applications, we can exercise similar (malicious) behaviors.

We design and implement PuppetDroid, an Android environment that supports both *manual application testing* (through a physical device or an emulator), to collect new stimulation traces, and *automatic application exercising*, which cleverly leverages previously recorded UI stimulation traces. PuppetDroid relies on the screenshots of an app to find similar apps, indexed using a fast metric known as perceptual hashing.

To evaluate PuppetDroid, first, we experimentally verified that manual exercising allows to stimulate malicious behaviors better than automatic techniques. Second, we validated our approach on 7,000 applications and found out that it can stimulate 12–24% more behaviors than state-of-the-art techniques. Interestingly, our system is able to unveil those corner behaviors that are difficult to exercise with a fully automatic approach (i.e., download an APK and execute it). Then, we show that our UI similarity technique is precise, scales, and has modest resource requirements. In summary:

- we propose a novel and orthogonal approach to exercise more behaviors during dynamic analysis of (malicious) mobile applications. Our approach is the first that takes the end users into play.
- We propose an original method to automatically exercise the UI of an unknown application re-using UI stimulation traces obtained from previously analyzed applications that present a similar layout.
- We implemented and evaluated our approach to demonstrate its feasibility and, more importantly, its effectiveness. Remarkably, we manually verified the outcome of each experiment.

2. BACKGROUND AND MOTIVATION

Many approaches have been proposed to analyze applications with the final goal of designing effective detection criteria [32, 17, 24, 6, 28, 38, 35, 14, 22]. To this end, program-analysis techniques used for traditional malware have been ported to Android (e.g., dynamic analysis, static analysis, taint tracking, symbolic execution), with their well-known, symmetric pros and cons. Static approaches can be hindered by obfuscated code, repackaging or dynamic payloads, two techniques widely used by modern malware. Symbolic execution (e.g., [33]) is promising yet resource intensive.

In spite of its efficiency and semantic richness, the main inherent limitation of dynamic analysis is its inability to obtain satisfactory code coverage: Dynamic analysis can examine the actions performed in an execution path only if that path is actually explored. This problem is particularly concerning because if a malware sample is not properly exercised, it may not expose its malicious behavior at all. Exercising mobile applications in a proper way, however, is not trivial, because of the highly interactive UI, which makes automatic exercising even harder than in conventional desktop scenarios.

State-of-the-art dynamic analysis approaches (e.g., [24]) incorporate automatic code-exercising and stimulation techniques. Other approaches leverage stress-test tools (e.g., Monkey [13]) or program analysis (e.g., SmartDroid [34], ActEVE [2]). Stress-test tools rely on pseudo-random generation of UI input events, which is simple to implement, but rather ineffective, since randomly stimulating UI elements displayed on the screen can hardly reproduce the typical usage of users. Approaches such as SmartDroid leverage static analysis to reconstruct the semantic of UI elements on the screen, and to find execution paths that expose malicious behaviors. Unfortunately, static analysis is ineffective against obfuscated samples, and the research tools will need a major upgrade with the introduction of new Android 4.4 runtime.

From our overview of the state of the art and related work in Section 7, we notice that previous work does not consider how the UI is exercised by a user. Interestingly, our experiments in Section 5.1 confirm our intuition that a human user is able to exercise certain behaviors that the state of the art code stimulation approach [24] fail to unveil.

Given the above motivations, we conclude that to take dynamic analysis of Android applications a step further, we need an orthogonal approach to stimulate the UI.

3. GOALS AND APPROACH OVERVIEW

Our first goal is to provide a sandboxed environment to safely perform manual tests on malicious applications and, at the same time, record user interaction with the UI of the application. Our second goal is to automatically exercise unknown applications, leveraging stimulation traces previously recorded on similar applications.

Our approach is to let applications run on a remote sandbox while users seamlessly interact with their UI as if they were running locally on their devices. More precisely, in **Phase 1 (Recording of stimulation traces)**, each sandbox uses a remote framebuffer protocol to collect *stimulation traces*, which represent the sequence of UI events performed by the user, as well as the list of UI elements actually stimulated during the test. Differently from previous work (e.g., [12]), we go beyond recording raw events from `/dev/input` and re-injecting them to another input device: We correlate such events to the respective UI elements (e.g., buttons, or other view objects), and collect information about the behaviors exhibited by the exercised applications, through dynamic analysis. As described in Section 4, this entails some challenges that we need to solve. From hereinafter, a *behavior* is a sequence of observable runtime events (e.g., system calls, API calls).

Manual stimulation on large datasets is clearly unfeasible. Thus, in **Phase 2 (Re-execution of stimulation traces)**, we leverage the collected stimulation traces to automatically exercise new applications, so as to increase the code covered during dynamic analysis. Our hypothesis is that by re-using stimulation traces we obtain better results (in terms of discovered behaviors) than with random UI exercisers. A naïve approach where we blindly try to exercise an application with *every* stimulation trace in our system is not accurate or efficient. Therefore, in **Phase 3 (Finding Similar Applications)**, we leverage the concept of *UI application similarity*. As summarized by the workflow in Figure 1, when a new sample is to be analyzed, we first look for similar (or equivalent) samples for which we have a stimulation trace. Then, we use only stimulation traces of the most similar known application. With our approach, calculating the similarity between two applications takes constant time and memory, whereas finding the most similar application to a given one, in a database of N applications, takes logarithmic time.

One could argue that the need of collecting a set of stimulation traces large enough to be useful may limit the scalability of our approach. However, we consider two factors. First, our system could attract the interest not only of security analysts, but also of normal users that want to safely try

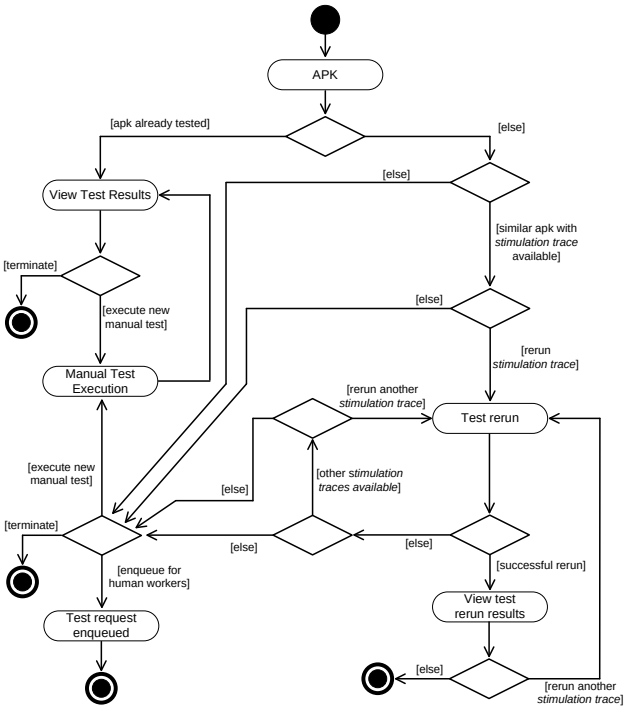


Figure 1: Workflow of our approach (Section 4).

potentially malicious applications they find on the web or in alternative markets. Secondly, we can leverage the accessibility of crowdsourcing services, like Amazon Mechanical Turk, to recruit human workers and generate new stimulation traces efficiently (being this a matter of engineering and deployment, we focus on our novel approach)

4. SYSTEM DETAILS

4.1 Phase 1: Recording of stimulation traces

We record the low-level input events generated while the user interacts with an application on his or her device. We developed an extended VNC client and server architecture through which this process happens transparently, with no changes in the way users interact with the UI of an application. For the client, we extended TightVNC, whereas we implemented the server on top of the Fastdroid libraries[2].

We translate the input events in a sequence of remote framebuffer (RFB)[3] *PointerEvent* or *KeyEvent* messages that are sent to the VNC server. For each event, we save the `timestamp` (according to the client), `event_type` (0 for touch events and 1 for keys), `action` (0 is "up", 1 is "down", 2 is "move"), `x_pos,y_pos` coordinates on the screen, and `key_code` (pressed button). Figure 2 shows an excerpt of a sample input events file generated by the VNC server.

Two similar applications may have some subtle UI differences that can make a re-execution test fail (e.g., slightly shifted buttons). Taking for example two distinct BaseBridge samples[4] from the Malware Genome Project [36], we notice that the main button of the second sample is slightly shifted. Exercising the second sample with the sequence of raw input

[2]`https://code.google.com/p/fastdroid-vnc/`
[3]`https://tools.ietf.org/rfc/rfc6143.txt`
[4]MD5s: 00c154b42fd483196d303618582420b89cedbf46, 73bb65b2431fefd01e0ebe66582a40e74928e053

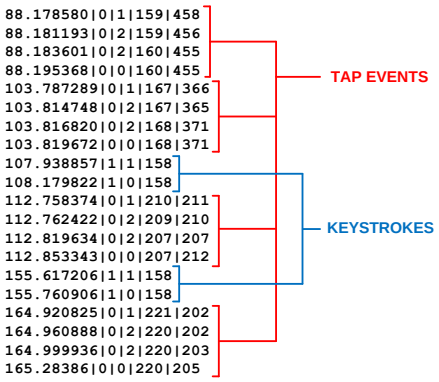


Figure 2: Excerpt of a sample input events file.

events recorded on the first sample (as one would do by using, for example, the approach in [12]), would cause an error.

To solve this, during recording we keep track of which view object (e.g., button identifier) consumed each input event during recording, in order to find that same view object during re-execution. For this, we rely on the `ViewServer`, which allows to "walk" the hierarchy[5] of displayed objects. More precisely, our VNC server performs the following steps when a new input event is received:

1. Process RFB *PointerEvent* message.
2. Send the `GET_FOCUS` command to the `ViewServer`, to get the name and hash code of the focused window (i.e., `Activity`).
3. Retrieve view hierarchy of the window sending `DUMPQ` command to `ViewServer`.
4. Search view hierarchy for the deepest-rightmost view object containing the coordinates of the input event.
5. Store the paths to the previously found view nodes.

We combine the collected information to extract the list of paths to the views that actually consumed the touch events. Moreover, as motivated in **Phase 2**, in case of touch events we log which activity has consumed each event, and the path to all the deepest nodes in the hierarchy that can consume the touch event. By combining this information with the coordinates of the touch events generated by the user we build the sequence of view objects stimulated. We also retain the relative position of the input event with respect to the view object, which is useful in **Phase 2**.

4.2 Phase 2: Re-execution of stimulation traces

For each event in the recorded sequence, we use the view object and ratio information (see Figure 3) to properly re-

[5]`http://developer.android.com/guide/topics/ui`

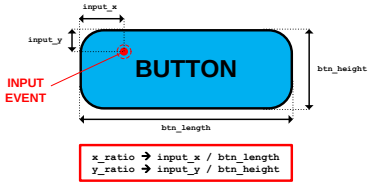


Figure 3: Input event relative position with respect to view object.

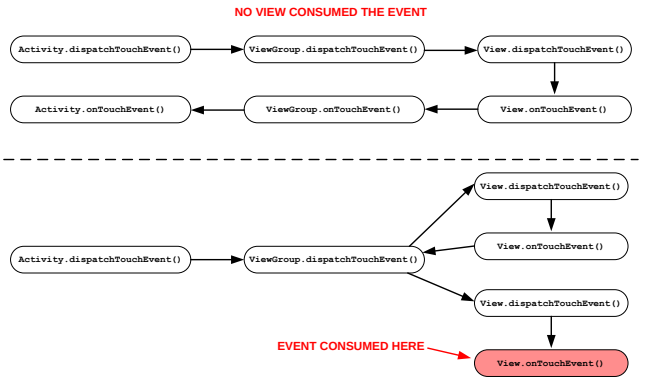


Figure 4: Examples of touch event management.

scale the horizontal and vertical coordinates. Then, we write the resulting event into the `/dev/input` device.

We treat touch events with special care to avoid the following rare corner case, which can occur if the view that receives the input event is not the view that eventually consumes it. More precisely, when a touch event is handled by the Android Touch System [26], the `Activity.dispatchTouchEvent()` method of the currently running Activity is called. This method dispatches the event to the root view in the hierarchy and waits for the result: If no view consumes the event, the Activity calls `onTouchEvent()` in order to consume itself the event before terminating. When a view object receives a touch event, the `View.dispatchTouchEvent()` is called: This method first tries to find an attached listener to consume the event, calling `View.OnTouchListener.onTouch()`, then tries to consume the event itself calling `View.onTouchEvent()`. If neither there is a listener nor the `onTouchEvent()` method is implemented, the event is not consumed and it flows back to the parent. When a `ViewGroup` receives a touch event, it iterates on its children views in reverse order and, if the touch event is inside the view, it dispatches the event to the child. If the event is not consumed by the child, it continues to iterate on its children until a view consumes the event. If the event is not consumed by any of its children, the `ViewGroup` acts as a `View` and tries to consume itself the event. Eventually, if it is not able to consume the event it sends back to the parent. Figure 4 shows two examples of touch events management: In the former, the event flows down through the hierarchy, and since it is not consumed by any view, it goes back to the `Activity`. In the latter, the event is consumed by the second `View` child of the `ViewGroup` object.

Our system avoids this corner case because it recorded, during **Phase 2**, which activity has consumed each event, and the path to all the deepest nodes in the hierarchy that can consume the touch event.

4.3 Phase 3: Finding Similar Applications

In case a stimulation trace for an application A is not available, after searching by MD5, we rely on visual similarity to find similar applications.

The goal of this phase is to find an application, B, for which a stimulation trace exists, and that has a UI similar to As. To this end we leverage the concept of visual similarity, implemented through perceptual hashing. Given an image in input, a perceptual hashing algorithm creates a metric fingerprint that is robust to image re-scaling, rotation, deformation, skew and compression. Thus, if two images are visually similar, their respective hashes, which are 64-bits unsigned integers, are very close. In particular, perceptually similar images have a hamming distance within bounds that can be reliably estimated[6], as we also show in Section 5.3.

In practice, to lookup a suitable stimulation trace for application A, we calculate the perceptual hash of its screenshots. Then, we look for B, an application which screenshots minimize the hamming distance from A's screenshots according to their respective perceptual hash. We pre-calculate the hashes of the known applications offline (which takes only 5.030453ms on average), and index them in a MVP tree [3], which allows lookup in logarithmic time.

If a screenshot is already available, which is very likely if the application is obtained from a market (e.g., screenshots are part of the app's metadata), our our system calculates the perceptual hash using the `ph_dct_imagehash` function of the `libphash` library. In case no screenshot is available, our system instantiates an emulator, installs the APK of A and leverages the `screencap` utility to take a screenshot once the application has started.

4.4 Implementation Details

The actual execution of the target application happens on the *server tier*, which receives the UI events, and records and proxies them to an instrumented Android Virtual Device (AVD) with the same screen size of the client. AVDs are concurrently instantiated for each new client. A VNC server instance is connected to each AVD screen to record the stimulation traces. For **Phase 2** the life cycle is almost the same, with the only differences that, instead of connecting VNC server, we inject the re-scaled and adapted input events into the running AVD. The devices associated to the touchscreen and to the keyboard are fixed and respectively are `/dev/input/event1` and `/dev/input/event2`.

We have tested PuppetDroid with the original AVD and DroidBox [28]. For our experiments we obtained access to CopperDroid[7] [24], which allows automatic dynamic- and stimulation-based behavioral analysis.

We patched the `ViewServer` [1] and stripped it down so as to collect only data useful to our purposes. This resulted in a 20–40x speedup over the original implementation.

5. EXPERIMENTAL EVALUATION

Our results shows that both manual exercising and re-execution of collected stimulation traces reach higher code coverage than the one obtained with automatic UI exercisers: We succeeded in stimulating more than the amount of behaviors stimulated by other exercising strategies. Moreover, we found some particular cases in which PuppetDroid succeeds in stimulating interesting malicious behaviors that are not exposed using automatic application exercising approaches.

In **Experiment 1** we verified that our stimulation approach led to a better stimulation compared to other automatic analysis approaches. For this, we compared the number of behaviors exercised with PuppetDroid vs. the ones exercised with automatic approaches (i.e., monkey) typically used in dynamic malware analysis frameworks, and vs. the system events stimulation strategy proposed in CopperDroid [24]. In **Experiment 2** we verified that the same stimulation trace can be reused on similar samples to exercise the same behaviors. For this, we compared the behaviors exercised

[6] `http://phash.org/docs/design.html`
[7] `http://copperdroid.isg.rhul.ac.uk`

on manually-stimulated APKs with the behaviors exercised on similar samples, and verified the outcome manually. In **Experiment 3** we verified that our UI similarity approach is accurate and efficient.

For dynamic analysis, we obtained access to the CopperDroid sandbox, which is convenient for our needs because (1) works at system-call, (2) incorporates a state-of-the-art stimulation approach, able to stimulate both statically and dynamically registered broadcast receivers, and (3) already provides a list of behaviors built by means of system calls.

5.1 Experiment 1: Manual UI exercising

5.1.1 Dataset

We used 15 APKs samples, 13 from the Android Malware Genome Project [36], and 2 from the Google Play store. The dataset is purposely small, because we performed multiple tests on each sample and manually inspected the output of each test in order to examine precisely the differences between different approaches. Therefore, we preferred focusing on a small dataset to perform a deeper analysis of each test result.

5.1.2 Experimental setup and procedure

For each sample in our dataset, we collect the system call traces during execution with four stimulation approaches:

- **NoStim**: without stimulation,
- **Copper**: with CopperDroid stimulation strategy,
- **Monkey** stimulation with an increasing number of input events (500, 1000, 2000, 5000).
- **Our**: an everyday Android user exercised the sample through PuppetDroid, without knowing the outcome of the other tests. We instructed the user to rely on his sole knowledge and try to use the application naturally, following anything the application asks, without thinking whether the action is dangerous or not.

For each couple of stimulation approaches A and B, we calculate the total stimulated behaviors by A and B, and the behaviors stimulated only by either A or B (set difference). In each test, we start a clean sandbox, install the APK sample, run it with the selected stimulation approach, and collect the system calls traces and the behavior lists.

5.1.3 Results

We calculated the average number of behaviors (total and distinct) observed with each stimulation approach, and the average number of distinct missed behaviors by each strategy (set difference). We repeated the experiment on the entire dataset, and then on each set of goodware and malware applications.

Figure 5 summarizes the results. From the bar chart on the top-left corner, we can see that the human-driven stimulation succeeds in stimulating more behaviors than any automatic approaches: we are able to exercise 112% of total behaviors and 124% of distinct behaviors more than the automatic stimulation. The same result holds regardless of whether the application is malicious or benign. The bar chart on the bottom-left corner confirms the above results regardless of whether the applications are benign (striped bars) or malicious (solid bars).

From the bar chart on the top-right corner, we can see that the other approaches miss many behaviors with respect

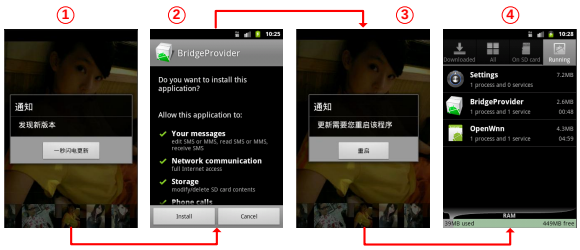


Figure 6: Experiment 1 (Case analysis): Steps to install BridgeProvider payloads: 1) Ask for application update; 2) Install payload; 3) Restart application; 4) Malicious service running on device.

to ours, whereas our technique misses a negligible amount of behaviors. From another perspective, PuppetDroid is able to stimulate 593% exclusive behaviors more in respect to monkey and 200% more i respect to the state of the art (CopperDroid). The bar chart on the bottom-right corner confirms the above results regardless of whether the applications are benign (striped bars) or malicious (solid bars). We analyze the results on malware and goodware samples separately.

Overall, our results confirm our hypothesis that PuppetDroid UI stimulation approach allows to obtain better results than automatic approaches during dynamic analysis.

Case analysis. A notable case that deserves detailed analysis is that of a malicious behavior exercised, and thus exposed, exclusively by our system. The malware sample under analysis is `com.keji.danti80`, belonging to BaseBridge malware family. BaseBridge is a trojan that, once installed, prompts the user with an upgrade dialog: if users accept to do so, the malware will install a second malicious application on the phone. This service communicates with a control server to receive instructions to perform unwanted activities (e.g., place calls or send messages to premium numbers). Meanwhile, the malware also blocks messages from the mobile carrier in order to prevent users from getting fee consumption updates: in this way all malicious activities are undertaken stealthily without the users' knowledge or consent. More details on this malware can be found in [21, 20, 27]. This case is very common and is used by malware authors to circumvent dynamic analysis with ineffective UI exercising.

Analyzing the sample with PuppetDroid we obtained the list of behaviors shown in Table 1. The underlined lines indicate a behavior that none of the other stimulation techniques were able to reveal. The malware writes another APK file, `xxx.apk`, on the filesystem. As a matter of fact, during the test, the application prompts the user to install a new malicious application, named BridgeProvider, to complete the update, as shown in Figure 6.

In conclusions, other stimulation approaches did not exercise the malware enough to make it reveal its true malicious behavior, with the consequent risk to consider the sample as safe. Instead, using PuppetDroid, the analyst is able to detect such a potential dangerous behavior and subsequently analyze in detail the functioning of the application.

Conclusions of Experiment 1: The results confirmed our intuition that automatic UI stimulation approaches can only exercise a subset of the (malicious) behaviors of a malware during dynamic analysis. Moreover, PuppetDroid approach based on human-driven UI exercising allows to reproduce typical victim interaction with the malware and to reach

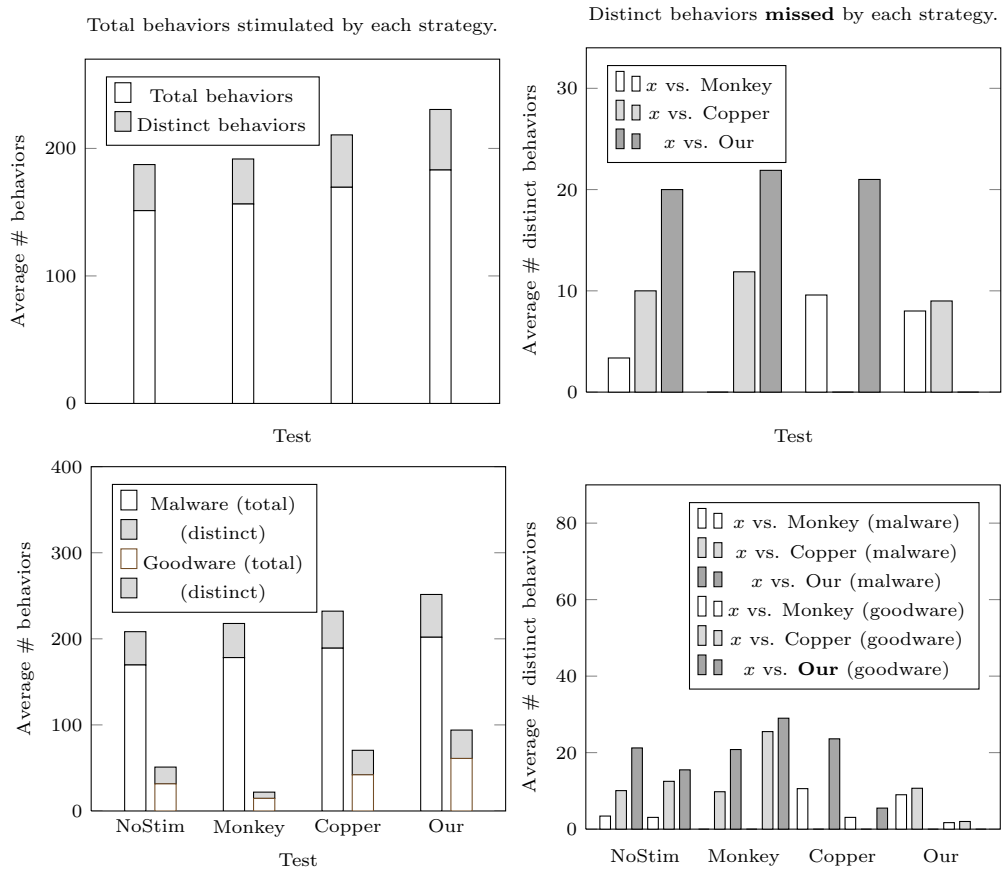


Figure 5: (left) Total behaviors stimulated on average by each strategy, and (right) behavior missed by each strategy. (bottom) breakdown of the above experiment on malware (m/w) and goodware (g/w).

Table 1: Experiment 1: List of behaviors extracted testing `com.keji.danti80` malware sample.

Hits	Beh.	Data
2	write	/sys/qemu_trace/process_name
2	write	/sys/qemu_trace/process_name
2	write	/sys/qemu_trace/process_name
1	mkdir	/data/com.keji.danti80/files
1	write	/data/com.keji.danti80/files/xxx.apk
2	write	/sys/qemu_trace/process_name
2	write	/sys/qemu_trace/process_name
1	mkdir	/data/com.sec.android.bridge/shared_prefs
1	unlink	/data/com.keji.danti80/files/xxx.apk
1	connect	host: 10.0.2.3, retval: 0, port: 53
1	ns_query	query_data: b3.8866.org. 1 1
1	connect	host: 221.5.133.18, retval: -115, port: 8080
1	write	/data/com.sec.android.bridge/shared_prefs/first_app_perfs.xml
3	unlink	/data/com.sec.android.bridge/shared_prefs/first_app_perfs.xml.bak
1	write	/data/com.keji.danti80/files/atemp.jpg
1	unlink	/data/com.keji.danti80/files/atemp.jpg
2	write	221.5.133.18 port:8080

Table 2: Experiment 2: Summary of the results (average values per test).

a) ManualVsRe-exec	
Manual test	201.85 (38.69 distinct)
Re-executed tests	230.20 (52.76 distinct)
Only in manual	24.00
Only in re-executed	25.00

c) AutomaticVsRe-exec				
Stimulated Behaviors				
No Stimulation	199.79 (39.73 distinct)			
Monkey	196.62 (39.61 distinct)			
CopperDroid	198.95 (41.84 distinct)			
Our	230.20 (52.76 distinct)			
Exclusive Behaviors				
Only in	**NoStim**	**Monkey**	**Copper**	**Our**
Monkey	4.35	0	8.37	8.82
Copper	6.54	8.05	0	7.74
Our	23.27	22.58	22.15	0

then higher code coverage.

5.2 Experiment 2: Automatic Re-execution

This experiment's goal is to verify the following novel hypothesis: If we succeed in exercising the (malicious) behaviors in a sample, the same stimulation should trigger behaviors in a similar sample.

In a preliminary experiment, we measured that the percentage of UI events successfully re-executed on a dataset of similar applications is 88.52%. This percentage is actually a conservative estimate. For example, suppose that we have a recording of a UI stimulation with 20 events: if PuppetDroid succeeds in re-executing 10 UI events but it is not able to find the correct view to inject the 11^{th} event, the re-execution is terminated. We then have a re-execution score of 50%.

In the reminder of this section we show the impact of such re-execution on the behaviors exposed during dynamic analysis.

5.2.1 Dataset

For this experimental evaluation we picked 13 malware

samples from the Android Malware Genome Project [36] used in **Experiment 1**, run **Phase 1** to record stimulation traces, and **Phase 3** to retrieve similar APKs from a repository of over 7,000 samples that also include Google Play and alternative markets.

5.2.2 Experimental setup and procedure

To verify if our re-execution approach is feasible, we need a way to evaluate the results of the re-executed tests. We follow four criteria:

a) ***ManualVsRe-exec:*** Compare the behaviors exercised with manual stimulation vs. the behaviors exercised automatically.

b) ***Re-execBehaviors:*** Verify if an interesting malicious behavior exhibited in the original sample is also exhibited during the re-execution on a similar application.

c) ***AutomaticVsRe-exec:*** Compare the behaviors exercised using automatic stimulation tools, as in **Experiment 1**, against the behaviors extracted during execution.

We structured each test as follows, for each of the 13 APK:

1. As in **Experiment 1**, we ask a user to manually test the application while PuppetDroid records UI stimulation during traces.
2. Search for the most similar APK using **Phase 3**.
3. Run **Phase 2** to automatically re-execute previously recorded UI stimulation on the similar application.
4. Test the similar application with automatic stimulation approaches:
   - **NoStim:** 1 test without stimulation,
   - **Monkey:** 20 tests using monkey,
   - **Copper:** 1 test using CopperDroid stimulation strategy.
5. Calculate the four evaluation criteria explained above.

For each sample in the dataset we performed one test.

5.2.3 Results

The results for ***a)*** and ***c)*** are summarized in Table 2 and are analyzed in the reminder of this section with the aid of Figure 7 and 8 respectively. The results for ***b)*** are analyzed in depth with the aid of Table 3 and a set of screenshots.

ManualVsRe-exec (Figure 7). We show the comparison between the total and distinct number of behaviors stimulated in the original, manual test vs. the average numbers of behaviors stimulated in re-executed tests. The rightmost plot shows the exclusive behaviors (i.e., those stimulated only during either strategy (manual and re-execution)).

One would expect behaviors extracted in the original test to be always more than those stimulated in re-executed tests. In some cases, this is not true, (e.g., in *Test3*) because we are comparing behaviors exercised in different, even if similar, applications: it is possible that an application similar to the one originally tested generates **more** behaviors even if less stimulated. For instance, when application A starts it generates 10 behaviors, whereas when application B, similar to A, starts it generates 20 behaviors. The same holds also for the UI stimulation, so clicking on a button of A we may obtain 2 behaviors, while clicking on the same button on B leads to 4 behaviors. Recall, however, that we are not simply counting the exercised behaviors: In this experiment we also evaluate *which* behaviors are exclusively exercised by each strategy.

Re-execBehaviors (Table 3). A notable case is that of a malicious behavior stimulated in the original sample, which is exercised during the re-execution on a similar application, too. We consider the application `com.keji.danti160`, belonging to BaseBridge malware family. We chose this sample because during the test it showed a behavior similar to the one shown by `com.keji.danti80`: when started, the application asks the user to update it and installs a malicious service, named `BridgeProvider`, on the phone. The list of behaviors extracted during the test is presented in Table 3 (left): underlined rows indicate the malicious actions executed by the application.

Scanning our sample repository with androsim, we found a sample, named `com.keji.danti161` very similar to `com.keji.danti160`. By re-executing the UI stimulation recorded with PuppetDroid on the application `com.keji.danti161`, we extracted the list of behaviors shown in Table 3 (right): underlined rows present the same malicious actions stimulated in the original test execution. This example illustrates that our approach can unveil behaviors hidden to otherwise automated tests.

AutomaticVsRe-exec (Figure 8). We now evaluate the stimulation obtained with re-execution compared with automatic stimulation approaches. Comparing the behaviors extracted from re-executed tests with the ones retrieved stimulating the same samples with Monkey and CopperDroid, we obtained the data shown in Figure 8. As we can see, the re-executed stimulation still allows to stimulate more behaviors than automatic approaches: in fact, using PuppetDroid re-execution, we are able to stimulate 116% of total behaviors and 130% of distinct behaviors more than automatic stimulation methodologies. Moreover, with re-execution, we stimulate 535% exclusive behaviors more than Monkey and 355% more than CopperDroid. It is also worth noting that this is a conservative estimate of re-execution effectiveness: as a matter of fact, our experimental data contain also cases in which re-execution promptly failed after test beginning.

Conclusions of Experiment 2. The results support our key intuition on the re-execution of UI stimulation traces: we demonstrated that if (malicious) behaviors are exercised during a manual test, it is quite likely that using the same stimulation over the UI of similar applications will lead them to show their behaviors during the analysis. Exercising the UI of an application with the re-execution of stimulation traces allows to expose more behaviors than automatic approaches.

5.3 Experiment 3: UI Similarity

This experiments' goal is to verify that using screenshot similarity as a mean to find apps with similar UI is a correct hypothesis. In addition, we verify that the approach of using perceptual hashing is time and memory efficient.

5.3.1 Dataset

We used a first dataset of screenshots that we created by executing one app at a time in an emulator and launching the `screencap` utility. We obtained 6,000 screenshots. This procedure took about 10 seconds per app, including the time required to install the APK. Considering that executing a full

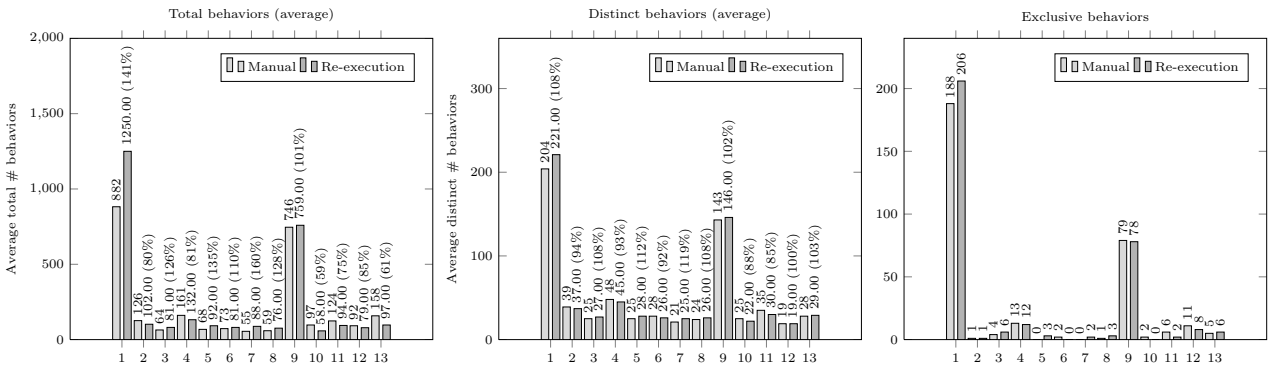


Figure 7: Experiment 2 (ManualVsRe-exec): Comparison of behaviors stimulated in the original, manual execution vs. the average total and distinct behaviors stimulated in re-executed tests. The third bar graph shows the exclusive behaviors exercised in either manual or re-executed tests.

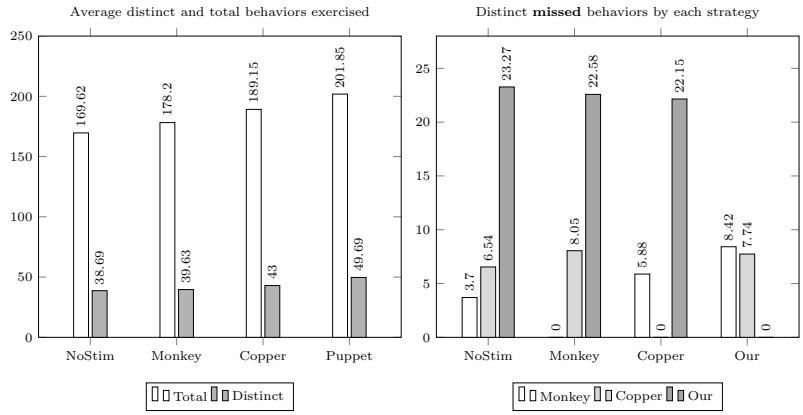


Figure 8: Experiment 2 (AutomaticVsRe-exec): Behaviors stimulated with re-execution in respect to behaviors extracted using automatic stimulation (left) and missed behaviors by each strategy (right).

Table 3: Experiment 2 (Re-execBehaviors): Behaviors found on `com.keji.danti160` (left) and re-executed automatically from PuppetDroid on a similar sample (right).

Original execution		Re-execution on a similar sample	
Data	**Beh. Data**	**Hits**	**Beh. Blob**
2	write /sys/qemu_trace/process_name	2	write /sys/qemu_trace/process_name
2	write /sys/qemu_trace/process_name	2	write /sys/qemu_trace/process_name
2	write /sys/qemu_trace/process_name	2	write /sys/qemu_trace/process_name
1	mkdir /data/com.keji.danti160/shared_prefs	1	mkdir /data/com.keji.danti161/shared_prefs
1	mkdir /data/com.keji.danti160/files	1	mkdir /data/com.keji.danti161/files
1	write /data/com.keji.danti160/files/xxx.apk	1	write /data/com.keji.danti161/files/xxx.apk
2	write /sys/qemu_trace/process_name	2	write /sys/qemu_trace/process_name
2	write /sys/qemu_trace/process_name	2	write /sys/qemu_trace/process_name
1	mkdir /data/com.sec.android.bridge/shared_prefs	2	mkdir /data/com.sec.android.bridge/shared_prefs
1	connect host: 10.0.2.3, retval: 0, port: 53	2	write /data/com.sec.android.bridge/shared_ prefs/first_app_perfs.xml
1	ns_query query_data: b3.8866.org. 1 1	1	unlink /data/com.sec.android.bridge/shared_ prefs/first_app_perfs.xml.bak
1	unlink /data/com.keji.danti160/files/xxx.apk	1	connect host: 10.0.2.3, retval: 0, port: 53
1	mkdir /data/com.keji.danti160/databases	1	ns_query query_data: b3.8866.org. 1 1
24	write /data/com.keji.danti160/databases/db.db	1	unlink /data/com.keji.danti161/files/xxx.apk
3	write /data/com.sec.android.bridge/shared_prefs/first_app_perfs.xml	1	connect host: 221.5.133.18, retval: -115, por t: 8080
2	unlink /data/com.sec.android.bridge/shared_prefs/first_app_perfs.xml.bak	1	mkdir /data/com.keji.danti161/databases
2	connect host: 221.5.133.18, retval: -115, por t: 8080	17	write /data/com.keji.danti161/shared_prefs/com.keji.danti161.xml
22	write /data/com.keji.danti160/shared_ prefs/com.keji.danti160.xml	16	unlink /data/com.keji.danti161/shared_prefs/com.keji.danti161.xml.bak
21	unlink /data/com.keji.danti160/shared_ prefs/com.keji.danti160.xml.bak	24	write /data/com.keji.danti161/databases/db.db
		4	write /data/system/dropbox/drop68.tmp

dynamic analysis in an instrumented environment takes time in the order of minutes, we consider this overhead negligible. We used also a second dataset of 16,000 screenshots that we obtained by crawling the blackmart[8] marketplace. The screenthots are part of each app metadata, as it happens in the majority of marketplaces. For both the datasets we saved the images in 8-bits JPG files at 288x480 to 319x480 square pixels (at most 221KB each).

5.3.2 *Experimental setup and results*

We ran our experiments on Xeon E5506 @ 2.13GHz with 6GB of RAM. We implemented **Phase 3** in C++ using the Boost UBLAS library, OpenMP, and the hamming distance `ph_hamming_distance` from the pHash library. In our experiments, we used 3 concurrent OpenMP threads to calculate the sparse distance matrix.

[8] `http://www.blackmartalpha.net/`

Table 4: Manual cluster analysis.

Classes	#clusters (%)	Homogeneity (%)	Avg. Size
Random sample from 420 clusters (6,000 dataset)			
1 (pure)	84 (85.71)	100.0	4.03
2	9 (9.1836)	67.30	5.11
3	2 (2.0408)	54.10	7.50
4	1 (1.0204)	86.10	9.00
5	1 (1.0204)	28.50	7.00
Random sample from 628 clusters (16,000 dataset)			
1 (pure)	190 (85.97)	100.0	2.17
2	24 (10.86)	53.00	2.17
3	2 (0.905)	53.00	6.50
4	3 (1.358)	33.00	4.67
5	1 (0.453)	25.00	8.00
6	1 (0.453)	29.00	7.00

To verify that our approach is time and memory efficient, we executed the `ph_dct_imagehash` function to calculate the hash of each image in the our larger dataset, which resulted in 5.030453ms on average, with a 2.172415ms standard deviation and less than 5 megabytes of main memory on a single core. As hashes can be indexed in proper data structures (e.g., MVP trees) that take into account metric distances, the time required to perform a k-nearest-neighbor search ($k = 1$) is also negligible as it grows logarithmically with the number of apps. We verified that the time required to lookup a similar app is minuscule with respect to the time required to run a full dynamic analysis.

Considered the low time and memory requirements, we were able to cluster both the datasets with DBSCAN [9], as demonstrated in Figure 9, and allows to further speedup the lookup phase if necessary. Moreover, in Figure 10 we show that the threshold on the hamming distance can be reliably chosen in an unsupervised fashion by taking into account the intra-cluster distance, inter-cluster distance, average cluster size and total number of clusters. We indeed observe that increasing the threshold above 16 (bits), the number of clusters drops significantly, while the average cluster size jumps from 2–6 elements to about half the size of the dataset. This indicates that using a threshold below 16 allows DBSCAN to find many small clusters, each with the similar apps, plus one noisy cluster of unpaired apps. This is also showed by the intra-cluster distance, which increases significantly at 16. We chose 10 as the threshold, and 2 as the minimum cluster size (as we want to find, at least, couples of similar apps).

In the Appendix we show that our approach can find interesting, non-obvious pairs of similar apps. To validate our approach we asked Mechanical Turk Master Qualified workers to analyze 321 randomly picked (i.e., `sort -R`) clusters and report the number and size of distinct classes of screenshots they found in each of them. With this we calculated each cluster's homogeneity as the size of the most frequent class over the cluster size. Ideally, homogeneity should be 100%, indicating 1 class of UI per cluster, which means a pure, perfectly formed cluster. We could not reliably use the name nor the MD5 of the application as a class label, because, as shown by previous work (see Section 7), many applications are actually repackaged versions or other applications. However, by randomly drawing 100 of 420 and 221 of 628 clusters we inspected 23.81% and 35.19% of the clusters respectively from the 6,000 and 16,000-images datasets. As Table 4 shows, our approach finds 85.71–85.97% pure clusters), whereas the reminders have a reasonably high homogeneity, except for some outliers.

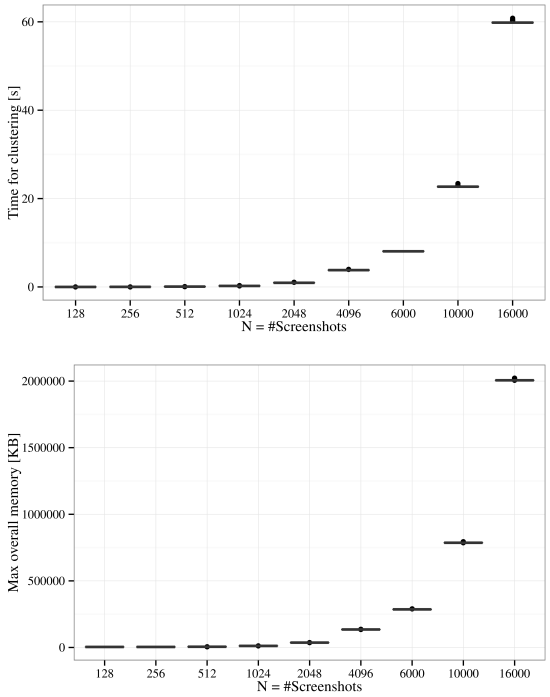


Figure 9: Experiment 3: Boxplot distribution of time (a) and memory (b) requirements (33 runs per each N) for clustering app screenshots. Time and memory for comparing 2 apps is always constant.

6. LIMITATIONS AND FUTURE WORK

The presence of a very small view object in the original sample layout that is not present in the layouts of similar applications is a corner case that can make our re-execution incomplete (see Appendix). For example, during the stimulation of the original sample, our user clicked on a link embedded in a TextView object. Re-executing the test on a similar application, the content of the TextView changed, with consequent vanishing of the link. Hence, clicking on the TextView in the original sample led to open a new window, while the same click on the similar application did not generate any transition, making the re-execution test fail.

To avoid this specific cases of dynamic layouts, our future work includes attaching semantic tags to each screenshot (e.g., list of known view objects visualized), so as to devise a similarity criterion that can recognize whether two layouts are very similar, yet with a significant tiny variation (e.g, absence of a single, small button). However, this creates the further challenge of deciding a threshold, because such a semantic similarity criterion cannot possibly be mapped on a metric space. Instead, our current method is simple and practical because it requires no threshold: We find the application that minimizes the distance, and we can do this because the features are metric.

7. RELATED WORK

Our work is related to dynamic analysis, similarity and UI exercising of Android applications.

Dynamic analysis. TaintDroid [7], integrated and extended by other analysis systems such as DroidBox [28] and Andrubis [17], extends Android to taint track privacy-sensitive resources and notify the user if such information

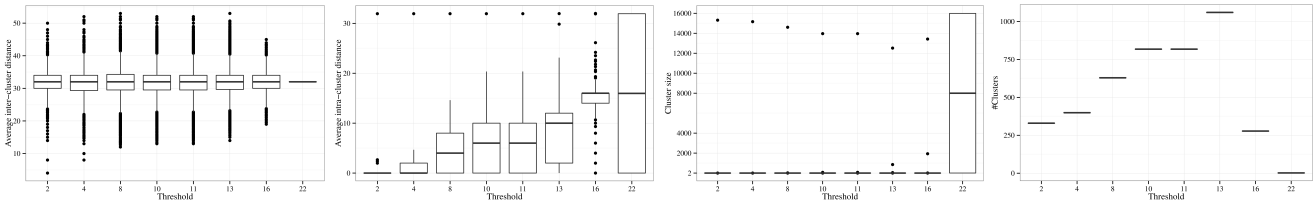


Figure 10: Experiment 3: Parameter estimation on 16,000 images. Larger versions in the Appendix.

leave the system via network, SMS, or else. Unfortunately, it prevents third-party apps from loading native libraries, and is version specific. DroidBox [28] extends TaintDroid with the ability to keep track of network traffic, sent SMS, phone calls, etc. DroidBox has recently been upgraded to APIMonitor, which works directly on the source code of applications (rather than on the code of the system).

Differently from previous approaches, CopperDroid [24], is an out-of-the-box dynamic analysis tool that relies on an instrumented version of the Android emulator to automatically reconstruct behaviors. By thoroughly inspecting syscalls and their arguments, CopperDroid performs an unified analysis of both low level (e.g., file writes) and high level (e.g., send an SMS) actions performed by an application. Furthermore, it uses an effective stimulation mechanism that increases the code coverage. Similarly, DroidScope is built on top of QEMU. Authors modified the translation phase from Android code to TCG, an intermediate representation used in QEMU, to insert extra instructions that enable fine-grained analyses. To reconstruct the two semantic levels of Android, (Linux and Dalvik), VMI is leveraged.

Application Similarity. Tools such as Androguard[9] assist reverse engineers in finding similar APKs, but have accuracy and scalability issues. Therefore, research in this direction is fairly active for different purposes. For example, [11, 4] use app similarity to find repackaged, ad-aggressive versions of applications distributed on alternative markets.

Juxtapp [15] recognizes whether applications contain known, flawed code, exhibit code reuse that indicates piracy, or are (repackaged) variants of known malicious apps. Juxtapp focuses on scalability, proposing a similarity metric that is suitable for map-reduce frameworks. Juxtapp requires 100 minutes of computation on 100 8-core machines with 64GB of RAM to analyze 95,000 distinct APKs. DNADroid [5] exploits the dependency graph to find pairs of matching methods to recognize plagiarized applications.

PuppetDroid differs substantially from previous work because it takes the UI into account. Our goal is not that of finding similar *code*, but to finding similar *interfaces*.

Exercising of Android applications. Dynodroid [18] uses an *"observe and execute"* approach (i.e., analyze the content of displayed UI elements and then generate tailored random input events). Dynodroid reaches the same code coverage obtained with Monkey, but with much less events.

SmartDroid [34] leverages static and dynamic analysis to extract a sequence of UI events that allow to stimulate suspicious behaviors. Static analysis is used to identify the invocations of sensitive methods. Then, sensitive paths from application's entry points to identified method invocations are built. Last, dynamic analysis is used to verify the validity of the paths previously found.

ACTEve [2] proposes an algorithm that leverages concolic execution [25] to automatically generate input events. It uses advanced subsumption and pruning algorithms to avoid the path explosion problem and, thus, is able to automatically generate test inputs that strive the execution flow of an application to get high code coverage. Despite its optimization, though, the overhead of this technique is high (i.e., hours) for malware analysis purposes.

Finally, RERAN [12] allows to record and replay low-level UI events directly reading from, and writing on, system input devices. This work uses an approach similar to the one used by PuppetDroid to inject input events, but it is limited to a mere re-execution of the original recorded touch events without offering any analysis of application UI.

PuppetDroid differs substantially from previous work because (1) takes human users into account, (2) introduces the concept of visual similarity and, at the same time, (3) binds low-level input events to view objects.

[9] `https://code.google.com/p/androguard/`

8. CONCLUSION

Dynamic analysis is facing new challenges with mobile malware. Mobile software (both goodware and malware) was born in a radically different ecosystem than traditional software, which includes, for instance, app marketplaces—the main distribution channel for malicious apps.

Because of this different nature, the victim's participation during the infection is essential, and greater than in traditional malware. We believe that orthogonal approaches to dynamic analysis, such as PuppetDroid, that strive to capture the user's actions, are an important research direction to pursue. Our experiments show that our hypotheses are true, and that human users can be effectively and efficiently included in the dynamic analysis workflow, also thanks to the availability and accessibility of crowdsourcing platforms.

This can potentially change the way we conduct dynamic analysis of mobile applications (from fully automatic, to scalable and collaborative): We believe that our system can attract the interest not only of security analysts but also of normal users that want to safely test potentially malicious applications.

Acknowledgements

This research has been partially funded under the EPSRC Grant Agreement EP/L022710/1 and by the FP7 project SysSec funded by the EU Commission under grant agreement no. 257007.

9. REFERENCES

[1] . Android testing patches project page. `https://code.google.com/p/android-app-testing-patches/`.

[2] S. Anand, M. Naik, H. Yang, and M. Harrold. Automated concolic testing of smartphone apps. In *Proc. of FSE*, 2012.

[3] T. Bozkaya and M. Ozsoyoglu. Indexing large metric spaces for similarity search queries. *ACM Transactions on Database Systems*, 24:361–404, 2002.

[4] H. Chen. Underground Economy of Android Application Plagiarism. In *Proceedings of the 1st International Workshop on Security in Embedded Systems and Smartphones (SESP)*, 2013.

[5] J. Crussell, C. Gibler, and H. Chen. Attack of the Clones: Detecting Cloned Applications on Android Markets. In *Proceedings of the 17th European Symposium on Research in Computer Security (ESORICS)*, 2012.

[6] W. Enck, P. Gilbert, B. Chun, L. Cox, J. Jung, P. McDaniel, and A. Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In *Proc. of USENIX OSDI*, 2010.

[7] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In *Proceedings of the 9th USENIX conference on Operating systems design and implementation*, OSDI'10, pages 1–6, Berkeley, CA, USA, 2010. USENIX Association.

[8] ESET Latin America Lab. Trends for 2013, Astounding growth of mobile malware. Technical report, ESET Latin America Lab, November 2012.

[9] M. Ester, H. peter Kriegel, J. S, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[10] R. Fedler, J. Schütte, and M. Kulicke. On the Effectiveness of Malware Protection on Android. Technical report, Fraunhofer AISEC, Berlin, 2013.

[11] C. Gibler, R. Stevens, J. Crussell, H. Chen, H. Zang, and H. Choi. AdRob: Examining the Landscape and Impact of Android Application Plagiarism. In *Proceedings of 11th International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2013.

[12] L. Gomez, I. Neamtiu, T. Azim, and T. Millstein. Reran: timing- and touch-sensitive record and replay for android. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, pages 72–81, Piscataway, NJ, USA, 2013. IEEE Press.

[13] Google Inc. UI/Application Exerciser Monkey. `http://developer.android.com/tools/help/monkey.html`.

[14] M. Grace, Y. Zhou, Q. Zhang, S. Zou, and X. Jiang. Riskranker: scalable and accurate zero-day android malware detection. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, MobiSys '12, pages 281–294, New York, NY, USA, 2012. ACM.

[15] S. Hanna, L. Huang, E. Wu, S. Li, C. Chen, and D. Song. Juxtapp: A scalable system for detecting code reuse among android applications. In *Proc. of DIMVA*, 2012.

[16] IDC. Apple Cedes Market Share in Smartphone Operating System Market as Android Surges and Windows Phone Gains, According to IDC. `http://www.businesswire.com/news/home/20130807005280/en/Apple-Cedes-Market-Share-Smartphone-Operating-System`, August 2013.

[17] I. S. S. Lab. Andrubis: A tool for analyzing unknown android applications.

[18] A. MacHiry, R. Tahiliani, and M. Naik. Dynodroid: An input generation system for android apps. In *Proceedings of the 2013 ACM Symposium on Foundations of Software Engineering*, FSE'13, 2013.

[19] F. Maggi, A. Valdi, and S. Zanero. AndroTotal: A Flexible, Scalable Toolbox and Service for Testing Mobile Malware Detectors. In *Proceedings of the 3rd Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices (SPSM)*, 2013.

[20] McAfee Inc. Virus Profile: Android/BaseBridge.G. `http://home.mcafee.com/virusinfo/virusprofile.aspx?key=665341`.

[21] Mobile Antivirus. New Android Trojan Detected, Called BaseBridge. `http://www.mobiantivirus.org/antivirus/basebridge.html`.

[22] J. Oberheide and C. Miller. Dissecting the Android's Bouncer. *SummerCon*, 2012. `http://jon.oberheide.org/files/summercon12-bouncer.pdf`.

[23] V. Rastogi, Y. Chen, and X. Jiang. DroidChameleon: Evaluating Android Anti-malware Against Transformation Attacks. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security (ASIACCS)*, 2013.

[24] A. Reina, A. Fattori, and L. Cavallaro. A system call-centric analysis and stimulation technique to automatically reconstruct android malware behaviors. In *Proceedings of the 6th European Workshop on System Security*, EUROSEC'13, April 2013.

[25] K. Sen, D. Marinov, and G. Agha. CUTE: a concolic unit testing engine for C. In *Proceedings of the Symposium on Foundations of Software Engineering*, Lisbon, Portugal, 2005.

[26] D. Smith. Mastering the android touch system. In *Proceedings of the 2012 Fourth Android Developer Conference*, AnDevConIV, 2012.

[27] Symantec Corporation. Android.Basebridge. `http://www.symantec.com/security_response/writeup.jsp?docid=2011-060915-4938-99`.

[28] The Honeynet Project. Droidbox. `https://code.google.com/p/droidbox/`.

[29] Trend Micro. Repeating History. Technical report, Trend Micro, January 2013.

[30] TrendLabs. Android under siege: Popularity comes at a price. Technical report, Trend Micro, Inc., 2013.

[31] T. Vidas and N. Christin. Sweetening Android Lemon Markets: Measuring and Combating Malware in Application Marketplaces. In *Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy (CODASPY)*, 2013.

[32] L. K. Yan and H. Yin. Droidscope: seamlessly reconstructing the os and dalvik semantic views for dynamic android malware analysis. In *Proceedings of the 21st USENIX conference on Security symposium*, Security'12, pages 29–29, Berkeley, CA, USA, 2012. USENIX Association.

[33] Z. Yang, M. Yang, Y. Zhang, G. Gu, P. Ning, and X. S. Wang. Appintent: analyzing sensitive data transmission in android for privacy leakage detection. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, CCS '13, pages

1043–1054, New York, NY, USA, 2013. ACM.

[34] C. Zheng, S. Zhu, S. Dai, G. Gu, X. Gong, X. Han, and W. Zou. Smartdroid: an automatic system for revealing ui-based trigger conditions in android applications. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*, SPSM '12, pages 93–104, New York, NY, USA, 2012. ACM.

[35] W. Zhou, Y. Zhou, X. Jiang, and P. Ning. Detecting repackaged smartphone applications in third-party android marketplaces. In *Proceedings of the second ACM conference on Data and Application Security and Privacy*, CODASPY '12, pages 317–326, New York, NY, USA, 2012. ACM.

[36] Y. Zhou and X. Jiang. Android malware genome project. `http://www.malgenomeproject.org/`.

[37] Y. Zhou and X. Jiang. Dissecting android malware: Characterization and evolution. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 95–109, Washington, DC, USA, 2012. IEEE Computer Society.

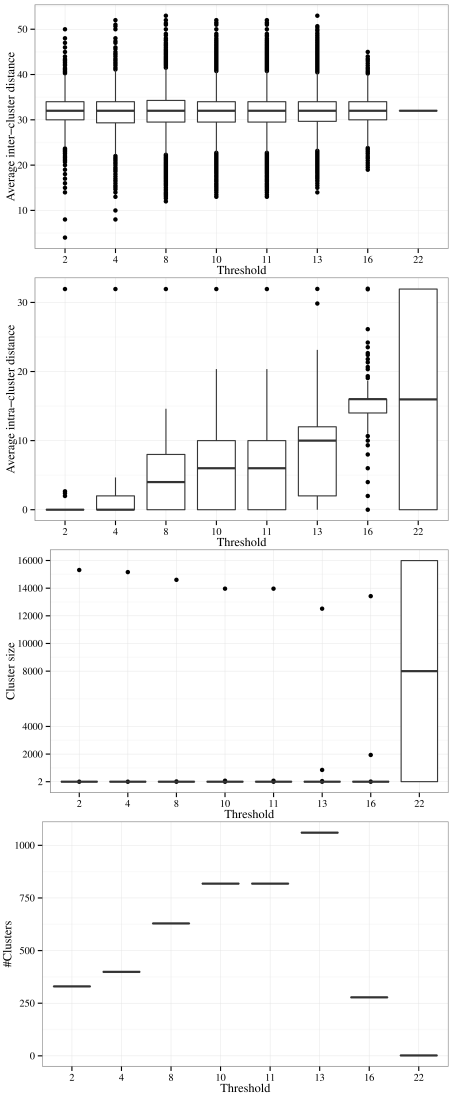
[38] Y. Zhou, Z. Wang, W. Zhou, and X. Jiang. Hey, you, get off of my market: Detecting malicious apps in official and alternative Android markets. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium*, NDSS'12, Feb. 2012.

APPENDIX

Sample Output of Phase 3

We show 6 sample clusters created by our approach, which highlight how it can find non obvious UI-similar applications.

Larger Version of Figure 10



Sample Corner Case in Phase 2

Example of re-execution failure due to the presence of particular UI elements. See Section 6.