



ANCHOR: locating android framework-specific crashing faults

Pingfan Kong¹ · Li Li²  · Jun Gao¹ · Timothée Riom¹ · Yanjie Zhao² · Tegawendé F. Bissyandé¹ · Jacques Klein¹

Received: 22 November 2020 / Accepted: 24 June 2021 / Published online: 12 July 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Android framework-specific app crashes are hard to debug. Indeed, the callback-based event-driven mechanism of Android challenges crash localization techniques that are developed for traditional Java programs. The key challenge stems from the fact that the buggy code location may not even be listed within the stack trace. For example, our empirical study on 500 framework-specific crashes from an open benchmark has revealed that 37 percent of the crash types are related to bugs that are outside the stack traces. Moreover, Android programs are a mixture of code and extra-code artifacts such as the Manifest file. The fact that any artifact can lead to failures in the app execution creates the need to position the localization target beyond the code realm. In this paper, we propose ANCHOR, a two-phase suspicious bug location suggestion tool. ANCHOR specializes in finding crash-inducing bugs outside the stack trace. ANCHOR is lightweight and source code independent since it only requires the crash message and the apk file to locate the fault. Experimental results, collected via cross-validation and in-the-wild dataset evaluation, show that ANCHOR is effective in locating Android framework-specific crashing faults. Finally, we put our empirical study results openly accessible at <https://github.com/anchor-locator/anchor>.

Keywords Android crash · Crashing fault · Fault localization

1 Introduction

App crashes are a recurrent phenomenon in the Android ecosystem (Wei et al. 2016). They generally cause damages to the app reputation and beyond that to the provider's brand (<https://developer.android.com/topic/performance/vitals/crash>).

✉ Li Li
li.li@monash.edu

Extended author information available on the last page of the article

Apps with too many crashes can even be simply uninstalled by annoyed users. They could also receive bad reviews which limit their adoption by new users. Too many apps crashes could also be detrimental to specific app markets that do not provide mechanisms to filter out low-quality apps concerning proneness to crash. The challenges of addressing Android app crashes have attracted attention in the research community.

Fan et al. (2018b) have recently presented insights on their large-scale study on framework-specific exceptions raised by open source apps. In more recent work, Kong et al. (2019) have proposed an automated approach to mine fix patterns from the evolution of closed-source apps (despite the lack of change tracking systems). Tan et al. (2018) further presented an approach to repair Android crashing apps. A common trait of all these crash-related studies is that the underlying approaches heavily rely on the generated stack traces to identify the fault locations. Although the state of the art is effective for many bugs, they are generally tailored to the generic cases where the stack traces provide relevant information for locating the bug. Unfortunately, there is a fair share of faults whose root causes may remain invisible outside the stack trace. Wu et al. (2014) have already reported this issue when locating crashing faults for general-purpose software. In the realm of Android, the phenomenon where the stack trace may be irrelevant for fault localization is exacerbated by two specificities of Android:

The Android system is supported by a callback-based and event-driven mechanism Each component in Android has its lifecycle and is managed by a set of callbacks. Every callback serves as a standalone entry point and root to a separate call graph. Yet, existing crash-inducing bug localization techniques for Java such as CrashLocator (Wu et al. 2014) assume a single entry point to compute certain metrics for the suspiciousness score of different methods. Additionally, since the Android system is event-driven, the invocation sequence to functions and callbacks is affected by non-deterministic user inputs or system events, making the stack trace unreliable for quick analyses.

The Android app package includes both code and resources that together form the program Android apps are more than just code. They are combinations of Java/Kotlin code, XML files, and resources (such as images and databases). The different component classes in an Android app are loosely coupled. They follow design principles like IoC (Inversion of Control) to receive the flow of control from a generic framework, herein, the Android framework. The Android framework analyzes metadata such as component layout information in XML files, and switches control by invoking the corresponding callback method of any component. In this way, the Android framework orchestrates functionalities in the apps. Therefore, an error by developers within an XML document can eventually lead to a runtime crash. Similarly, it is important to note that crashes can occur due to other concerns such as the arrangements of app resources, use of deprecated APIs [e.g., due to version incompatibility (Li et al. 2018)], omissions in permission requests, etc. Typical such errors, which occur outside of code pointed to by stack traces, will cause either developers or Automatic Program Repair (APR) tools (e.g., Tan et al. (2018)) to pointlessly devote time in attempting to fix the code.

This paper Our work aims at informing the research community on the acute challenges of debugging framework-specific crashes. To that end, we propose to perform an empirical study that investigates the share of crashes that cannot be located by current localization approaches. Following this study, we present a new approach to locate faults, aiming at covering different categories of root cause locations. Overall, we make the following contributions:

- We present the results of an empirical study performed on a set of 500 app crashes retrieved from the ReCBench dataset (Kong et al. 2019). A key finding in this study is that we were able to identify that 37% crash root causes are associated with crash cases where the stack trace is not directly relevant for fault localization.
- We propose ANCHOR, a tool-supported approach for locating crashing faults. ANCHOR unfolds in two phases and eventually yields a ranked list of location candidates. The first phase applies a classification algorithm to categorize each new crash into a specific category. Depending on this category, a dedicated localization algorithm is developed in the second phase. ANCHOR currently implements 3 localization algorithms that eventually generate a ranked list of buggy methods (when the bug is in the code) or resource types (when it is outside of code).
- We performed 5-fold cross-validation on the 500 crash cases to assess the effectiveness of ANCHOR in placing the crashing fault location in the top of its ranked list of suggestions. ANCHOR exhibited an overall MRR (Mean Reciprocal Rank) metric value of 0.85. An analysis of the open dataset of crashed open-source Android apps further shows that our method scales to new app crashes.

The rest of this paper is organized as follows. Section 2 introduces background details on Android app crashes and callback-based event-driven mechanisms. Section 3 revisits the motivating example by the previous work (Tan et al. 2018) and demonstrates why research in crash localization has standing challenges. Section 4 discusses the findings of our empirical study and explores the insights that can be leveraged for a new approach. Section 5 presents ANCHOR. We describe experimental setup in Sect. 6 and approach evaluation in Sect. 7. We bring further discussion in Sect. 8. Threats to validity are acknowledged in Sect. 9 and related work is presented in Sect. 10. Finally, Sect. 11 concludes the paper.

2 Background

In this section, we introduce the important concepts related to this paper.

2.1 Android App crash stack trace

Like all Java¹ based software, when Android apps crash, they can dump execution traces which include the exception being thrown, a crash message, and most importantly, a stack trace of a callee-caller chain starting from the *Signaler*, i.e., the method that initially constructed and threw the exception object. Figure 1 is an example of stack trace for the crash of the app *Sailer's Log Book*. This app helps sailors to keep their logbook accurate and up-to-date. On the first line, the exception *IllegalArgumentException* is thrown. On the second line, the log system reports message "recursive entry to executePendingTransactions". Starting from the third line, the *Signaler* of the stack trace is listed: it is this framework method that instantiates the exception type, composes the log message and throws it to its caller to handle. On Lines 4-5 that are also marked in grey, there are other two methods that continue to pass on the exception. Line 5 holds the *API*, which is the only framework method in this stack trace that is visible to the developer. Since the crash happens directly due to invocation to it, we call it the *Crash API*. Line 6 is the developer method that invoked this API. Line 7 is the developer implementation of the callback, inherited from the superclass of the Android framework. Android framework decides, based on certain conditions and system/user events, when to invoke this method, and what parameter values to pass in. Lines 8-9 are part of the Android framework core that is, again, not accessible to developers.

The crash stack trace is often the first thing that developers want to examine when a crash is reported (Kim et al. 2011). Even when it is not given, developers would reproduce and retrieve them. Intuitively, the crash arises from mistakes in the developer methods, e.g., Lines 6–7 in Fig. 1. Particularly, the *Crash method* that directly invoked the *Crash API*. Our empirical study in Sect. 4 shows that this intuition is

Exception	IllegalArgumentException		1
Message	Recursive entry to executePendingTransactions.		2
Framework	android.app.FragmentManagerImpl.executePendingActions	Signaler	3
	android.app.FragmentManagerImpl.executePendingTransactions		4
	android.app.FragmentManagerImpl.popBackStackImmediate	Crash API	5
Developer	com.poterion.logbook.activities.MainActivity.show	Crash method	6
	com.poterion.logbook.fragments.SettingsFragment.onSharedPreferencesChanged	Callback	7
Framework	:		
	com.android.internal.os.ZygoteInit\$MethodAndArgsCaller.run		8
	com.android.internal.os.ZygoteInit.main	Core	9

Fig. 1 Crash stack trace of app Sailer's log book

¹ Kotlin has also been widely used in recent years as an alternative for Android app development, it is designed to fully interoperate with Java.

correct, that 63% of the total crash types are in the stack trace. However, in the rest of this section, we will introduce the speciality of Android that may lead to the rest 37%.

2.2 Callback-based and event-driven mechanism

Unlike most Java programs, Android apps do not have the *main* method from which the programs start their execution (Arzt et al. 2014). Android apps comprise 4 basic component types: Activity, Broadcast Receiver, Service, and Content Provider. Each basic component instance can be started on their own. The lifecycle of such components are managed by callback methods (e.g., Line 7 in Fig. 1). The callback methods are declared in the component’s base class (as part of the Android framework), inherited by the developer-defined subclass, and maybe overridden by the developers. The Android framework core, based on the user inputs and system environments, decides when to invoke the callbacks and what parameter values to pass in. Each callback is standalone, and in general Android does not encourage developers to invoke those callbacks from their self-defined methods, unless these methods are callbacks overriding and invoking their super. As a result, existing static call graph based fault localization techniques (Wu et al. 2014) for Java programs can not be simply reused, since they assume a single *main* method and need to compute weighing scores based on the graph. There are, however, works (Yang et al. 2015; Oceau et al. 2015) that have invented methods to track the control flows or data flows and tried to build the callback connections. These proposed approaches are either computationally expensive or confined in limited number of component classes, and does not scale to all scenarios. Other approaches like (Li et al. 2016a; Arzt et al. 2014) create a dummy main to invoke all callbacks in order to reuse Java based analysis tools, but this method discarded the relation among callbacks, which is crucial to estimate the possibility of a method containing the real bug.

Figure 2 exemplifies the difference of call graphs between general Java program (left) and Android app (right). The filled circles represent the developer methods in the stack trace, while the non-filled circles represent developer methods outside the stack trace. The partially filled circles represent the *Crash method* that invokes the *Crash API*. Generally, the buggy method is the *Crash method*. However, as shown

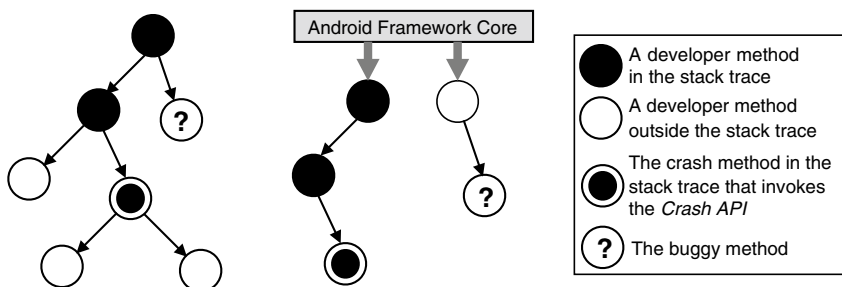


Fig. 2 Call graph comparison between general java program (left) and android app (right), inspired from Wu et al. (2014)

in our empirical study, it appears that the buggy method (the circle filled with question mark in Fig. 2) is not connected to the *Crash method*. A traditional Java program static call graph based approach such as CrashLocator (Wu et al. 2014) will be able to locate this buggy method only if the buggy method is "close enough" to the generated call graph (roughly speaking they generate an extended call graph leveraging the stack trace). However, on the right, in the case of Android apps, the buggy method could be in a separate call graph because of callback methods that are invoked by the Android framework. Such cases will be missed by approaches such as CrashLocator (Wu et al. 2014) that only detects buggy methods captured by its extended call graph, but does not consider callback methods.

2.3 Android APK file format

Android apps are distributed in a package file format with extension ".apk". It is a compressed folder containing code, resources, assets, certificates, and manifest file. All of these files are crucial to the expected good functioning of the apps. Therefore, some crashes may be induced when there are problems with these files.

2.3.1 Android manifest file

Every app project needs to have an AndroidManifest.xml file at the root of the project source set (LLC 2020a). Along with the package name and components of the app, this file also declares the permissions that the apps needs, as well as the hardware and software features that the app requires.

2.3.2 Android component layout description file

Android component layout description files are also crucial to the execution of the app. E.g., Listing 1 is the layout file of the main Activity of an Android app *Transistor*. In this file, a child fragment is defined and described. The attribute *android:id* defines the layout file name to be inflated for the fragment, the attribute *android:name* gives the full name of the user defined Fragment class. When the main Activity is being created, the Android framework scans this layout file, and invokes a series of relevant callbacks on this Fragment to draw it along with the main Activity.

Listing 1: Main Activity Layout File of app Transistor.

```
<?xml version="1.0" encoding="utf-8"?>
<fragment xmlns:android="http://schemas.android.com/apk/res/android"
    xmlns:tools="http://schemas.android.com/tools"
    android:id="@+id/fragment_main"
    android:name="org.y20k.transistor.MainActivityFragment"
    android:layout_width="match_parent"
    android:layout_height="match_parent"
    tools:layout="@layout/fragment_main" />
```

3 Motivating example

We further illustrate the challenges of locating faults outside Android app stack traces by revisiting an example that was used to motivate a previous work on Android app crash automatic repairing by Tan et al. (2018). *Transistor*² is a popular online radio streaming app. We showed its partial resources in Sect. 2.3.2. However, it was reported that following the input sequence in Fig. 3, the app will crash.

Listing 2: Crash Message of *Transistor*.

```

1 java.lang.IllegalStateException:
2 MainActivityFragment{e7db358} not attached to Activity
3 at ...MainActivityFragment.startActivityForResult(Fragment.java:925) (Crash API)
4 at ...agment.selectFromImagePicker(MainActivityFragment.java:482) (Crash method)
5 at ...k.transistor.MainActivityFragment.access$500(MainActivityFragment.java:58)
6 at ...transistor.MainActivityFragment$6.onReceive(MainActivityFragment.java:415)

```

The crash message filtered out from logcat is shown in Listing 2. It appears that invoking the *startActivityForResult* API on the *MainActivityFragment* (line 3) throws an unhandled *IllegalStateException* (line 1), and the Android system reports that the fragment is not attached to the hosting activity (line 2). By inspecting the source code of Android framework of the *Crashed API* (line 3), we see that the *startActivityForResult* method of the fragment instance attempts to invoke its context's (i.e., its host *Activity*'s) API with the same name *startActivityForResult*. This invocation is guarded by an if-clause, which checks whether the fragment is still attached to the host *Activity*. If not, however, the *IllegalStateException* will be thrown.

Listing 3: Fix from Tan et al.

```

new BroadcastReceiver(){
    onReceive(...){ ...
+   if(getActivity()!=null)
        startActivityForResult(pickImageIntent,REQUEST_LOAD_IMAGE);}}

```

Biased by the assumption that the fault should only be in the developer methods in the stack trace (lines 4–6), Tan et al. (2018) proposed to amend the *Crash method* (line 4). Listing 3 shows their fix. Their fix applies a checker on invocation

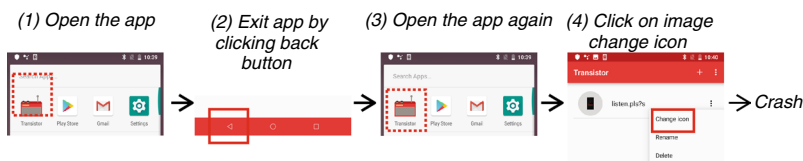


Fig. 3 Crash of transistor

² <https://github.com/y20k/transistor/issues/21>.

to *startActivityForResult*, which will not be executed if value of *getActivity* is null (i.e., when the fragment is no longer attached to its hosting *Activity*). As a result, the app avoids crashing. This fix indeed prevents the exception. However, it is not explainable: applying the checker not only prevents the crash, but it should also prevent opening the *SelectImageActivity* as designed for. Due to this paradox, we have a good reason to suspect that the true bug location is still hidden.

Transistor's developer, who is also dedicated in debugging in the stack trace, proposed a fix on her/his own in Listing 4. Realizing that the *Fragment* lost its reference to the host *Activity*. The developer declared a variable *mActivity* to hold the reference. Then in the *Crash method* (line 4 in Listing 2), she/he switched the invocation of the *startActivityForResult API* from *Fragment* to *mActivity*.

Listing 4: Fix from Developer.

```
+ mActivity = getActivity(); ...
new BroadcastReceiver(){
    onReceive(...){ ...
-   startActivityForResult(pickImageIntent,REQUEST_LOAD_IMAGE);
+   mActivity.startActivityForResult(pickImageIntent,REQUEST_LOAD_IMAGE);}}
```

This fix also bypassed the crash, but it causes regression. After the final step in Fig. 3, if the user clicks on the back button two more times, the app should have first returned to the *MainActivity*, then back to the home screen. Instead, it opens another *SelectImageActivity*. In the issue tracking, the developer admits that she/he had no idea of how to fix it. While after several months, the bug "fixed" itself, which she/he described as "scary". Even Tan et al. failed to explain the cause of this regression.

Based on the understanding of Android' callback-based mechanism introduced in Sect. 2.2, we suspect that the bug may not exist in the stack trace. We confirmed our fix shown in Listing 5. This fix is reported to the developer and we received positive feedback in the issue tracking, as can be verified in *Transistor*'s repository given above.

Listing 5: Fix Inspired by this Article.

```
MainActivityFragment extends Fragment{
    onDestroy(){
+   super.onDestroy();
+   LocalBroadcastManager.getInstance(mApplication).unregisterReceiver(
        imageChangeRequestReceiver,imageChangeRequesIntentFilter);}}
```

We broaden the search for the bug outside the stack trace. Noticing the crash originated from the *onReceive* callback (cf. line 6 in Listing 2), we examine the lifecycle of this *BroadcastReceiver* object. We found that it is registered in the *onCreate* callback of *MainActivityFragment*, but never unregistered in its counterpart callback *onDestroy*. As a result, after Step 2 (cf. Fig. 3), the registered *BroadcastReceiver* and its host *MainActivityFragment* are

leaked in the memory. In Step 4, the callbacks of the leaked objects are stealthily invoked by the Android framework and eventually caused the *IllegalStateException*. Knowing the true cause of the crash, it is not difficult to explain the paradox of Tan et al.'s fix and the regression caused by the developer's fix. However, given the page limit, we put detailed reasoning online at <https://anchor-locator.github.io>.

Hint The fault locations in Android apps may: (1) Be outside the stack trace; (2) Be even outside the call graph extended from the stack trace; (3) Not even "exist" in the code, i.e., they are inherited methods without visible code snippets. Locating such faults may require tremendous efforts. Fixes based on incorrect localization may even cause regression.

4 Empirical study on fault locations

In this section, we present the results of an empirical study that we performed on a set of 500 app crashes retrieved from the ReCBench dataset (Kong et al. 2019). This study aims at assessing to what extent the locations of crashing faults reside outside the stack trace.

4.1 Dataset construction

We extract our dataset from ReCBench, an open dataset proposed by Kong et al. (2019) in 2019. ReCBench has been built by running hundreds of thousands of Android apps downloaded from various well-known Android markets (Allix et al. 2016; Li et al. 2017b). In addition to a collection of crashed Android apps focusing on framework-specific crashes³, ReCBench offers the possibility to collect crash log messages and scripts to reproduce the crashes. Today, ReCBench contains more than 1000 crashed apps (still growing). For our empirical study, we focus on crashed apps for which:

- First, the stack trace of the crash contains at least one developer method. This is a requirement to be able to start an exploration process to find the crash root cause.
- Second, since we specifically target the crashes induced by Android APIs, the *Signaler* must be Android-specific.

The reason why we only choose bug reports with higher priority are two-fold. First, we want to target bugs that are most difficult for developers to locate. Second, we need to limit the number of crashing cases for manual validation. After applying

³ Android framework methods are not visible or understandable to general developers, hence greater challenge is acknowledged for locating framework-specific crashes compared to developer-written methods. Fan et al. (2018b), Kong et al. (2019).

these two rules, we randomly selected 500 crashed apps from the remaining apps. The dataset is publicly accessible at:

<https://github.com/anchor-locator/anchor>.

4.2 Ground truth and results

We manually inspect all the 500 crashed apps to understand the reason behind the crashes and to create our ground truth. We perform this manual inspection following a similar protocol discussed in the large scale analysis of Android framework-specific app crashes (Fan et al. 2018b). First, we group crashes into *buckets*. Specifically, if two crash cases have identical framework crash sub-trace, they will be put into the same bucket. Second, for each bucket, we turn to the Android official API reference ⁴ as well as online discussion forums like StackOverflow ⁵ and GitHub ⁶ to understand the root cause. Third, we analyze the code of each crashed apk combining the root cause to locate the true bug locations. Note that unlike (Fan et al. 2018b), source code is not available for the crashed apks in our dataset. We therefore leverage the CodeInspect (Fraunhofer 2020) tool. CodeInspect is an Integrated Development Environment (IDE) that transforms the app apk's Dalvik bytecode into the Jimple (Bartel et al. 2012) format and better visualize for human comprehension.

Each of the crashed apps has been categorized into one of the following categories:

- Category A groups the crashed apps for which the buggy method (i.e., the root cause) is one of the developer methods present in the stack trace;
- Category B groups the crashed apps for which the buggy method is not present in the stack trace, but still in the code.
- Category C groups the crashed apps for which the crash arises from non-code reasons.

The above partition is one out of many alternatives, e.g., one can also separate bugs based on whether they are concurrent (Wang et al. 2018; Bielik et al. 2015; Li et al. 2016b; Tang et al. 2016; Maiya et al. 2014). We later show in Sect. 5.2 how this partition helps with building our localization tool. Table 1 summarizes the outcome of the empirical study. It appears that for 89 (49+40) crashed apps (representing 18% of the total cases), the crashing fault location is not in any of the developer methods present in the stack trace. The respective numbers of Categories B and C are close, with 49 cases in Category B and 40 cases in Category C. The last two columns in Table 1 present the number of *buckets* per category. Overall, there are 105 types of crashes (i.e., buckets) in the dataset. The percentage of types of crashes in Categories B and C are 16% and 21%, respectively. In total, there are 37% of buckets whose buggy reasons are not shown in the stack traces. Each unique framework

⁴ <https://developer.android.com/reference>.

⁵ <https://stackoverflow.com>.

⁶ <https://github.com>.

Table 1 Categories of fault locations in android apps

Category	Stack trace	Code	Case#	Percent (%)	Bucket#	Percent (%)
A	In	Yes	411	82	66	63
B	Out	Yes	49	10	17	16
C	Out	No	40	8	22	21
Total	–	–	500	100	105	100

Bold indicates the percentage of crash cases where the stack traces are not directly relevant for localization

crash sub-trace suggests a unique type of crash-inducing bug. Therefore, considering crash types encountered per the same number of cases (buckets#/case#) in each category, more debugging effort will be needed for Categories B and C than in Category A.

Hint 18% of the crashes are due to bugs for which the location is outside the stack trace. A significant number of root causes (buckets), i.e., 37% (16%+21%), are associated with cases where the stack trace is not directly relevant for localization. In even 21% of the cases, the root causes are not located in the code.

We now detail each category in the rest of this Section.

4.3 Category A: in stack trace

Category A includes all crash cases whose bugs reside in one of the developer methods present in the stack trace. Most crashes in our dataset fall into this category. It is expected that by default, developers start their debugging process with the methods present in the stack trace (Jiang et al. 2010; Schroter et al. 2010; Sinha et al. 2009; Indi et al. 2016). The automatic crash-inducing bug repairing tool named Droix (Tan et al. 2018) implements a locator, by assuming that the *Crash method* is the bug location in all scenarios. However, we also notice that the true crashing fault may reside in other developer methods, in particular when moving downward in the stack trace. An example of such a case is when the caller methods pass an incorrect instance to the crashed developer methods. Generally, much less effort is needed in locating faults in this category. Since the number of suspected methods is limited and their names are already known. Therefore they are not the focus of this paper.

4.4 Category B: out of stack trace, in the code

It has drawn attention to researchers that Java program crashes can be caused by methods that are not listed in stack traces. Approaches like CrashLocator (Wu et al. 2014) broadens the search for such faulty methods in extended call graphs from stack traces. We demonstrate in the rest of this section why this broadened search is not enough for Android apps. There are in total 49 cases in this category, each crashed from wrongly handling a framework *API*. Based on the type of the framework *API*

(call-in or callback), we further categorize them into two sub-categories: (1) Mis-used Call-In APIs and (2) Non-Overridden Callback APIs.

4.4.1 Type 1: misused call-in APIs (44 cases out of 49)

This first type groups crashing faults caused by the misuse of call-in APIs. This means that the bug leading to a crash is due to a buggy explicit invocation of an API from a developer method. Moreover, this invocation is often performed from another implemented callback, other than the callback in the stack trace. Since both callback methods are triggered by the framework, it is unlikely that an extended call graph can cover such methods (cf. Fig. 2).

Listing 6: Bug Explanation to app Geography Learning.

```
public class MainActivity extends Activity{
    onCreate(...){
        try{bindService(intent,serviceConnection,integer);/*Bug Location*/
        }...}...
    onDestroy(){unbindService(serviceConnection);/*Crash location*/}}
```

Listing 6 gives a concrete example. This example is extracted from an app named *Geography Learning* which helps users to remember geography knowledge in a quiz game format. When the *MainActivity*⁷ of this app is launched, the callback method `onCreate` is automatically triggered by the Android framework. Then, this `onCreate` method invokes the `bindService` API to bind to *Service*. *Service* is one of the four basic components of Android, and wrongly handling of *Service* is not uncommon (Song et al. 2019) in Android app development. When the user exits the *MainActivity*, the Android Framework invokes the `onDestroy` callback method and tries to unbind the *Service* bound in the `onCreate` method. Thereafter, the app crashes with the exception type *IllegalArgumentException*. Analysing the message which says: “*Service not registered: com.yamlearning. geographylearning.e.a.e@29f6021*”, we understand that the *Service* has not been bound. In the method body of the overridden `onCreate` callback, we found that the invocation to API `bindService` was misused. Indeed, `bindService` is surrounded by a try-catch clause, and another statement preceding this API invocation threw an exception which redirects the execution flow to the catch block, skipping the invocation to `bindService`.

Out of a total of 49 cases in Category B, 44 falls into this sub-category.

4.4.2 Type 2: non-overridden callback APIs (5 cases out of 49)

This second type includes crashes caused by the non-overridden callback APIs. Callbacks, or call-afters, are APIs that are invoked when the Android framework decides so, based on certain system environment change and/or user actions.

⁷ The *Main Activity* of an app is the first screen shown to the user when launched.

Callbacks are inherited, when developers define classes that are subclassing Android base component classes. Developers are often required to override certain, although not all, callback APIs. Forgetting to handle these callbacks may cause apps to crash immediately. Moreover, these crashes may often seem flaky, since its reproduction requires re-establishing the same system environments and/or repeating user action sequences. Existing Java crash locators fail to spot such bugs with two reasons: (1) These callback APIs are not in the extended call graphs of stack traces; (2) The method snippets in developer-defined codes do not exist, so are easily missed.

Listing 7 shows an example of this crash type. The app *Fengshui Master* is a Chinese fortune teller app. The app crashes when trying to get a reference to the writable database. However, when the app crashes, the exception *SQLiteDatabaseException* is triggered with a message claiming “not able to downgrade database from version 19 to 17”.

Listing 7: Bug Explanation to Android app Fengshui Master.

```
public class com.divination1518.f.s{
    a(..){sqliteOpenHelper.getWritableDatabase();/*Crash location*/}
public class com.divination1518.g.p extends SQLiteOpenHelper{ ...
+ onDowngrade(..){...}/*Bug Location*/}
```

According to the Android documentation, the app developer needs to implement the callback method *onDowngrade* in the self-defined subclass of *SQLiteOpenHelper*. This callback method will be invoked when the database in storage has a higher version than that of the system distribution. Failing to override this callback API immediately crashes the app. Note that the motivating example (cf. Sect. 3) also falls into this sub-category. Given the stealthiness of this fault type, it is particularly difficult, even for a human developer, to spot the bug reason without being very familiar with the Android official documentation. Out of a total of 49 cases in Category B, 5 falls into this sub-category.

Note that we use api_h to denote the wrongly handled API (call-in API or callback API) for cases of Category B. This denotation is later needed for Sect. 5.2.2.

4.5 Category C: out of stack trace, out of code

As introduced in Sect. 2.3, except code, an Android apk also contains resources, assets, certificate, and manifest. They are critical to the functioning of the app. As a result, mistakes in those files may also cause crashes. Table 2 gives a summary of the buggy locations outside of code. As illustrated, eleven cases of crashes originate from the *Manifest.xml* file. Most cases in this type are because the permissions are not properly declared in the manifest. Resources include specifically files with “.xml” extension (excluding the *Manifest.xml* file). An Android app uses these resource files to store the layout and pieces of information like string values. If the required resource is missing or wrong, then the app will crash. Assets are the large files, like fonts, images, bitmaps. Assets should be put in the correct directory. If the Android framework is not able to find them and it will crash the app.

Table 2 Crash causes of categorie C

Sub-category	Manifest	Hardware	Asset	Resource	Firmware
Cases	11	5	4	2	18

Aside from the files inside the apk, some constraints put forward by the device's hardware and firmware, i.e., the framework may also cause the app to crash. For example, the Android billing service can only be tested on real devices, if, however, tested on emulators, the app crashes (LLC 2020b). Also, since Android is quickly updated with new features and designs, old apps may crash on newly distributed devices, due to reasons like deprecated APIs and new security protocols established by the framework. Developers should generally redesign the relevant functionalities, therefore no single buggy location can be decided.

5 Ranking suspicious locations

To help developers identify the true fault locations when debugging app crashes, including faults that reside outside the stack traces, we propose ANCHOR. ANCHOR is a fault location recommendation system based on a two-phase approach. In the first phase, ANCHOR categorizes a given crash in one of the three categories (A, B, or C) with a classification system. Then, in the second phase, according to the decided category, ANCHOR each adopts a unique algorithm to suggest a rank of locations that are suspected to contain the true faults. The rest of this section describes Phase 1 and Phase 2 in more detail.

5.1 Phase 1: categorization

The first phase aims at assigning a new crash to one of the three categories (A, B, or C). We start by trying to develop a rule-based algorithm that summarizes rules from augmenting the stack traces. However, as Fan et al. (2018a)'s experiments suggest, there are at least thousands of unique stack traces which the crash log may contain. Summarizing rules from such a great number of stack traces is extremely effort-consuming and error-prone. Even if such a rule-based algorithm can be developed, it requires constant manual updates when new stack traces are reported. Therefore, we consider using machine learning algorithms to achieve the goal. We use the Naïve Bayes algorithm (Rish et al. 2001) for the categorization. Naïve Bayes is one of the most effective and competitive algorithms in text-based classification. It is widely used for spam detection (Metsis et al. 2006; Yang et al. 2006), weather forecasting (Walton et al. 2010), etc. It is especially suitable in the scenario when the training set does not contain a large number of records (Huang and Li 2011), e.g., our empirical dataset contains merely 500 manually constructed records.

To construct a vector for each crash record, we feature words extracted from the crash message. The value of each feature dimension is binary, indicating whether a word exists or not in the message. More specifically, we extract three parts from the crash message: (1) The exception type, which is a Java class (e.g., `IllegalArgumentException`); (2) The exception message, which briefly describes the reason of the crash, e.g., line 2 in Fig. 1; (3) The top framework stack frames, each being a Java method, e.g., lines 3–5 in Fig. 1. For (1) and (3), we use “.” as the word separator, for (2), we use space as the separator. To avoid overfitting and to save computing resources, we do not need the entirety of the vocabulary to build the vector. In Sect. 6.4, we further discuss how many words are necessary.

With this categorization system, each new crash will firstly be categorized as a type of “A”, “B” or “C” before being processed in Phase 2.

5.2 Phase 2: localization

The goal of this phase is to provide a rank of potential bug locations (in descending order of suspiciousness), in the form of developer methods when the bug is in the code (i.e., Categories A and B) and of sub-categories when the bug is not in the code (i.e., Category C). Before presenting in the following sub-sections 3 standalone algorithms, one for each category, we explain how we compute a similarity score between two crashes. This similarity score is used in both Categories B and C localization algorithms.

Similarity between two Crashes We quantify the similarity between two crashes C_1 and C_2 by computing the edit similarity between their crash messages as presented in Equation 1:

$$Sim_{C_1, C_2} = Edit_Sim(seq_{C_1}, seq_{C_2}) = 1 - \frac{Lev(seq_{C_1}, seq_{C_2})}{\max(len(seq_{C_1}), len(seq_{C_2}))} \quad (1)$$

In this equation, Sim_{C_1, C_2} represents the similarity between two crashes C_1 and C_2 . seq_{C_i} is the sequence of framework stack frames in a crash message C_i , e.g., lines 3–5 in Fig. 1. Sim_{C_1, C_2} is computed by considering the *Edit Similarity* between the sequences seq_{C_1} and seq_{C_2} . $Lev(seq_{C_1}, seq_{C_2})$ is the *Levenshtein distance* (Qin et al. 2011) of the two sequences. It equals the minimum number of single stack frame edits required to change seq_{C_1} to seq_{C_2} . The intuition here is that when two crashes share similar bug reasons, they tend to share framework stack frames, although not necessarily in identical sequence. Equation 1 then normalizes the distance and takes one’s complement to quantify such similarity.

5.2.1 Category A: in stack trace

Since the crash is assigned to Category A, it indicates that the buggy method is one of the developer methods in the stack trace. We inherit the intuition from Tan et al. (2018), that if the developer method is closer to the *Crash API* in the stack trace, there is a higher chance that it contains the true fault. Therefore, we can obtain the

rank without changing the order of the developer methods in the stack trace. For example, in Fig. 1, methods on line 6 and line 7 are respectively ranked first and second.

5.2.2 Category B: out of stack trace, in the code

When the crash is classified into Category B, it indicates that the buggy developer method is not in the stack trace, but still in the code. As discussed in Sect. 4.4, the buggy method should either be a developer method that misused a call-in API, or a callback API that has not been overridden. In the remainder of this section, we will note api_h this API (call-in API or callback API) that has been wrongly handled (cf. Sect. 4.4). To infer a ranked list of potentially buggy methods, we propose Algorithm 1. The overall idea is, starting from each developer method in the stack trace, in addition to examining the developer methods (1) in the extend call graph, we also examine those that either (2) control the Android components' lifecycles, or (3) are involved in the data flow of the crash. The computation of the suspiciousness score follows the same intuition as explained in Sect. 5.2.1.

First of all, Algorithm 1 requires three input data: (1) *crash*, the crash under study; (2) *ST*, which is the list of developer methods contained in the stack trace, e.g. lines 6-7 in Fig. 1; (3) api_h , the wrongly handled API, which is approximated as the associated wrongly handled API of the most similar crash present in Category B of our empirical dataset. More formally, let be $Crash_B$ the set of all the crashes in Category B. We identify the most similar crash $crash_{sim}$ by following Equation 2. Since their crash reasons are the most similar, it is with the highest possibility that both have wrongly handled the same API.

$$Sim_{crash,crash_{sim}} = \max(Sim_{crash,crash_b}), crash_b \in Crash_B \quad (2)$$

The algorithm starts with retrieving a set of developer methods S from the entire apk that has invoked the api_h (line 1). The outmost for-loop (lines 2–19) loops over each stack frame sf in the stack trace *ST*. Then based on the type of the api_h , there are two sub-routines: (a) when api_h type is “call-in” (lines 4–11); (b) when api_h type is “call-back” (lines 13-17). Next we discuss both sub-routines in detail.

Data: *crash*: the crash to resolve
Data: *ST*: List of developer methods in stack trace of *crash*
Data: *api_h*: *Wrongly handled API*
Result: *R*: Rank of suspicious developer methods

```

1: S ← Developer methods that invoke apih;
2: for sf ∈ ST do
3:   if apih type “call-in” then
4:     for s ∈ S do
5:       for am ∈ AM do
6:         if s links am then
7:           s.score+ =  $\frac{1}{d}$ 
8:         end if
9:       end for
10:    end for
11:    R ← S.sort()
12:  else if apih type “callback” then
13:    for nc ∈ NC do
14:      if nc inherits apih then
15:        R.put(nc)
16:      end if
17:    end for
18:  end if
19: end for

```

Algorithm 1: Localization Algorithm for Category B

Sub-routine for type “call-in” is a for-loop (lines 4–11) that loops over each method *s* in *S*. We then loop over (lines 5–9) all *Active Methods* (*AM*) declared in the same class as *sf*, where *Active Methods* are methods having actual code snippets in the Java class files, not including the inherited ones. The function *links* (line 6) checks 3 sub-conditions: (1) if *s* is invoked by *am*, or (2) if *s* and *am* are declared in the same Java class or (3) if an instance of the declaring class of *s* has been passed to *am* as a parameter. Sub-condition (1) checks if *s* is in the extended call graph of *am*, same as locators like (Wu et al. 2014). Sub-condition (2) implies that *s* is a callback method that involves controlling the component lifecycle as *am* does. Sub-condition (3) implies potential data flow between *s* and *am*. When the condition holds true in line 6, a score is added for *s* (line 7). Here *d* is the distance between *sf* and *Crashed API* in the stack trace. It reflects on the same intuition in Sect. 5.2.1.

Sub-routine for type “callback” is implemented with a for-loop (lines 13–17) that loops over all the inherited Non-overridden Callback (*NC*) of the class where *sf* is declared. If *nc* inherits from *api_h* (line 14), it implies that overriding it may fix the problem, therefore *nc* will be added to the rank *R* (line 15). With the same intuition in Sect. 5.2.1, this sub-routine is designed so that when *sf* is closer to *Crashed API* in the stack trace, *nc* is in the higher location in the rank.

Algorithm 1 addresses the concerns in the empirical study (cf. Sect. 4.4). It can further locate faulty methods that are not in the extended call graphs, or even methods without actual code snippets.

5.2.3 Category C: out of stack trace, out of code

Figure 4 describes the localization process for crashes that have been classified into Category C. To infer a ranked list of potentially buggy locations, this process computes a suspiciousness score for each location. Since the true fault locations in Category C are not in the code, the locations in this ranked list are sub-categories (e.g. manifest, asset, etc.).

With any new crash, we start the process by computing the similarity score $Sim_{crash, crash_c}, crash_c \in Crash_C$. Here $Crash_C$ is the set of all the crashes of Category C in the empirical dataset. In Fig. 4, the similarity scores are denoted as Sim_{caseID} for short. We then take an average of Sim_{caseID} over the same sub-categories. Sub-categories with higher similarity scores take higher positions in the Rank.

6 Experimental setup

This section clarifies the research questions, the metrics used to assess ANCHOR, and the parameter values involved.

6.1 Research questions

We empirically validate the performance of ANCHOR by investigating the following research questions:

- *RQ1* To what extent is the categorization strategy effective?
- *RQ2* To what extent are the localization algorithms reliable?

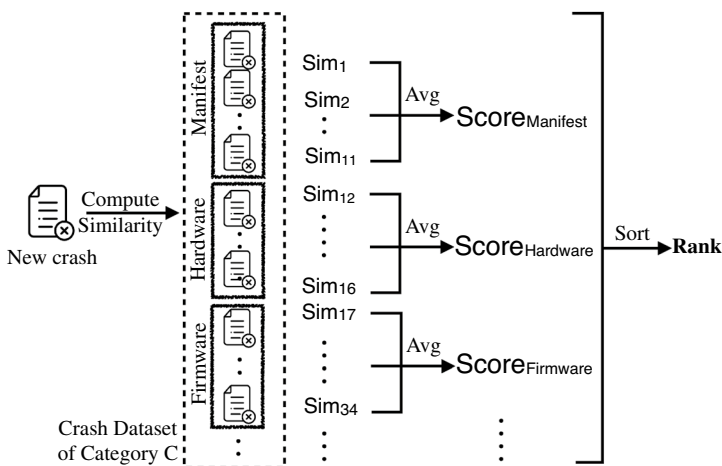


Fig. 4 Localization process for category C

- *RQ3* What is the overall performance of ANCHOR ?
- *RQ4* How does ANCHOR perform on crashes in the wild?

6.2 Metrics

Crash localization is a recommendation problem. To measure the performance of ANCHOR, we rely on rank-aware metrics, which are widely used in information retrieval communities and have been previously used to evaluate crash localization techniques (Wu et al. 2014).

Recall@k The percentage of crash cases whose buggy functions appear in top k locations. A higher score indicates better performance of ANCHOR .

MRR (Mean Reciprocal Rank): The mean of the multiplicative inverse of the rank of the first correct location. As defined in Equation 3, $Rank_i$ is the rank for the i^{th} crash case, in a set of crash cases E . A high value of MRR means developers on average need to examine fewer locations in the rank, and therefore, a better performance (Shi et al. 2012).

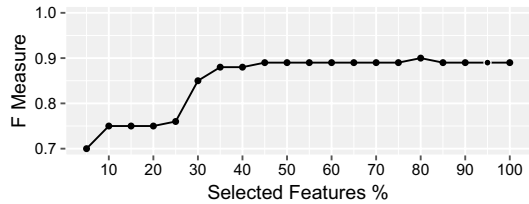
$$MRR = \frac{1}{|E|} \sum_{i=1}^{|E|} \frac{1}{Rank_i} \quad (3)$$

6.3 Cross-validation

We perform 5-fold cross-validation over the empirical dataset of 500 sample crashes. The dataset is randomly divided into 5 subsets of 100 sample crashes: 5 experiments are then carried where every time a specific subset of 100 is used as “test” data while the remaining subsets containing the rest 400 cases are merged to form “training” dataset. The computed performance metrics are then summed over the 5 folds.

6.4 Feature selection

In the empirical dataset, the vocabulary contains 1108 unique words. To avoid overfitting, we select only a portion of them for Phase 1. We use the χ^2 test for each word (Miller and Siegmund 1982). A higher value of χ^2 indicates a stronger correlation between the word and the category. Figure 5 shows the relation between the F Measure of Phase 1 and the percentage of words chosen (ranked in descending order by χ^2 values). We can see that with the top 50% of the features, the overall performance already stabilizes. We then always use top 50% of the words in the vocabulary.

Fig. 5 F measure versus selected features

7 Experimental results

7.1 RQ1: effectiveness of categorization

We use our ground truth of 500 crashes to assess the performance of ANCHOR in the first phase of the approach, namely the categorization. We provide in Table 3 the confusion matrix as well as the precision and recall of our experimental results. ANCHOR yields a very high precision for predicting crashes in Category A, reaching 0.96. The precision for crashes in Categories B and C are comparably lower, at 0.65 and 0.60, respectively. In terms of recall, the approach is effective for Category A (0.91), Category B (0.82), and Category C (0.75). Overall, ANCHOR is successful in categorizing 444 out of 500 crash samples (88.8%).

Answer to RQ1 ANCHOR is overall effective in categorizing new crash samples. However, there is still room of improving the precision when predicting samples in Categories B and C.

7.2 RQ2: effectiveness of localization

Since the two phases of ANCHOR are loosely coupled (i.e., each phase is an independent module that can be improved or replaced individually, as long as the in/output format is maintained), it may be insightful to investigate the performance of Phase II when crashes are previously perfectly categorized. In this way, the evaluation of Phase II is not affected by the performance of Phase I.

Table 3 Effectiveness of categorization (phase 1)

	Actual			Total
	A	B	C	
Predicted as category A	374	6	8	388
Predicted as category B	20	40	2	62
Predicted as category C	17	3	30	50
Total	411	49	40	500
	Precision	Recall		
Category A	0.96	0.91		
Category B	0.65	0.82		
Category C	0.60	0.75		

Table 4 Localization performance

Category	Recall@1	Recall@5	Recall@10	MRR
A	0.97 (400/411)	0.99 (406/411)	0.99 (407/411)	0.98
B	0.39 (19/49)	0.61 (30/49)	0.63 (31/49)	0.48
C	0.78 (31/40)	1.00 (40/40)	1.00 (40/40)	0.85
Total	0.90 (450/500)	0.95 (476/500)	0.96 (478/500)	0.92

To evaluate the localization phase of ANCHOR, we consider sample crashes for each category and assess the rank localization yielded by the specific algorithm developed for that category. Table 4 summarizes the Recall@k (with $k \in \{1, 5, 10\}$) and MRR.

To make sure the evaluation of Phase 2 is not affected by the outcome of Phase 1, we propose to assess the performance of localization with the assumption of perfect categorization.

For cases in Category A, the true fault location can almost always be found at the top of the rank. The high value of MRR at 0.98 confirms the intuition in Sect. 5.2.1 that it takes much less effort in finding fault location for Category A. For cases in Category B, the recall@1 starts at 0.39 and increased substantially for recall@5 at 0.61. One more case is successfully located with recall@10 at 0.63. The overall MRR is 0.48. Given the fact that the search space is vast (there can be tens of thousands of developer methods in the apk), Algorithm 1 demonstrates decent performance. For most cases in Category C, the true sub-category of the fault location can be found topmost, with the MRR at 0.85.

Answer to RQ2 The localization algorithms (Phase 2) of ANCHOR are reasonably effective for suggesting the correct fault location. ANCHOR shows descent performance even when challenged by the vast search space for crashes in Category B.

7.3 RQ3: overall performance of ANCHOR

Table 5 summarizes the overall performance of ANCHOR combining Phase 1 and 2. The MRR of all 3 categories slightly dropped, since some cases are mis-categorized in Phase 1. Clearly, the overall performance is affected by Phase 1. However, since the two phases in ANCHOR are loosely coupled, we envisage improvements of overall performance in the future when better classifiers are proposed.

Table 5 Overall performance of ANCHOR

Category	Recall@1	Recall@5	Recall@10	MRR
A	0.90 (370/411)	0.91 (373/411)	0.91 (373/411)	0.90
B	0.37 (18/49)	0.59 (29/49)	0.61 (30/49)	0.46
C	0.72 (29/40)	0.75 (30/40)	0.75 (30/40)	0.73
Total	0.83 (417/500)	0.86 (432/500)	0.87 (433/500)	0.85

Answer to RQ3 ANCHOR is an effective approach for locating crashing faults when they are in/outside stack traces, even outside code. Better performance is guaranteed when categorization (Phase 1) is further improved.

7.4 RQ4: performance in the wild

The heuristics based on which ANCHOR is built may be biased by the empirical dataset. To mitigate this threat, we assess the effectiveness of ANCHOR with a dataset selected in the wild. We want to verify to what extent ANCHOR can be generalized. We leverage the independent dataset prepared by Fan et al. (2018b) who thoroughly (by crawling the entire GitHub) and systematically (by applying strict criteria) collected 194 crashed apks from open-source Android repositories. Before evaluation, we apply the constraint rules of Sect. 4.1, and focus on the 69 relevant crash cases that could be identified. Note that this dataset contains true fault locations already verified by the app developers. Since the cases in the dataset are from a wide time span (2011–2017), the partition is randomly decided on normal distribution over the year of app release.

Table 6 shows the confusion matrix, as well as the precision and recall of Phase 1 (categorization) on this independent dataset. The precision for all categories is high, reaching 0.98 (54/55), 0.67 (6/9), and 0.80 (4/5) respectively. The recalls are also high, at 0.93 (54/58) for A, 0.86 (6/7) for B, and a perfect 1.00 (4/4) for C.

Table 7 provides measures for the overall performance. To compute the similarity scores which are required to locate the bug related to crashes from Categories B

Table 6 Categorization on an independent dataset

	Actual			Total
	A	B	C	
Predicted as category A	54	1	0	55
Predicted as category B	3	6	0	9
Predicted as category C	1	0	4	5
Total	58	7	4	69
	Precision	Recall		
Category A	0.98	0.93		
Category B	0.67	0.86		
Category C	0.80	1.00		

Table 7 Recall@k and MRR on an independent dataset

Category	Recall@1	Recall@5	Recall@10	MRR
A	0.72 (42/58)	0.93 (54/58)	0.93 (54/58)	0.81
B	0.43 (3/7)	0.43 (3/7)	0.43 (3/7)	0.43
C	0.25 (1/4)	1.00 (4/4)	1.00 (4/4)	0.40
Total	0.67 (46/69)	0.88 (61/69)	0.88 (61/69)	0.74

and C, we use the crash records from the empirical dataset. The recalls and MRR in Category A remain high. As for Category B, ANCHOR is able to yield recall@k values and MRR of 0.43 when suggesting fault locations. As for Category C, the total MRR is at 0.43, suggesting more stack traces in Category C might be the key for better performance.

Answer to RQ4 The evaluation on an independent dataset shows that ANCHOR can be generalized. ANCHOR is a milestone in this respect that it considers various crashing location cases. However, a community effort is still required to construct a representative dataset of crashes to push forward the state of the art in crashing fault localization.

8 Discussion

8.1 Comparing ANCHOR with other locators

Along with their empirical analysis of Android app crashes, Fan et al. (2018b); Su et al. (2020) mentioned a prototype crashing fault locator: ExLocator. Unfortunately, since the tool has not been publicly released, we could not directly compare it against ANCHOR. We note, based on its description, however, that ExLocator has a limited usage scenario since it focuses on only 5 exception types. Fan et al. (2018a) also studied one specific type of exception locating. CrashLocator (Wu et al. 2014) can also locate faults outside the stack trace. However, CrashLocator needs to abstract patterns from a great number of repeated crashes of the same project. Unfortunately, for both datasets presented in this paper, this requirement is not satisfied. Moreover, CrashLocator requires source code and change tracking of the projects, unavailable for our empirical dataset. Therefore, we are not able to apply CrashLocator.

Although direct comparison in terms of effectiveness is not possible in this scenario, we can compare the applicability. ANCHOR is considered to have a wider application range compared to ExLocator, i.e., it can be applied to all exception types, and considered to be more lightweight and source code independent compared to CrashLocator, i.e., it requires only the crash message and the apk.

8.2 Developer effort for locating bugs

In the motivating example, we demonstrated why locating buggy methods outside the stack trace can be challenging. We also want to measure the effort that developers put in locating such bugs. In Fan et al.'s dataset, each crash is documented with its duration, i.e., the time between the issue creation and its official closure by the developers. For bugs in the stack trace, it takes developers 26 days on average to close the issues. For bugs outside the stack trace, it drastically increases to 41 days. The ratio is $41/26=158\%$. Although it may not always be precise to measure effort in terms of issue duration, this would confirm our observation to some extent.

8.3 Improving locating for category B

In RQ3, we see that ANCHOR has the lowest performance for locating crashes in Category B. This is because Android apks may contain thousands of developer methods. CrashLocator (Wu et al. 2014) also reports similar overall MRR for large Java projects. Despite this challenge, we revisited some unsatisfying rankings and concluded two things that can be done to improve for Category B. First, the api_h (c.f Sect. 5.2.2) inferred could be incorrect. As a result, the rank never contains the true bug location. Computing api_h correctly calls for a larger and more sophisticated set of crash samples with ground truth. Because there's a higher possibility that such a set may contain samples sharing a similar root cause with the crash-under-study. Second, the true buggy location may appear low in the rank. This may be because the weight assigned to them is comparatively low. Refining the weight computation (c.f Line 7, Algorithm 1) may result in better performance.

8.4 Locating ICC- and IPC-related crashes

The Android framework adopts the *Intent* objects to facilitate ICC (Inter-Component Communication) and IPC (Inter-Process Communication) in order to provide a message passing mechanism for data exchange among components and even among apps (Arzt et al. 2014; Li et al. 2015). However, apps may also crash because of malformed *Intent* objects (Maji et al. 2012). When the crash arises from IPC, the true fault location may reside in other apps. In such a scenario, ANCHOR is not appropriate, since it only looks for fault locations inside the current apk. When the crash arises from the ICC, ANCHOR may be able to find the fault location. For example, one common exception in many crashes in our empirical dataset is *ActivityNotFoundException*. There are often two scenarios. First, the *Intent* is implicit, but there is no *Activity* in the current device/emulator that matches the *Intent*'s filter. Second, the *Intent* is explicit, but the specified destination *Activity* of the *Intent* is not available in the device/emulator. In both scenarios, ANCHOR tends to categorize the crash in Category A and effectively finds the true location. When other fields of the *Intent* is malformed and the true bug location is outside the stack trace, it is also possible that ANCHOR finds the buggy method by following Algorithm 1, i.e., by correctly computing api_h and putting the buggy method which misused api_h in the rank.

8.5 Generalization of ANCHOR

Android Remote Method Invocation (RMI) (Kang et al. 2016) allows Android apps to invoke methods from other devices running Android. Since ANCHOR only searches for bug locations inside Android apk files, if an app crashes from remotely invoked methods, ANCHOR will not be able to locate them. Apps running on other mobile platforms like iOS also suffer from app crashes, even those with billions of downloads (<https://www.theverge.com/21319817/facebook-ios-sdk-bug-fix-spotify-pinterest-tinder-crash-iphone-ios>). Similarly, a crash log containing the stack trace is also

available for crashed iOS apps. However, locating for such crashes is also challenged by bugs outside the stack trace (https://developer.apple.com/documentation/xcode/diagnosing_and_resolving_bugs_in_your_running_app). Therefore, it is possible to apply ANCHOR to locate crash-inducing bugs for iOS apps. Although some details of the algorithm need to be reconsidered, e.g., the core development language that iOS apps use is Swift, instead of Java or Kotlin for Android apps.

9 Threats to validity

9.1 Internal threats

9.1.1 Ground truth

In the empirical study presented in Sect. 4, we have manually built the ground truth of buggy locations that we made available to the community. Although we have tried our best to perform this manual inspection with the help of (1) the Android official documentation, (2) programmer information exchanging forums like Stack-Overflow or GitHub, (3) tools such as Soot or CodeInspect, there is no guarantee that all buggy locations we retrieved are the true causes for the crashes. This might affect the conclusions we draw from this dataset and the answers to RQ1-RQ3.

9.1.2 Taxonomy

By dividing bug locations into “code” and “non-code” (Category C), and by further dividing “code” into “in stack trace” (Category A) and “outside stack trace” (Category B), we are confident that our taxonomy guarantees exhaustion. However, our taxonomy is not the only option. Also, more fine-grained divisions can be included in this taxonomy, e.g., if the Android framework supports new non-code features in future versions, there might be more sub-categories in Category C.

9.2 External threats

9.2.1 Datasets

We extracted our dataset from the open benchmark ReCBench built by Kong et al. (2019). Although the large dataset they propose contains diverse apks collected from various popular app markets such as Google Play (ensuring a good diversity of apps), the collected crash cases are retrieved by testing apks with only two testing tools. Therefore, the yielded crashes could not be representative of the whole spectrum of crashes present in the Android ecosystem. Similarly, the dataset proposed by Fan et al. (2018b) is extracted from open source Android app GitHub repositories only. Moreover, they have applied certain rules for collecting the crashed cases, e.g., they extract only crash bugs that have been closed by repository maintainers. The

potential limitations with both datasets may affect the effectiveness we have shown in RQ1-RQ4.

9.2.2 Android framework evolution

Android framework is fast evolving (<https://source.android.com/setup/start/build-numbers>). New APIs are proposed in every version. Old APIs maybe deprecated or may have their logic renewed. This would result in new crashing faults previously unseen. However, ANCHOR is able to evolve with these new changes. First, for the categorization phase, we can include more crashing logs and the crashed apks retrieved from running apps on newer Android framework versions. Like all Machine Learning-based algorithms, our categorization phase suffers from concept drifting, therefore, it is necessary to retrain the model with new data. Second, such new crash logs also benefit the localization of crashes in Category B (e.g., new misused APIs and new non-overridden callbacks) and in Category C (e.g., new sub-categories).

10 Related work

A recent survey by Wong et al. (2016) marks the activity of identifying the locations of faults in a program to be most tedious, time-consuming, and expensive, yet equally critical. Therefore, lots of techniques have been proposed attempting to ease the work of finding the fault locations. Although we did not find a dedicated tool for identifying locations in Android apps, there are some approaches proposed for general software programs. For example, Wu et al. proposed CrashLocator (Wu et al. 2014) to score and rank suspicious locations that have caused program crashes. CrashLocator suggests that the buggy methods can be located in the static call graphs extended from the stack traces. However, it is not suitable to work on programs with multiple entry points and separate call graphs such as Android apps. Moreover, its scoring factors, which require source code and change histories, also limit its application scope to Android apps, for which most of them are released in a closed way (i.e., no change histories). Gu et al. (2019) proposed another approach called CraTer that adopts information retrieval techniques to predict whether the real fault resides inside the stack traces. However, CraTer is not able to suggest the actual buggy location. BugLocator (Zhou et al. 2012) applies a revisited Vector Space Model (rSVM) to retrieve relevant files for fixing a bug on a large number of bug reports. However, its granularity falls in file level, which still requires human verification for more fine-grained location identification. Wong et al. (2014) build their work on top of BugLocator (Zhou et al. 2012) and leveraged stack trace to improve the approach and indeed achieved better performance. Fan et al. (2018b) briefly describes a fault localization prototype ExLocator for Android apps. ExLocator only supports 5 exception types and has a limited usage scenario. APEChecker (Fan et al. 2018a) manifests async programming errors and can statically spot faults for this specific crash inducing error. Furthermore, in the community of Automatic Program Repair (APR), statement-level fault localization is often among the first few steps.

Researchers have improved it in various aspects (Abreu et al. 2007; Koyuncu et al. 2019; Abreu et al. 2009a; Jones and Harrold 2005; Abreu et al. 2009b; Wang et al. 2015; Lal and Sureka 2012).

Many research works have been proposed to address Android app crashes in recent years. For example, Fan et al. (2018b) performed a large scale analysis on framework-specific Android app crashes. They have invented the grouping techniques to group the Android app crash cases into buckets to study similar root causes based on each bucket. Researchers have also spent efforts attempting to automatically reproduce the reported crashes (Li et al. 2017a; Martin et al. 2016). Indeed, to achieve this purpose, Zhao et al. have proposed ReCDroid (Zhao et al. 2019), which applies a combination of natural language processing (NLP) and dynamic GUI exploration to reproduce given crashes. Gómez et al. (2016) proposed another approach for reproducing crashes by providing sensitive contexts. Moran et al. (2016) further presented a prototype tool called CrashScope, aiming at generating an augmented crash report to automatically reproduce crashes on target devices. Researchers have gone one step deeper to propose automated tools to automatically fix such identified crashes. Indeed, Tan et al. (2018) have proposed an automatic repairing framework named Droix for crashed Android apps. Droix adopts 8 manually constructed fixing patterns on crashed Android apps to generate app mutants and suggest one that fixes the crash. Following this work, Kong et al. (2019) present to the community an automatic fix pattern generation approach named CraftDroid for fixing apps suffering from crashes.

The special Android callback-based mechanism and its effect have drawn the attention of many researchers with the ever-booming of Android devices. Yang et al. (2015) targets the even-driven and multi-entry point issue of Android apps, and proposed a program representation that captures callback sequences by using context-sensitive static analysis of callback methods. Flowdroid (Arzt et al. 2014) targets at exposing privacy leakages on Android phones. It establishes a precise model of the Android lifecycle, which allows the analysis to properly handle callbacks invoked by the Android framework. Relda2 (Wu et al. 2016) is a light-weight and precise static resource leak detection tool based on Function Call Graph (FCG) analysis, which handles the features of the callbacks defined in the Android framework. Together with other existing works like (Yang and Yang 2012; Li et al. 2016a), they all dealt with Android callback-based mechanism in various manners. Although these works are different from ours, their approach in handling lifecycle and callback methods could be borrowed to enhance our approach towards better dealing with Category B crashes.

11 Conclusions

In this work, we performed an empirical study. This study shows that 37% crash types are related to bugs that are outside the stack traces, which imposes challenges to the localization problem. We then proposed ANCHOR, a two-phase categorization and localization tool that is able to generate a ranked list of bug locations for

developers to examine. The effectiveness of ANCHOR is assessed with both this empirical dataset and an in-the-wild scenario on a third-party dataset. Our work brings inspiring insights into the crashing faults localization problem for Android apps and calls for more attention from both the developers and the research community.

Acknowledgements This work was supported by the Fonds National de la Recherche (FNR), Luxembourg, under CORE projects CHARACTERIZE C17/IS/11693861 and RECOMMEND 15/IS/10449467, and AFR-PhD project 11620657.

References

- Abreu, R., Zoetewij, P., Van Gemund, A.J.: On the accuracy of spectrum-based fault localization. In: Testing: Academic and Industrial Conference Practice and Research Techniques-MUTATION (TAI-CPART-MUTATION 2007), IEEE, pp. 89–98 (2007)
- Abreu, R., Zoetewij, P., Golsteijn, R., Van Gemund, A.J.: A practical evaluation of spectrum-based fault localization. *J. Syst. Softw.* **82**(11), 1780–1792 (2009a)
- Abreu, R., Zoetewij, P., Van Gemund, A.J.: Spectrum-based multiple fault localization. In: 2009 IEEE/ACM International Conference on Automated Software Engineering, IEEE, pp. 88–99 (2009b)
- Allix, K., Bissyande, T.F., Klein, J., Traon, Y.L.: AndroZoo: Collecting Millions of Android Apps for the Research Community. In: 2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR), pp 468–471, (2016) <https://doi.org/10.1109/MSR.2016.056>
- Arzt, S., Rasthofer, S., Fritz, C., Bodden, E., Bartel, A., Klein, J., Le Traon, Y., Octeau, D., McDaniel, P.: Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *ACM Sigplan Not. ACM* **49**, 259–269 (2014)
- Bartel, A., Klein, J., Le Traon, Y., Monperrus, M.: Dexpler: converting android dalvik bytecode to jimple for static analysis with soot. In: Proceedings of the ACM SIGPLAN International Workshop on State of the Art in Java Program analysis, pp 27–38 (2012)
- Bielik, P., Raychev, V., Vechev, M.: Scalable race detection for android applications. *ACM SIGPLAN Not.* **50**(10), 332–348 (2015)
- Fan, L., Su, T., Chen, S., Meng, G., Liu, Y., Xu, L., Pu, G.: Efficiently manifesting asynchronous programming errors in android apps. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, pp 486–497 (2018a)
- Fan, L., Su, T., Chen, S., Meng, G., Liu, Y., Xu, L., Pu, G., Su, Z.: Large-scale analysis of framework-specific exceptions in android apps. In: Proceedings of the 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), IEEE, pp 408–419 (2018b)
- Fraunhofer (2020) Codeinspect tool of fraunhofer. <https://codeinspect.sit.fraunhofer.de/>
- Gómez, M., Rouvoy, R., Adams, B., Seinturier, L.: Reproducing context-sensitive crashes of mobile apps using crowdsourced monitoring. In: International Conference on Mobile Software Engineering and Systems (2016)
- Google (2020) Crashes | android developers. <https://developer.android.com/topic/performance/vitals/crash>
- Gu, Y., Xuan, J., Zhang, H., Zhang, L., Fan, Q., Xie, X., Qian, T.: Does the fault reside in a stack trace? assisting crash localization by predicting crashing fault residence. *J. Syst. Softw.* **148**, 88–104 (2019)
- Huang, Y., Li, L.: Naive bayes classification algorithm based on small sample set. In: 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, IEEE, pp 34–39 (2011)
- Inc A: Diagnosing and resolving bugs in your running app. https://developer.apple.com/documentation/xcode/diagnosing_and_resolving_bugs_in_your_running_app
- Indi, T.S., Yalagi, P.S., Nirgude, M.A.: Use of java exception stack trace to improve bug fixing skills of intermediate java learners. In: 2016 International Conference on Learning and Teaching in Computing and Engineering (LaTICE), IEEE, pp 194–198 (2016)
- Jiang, S., Zhang, H., Wang, Q., Zhang, Y.: A debugging approach for java runtime exceptions based on program slicing and stack traces. In: 2010 10th International Conference on Quality Software, IEEE, pp 393–398 (2010)

- Jones, J.A., Harrold, M.J.: Empirical evaluation of the tarantula automatic fault-localization technique. In: Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering, ACM, pp 273–282 (2005)
- Kang, H., Jeong, K., Lee, K., Park, S., Kim, Y.: Android rmi: a user-level remote method invocation mechanism between android devices. *J. Supercomput.* **72**(7), 2471–2487 (2016)
- Kim, D., Wang, X., Kim, S., Zeller, A., Cheung, S.C., Park, S.: Which crashes should i fix first?: Predicting top crashes at an early stage to prioritize debugging efforts. *IEEE Trans. Softw. Eng.* **37**(3), 430–447 (2011)
- Kong, P., Li, L., Gao, J., Bissyandé, T.F., Klein, J.: Mining android crash fixes in the absence of issue-and change-tracking systems. In: International Symposium on Software Testing and Analysis (ISSTA), ACM, pp 78–89 (2019)
- Koyuncu, A., Liu, K., Bissyandé, T.F., Kim, D., Monperrus, M., Klein, J., Le Traon, Y.: ifixr: bug report driven program repair. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ACM, pp 314–325 (2019)
- Lal, S., Sureka, A.: A static technique for fault localization using character n-gram based information retrieval model. In: Proceedings of the 5th India Software Engineering Conference, pp 109–118 (2012)
- Li, L., Bartel, A., Bissyandé, T.F., Klein, J., Le Traon, Y., Arzt, S., Rasthofer, S., Bodden, E., Octeau, D., McDaniel, P.: Iccta: Detecting inter-component privacy leaks in android apps. In: Proceedings of the 37th International Conference on Software Engineering-Volume 1, IEEE Press, pp 280–291 (2015)
- Li, L., Bissyandé, T.F., Octeau, D., Klein, J.: Droidra: Taming reflection to support whole-program analysis of android apps. In: The 2016 International Symposium on Software Testing and Analysis (ISSTA 2016) (2016a)
- Li, L., Bissyandé, T.F., Papadakis, M., Rasthofer, S., Bartel, A., Octeau, D., Klein, J., Le Traon, Y.: Static analysis of android apps: a systematic literature review. *Inf. Softw. Technol.* **88**, 67–95 (2017a)
- Li, L., Gao, J., Hurier, M., Kong, P., Bissyandé, T.F., Bartel, A., Klein, J., Le Traon, Y.: Androzoo++: Collecting millions of android apps and their metadata for the research community. (2017b) arXiv preprint [arXiv:170905281](https://arxiv.org/abs/170905281)
- Li, L., Bissyandé, T.F., Wang, H., Klein, J.: Cid: Automating the detection of api-related compatibility issues in android apps. In: The ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2018) (2018)
- Li, Q., Jiang, Y., Gu, T., Xu, C., Ma, J., Ma, X., Lu, J.: Effectively manifesting concurrency bugs in android apps. In: 2016 23rd Asia-Pacific Software Engineering Conference (APSEC), IEEE, pp 209–216 (2016b)
- LLC G: Codenames, tags, and build numbers. <https://source.android.com/setup/start/build-numbers>
- LLC G App manifest overview. (2020a) <https://developer.android.com/guide/topics/manifest/manifest-intro>, Accessed: 2020-01-26
- LLC G Test google play billing. (2020b) https://developer.android.com/google/play/billing/billing_testing.html, Accessed: 2020-01-26
- Maiya, P., Kanade, A., Majumdar, R.: Race detection for android applications. *ACM SIGPLAN Not.* **49**(6), 316–325 (2014)
- Maji, A.K., Arshad, F.A., Bagchi, S., Rellermeyer, J.S.: An empirical study of the robustness of inter-component communication in android. In: IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2012), IEEE, pp 1–12 (2012)
- Martin, W., Sarro, F., Jia, Y., Zhang, Y., Harman, M.: A survey of app store analysis for software engineering. *IEEE Trans. Softw. Eng.* **43**(9), 817–847 (2016)
- Metsis, V., Androustopoulos, I., Paliouras, G.: Spam filtering with naive bayes-which naive bayes? *CEAS Mountain View CA* **17**, 28–69 (2006)
- Miller, R., Siegmund, D.: Maximally selected chi square statistics. *Biometrics*, pp. 1011–1016 (1982)
- Moran, K., Linares-Vásquez, M., Bernal-Cárdenas, C., Vendome, C., Poshvanyk, D.: Automatically discovering, reporting and reproducing android application crashes. In: Software Testing, Verification and Validation (ICST), 2016 IEEE International Conference on, IEEE, pp 33–44 (2016)
- Octeau, D., Luchau, D., Dering, M., Jha, S., McDaniel, P.: Composite constant propagation: Application to android inter-component communication analysis. In: Proceedings of the 37th International Conference on Software Engineering-Volume 1, IEEE Press, pp 77–88 (2015)

- Qin, J., Wang, W., Lu, Y., Xiao, C., Lin, X.: Efficient exact edit similarity query processing with the asymmetric signature scheme. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pp 1033–1044 (2011)
- Rish, I., et al.: An empirical study of the naive bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol 3, pp 41–46 (2001)
- Schroter, A., Schröter, A., Bettenburg, N., Premraj, R.: Do stack traces help developers fix bugs? In: 2010 7th IEEE Working Conference on Mining Software Repositories (MSR 2010), IEEE, pp 118–121 (2010)
- Shi Y, Karatzoglou A, Baltrunas L, Larson M, Oliver N, Hanjalic A (2012) Clmf: learning to maximize reciprocal rank with collaborative less-is-more filtering. In: Proceedings of the sixth ACM conference on Recommender systems, pp 139–146
- Sinha, S., Shah, H., Görg, C., Jiang, S., Kim, M., Harrold, M.J.: Fault localization and repair for java runtime exceptions. In: Proceedings of the eighteenth international symposium on Software testing and analysis, pp 153–164 (2009)
- Song, W., Zhang, J., Huang, J.: Servdroid: detecting service usage inefficiencies in android applications. In: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp 362–373 (2019)
- Su, T., Fan, L., Chen, S., Liu, Y., Xu, L., Pu, G., Su, Z.: Why my app crashes understanding and benchmarking framework-specific exceptions of android apps. In: IEEE Transactions on Software Engineering (2020)
- Tan, S.H., Dong, Z., Gao, X., Roychoudhury, A.: Repairing crashes in android apps. In: Proceedings of the 40th International Conference on Software Engineering, ACM, pp 187–198 (2018)
- Tang, H., Wu, G., Wei, J., Zhong, H.: Generating test cases to expose concurrency bugs in android applications. In: Proceedings of the 31st IEEE/ACM international Conference on Automated software engineering, pp 648–653 (2016)
- Verge T: How to stop facebook's ios bug from crashing spotify, pinterest, tinder, and other apps. <https://www.theverge.com/21319817/facebook-ios-sdk-bug-fix-spotify-pinterest-tinder-crash-iphone-ios>
- Walton, N.A., Poynton, M.R., Gesteland, P.H., Maloney, C., Staes, C., Facelli, J.C.: Predicting the start week of respiratory syncytial virus outbreaks using real time weather variables. BMC Med. Inform. Dec. Mak. **10**(1), 68 (2010)
- Wang, J., Jiang, Y., Xu, C., Li, Q., Gu, T., Ma, J., Ma, X., Lu, J.: Aatt+: effectively manifesting concurrency bugs in android apps. Sci. Comput. Program. **163**, 1–18 (2018)
- Wang, Q., Parnin, C., Orso, A.: Evaluating the usefulness of ir-based fault localization techniques. In: Proceedings of the 2015 International Symposium on Software Testing and Analysis, pp 1–11 (2015)
- Wei, L., Liu, Y., Cheung, S.C.: Taming android fragmentation: Characterizing and detecting compatibility issues for android apps. In: Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, ACM, pp 226–237 (2016)
- Wong, C.P., Xiong, Y., Zhang, H., Hao, D., Zhang, L., Mei, H.: Boosting bug-report-oriented fault localization with segmentation and stack-trace analysis. In: 2014 IEEE International Conference on Software Maintenance and Evolution, IEEE, pp 181–190 (2014)
- Wong, W.E., Gao, R., Li, Y., Abreu, R., Wotawa, F.: A survey on software fault localization. IEEE Trans. Softw. Eng. **42**(8), 707–740 (2016)
- Wu, R., Zhang, H., Cheung, S.C., Kim, S.: Crashlocator: locating crashing faults based on crash stacks. In: Proceedings of the 2014 International Symposium on Software Testing and Analysis, ACM, pp 204–214 (2014)
- Wu, T., Liu, J., Xu, Z., Guo, C., Zhang, Y., Yan, J., Zhang, J.: Light-weight, inter-procedural and call-back-aware resource leak detection for android apps. IEEE Trans. Softw. Eng. **42**(11), 1054–1076 (2016)
- Yang, S., Yan, D., Wu, H., Wang, Y., Rountev, A.: Static control-flow analysis of user-driven callbacks in android applications. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, IEEE, vol 1, pp 89–99 (2015)
- Yang, Z., Yang, M.: Leakminer: Detect information leakage on android with static taint analysis. In: 2012 Third World Congress on Software Engineering, IEEE, pp 101–104 (2012)
- Yang, Z., Nie, X., Xu, W., Guo, J.: An approach to spam detection by naive bayes ensemble based on decision induction. In: Sixth International Conference on Intelligent Systems Design and Applications, IEEE, vol 2, pp 861–866 (2006)

- Zhao, Y., Yu, T., Su, T., Liu, Y., Zheng, W., Zhang, J., Halfond, W.G.: Reccroid: automatically reproducing android application crashes from bug reports. In: Proceedings of the 41st International Conference on Software Engineering, IEEE Press, pp 128–139 (2019)
- Zhou, J., Zhang, H., Lo, D.: 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), IEEE, pp 14–24 (2012)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Pingfan Kong¹ · Li Li²  · Jun Gao¹ · Timothée Riom¹ · Yanjie Zhao² · Tegawendé F. Bissyandé¹ · Jacques Klein¹

Pingfan Kong
pingfan.kong@uni.lu

Jun Gao
jun.gao@uni.lu

Timothée Riom
timothee.riom@uni.lu

Yanjie Zhao
yanjie.zhao@monash.edu

Tegawendé F. Bissyandé
tegawende.bissyande@uni.lu

Jacques Klein
jacques.klein@uni.lu

¹ University of Luxembourg, Esch-sur-Alzette, Luxembourg

² Monash University, Clayton, Australia