. Last visited: 11/11/2020.

[36] M. Pan, T. Xu, Y. Pei, Z. Li, T. Zhang, and X. Li. 2019. GUI-Guided Repair of Mobile Test Scripts. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. 326–327.

[37] Minxue Pan, Tongtong Xu, Yu Pei, Zhong Li, Tian Zhang, and Xuandong Li. 2020. GUI-Guided Test Script Repair for Mobile Apps. *IEEE Transactions on Software Engineering* 5589, c (2020), 1–1. https://doi.org/10.1109/tse.2020.3007664

[38] F. Pastore, L. Mariani, and G. Fraser. 2013. CrowdOracles: Can the Crowd Solve the Oracle Problem?. In *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation*. 342–351.

[39] Fabrizio Pastore, Leonardo Mariani, Alberto Goffi, Manuel Oriol, and Michael Wahler. 2012. Dynamic Analysis of Upgrades in C/C++ Software. In *Proceedings of the 2012 IEEE 23rd International Symposium on Software Reliability Engineering (ISSRE '12)*. IEEE Computer Society, USA, 91–100. https://doi.org/10.1109/ISSRE.2012.9

[40] Fabrizio Pastore, Leonardo Mariani, Antti E. J. Hyvärinen, Grigory Fedyukovich, Natasha Sharygina, Stephan Sehestedt, and Ali Muhammad. 2014. Verification-Aided Regression Testing. In *Proceedings of the 2014 International Symposium on Software Testing and Analysis (ISSTA 2014)*. Association for Computing Machinery, New York, NY, USA, 37–48. https://doi.org/10.1145/2610384.2610387

[41] Bill Phillips and Brian Hardy. 2013. *Android Programming: The Big Nerd Ranch Guide* (1st ed.). Big Nerd Ranch.

[42] Andreas Rau, Jenny Hotzkow, and Andreas Zeller. 2018. Efficient GUI Test Generation by Learning from Tests of Other Apps. In *Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings (ICSE '18)*. Association for Computing Machinery, New York, NY, USA, 370–371. https://doi.org/10.1145/3183440.3195014

[43] Gregg Rothermel and Mary Jean Harrold. 1997. A Safe, Efficient Regression Test Selection Technique. *ACM Trans. Softw. Eng. Methodol.* 6, 2 (April 1997), 173–210. https://doi.org/10.1145/248233.248262

[44] Atanas Rountev and Dacong Yan. 2014. Static Reference Analysis for GUI Objects in Android Software. In *Proceedings of Annual IEEE/ACM International Symposium on Code Generation and Optimization (CGO '14)*. Association for Computing

Machinery, New York, NY, USA, 143–153. https://doi.org/10.1145/2581122.2544159

[45] K. Rubinov and L. Baresi. 2018. What Are We Missing When Testing Our Android Apps? *Computer* 51, 4 (April 2018), 60–68. https://doi.org/10.1109/MC.2018.2141024

[46] Barbara G. Ryder and Frank Tip. 2001. Change Impact Analysis for Object-Oriented Programs. In *Proceedings of the 2001 ACM SIGPLAN-SIGSOFT Workshop on Program Analysis for Software Tools and Engineering (PASTE '01)*. Association for Computing Machinery, New York, NY, USA, 46–53. https://doi.org/10.1145/379605.379661

[47] Ripon K. Saha, Lingming Zhang, Sarfraz Khurshid, and Dewayne E. Perry. 2015. An Information Retrieval Approach for Regression Test Prioritization Based on Program Changes. In *Proceedings of the 37th International Conference on Software Engineering - Volume 1 (ICSE '15)*. IEEE Press, 268–279.

[48] S. Shamshiri, G. Fraser, P. Mcminn, and A. Orso. 2013. Search-Based Propagation of Regression Faults in Automated Regression Testing. In *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation Workshops*. 396–399. https://doi.org/10.1109/ICSTW.2013.51

[49] Aman Sharma and Rupesh Nasre. 2019. QADroid: Regression Event Selection for Android Applications. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2019)*. Association for Computing Machinery, New York, NY, USA, 66–77. https://doi.org/10.1145/3293882.3330550

[50] F. Song, Z. Xu, and F. Xu. 2017. An XPath-Based Approach to Reusing Test Scripts for Android Applications. In *2017 14th Web Information Systems and Applications Conference (WISA)*. 143–148.

[51] STH blog editors. [n.d.]. 10 Most Popular Crowdsourced Testing Companies in 2020. https://www.softwaretestinghelp.com/crowdsourced-testing-companies/. Last visited: 07/07/2020.

[52] Ting Su, Guozhu Meng, Yuting Chen, Ke Wu, Weiming Yang, Yao Yao, Geguang Pu, Yang Liu, and Zhendong Su. 2017. Guided, Stochastic Model-based GUI Testing of Android Apps. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017)*. ACM, New York, NY, USA, 245–256. https://doi.org/10.1145/3106237.3106298

[53] L. D. Toffola, C. Staicu, and M. Pradel. 2017. Saying 'Hi!' is not enough: Mining inputs for effective test generation. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 44–49. https://doi.org/10.1109/ASE.2017.8115617

[54] Porfirio Tramontana, Domenico Amalfitano, Nicola Amatucci, and Anna Rita Fasolino. 2019. Automated functional testing of mobile applications: a systematic mapping study. *Software Quality Journal* 27, 1 (2019), 149–201. https://doi.org/10.1007/s11219-018-9418-6

[55] András Vargha and Harold D. Delaney. 2000. A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* 25, 2 (2000), 101–132. https://doi.org/10.3102/10769986025002101 arXiv:https://doi.org/10.3102/10769986025002101

[56] Wenyu Wang, Dengfeng Li, Wei Yang, Yurui Cao, Zhenwen Zhang, Yuetang Deng, and Tao Xie. 2018. An Empirical Study of Android Test Generation Tools in Industrial Cases. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE 2018)*. ACM, New York, NY, USA, 738–748. https://doi.org/10.1145/3238147.3240465

[57] Yan Wang, Hailong Zhang, and Atanas Rountev. 2016. On the Unsoundness of Static Analysis for Android GUIs. In *Proceedings of the 5th ACM SIGPLAN International Workshop on State Of the Art in Program Analysis (SOAP 2016)*. Association for Computing Machinery, New York, NY, USA, 18–23. https://doi.org/10.1145/2931021.2931026

[58] M. Wen, R. Wu, and S. Cheung. 2016. Locus: Locating bugs from software changes. In *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 262–273.

[59] Haowei Wu, Hailong Zhang, Yan Wang, and Atanas Rountev. 2019. Sentinel: generating GUI tests for sensor leaks in Android and Android wear apps. *Software Quality Journal* (2019). https://doi.org/10.1007/s11219-019-09484-z

[60] Minzhi Yan, Hailong Sun, and Xudong Liu. 2014. ITest: Testing Software with Mobile Crowdsourcing. In *Proceedings of the 1st International Workshop on Crowd-Based Software Development Methods and Technologies (CrowdSoft 2014)*. Association for Computing Machinery, New York, NY, USA, 19–24. https://doi.org/10.1145/2666539.2666569

[61] Shengqian Yang, Haowei Wu, Hailong Zhang, Yan Wang, Chandrasekar Swaminathan, Dacong Yan, and Atanas Rountev. 2018. Static Window Transition Graphs for Android. *International Journal of Automated Software Engineering* 25, 4 (Dec. 2018), 833–873.

[62] Shengqian Yang, Hailong Zhang, Haowei Wu, Yan Wang, Dacong Yan, and Atanas Rountev. 2015. Static Window Transition Graphs for Android. In *IEEE/ACM International Conference on Automated Software Engineering*. 658–668.

[63] K. C. Youm, J. Ahn, J. Kim, and E. Lee. 2015. Bug Localization Based on Code Change Histories and Bug Reports. In *2015 Asia-Pacific Software Engineering Conference (APSEC)*. 190–197.

[64] R. N. Zaeem, M. R. Prasad, and S. Khurshid. 2014. Automated Generation of Oracles for Testing User-Interaction Features of Mobile Apps. In *2014 IEEE Seventh International Conference on Software Testing, Verification and Validation*. 183–192. https://doi.org/10.1109/ICST.2014.31

[65] J. Zhou, H. Zhang, and D. Lo. 2012. Where should the bugs be fixed? More accurate information retrieval-based bug localization based on bug reports. In *2012 34th International Conference on Software Engineering (ICSE)*. 14–24.