A Machine-learning Approach for Classifying and Categorizing Android Sources and Sinks

Siegfried Rasthofer & Steven Arzt
Secure Software Engineering Group
EC SPRIDE, Technische Universität Darmstadt
{firstname.lastname}@ec-spride.de

Eric Bodden
Secure Software Engineering Group
Fraunhofer SIT & Technische Universität Darmstadt
eric.bodden@sit.fraunhofer.de

***Abstract*—Today's smartphone users face a security dilemma: many apps they install operate on privacy-sensitive data, although they might originate from developers whose trustworthiness is hard to judge. Researchers have addressed the problem with more and more sophisticated static and dynamic analysis tools as an aid to assess how apps use private user data. Those tools, however, rely on the manual configuration of lists of *sources* of sensitive data as well as *sinks* which might leak data to untrusted observers. Such lists are hard to come by.**

We thus propose SuSi, a novel machine-learning guided approach for identifying sources and sinks directly from the code of any Android API. Given a training set of hand-annotated sources and sinks, SuSi identifies other sources and sinks in the entire API. To provide more fine-grained information, SuSi further categorizes the sources (e.g., unique identifier, location information, etc.) and sinks (e.g., network, file, etc.).

For Android 4.2, SuSi identifies hundreds of sources and sinks with over 92% accuracy, many of which are missed by current information-flow tracking tools. An evaluation of about 11,000 malware samples confirms that many of these sources and sinks are indeed used. We furthermore show that SuSi can reliably classify sources and sinks even in new, previously unseen Android versions and components like Google Glass or the Chromecast API.

I. Introduction

Current smartphone operating systems, such as Android or iOS, allow users to run a multitude of applications developed by many independent developers available through various app markets. While this flexibility is very convenient for the user, as one will find a suitable application for almost every need, it also makes it hard to determine the trustworthiness of these applications.

Smartphones are widely used to store and process highly sensitive information such as text messages, private and business contacts, calendar data, and more. Furthermore, while a large variety of sensors like GPS allow a context-sensitive user experience, they also create additional privacy concerns if used for tracking or monitoring.

To address this problem, researchers have proposed various analysis tools to detect and react to data leaks, both statically [1]–[13] and dynamically [14]–[17]. Virtually all of these tools are configured with a privacy policy, usually defined in terms of lists of *sources* of sensitive data (e.g., the user's current location) and *sinks* of potential channels through which such data could leak to an adversary (e.g., a network connection). As an important consequence, no matter how good the tool, it can only provide security guarantees if its list of sources and sinks is complete. If a source is missing, a malicious app can retrieve its information without the analysis tool noticing. A similar problem exists for information written into unrecognized sinks.

This work focuses on Android. As we show, existing analysis tools, both static and dynamic, focus on a handful of hand-picked sources and sinks, and can thus be circumvented by malicious applications with ease. It would be too simple, though, to blame the developers of those tools. Android's version 4.2, for instance, comprises about 110,000 public methods, which makes a manual classification of sources and sinks clearly infeasible. Furthermore, each new Android version includes new functionality (e.g., NFC in Android 2.3 or Restricted Profiles in the brand-new Android 4.3) which often also leads to new sources and sinks. This shows that a manual identification of sources and sinks is impractical. It would impose a high workload on the analyst and would have to be done again for every new Android version. Additionally, hand-picking is an error-prone task.

We therefore propose SuSi, an automated machine-learning guided approach for identifying sources and sinks directly from the code of an Android API. We have identified both semantic and syntactic features to train a model for sources and sinks on a small subset of hand-classified Android API methods. SuSi can then use this model to classify arbitrarily large numbers of previously unknown Android API methods. In the Android 4.2 operating system, SuSi finds *several hundred sources and sinks*, only a small fraction of which were previously known from the scientific literature or included in configurations of available analysis tools.

While SuSi is not able to identify each and every source or sink, it resembles a practical best-effort solution that solves the problem to a large extent, which is a substantial improvement over existing hand-picked sets. In cross-validation, SuSi achieves a precision and recall of over 92%, which means

NDSS '14, 23-26 February 2014, San Diego, CA, USA
Copyright 2014 Internet Society, ISBN 1-891562-35-5
http://dx.doi.org/10.14722/ndss.2014.23039

that the use of SuSi to identify sources and sinks greatly reduces (though not completely eliminates) the risk of missing sensitive data flows when used for configuring information-flow tracking tools. To evaluate how well SuSi predicts sources and sinks outside the training set, we applied SuSi's model to the Google Mirror API [18] that can be used for the communication between Google Glass [19] and an Android smartphone. We also applied SuSi to the Google Cast API [20] which is used for screen-sharing between smartphones and televisions, in particular using the new Chromecast device. Manual validation of SuSi's results on these new APIs shows an average precision and recall of over 98% for both Google Cast and Google Mirror. An evaluation of 11,000 malware samples from Virus Share [21] shows that malware does cause data leaks using Android API methods recognized as sources or sinks by SuSi, but missed by existing static and dynamic taint analysis tools including TaintDroid [14] or SCanDroid [5] (more details in Section V-C).

SuSi is the first dedicated approach to detecting sources and sinks. Due to missing lists of sources and sinks, some code-analysis approaches (for instance LeakMiner [3]) so far consider those methods as sources and sinks that require a permission to execute. These methods can be identified using a permission map which can be created either statically [22], [23] or dynamically [24]. As we show in this work, permission lists are a less than optimal heuristic for detecting sources and sinks: many methods called in the control-flow of permission checks are neither sources nor sinks, and even worse some calls to methods that are sources or sinks are not protected by permission checks. As an example, SuSi identifies as source the unprotected *getNetworkOperatorName()* method in the *TelephonyManager* class, which returns the name of the network operator or carrier. Our study reveals malware samples that use this method for reading out the network operator name and sending it to a malicious server. Furthermore, permission checks are scattered over several layers of the Android operating system. The *Internet* permission, for instance, is checked in native code while most other permissions are enforced in the middleware. Fortunately, as our work shows, the implementation of the Android API on the middleware layer reveals clues that help identify sources and sinks much better than by just using permission checks.

Awareness of sources and sinks is highly useful but if a leak is found, the user often desires additional information on *what* information has leaked *where*, for instance location information to the Internet. SuSi thus further classifies the identified sources and sinks into 12 source categories and 15 sink categories. The categorization shows that there is often more than one way to retrieve a certain piece of data, and that there are multiple ways to send it out to an attacker since all categories contain more than a single method.

This paper presents the following original contributions:

- a practical and precise definition of data sources and sinks in Android applications,
- an automated, machine-learning based approach for identifying data source and sink methods in the Android framework, even in case of new, previously unseen Android versions and variants,
- a classifier for data source and sink methods into semantic categories like network, files, contact data, etc., and
- a categorized list of sources and sinks for different Android versions, as well as the Google Mirror and Google Cast APIs. The list can be directly used by existing static and dynamic analysis approaches.

Our complete implementation is available as an open-source project at:

https://github.com/secure-software-engineering/SuSi

The remainder of this paper is structured as follows. Section II presents a motivating example, while Section III gives a precise definition of the notions of sources and sinks. Section IV presents the classifiers, which we evaluate in Section V. Section VI discusses other sources of sensitive information that are not directly related to method calls. In Section VII we give an overview of related work. Section VIII concludes.

II. Motivating Example

As mentioned earlier, comprehensive lists of sources and sinks are hard to come by. As a consequence, lists of sources and sinks known from the scientific literature [4], [5], [14] only contain a few well-known Android API methods for obtaining and sending out potentially sensitive information. (Section VI gives detailed information about the current state of the art.) However, there are often multiple ways to achieve the same effect. Developers of malicious applications can thus choose less well known sources and sinks to circumvent analysis tools. Let us assume an attacker is interested in obtaining the user's location information and writing it to a publicly accessible file on the internal storage without being noticed by existing program-analysis approaches.

Listing 1 shows an example that attempts to disguise a data leak by using less common methods for both the source and the sink. In our scenario, we have two source methods. Firstly, line 9 calls *getCid()*, returning the cell ID. Line 11 then calls *getLac()*, returning the location area code. Both pieces of data in combination can be used to uniquely identify the broadcast tower servicing the current GSM cell. While this is not an exact location, it nevertheless provides the approximate whereabouts of the user. In line 12 the code checks for a well-known cell-tower ID in Berlin, Germany. An actual malicious app would perform a lookup in a more comprehensive list. Finally, the code needs to make the data available to the attacker. The example creates a publicly accessible file on the phone's internal storage, which can be accessed by arbitrary other applications without requiring any permissions. Instead of employing Java's normal file writing functions, the code uses a little-known Android system function (line 17) which SuSi identifies as a "FILE" sink but which is normally hidden from the SDK: the *FileUtils.stringToFile* function can only be used if the application is compiled against a complete platform JAR file obtained from a real phone, as the *android.jar* file supplied with the Android SDK does not contain this method. Nevertheless, the example application runs on an unmodified stock Android phone.

```
void onCreate() {
TelephonyManager tm; GsmCellLocation loc;
// Get the location
tm = (TelephonyManager) getContext().
    getSystemService
        (Context.TELEPHONY_SERVICE);
loc = (GsmCellLocation)
    tm.getCellLocation();

//source: cell-ID
int cellID = loc.getCid();
//source: location area code
int lac = loc.getLac();
boolean berlin = (lac == 20228 && cellID
    == 62253);

String taint = "Berlin: " + berlin + " ("
    + cellID + " | " + lac + ")";
String f = this.getFilesDir() +
    "/mytaintedFile.txt";
//sink
FileUtils.stringToFile(f, taint);
//make file readable to everyone
Runtime.getRuntime().exec("chmod 666 "+f);
}
```

Listing 1. Android Location Leak Example

This example is, at least for the source methods, a representative example for malware [21] we inspected. We have tested this example with publicly-available static and dynamic taint analysis tools including Fortify SCA [4], SCanDroid [5], IBM AppScan [13] and TaintDroid [14] and confirmed that none of these tools detected the leak. This shows how important it is to generate a comprehensive list of sources and sinks for detecting malicious behavior in deceptive applications. SuSi discovers and classifies appropriately all sources and sinks used in the example.

III. Definition of Sources and Sinks

Before one can infer sources and sinks, one requires a precise definition of the terms “source” and “sink”. Several publications in the area of taint and information-flow analysis discuss sources and sinks, but all leave open the precise definitions of these terms. For instance, Enck et al. [14] define sinks informally as “data that leaves the system” which is, however, too imprecise to train a machine-learning based classifier; such classifiers are only as good as their training data.

Taint and information-flow analysis approaches track through the program the flow of *data*. Sources are where such data flows enter the program and sinks are where they leave the program again. This requires us to first define *data* in the context of data flows in Android applications.

Definition 1 (Data): A piece of data is a value or a reference to a value.

For instance, the IMEI in mobile applications is a piece of data, as would be the numerical value 42. We also treat as data, for instance, a database cursor pointing to a table of contact records, since it directly points to a value and is thus equivalent in terms of access control.

In taint tracking, one monitors the flow of data between resources such as the file system or network. Conversely, due to Android’s app isolation, data that is simply stored in the app’s address space is not of interest. Before one can define sources and sinks, one must therefore define the notion of a resource method. Mobile operating systems like Android enable applications to access resources using predefined methods. While one could also imagine fields being used for resource access, we found this not to be the case with Android.

Definition 2 (Resource Method): A resource method reads data from or writes data to a shared resource.

For instance, the operating system method for reading the IMEI (*getDeviceId()* in class *TelephonyManager*) is a resource method. In this case, the phone’s hardware itself is the resource as the IMEI is branded into the silicon. The *sendTextMessage()* method in class *SmsManager* is a resource method for sending text messages to a specific phone number. The resource is the GSM network.

Note that a writing resource method does not necessarily need a reading counterpart. In our definition, there is no restriction on how the data is shared. A writing resource method might, for instance, send out data over the network (which is a resource). Though another application cannot directly obtain this data through a simple method call, the data can easily be sniffed from the network and is thus shared. Data leaving the phone is thus always considered shared.

After defining *data* and *resource methods* we can now define sources and sinks in the context of Android applications:

Definition 3 (Android Source): Sources are calls into resource methods returning non-constant values into the application code.

The *getDeviceId()* resource method is an Android source. It returns a value (the IMEI) into the application code. The IMEI is considered non-constant as the method returns a different value on every phone. Looking at the source code alone does not reveal this value. In contrast, a function that just reads a fixed constant from a database is a resource method but, by our definition, is not an Android source.

Note that our definition of sources does not make any restrictions on whether the data obtained from a source is actually *private*. SuSi will thus, at first, report sources of non-private data as well. However, in a second step SuSi then applies a further categorization which partitions sources into different categories of private data. This partitioning includes a class NO_CATEGORY, which represents sources of non-private data, which privacy-analysis tools can ignore. Details will be given in Section IV.

Definition 4 (Android Sinks): Sinks are calls into resource methods accepting at least one non-constant data value from the application code as parameter, if and only if a new value is written or an existing one is overwritten on the resource.

The *sendTextMessage()* resource method is an Android sink as both the message text and the phone number it receives are possibly non-constant. On the other hand, the *reboot* method in the *PowerManager* class, for instance, just receives a kernel code for entering special boot modes which must be part of a predefined set of supported flags. This method is thus only a resource method (the data is written into the kernel log), but not an Android Sink. We require this restriction on constant values for methods which do not introduce any new information into the calling application in the case of sources, or do not directly leak any data across the application boundary in the case of sinks. The values at calls to such methods are of a purely technical kind (e.g., system constants, network pings etc.) and not of interest to typical analysis tools. Note that our definition also excludes some implicit information flows. This is a design choice. For instance, in our approach the vibration state of the phone is not considered a single-bit resource, even though it could theoretically be observed and would then be "shared".

A malicious app can try to access private information not only through calls to the official Android framework API but also through calls to code of pre-installed apps. For instance, the default email application provides a readily-available wrapper around the *getDeviceId()* function. This app is pre-installed on every stock Android phone, which gives a malicious app easy access to the wrapper: the app just instructs the Android class loader to load the respective system APK file and then instantiates the desired class. To cover such cases, our approach does not only analyze the framework API but the pre-installed apps as well. (We use a Samsung Galaxy Nexus with Android 4.2.). In other words, our analysis boundary is between a (potentially malicious) user application and all components pre-installed on the device.

IV. CLASSIFICATION APPROACH

In this section, we explain the details of SuSi, our machine-learning approach to automatically identify sources and sinks corresponding to the definitions given in Section III. We address two classification problems. For a given unclassified Android method, SuSi first decides whether it is a *source*, a *sink*, or *neither*. The second classification problem refines the classification of sources and sinks identified in the first step. All methods previously classified as *neither* are ignored. For an uncategorized source or sink, SuSi determines the most likely semantic category it belongs to. In our design, every method is assigned to exactly one category.

Section IV-A gives a short introduction to machine learning. Section IV-B then presents the general architecture of SuSi, while Section IV-C discusses the features SuSi uses to solve its classification problems. Section IV-D gives more details on one particularly important family of features which deals with data flows inside the methods to be classified. In Section IV-E we show how the semantics of the Java programming language can be exploited to artificially generate further annotated training data.

A. Machine Learning Primer

SuSi uses *supervised learning* to train a classifier on a relatively small subset of manually-annotated training examples. This classifier is afterwards used to predict the class of an arbitrary number of previously unseen test examples. Classification is performed using a set of features. A feature is a function that associates a training or test example with a value, i.e., evaluates a certain single domain-specific criterion for the example. The approach assumes that for every class there is a significant correlation between the examples in the class and the values taken by the feature functions.

ID	Experience	Alcohol	Phone No	Accident
T1	5 yrs	0.6	1234	yes
T2	11 yrs	0.4	45646	yes
T3	7 yrs	0.2	76546	yes
T4	4 yrs	0.0	54645	no
T5	10 yrs	0.2	78354	no
C1	6 yrs	0.1	6585	?
C2	12 yrs	0.55	67856	?

TABLE I. CLASSIFICATION EXAMPLE ON DRUNK DRIVING

As a simple example, consider the problem of estimating the risk of a driving accident for an insurance company. We may identify three features: years of experience, blood alcohol level and the driver's phone number. Assume the learning algorithm deduces that a higher level of experience is negatively correlated with the accident rate, while the alcohol level is positively correlated and the phone number is completely unrelated. The impact of a single feature on the overall estimate is deduced from its value distribution over the annotated training set. If there are many examples with high-alcohol accidents, then this feature will be given a greater weighting than the years of experience. However, if there are more accidents of inexperienced drivers in the training set than alcohol-related issues, the classifier will rank the experience feature higher.

The classifier works on a matrix, organized with one column per feature and one row per instance. Table I shows some sample data. An additional column indicates the class and is only filled in for the training data. In our example, this column would indicate whether or not an accident took place. The first five rows are training data, the last two rows are test records to be classified.

In this example, a simple rule-based classifier would deduce that all reports with alcohol levels larger than 0.2 also contained accidents, so C2 would be classified as *accident:yes*. However, since the converse does not hold, further reasoning is required for C1. Taking the experience level into account, there are two records of inexperienced drivers with levels of 0.2 or below in our test set: one with an accident and one without. In this case, the classifier would actually pick randomly, since both *accident:yes* and *accident:no* are equally likely. A probabilistic classifier could also choose *accident:yes* because accidents are more likely for inexperienced drivers (two out of three with five years of experience or less in this test data set) in general. This demonstrates that results can differ depending on the choice of the classifier.

As a concrete classifier, we use *support vector machines* (SVM), a margin classifier, more precisely the *SMO* [25] implementation in Weka [26] with a linear kernel. We optimize for minimal error. The basic principle of an SVM is to represent training examples of two classes (e.g., "sink" and "not a sink") using vectors in a vector space. The algorithm then tries to find a hyper-plane separating the examples. For a new, previously unseen test example, to determine its estimated

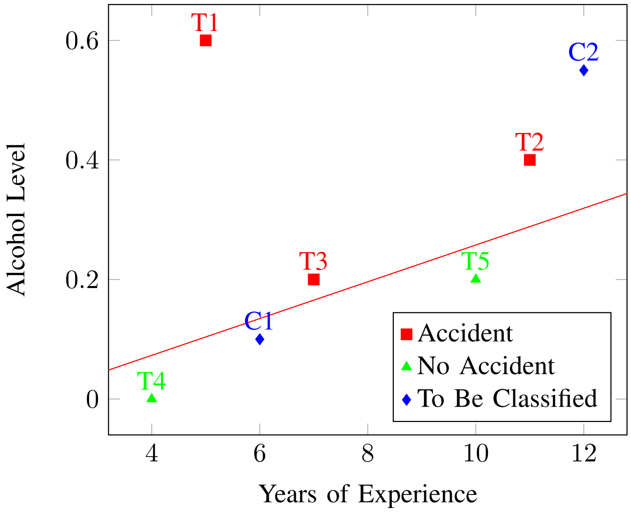

Fig. 1. SMO Classification Example

class, it checks on which side of the hyper-plane it belongs. In general, problems can be transformed into higher-dimensional spaces if the data is not linearly separable, but this did not prove necessary for any one of our classification problems.

Figure 1 shows an SMO diagram for Table I. We have not included the phone number feature since it is unrelated to the probability of an accident. The red line shows a projection of the hyper-plane. In this example, the SMO detects that all points above the line are positive examples (i.e., records of accidents), and all points below are negative ones (i.e., no accident). C2 would thus be classified as an accident, just as with the simple rule-based classifier above, but C1 would now definitely be classified as non-accident because it lies below the line.

SMO is only capable of separating two classes. However, in SuSi, we have three classes in the first problem (source/sink/neither) and a lot more in the second one (the categorization). We solve the problem with a one-against-all classification, a standard technique in which every possible class is tested against all other classes packed together to find out whether the instance corresponds to the current single class or whether the classification must proceed recursively to decide between the remaining classes.

We also evaluated other classification algorithms based on different principles, for instance Weka's J48 rule learner, which implements a pruned C4.5 decision tree [27]. The main problem with a rule set is its lack of flexibility. While many source-method names, for instance, start with *get*, this is not the case for *all* source methods. On the other hand, not all methods that start with *get* are actually sources. Since this rule of thumb is correct most of the time, however, a rule tree would usually include a rule mapping all *get* methods to sources and only perform further checks if the method name has a different prefix. With an SVM, such aspects that are usually correct, but not always, can be expressed more appropriately by shifting the hyper-plane used for separation.

Probabilistic learning algorithms like Naive Bayes [28] produced very imprecise results. This happens because our classification problem is *almost* rule-based, i.e., has an almost fixed semantics. The variance is simply not large enough to justify the imprecision introduced by probabilistic approaches which are rather susceptible to outliers.

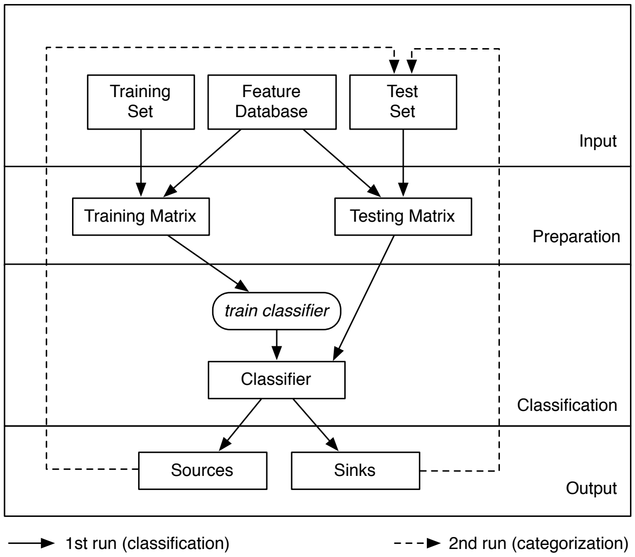

Fig. 2. Machine learning approach

B. Design of the Approach

Figure 2 shows SuSi's overall architecture. It includes four different layers: *input*, *preparation*, *classification*, and *output*. The square elements denote objects, while the round elements represent actions. We run two rounds: One for classifying methods as *sources*, *sinks*, or *neither*, and one for categorizing them. Solid lines denote the data flow within SuSi. The two dashed lines denote the initialization of the second round. The general process is the same for both rounds. For the categorization, SuSi just takes the outputs of the classification as test data inputs. More precisely, SuSi categorizes separately those methods it has previously identified as sources or sinks and disregards those it classified as neither.

SuSi starts with the input data for the first classification problem, i.e., for identifying sources and sinks. This data consists of the Android API methods to analyze. These methods can be separated into a set of training data (hand-annotated training examples) and a set of test data for which we do not know whether a method is a source, sink or neither. The set of training data is much smaller than the set of unknown test data, in our case only roughly 0.7% for the classification and about 0.4% for the categorization. Beside the API methods we need a database of features, both for the classification and categorization. The features are different for classification and categorization. See Section IV-C for details.

As described in in Section IV-A, a supervised learning approach requires two matrices. The first one is built by evaluating the features on the set of hand-annotated training data, the second one by applying the same feature set as well to the test data yet to be classified (*preparation* step). SuSi then uses the first matrix to train the classifier (*classification* step), which afterwards decides on the records in the test matrix (*output* step).

While there are a few methods in the Android library that are both sources and sinks (such as some of the *transceive* methods of the NFC implementation), their scarcity stops us from establishing a fourth category "*both*", even though in theory such a category might sound sensible. Classifying a sufficient amount of training data for a machine learning approach would be equal to classifying almost all transceiver methods. Respectively, we treat such methods as either sources or sinks. This decision affects both the training data and the classifier's results.

In a second step, SuSi *categorizes* the sources and sinks set. In this step, SuSi separately considers the sources and sinks determined in the first step as new test sets (dashed arrows). Note that methods classified as *neither* are ignored at this point. SuSi also requires new training data for the second classification problem. To provide such data, we hand-annotated a subset of the Android sources and sinks with semantic categories related to the mobile domain. We furthermore chose different kinds of features for the feature database as explained in Section IV-C. We chose 12 different kinds of source-categories that we identified as being sufficiently meaningful for the different Android API methods: *account*, *bluetooth*, *browser*, *calendar*, *contact*, *database*, *file*, *network*, *nfc*, *settings*, *sync*, and *unique-identifier*. For the sinks, we defined 15 different kinds of categories: *account*, *audio*, *browser*, *calendar*, *contact*, *file*, *log*, *network*, *nfc*, *phone-connection*, *phone-state*, *sms/mms*, *sync*, *system*, and *voip*. For the purpose of compiling our training data, if a method is not relevant or does not fit in any of the identified categories, it is annotated as belonging to the special *no-category* class. If one wants to add a new category, one simply has to create new features for the feature database and randomly annotate the corresponding API methods. Our approach then automatically uses the new feature for the generation of the categorized sources and/or sinks. The subsequent steps as shown in Figure 2 are equal to the ones for the classification. The final output consists of two files, one for the categorized sources and one for the categorized sinks.

Note that some of these categories refer to data being managed by applications, not the operating system itself. One example are contacts: The system provides a data interface to make sure that there is a uniform way of obtaining contacts for all applications that require them, e.g., travel planners, or calendars sending invitations. Additionally, Android contains system applications providing default implementations of these interfaces, so there are methods which are available on every Android phone and which can be called in order to obtain private data. Therefore, we include categories for such methods, despite them not being part of the operating system as such.

Since we have different categories for sources and sinks, their categorization comprises two distinct classification problems: one for sources and one for sinks. Though they share the same feature set (see Section IV-C), both are solved independently of each other. Thus, quite naturally, the resulting correlations might differ significantly, as some features might be more relevant to distinguish different kinds of sources than different kinds of sinks, and vice versa.

C. Feature Database

We used a set of 144 syntactic and semantic features for classifying the Android methods. A single feature alone does not usually give enough information to decide whether a given Android method is a source, a sink or neither. However, all features in combination can be used to train a highly precise classifier. The same holds for the second classification problem in which we need to find categories for our sources and sinks.

One main reason for why these features work is that many developers of the Android framework do in fact follow a certain regular coding style, or duplicate parts of one method's implementation when implementing another. These social aspects of software development lead to a certain degree of regularity and redundancy in the code base, which a machine-learning approach such as ours can discover and take advantage of.

Though we have a large number of distinct features, most of them are instances of the same parameterized class. For example, the "method name starts with" feature class has instances "method name starts with *get*", "method name starts with *put*", and so on. For identifying sources and sinks, SuSi uses the following classes of features:

- **Method Name:** The method name contains or starts with a specific string, e.g., "get", which can be an indicator for a source.
- **Method has Parameters:** The method has at least one parameter. Sinks usually have parameters, while sources might not.
- **Return Value Type:** The method's return value is of a specific type. A returned cursor, for instance, hints at a source, while a method with a void return value is rarely ever a source.
- **Parameter Type:** The method receives a parameter of a specific type. This can either be a concrete type or all types from a specific package. For instance, a parameter of type *java.io.** hints at a source or a sink.
- **Parameter is an Interface:** The method receives a parameter of an interface type. This is often the case with methods that register callbacks. Note that such methods are neither sources nor sinks according to our definition, since they do not perform any actual operation on the data itself.
- **Method Modifiers:** The method is static/native/etc. Static methods are usually neither sources nor sinks, with some exceptions. Additionally, sources and sinks are usually public.
- **Class Modifiers:** The method is declared in a protected/abstract etc. class. Methods in protected classes are usually neither sources nor sinks.
- **Class Name:** The method is declared in a class whose name contains a specific string, e.g., *Manager*.
- **Dataflow to Return:** The method invokes another method starting with a specific string (e.g. *read* in the case of a source). The result of this call flows into the original method's return value. This hints at a source.
- **Dataflow to Sink:** One of the method's parameter flows into a call to some other method starting with

a specific string, e.g., *update*, which would suggest a sink.

- **Data Flow to Abstract Sink:** One of the method's parameter flows into a call to an abstract method. This is a hint for sink as many command interfaces on the hardware abstraction layers are built on top of abstract classes.
- **Required Permission:** Invoking the method requires a specific permission. There is one such feature for every permission declared in the Android API. We were only able to use this feature on the approximately 12,600 methods for which we had permission annotations from the PScout [22] list.

Some features, in particular "Method Name", might sound naïve at first, but it turns out that such syntactic features are among the ones that correlate the strongest with sources and sinks. Of course, their effect is only positive in combination with other features; one could not, for instance, detect sources by *only* looking at prefixes of method names.

All our features can assume one of three values: "True" means that the feature applies, i.e., a method does indeed start with a specific string. "False" means that the feature does not apply, i.e., the method name does not have the respective prefix. "Not Supported" means that the feature cannot be decided for this specific method. The latter can happen if, for example, the feature needs to inspect the method body, but no implementation is available in the current Android version's platform JAR file.

The details of our dataflow features are explained in Section IV-D. SUSI's features for *categorizing* sources and sinks can be grouped as follows:

- **Class Name:** The method is declared in a class whose name contains a specific string, e.g., *Contacts*.
- **Method Invocation:** The method directly invokes another method whose fully-qualified name starts with a specific string, e.g., *com.android.internal.telephony* for Android's internal phone classes. This feature does not consider the transitive closure of calls starting at the current method.
- **Body Contents:** The method body contains a reference to an object of a specific type, e.g. *android.telephony .SmsManager* for the *SMS_MMS* category).
- **Parameter Type:** The method receives a parameter of a specific type (similar feature as for the classification problem with different instances).
- **Return Value Type:** The method's return value is of a specific type, e.g., *android.location.Country* for regional data.

Note that we do not use permission-based features for the categorization, since many methods require permissions for internal functionality not directly related to their respective category. For instance, a backup method requests many permissions, but does not necessarily give out all of the data it accesses using these permissions if it only creates an internal save point that can be restored later. The permission list alone thus does not directly relate to the method's category.

It becomes apparent that semantic features are much more suitable for identifying sources and sinks than for categorizing them. On the source-code level, Android's sources and sinks share common patterns which can be exploited by our dataflow feature. For finding categories, however, there seems to be no such technical distinction and SUSI must rather rely on syntactical features such as class and method names.

D. Dataflow Features

As we found through empirical evaluation, considering a method's signature and the syntax of its method body alone is insufficient to reliably detect sources and sinks. With such features alone we were unable to obtain a precision or recall higher than about 60%. It greatly helps to take the data flows inside the method into consideration as well. Recall from our definitions in Section III that sources must read from and sinks must write to resources.

To analyze data flows, we originally experimented with a highly precise (context-, flow- and object-sensitive) data-flow analysis based on Soot [29], but found out that this did not easily scale to the approximately 110,000 methods of the Android SDK. Computing precise call graphs and alias information simply took too long to be practical. We thus changed to a much more coarse-grained intra-procedural approximation (also based on Soot[1]) which runs much faster whilst remaining sufficiently precise for the requirements of our classification. Keep in mind that the result of the data-flow analysis is only used as one feature out of many. Thus, it suffices if the analysis is somewhat precise, i.e., produces correct results with just a high likelihood.

Our data-flow features are all based on taint tracking inside the Android API method m to be classified. Depending on the concrete feature, we support the following analysis modes:

- Treat all parameters of m as sources and calls to methods starting with a specific string as sinks. This can hint at m being a sink.
- Treat all parameters of m as sources and calls to abstract methods as sinks. This can hint at m being a sink.
- Treat calls to specific methods as sources (e.g. ones that start with ”read”, ”get”, etc.) and the return value of m as the only sink. This can hint at m being a source. Optionally, parameter objects can also be treated as sinks.

Based on this initialization, we then run a fixed-point iteration with the following rules:

- If the right-hand side of an assignment is tainted, the left-hand side is also tainted.
- If at least one parameter of a well-known *transformer* method is tainted, its result value is tainted as well.
- If at least one parameter of a well-known *writer* method is tainted, the object on which it is invoked is tainted as well.
- If a method is invoked on a tainted object, its return value is tainted as well.

[1] We take the *android.jar* built from the OS and the system applications on a real phone (Galaxy Nexus running Android 4.2) as input for Soot.

- If a tainted value is written into a field, the whole base object becomes tainted. For arrays, the whole array becomes tainted respectively.

When the first source-to-sink connection is found, the fixed-point iteration is aborted and the dataflow feature returns "True" for the respective method to which it was applied. If the dataflow analysis completes without finding any source-to-sink connections, the feature returns "False".

While such an analysis would be too imprecise for a general-purpose taint analysis, it is very fast and usually reaches its fixed point in less than three iterations over the method body. Since the analysis is intra-procedural, its runtime is roughly bounded by the number of statements in the respective method.

Instead of using fixed initialization rules as explained above, one can also first run the machine learning algorithm wihout the data flow feature enabled, and then initialize the data flow feature with the results of this preliminary round. This method can be applied incrementally until a fixed point is reached. We plan to investigate the tradeoffs involved with this method in the future.

E. *Implicit Annotations for Virtual Dispatch*

SuSi's implementation is based on Weka, a generic machine-learning tool, which has no knowledge about the language semantics of Java. However, we found that when annotating methods to obtain training data it would be beneficial to propagate method annotations up and down the class hierarchy in cases in which methods are inherited. Such a propagation models the semantics of virtual dispatch in Java. We thus extended SuSi such that if encountering an annotated method *A.foo*, the annotation is implicitly carried over also to *B.foo* in case *B* is a subclass of *A* that does not override *foo* itself, thus inheriting the definition in *A*. Similarly, if *B.foo* were annotated, but not *A.foo*, we would copy the annotation in the other direction.

For our subset of 12,600 methods with permission annotations taken from the PScout list [22], SuSi was able to automatically create implicit annotations for 305 methods. After loading the remaining methods of the Android API to get our full list of 110,000 methods, SuSi was able to automatically annotate another 14 methods.

V. Evaluation

Our evaluation considers the following research questions:

RQ1 Can SuSi be used to effectively find sources and sinks with high accuracy?

RQ2 Can SuSi be used to categorize the found sources and sinks with high accuracy?

RQ3 Which kind of sources and sinks are used in malware apps?

RQ4 How do the sources and sinks change during different Android versions? Can SuSi be used to identify sources and sinks in new, previously unseen Android versions?

RQ5 How complete are the lists of sources and sinks distributed with existing Android analysis tools and how do they relate to SuSi's outputs?

The following sections address these questions in order.

A. *RQ1: Sources and Sinks*

To assess the precision and recall of SuSi on our training data, we applied a ten fold cross validation and report the results in Section V-A1. Since the test data used for the cross validation is picked randomly, the results of the cross validation usually carry over to the complete classification performance on unknown training sets if the test set was sufficiently representative. To confirm that this actually holds, we manually evaluated the source and sink lists SuSi generated for the Google Mirror and Google Cast APIs and report the results in Section V-A2. The Google Cast API is used for the communication between an Android-based smartphone and Google's Chromecast device [20]. The Google Mirror API links an Android device to Google Glass [19]. We chose these two APIs to show that SuSi is actually able to efficiently handle even previously unseen Android or Java APIs. Note that neither API is included in the base Android system. Secondly, both APIs include methods that handle personal data, such as location or network information. To the best of our knowledge no taint analysis tool has considered these APIs yet. Thirdly, the APIs are of manageable size, making a complete manual validation of SuSi's results practical.

1) Cross Validation: We envision SuSi to be used as an automated approach in which experts like ourselves hand-annotate parts of the Android API and then use SuSi to automatically extrapolate these annotations to larger parts of the API. Of course, such an approach only makes sense if the extrapolation is meaningful, which is equivalent to delivering a high precision and recall. Measuring precision and recall is hard in this setting, as one has no gold standard to work with: there is no correctly pre-annotated Android API with which one could compare SuSi's results. Thus, as a best-effort solution we hand-annotated a subset of the Android API ourselves (details below) and then used these methods both as training and test data in a ten-fold cross validation [30] which is the standard approach for evaluating machine-learning techniques. It works by randomly dividing all training data into 10 equally-sized buckets, training the classifier on nine of them, and then classifying the remaining bucket. The process is repeated 10 times, omitting another bucket from training each time. In the end, SuSi reports the average precision and recall. For each class *c*, precision is the fraction of correctly classified elements in *c* within all elements that *were* assigned to *c*. If precision is low it means that *c* was assigned many incorrect elements. Recall is defined as fraction of correctly classified elements in *c* within all elements that *should have been* assigned to *c*. If recall is low it means that *c* misses many elements.

Table II shows the results of this ten-fold cross validation over our training set of 779 methods randomly picked from the PScout subset [22] of about 12,600 methods. The training set contains 13% *source*-, 22% *sink*- and 65% *neither*-annotations. We started with this subset as it provided mappings between methods and required permissions and thus enabled us to also use Android permissions as features for our classifier. The averages we report in our tables are taken from Weka's output. They are weighted with the number of examples in the respective class. Also note that, since our training set is randomly picked, the precision and recall should carry over to the entire Android API with high probability.

Our final results for the source/sink classification had to be

Category	Recall [%]	Precision [%]
Sources	92.3	89.7
Sinks	82.2	87.2
Neither	94.8	93.7
Weighted Average	91.9	91.9

TABLE II. SOURCE/SINK CROSS VALIDATION PSCOUT

computed without any permission features, though, since we do not have permission associations for the complete Android API[2]. For assessing the impact of the permission feature, we ran the PScout subset again with the permission feature disabled, yielding the results shown in Table III. Interestingly, the average precision and recall are almost the same with the permission feature and without. The impact of the permission feature is apparently low enough for not having to worry about the lack of permission information when analyzing the complete Android 4.2 API. Conversely, the results also indicate that permissions alone are not a good indicator for identifying sources or sinks.

Category	Recall [%]	Precision [%]
Sources	90.5	91.3
Sinks	86.0	88.8
Neither	95.2	94.4
Weighted Average	92.8	92.8

TABLE III. SOURCE/SINK CROSS VALIDATION PSCOUT WITHOUT PERMISSION FEATURE

We evaluated SUSI on an extended test set obtained using the implicit-annotation technique explained in section Section IV-E. With this technique, classifications for a method are copied to all other methods that would lead to the same code being executed according to the semantics of virtual method dispatch in Java. SUSI again shows an average recall and precision of more than 92%, see Table IV. The results are not exactly equal because some of our features consider not just a method's definition but also its container, e.g., the name of the class the method resides in. The fact that SUSI obtains similar results despite these differences is a good indicator of inherent consistency in the results as it shows that semantically equal methods (i.e., ones that have not been overwritten and are thus exposed as-is) are also recognized equally.

Category	Recall [%]	Precision [%]
Sources	89.6	88.0
Sinks	84.7	90.8
Neither	95.2	93.6
Weighted Average	92.3	92.3

TABLE IV. SOURCE/SINK CROSS VALIDATION WITH IMPLICIT ANNOTATIONS

The classifier takes about 26 minutes to classify the complete Android 4.2 API on a MacBook Pro computer running MacOS X version 10.7.4 on a 2.5 GHz Intel Core i5 processor and 8 GB of memory.

As explained in Section IV-A, we experimented with various classification algorithms, and found that SMO performed best. In Table V, we compare the weighted average precision for SMO, J48, and Naive Bayes, the most well-known representatives of their respective families of classifiers (margin, rule-based and stochastic classifier, respectively). The results were computed on the extended training set obtained through the implicit-annotation technique. The permission feature was not used.

Category	Recall [%]	Precision [%]
ACCOUNT	100.0	100.0
BLUETOOTH	83.3	100.0
BROWSER	83.0	100.0
CALENDAR	100.0	100.0
CONTACT	95.0	100.0
DATABASE	50.0	100.0
FILE	75.0	100.0
NETWORK	83.3	83.3
NFC	100.0	100.0
SETTINGS	75.0	85.7
SYNC	100.0	100.0
UNIQUE_IDENTIFIER	88.9	100.0
NO_CATEGORY	95.7	62.9
Weighted Average	88.7	89.6

TABLE VI. SOURCE CATEGORY CROSS VALIDATION

2) Validating SUSI*'s Source/Sink Output:* The output of SUSI's first phase is a list of sources and a separate list of sinks. In this section we verify that the precision and recall of the cross validation in Section V-A1 is representative for SUSI's actual output. Since manually verifying the outputs for the complete Android API is infeasible, we concentrate on two APIs: The Google Cast API and the Google Mirror API.

Our manual validation of the Google Cast API results in a precision of 96% and a recall of 99% for the sources and a precision of 100% and recall of 88% for the sinks. The somewhat lower recall for the sinks is due the fact this API has only 18 sinks, out of which 16 were detected. The Google Mirror API yields a precision of 100% and a recall of 97% for the sources and a precision of 100% and recall of 94% for the sinks. In result it seems that one can be rather optimistic: at least for these APIs the precision and recall are even higher than the ones obtained through cross validation (cf. Section V-A1).

B. RQ2: Categories for Sources and Sinks

For evaluating the categorization of sources and sinks, we used similar techniques like the ones used for assessing the identification of sources and sinks in Section V-A. However, recall that only methods identified as sources or sinks in the first step get categorized by SUSI.

1) Cross Validation: We use ten-fold cross validation on our training data to assess the quality of our categorization. For this task, we do not use the permission feature, but do apply the implicit annotation technique from Section IV-E. Table VI shows the cross-validation results for categorizing the sources, while Table VII contains those for the sinks.

While SUSI achieves a very high precision and recall for most of the categories, the results for a few categories (e.g. Bluetooth) are considerably worse. These categories are rather small, i.e., randomly picking training methods from the overall

[2] The available permission lists including PScout are incomplete since they exclude permissions enforced through calls to native code.

Classifier	**Avg. Recall**			**Avg. Precision**		
	Class. [%]	Source Cat. [%]	Sink Cat. [%]	Class. [%]	Source Cat. [%]	Sink Cat. [%]
Margin (SMO)	92.3	88.8	88.4	92.3	89.7	90.4
Rule-Based (J48)	89.5	81.0	80.2	89.4	81.6	77.4
Probabilistic (Naive Bayes)	86.9	61.5	46.6	87.1	61.7	36.1

TABLE V. SOURCE/SINK CLASSIFIER COMPARISON

Category	**Recall** [%]	**Precision** [%]
ACCOUNT	85.7	100.0
AUDIO	100.0	100.0
BROWSER	50.0	100.0
CALENDAR	100.0	100.0
CONTACT	91.7	100.0
FILE	60.0	100.0
LOG	100.0	71.4
NETWORK	72.7	88.9
NFC	100.0	100.0
PHONE_CONNECTION	75.0	85.7
PHONE_STATE	100.0	100.0
SMS_MMS	96.3	100.0
SYNC	80.0	100.0
SYSTEM	80.6	89.3
VOIP	66.7	100.0
NO_CATEGORY	97.1	70.2
Weighted Average	85.7	88.0

TABLE VII. SINK CATEGORY CROSS VALIDATION

set of 110,000 Android 4.2 API methods yields only few entries belonging to such categories. Respectively, there is not much material to train the classifier on. Annotating more data (recall that we only have category annotations for 0.4% of all methods) would certainly improve the situation.

Categories can be ambiguous in some cases. A method to set the MSIDN (the phone number to be sent out when placing a call) could for instance be seen as a system setting (category SETTINGS), but could also be considered a UNIQUE_ID. In such cases, we checked the classifier's result and updated our training data if a misclassification was to due semantic ambiguity, i.e., the result would be right in both categories. Categories that ended up empty or almost empty due to such shifts were removed.

Categorizing the sources took about 6 minutes on our test computer. The sinks were classified in about 3 minutes.

2) Validating SUSI*'s Categorized Source/Sink Output:* Manually evaluating the categorized sources and sinks for the Google Cast and Google Mirror APIs shows a precision and recall of almost 100% . The precision and recall for the Google Cast API are 100% for both sources and sinks. For sources in the Google Mirror API the precision is 98% and the recall is 100%. For sinks, both precision and recall are 100%. This shows that the results from Section V-A2 also carry over to the categorization.

C. RQ3: Sources and Sinks in Malware Apps

It is an important question to ask whether existing malware apps already use sources and/or sinks discovered by SUSI but not currently recognized by state-of-the-art program-analysis tools. To address this question, we selected about 11,000 malware apps from Virus Share [21] and analyzed which kinds of sources and sinks these malware samples use. Unsurprisingly, as already found by different researchers [4], [31], [32] current malware is leaking privacy information such as location information or the address book.

Interestingly, however, these samples do not only use the standard source and sink methods commonly known to literature, but also such ones not detected by popular program analysis tools (see Section V-E). In total, the samples revealed usage of more than 900 distinct source methods, all of which can be used to obtain privacy-sensitive information. Furthermore, the samples leak data through more than 500 distinct sink methods. The `getLac()` and `getCid()` methods used in our motivating example (see Section II) are two of the most commonly used methods in the LOCATION_INFORMATION category. This is partly related to the fact that both are called in the Google Maps Geolocation API [33], which is used in the respective malware samples. Another example is the `getMacAddress()` method in the `WifiInfo` class that SUSI categorizes as NETWORK_INFORMATION. This method is among the most often called methods in this category and is not treated as a source by many tools either. By manual analysis of different malware samples, we found that these source methods are not just called, but their privacy-sensitive return values are indeed leaked to a remote web server.

Since approaches such as LeakMiner [3] create their source and sink lists from a permission map, we also analyzed whether malware samples exploit source methods that do not need a permission. Examples of such methods are `getSimOperatorName()` in the `TelephonyManager` class (returns the service provider name), `getCountry()` in the `Locale` class, and `getSimCountryIso` in the `TelephonyManager` class (both return the country code), all of which are correctly classified by SUSI. By manually analyzing the malware samples, we found that these methods are used frequently and that this data is actually leaked to web servers. This confirms that approaches which solely rely on the permission map for inferring sources and sinks miss data leaks in real-world malware samples.

SUSI's categorized output of sources and sinks for Android 4.2 (see Section V-A1) includes a lot of methods which return privacy-sensitive information, such as the IMEI. SUSI found that there is not only one way of accessing such information (e.g. via `getDeviceID` for the IMEI). Instead, there are plenty of wrapper methods in internal Android classes or pre-installed apps that return the same value. One example would be the internal `GSMPhone` class or the pre-installed email-application which contains a `getDeviceId()` method for returning the IMEI. These methods can only be called using explicit class

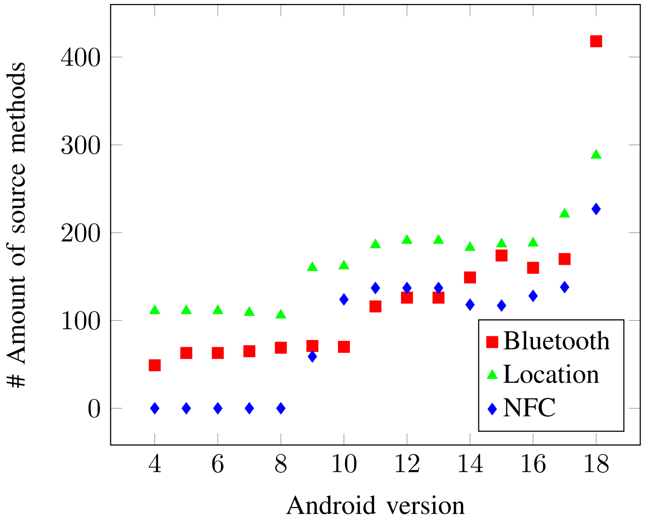

Fig. 3. Amount of source methods for bluetooth, location and NFC information in different Android versions

loading and reflection, but still work on an unmodified stock Android phone. We analyzed the malware samples for this obfuscation technique but found no sample that actually tries to obtain personal data through such methods. Furthermore, we did also not find methods for sinks which are not so well known as shown in the motivating example (cf. Section II). However, we expect such advanced techniques to become more prevalent when security tools evolve, for instance by incorporating the results of this paper, and thus more effectively detect the easier cases.

D. RQ4: Changes during Android Versions

To assess how well SUSI can deal with previously unseen versions of the Android operating system, we compared the categorized source and sink lists generated for a selection of different Android releases. Figure 3 shows the number of sources found for API versions 4 (Android 1.6) to 18 (Android 4.3). We here focus on the *bluetooth*, *location*, and *NFC* categories, as they nicely demonstrate how Android was extended over the various versions. From the figure one can clearly deduce that new sources are introduced with every version. This is yet another motivation to use a tool-supported approach like SUSI's to discover sources and sinks.

The distribution of the number of source methods for location information shows three different jumps, namely between versions 8 and 9, between 16 and 17, and between 17 and 18. This is due to major changes in the Android location APIs [34]–[36]. The same holds for the jumps in the number of bluetooth information sources between versions 17 and 18, where new source-bearing classes where added to the `android.bluetooth` API. One also clearly sees that NFC was added to Android in API version 9 [34]. There are some cases in which the number of sources is decreased from one version to the next, e.g. between versions 4, 5, 6, 7 and 8 for location. This is related to minor changes in the API. The cross-validation results on the different Android versions were effectively the same as reported for version 4.2 in Sections V-A2 and V-B.

Our results show that SUSI detects the changes in different API versions very well. It reliably finds new sources and sinks that were added to the Android platform and thus provides a much higher level of coverage than available lists assembled by hand. Note that for completely new, previously unanticipated APIs that should yield a new category, SUSI obviously cannot anticipate this category either. In such cases one can easily open a new category, though, by annotating by hand a few examples that fall into this category. This is exactly how we formed categories in SUSI's training set.

E. RQ5: Existing lists of sources & sinks

In this section we assess to what extent current static [2]–[7], [9], [12], [13] and dynamic [14], [15] code analysis approaches could benefit from our categorized sources/sinks list. As our results show, SUSI finds all the sources and sinks these previous approaches mention, plus many others which the community was previously unaware of, including some of which are actually being used by malware. Most of the code-analysis tools were not publicly available, precluding one from directly comparing their source and sink lists to SUSI's [2], [3], [6], [7], [9], [15]. For those approaches we thus estimated the lists from their research papers.

Mann et al. [9] mention a few concrete source and sink methods. This hand-picked list is only a fraction of the one produced by SUSI. The taint-tracking tool CHEX [2] uses a list of 180 semi-automatically collected sources and sinks. Unfortunately, this list is not publicly available and the paper does not explain how the semi-automatic approach works. The authors do mention that their list is based on the Android permission map by Porter Felt et al. [37] but also argue that this list is insufficient. LeakMiner [3] uses the Android permission map to identify sources and sinks. From this map it filters out all methods an application is not allowed to use. However, this leaves open how the tool actually identifies the *relevant* sources and sinks in the remaining method set. Furthermore, if all methods not requiring a permission are filtered, some sensitive data might be overlooked as we have shown. ScanDal [6] and AndroidLeaks [7] do not provide concrete lists of source and sink methods. The publications only provide categories (e.g., location information, phone identifier, internet, etc.), which are also covered by our automatic categorization. Aurasium [15] shifts the problem of identifying sources and sinks by intercepting calls at the system level, i.e., between the native Android libraries and the standard Linux system libraries. While this reduces the number of methods to consider, it makes it harder to reconstruct higher-level semantics, and is failure-prone in case of Android version upgrades. Due to this design, the sources and sinks considered by Aurasium are incomparable to SUSI's results.

Three different taint-analysis approaches were publicly available to us: The dynamic taint analysis tool TaintDroid [14], an approach based on DeD by Enck et al. [4], and SCanDroid [38]. TaintDroid does not specify the high-level API calls as sources or sinks. Instead, it uses the smaller set of lower-level internal system methods called by those, an approach somewhat comparable to Aurasium. However, this again raises the problem of reconstructing the higher-level context from lower-level calls. The type of data leaked can thus be imprecise. The DeD approach works by decompiling the Android bytecode into Java bytecode which is then used as input for the commercial Fortify SCA [12] static code-analysis suite. Fortify can be configured with rules for defining sources and sinks. Enck et al. created such rules and made them publicly available [39]. The list contains about 100 Android sources and 35 Android sinks, all of which are also included in SuSi's source and sink lists. For SCanDroid, we extracted the source and sink specifications from the source code (version of April 2013). The resulting list appears hand-picked and is fully covered by SuSi's output.

For evaluating the completeness of the source and sink lists contained in these three tools, we analyzed the most frequently referenced source and sink methods in the malware samples from Section V-C. Table VIII shows that the three tools treat only a few of the methods as a sources or sinks respectively. To assess TaintDroid, we created a separate app for every source and sink in the table. For a source, the respective data is obtained and then leaked via the network (note that the network connection is treated as a sink by TaintDroid). For the sinks we used the well-known `getLongitude()` method as a source (which is treated as a source by TaintDroid) and also created one app per sink. We ran all of our apps on a phone with Cyanogenmod 10 [40] containing TaintDroid for Android 4.1. The results of our evaluation are shown in Table VIII.

Table VIII shows that the source and sink lists of the three tools are missing some important methods such as one returning the Wifi MAC-address which enables a phone to be uniquely identified. All three tools also miss the method for obtaining the list of accounts (mail, Exchange, social networks, etc.) registered in the phone.

We also found that TaintDroid over-approximates the list of sources and sinks, leading to over-tracking, for instance by tainting the result value of all methods in the *TelephonyManager* class, including the result of *toString()*, which is just the Java object ID (default implementation inherited from *java.lang.Object*). We thus argue that automatically inferring higher-level API methods as provided by our approach would improve tools like TaintDroid as this would allow one to more easily categorize and differentiate various types of sources and sinks.

In total, the results of our evaluation show that obtaining a complete list of sensitive sources and sinks is difficult and SuSi's automatically generated list of categorized sources and sinks can be used to improve this situation.

We also examined well-known commercial tools for static code analysis such as Fortify SCA [12] by HP and IBM AppScan Source [13]. As we found, by default these tools provide lists that are rather incomplete. However, both provide an easy way to integrate new sources and sinks to be considered by the analysis. This shows that these tools shift the problem of defining sources and sinks to the analyst, who still needs to obtain such a list from somewhere. SuSi can help to provide more comprehensive defaults.

VI. Sources Not Considered by SuSi

```
NmeaListener mylistener = new
    NmeaListener() {
  public void onNmeaReceived(long arg0,
      String nmea) {
    if (nmea.startsWith("$GPGLL")) {
      String[] data = nmea.split(",");
      Log.d("Loc", "Longitude: "
        + data[3]} + data[4]
        + ", Latitude: " + data[1] +
            data[2]);
    }
  }
};
LocationManager lm = (LocationManager)
    this.getSystemService(LOCATION_SERVICE);
lm.addNmeaListener(mylistener);
// Just to start GPS, no data from this
    callback is ever used
lm.requestLocationUpdates
    (LocationManager.GPS_PROVIDER, 0, 0,
    new LocationListener() { ... });
```

Listing 2. Android Location via NMEA Data

SuSi works well when it comes to classifying sources and sinks based on their structural similarity to other sources, respectively sinks. In practice, this seems to work well for sources that return data from method calls and sinks that obtain data through parameters. Android offers other less prevalent sources and sinks, however, which cannot be easily classified through machine learning which we will show in this section.

Applications can implement callback methods and receive data from the operating system through the parameters of these methods. This is commonly used to, e.g., obtain the location in an Android application. In an attempt to avoid detection, the app could however register the callback with *onNmeaReceived* instead of the well-known *onLocationChanged* method and then parse the raw GPS data (the NMEA records) as shown in Listing 2 to get the same data. This shows, that a complete list of callback methods is required for finding all data leaks. SuSi cannot currently find such callbacks due to our definition of sources. The number of callback interfaces in the Android operating system is however sufficiently small for manual inspection. All callback handlers are defined using a small set of well-known and documented interfaces. Static analysis thus aid their detection by finding methods taking these interfaces as parameters. This approach scales well and does not introduce an unreasonable number of false positives as shown in [41].

Android defines layout controls through XML files. In the source code, they can be accessed by passing the respective identifier to the system's *findViewById* function. Depending on the ID that is passed, this function can return, for instance, a reference to a password field or to a button with a constant label. Thus, depending on the ID, the method can or can not be a source. Since calls to this function are present in almost every Android app, a precise analysis must model the Android resource system. If UI sources are restricted to password fields (the default in FlowDroid [41]), the analysis scales well in terms of precision. Regarding every input field as a source, on the other hand, can lead to a substantial number of false positives. A more fine-grained tradeoff might be possible by exploiting knowledge about the app's expected behavior.

VII. Related Work

Our work was originally inspired by Merlin [42], a probabilistic appraoch that uses a potentially incomplete specification

Method	Description	TaintDroid	SCanDroid	DeD
android.bluetooth.BluetoothAdapter.getAddress()	Returns the hardware address of the local Bluetooth adapter.	no	no	no
android.net.wifi.WifiInfo.getMacAddress()	Returns the MAC address of the Wifi interface.	no	no	no
java.util.Locale.getCountry()	Returns the country code for the phone's locale.	no	no	no
android.net.wifi.WifiInfo.getSSID()	Returns the SSID of the current 802.11 network.	no	no	no
android.telephony.gsm.GsmCellLocation.getCid()	Returns the GSM cell id.	no	no	no
android.telephony.gsm.GsmCellLocation.getLac()	Returns the GSM location area code.	no	no	no
android.location.Location.getLongitude()	Returns the longitude in degrees.	yes	yes	yes
android.location.Location.getLatitude()	Returns the latitude in degrees.	yes	yes	yes
android.accounts.AccountManager.getAccounts()	Returns all accounts of any type registered on the device as a list.	no	no	no
java.util.Calendar.getTimeZone()	Returns the time zone.	no	no	no
android.telephony.TelephonyManager.getDeviceId()	Returns the unique device ID.	yes	no	yes
android.telephony.TelephonyManager.getSubscriberId()	Returns the unique subscriber ID.	yes	no	yes
android.telephony.TelephonyManager.getLine1Number()	Returns the phone number of the device.	yes	no	yes
android.telephony.TelephonyManager.getSimSerialNumber()	Returns the serial number of the SIM.	yes	no	yes
android.provider.Browser.getAllBookmarks()	Returns a cursor pointing to a list of all the bookmarks.	yes	no	no
android.telephony.SmsManager.sendTextMessage()	Send a text based SMS.	yes	yes	yes
android.util.Log.d()	Sends a debug log message.	no	no	yes
java.net.URL.openConnection()	Returns a URLConnection instance that represents a connection to the remote object referred to by the URL.	yes	no	no

TABLE VIII. DETECTION OF MOST FREQUENTLY USED SOURCES AND SINKS IN MALWARE SAMPLES [21] IN DIFFERENT ANALYSIS TOOLS

of sources, sinks and sanitizers to produce a more complete one. Livshits et al.'s approach is based on a *propagation graph*, a representation of the interprocedural data flow in the program where probabilistic inference rules are applied. Their specifications are based on *string-related* vulnerabilities, such as cross-side-scripting vulnerabilities or sql-injections. SUSI in comparison to Merlin does not need any information about the client program or application. It instead analyzes the Android framework code alone to generate a list of categorized sources and sinks. Furthermore, purely string-based approaches fit a web application scenario, while SuSi focuses on privacy-related aspects of Android where data is usually not of type *string* (e.g., the longitude and latitude information is of type *double*).

Privacy violations through leaks of sensitive data in Android applications are well known in the community. To protect the user's privacy, different kinds of taint-tracking approaches have been proposed, both static [1]–[13] and dynamic [14], [15], [17]. As already described in Section I, such approaches are only as good as the source and sink lists they are configured with. In Section V-E we have shown that all approaches we have evaluated only consider a few sensitive methods for sources and sinks. With the support of our categorized list of sources and sinks, we argue that all of them could be improved to detect more data leaks that are a security problem for the mobile device user.

More generic policy enforcement approaches such as AppGuard [16] also require comprehensive lists of sensitive information sources. AppGuard, for instance, provides the user with the ability to revoke permissions after app-installation time. The implementation inserts additional permission checks into the application (not the framework). This requires the identification of relevant methods at the API level for which such checks are required. Our list of sources and sinks includes many methods that require permissions and access sensitive information (e.g., phone identifier, location information, etc.) but are not considered by AppGuard (evaluated version 1.0.3).

Applying machine learning for security has already been done for automatic spam detection [43] or anomaly detection in network traffic [44]). Sarma et al. [45] and Peng et al. [46] successfully used various machine-learning approaches to detect malicious Android applications. MAST [47] is a machine-learning approach based on Multiple Correspondence Analysis (MCA) for automatically identifying malicious applications from various Android markets. The tool aims at ranking apps for inspection by a human security analyst, thereby giving priority to those applications that look suspicious. For classifying sources and sinks, we use SMO instead of MCA since MCA requires a logical ordering of records which is not applicable to our scenario. SUSI instead works on discrete and independent classes.

VIII. CONCLUSIONS

In this paper, we have shown that privacy-enhancing technologies for Android are threatened by the fact that they come with largely incomplete lists of sources and sinks of private information, thereby allowing attackers to circumvent their measures with ease. We have presented SuSi, a novel automated machine-learning guided approach for identifying sources and sinks in the Android framework and pre-installed apps. The approach is capable of automatically categorizing findings according to the type of data being processed, for instance to distinguish between sources providing unique identifiers and sources providing file data.

A ten-fold cross validation showed our approach to have an average precision and recall of more than 92%. On Android 4.2, SuSi finds hundreds of sources and sinks. A manual comparison with existing hand-written (categorized) lists shows that, while SuSi finds all sources and sinks of the existing lists it also finds many more that were previously unknown, thus greatly reducing the risk for analysis tools to miss privacy violations. We showed that these previously missed sources and sinks are already used in existing malware samples which are thus not detected by state-of-the-art analysis tools. Furthermore, we showed that current approaches based on permission checks alone are inadequate as permission checks are, contrary to popular belief, not a good indicator for a method's relevance. Additionally, we have shown that new versions of the Android operating system come with new sources and sinks. While static hand-crafted lists usually do not contain such methods, as the manual effort for keeping the lists current is impractically high, SuSi can automatically infer them whenever a new Android version is released.

As future work, we aim to apply our approach to interfaces for automatically finding and classifying sensitive callbacks. We also want to further investigate how our approach can be applied to other environments than Android, e.g., J2EE. We are confident that the same concepts can also be applied to identify sources and sinks in other procedural programming languages such as C#, C++ or PHP.

Acknowledgements: We would like to thank Christian Wirth of the Knowledge Engineering Group at TU Darmstadt for his support in the field of machine learning and for his support with WEKA. Thanks to Damien Octeau for helping us to identify sources and sinks used by TaintDroid. This work was supported by a Google Faculty Research Award, by the BMBF within EC SPRIDE, by the Hessian LOEWE excellence initiative within CASED, and by the DFG within the project RUNSECURE, which is associated with the DFG priority program 1496 "Reliably Secure Software Systems - RS^3".

References

[1] J. Hoffmann, M. Ussath, M. Spreitzenbarth, and T. Holz, "Slicing Droids: Program Slicing for Smali Code," in *Proceedings of the 28th Symposium On Applied Computing*, ACM, Ed., 2013, pp. 0–0.

[2] L. Lu, Z. Li, Z. Wu, W. Lee, and G. Jiang, "Chex: statically vetting android apps for component hijacking vulnerabilities," in *Proceedings of the 2012 ACM conference on Computer and communications security*, ser. CCS '12. New York, NY, USA: ACM, 2012, pp. 229–240. [Online]. Available: http://doi.acm.org/10.1145/2382196.2382223

[3] Z. Yang and M. Yang, "Leakminer: Detect information leakage on android with static taint analysis," in *Third World Congress on Software Engineering (WCSE 2012)*, 2012, pp. 101–104.

[4] W. Enck, D. Octeau, P. McDaniel, and S. Chaudhuri, "A study of android application security," in *Proceedings of the 20th USENIX conference on Security*, ser. SEC'11. Berkeley, CA, USA: USENIX Association, 2011, pp. 21–21. [Online]. Available: http://dl.acm.org/citation.cfm?id=2028067.2028088

[5] A. P. Fuchs, A. Chaudhuri, and J. S. Foster, "Scandroid: Automated security certification of android applications," *Manuscript, Univ. of Maryland, http://www. cs. umd. edu/˜ avik/projects/scandroidascaa*, 2009.

[6] J. Kim, Y. Yoon, K. Yi, and J. Shin, "ScanDal: Static analyzer for detecting privacy leaks in android applications," in *MoST 2012: Mobile Security Technologies 2012*, H. Chen, L. Koved, and D. S. Wallach, Eds. Los Alamitos, CA, USA: IEEE, May 2012. [Online]. Available: http://ropas.snu.ac.kr/scandal/

[7] C. Gibler, J. Crussell, J. Erickson, and H. Chen, "Androidleaks: automatically detecting potential privacy leaks in android applications on a large scale," in *Proceedings of the 5th international conference on Trust and Trustworthy Computing*, ser. TRUST'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 291–307. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-30921-2_17

[8] L. Batyuk, M. Herpich, S. A. Camtepe, K. Raddatz, A.-D. Schmidt, and S. Albayrak, "Using static analysis for automatic assessment and mitigation of unwanted and malicious activities within android applications," in *Proceedings of the 2011 6th International Conference on Malicious and Unwanted Software*, ser. MALWARE '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 66–72. [Online]. Available: http://dx.doi.org/10.1109/MALWARE.2011.6112328

[9] C. Mann and A. Starostin, "A framework for static detection of privacy leaks in android applications," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, ser. SAC '12. New York, NY, USA: ACM, 2012, pp. 1457–1462. [Online]. Available: http://doi.acm.org/10.1145/2231936.2232009

[10] Z. Zhao and F. C. C. Osorio, ""trustdroid;": Preventing the use of smartphones for information leaking in corporate networks through the used of static analysis taint tracking," in *MALWARE*, 2012, pp. 135–143.

[11] E. Chin, A. P. Felt, K. Greenwood, and D. Wagner, "Analyzing inter-application communication in android," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, ser. MobiSys '11. New York, NY, USA: ACM, 2011, pp. 239–252. [Online]. Available: http://doi.acm.org/10.1145/1999995.2000018

[12] "Fortify 360 source code analyzer (sca)," Apr. 2013, http://www8.hp.com/us/en/software-solutions/software.html?compURI=1214365#.UW6CVKuAtfQ.

[13] "Ibm rational appscan," Apr. 2013, http://www-01.ibm.com/software/de/rational/appscan/.

[14] W. Enck, P. Gilbert, B. gon Chun, L. P. Cox, J. Jung, P. McDaniel, and A. Sheth, "Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones," in *OSDI*, 2010, pp. 393–407.

[15] R. Xu, H. Saïdi, and R. Anderson, "Aurasium: practical policy enforcement for android applications," in *Proceedings of the 21st USENIX conference on Security symposium*, ser. Security'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 27–27. [Online]. Available: http://dl.acm.org/citation.cfm?id=2362793.2362820

[16] M. Backes, S. Gerling, C. Hammer, M. Maffei, and P. von Styp-Rekowsky, "Appguard: enforcing user requirements on android apps," in *Proceedings of the 19th international conference on Tools and Algorithms for the Construction and Analysis of Systems*, ser. TACAS'13. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 543–548. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-36742-7_39

[17] J. Jeon, K. K. Micinski, J. A. Vaughan, A. Fogel, N. Reddy, J. S. Foster, and T. Millstein, "Dr. android and mr. hide: fine-grained permissions in android applications," in *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*, ser. SPSM '12. New York, NY, USA: ACM, 2012, pp. 3–14. [Online]. Available: http://doi.acm.org/10.1145/2381934.2381938

[18] "Google mirror api," aug 2013, https://code.google.com/p/google-api-java-client/wiki/APIs#Google_Mirror_API.

[19] "Google glass," aug 2013, https://developers.google.com/glass/.

[20] "Google cast," aug 2013, https://developers.google.com/cast.

[21] "Virus share," aug 2013, http://virusshare.com/.

[22] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie, "Pscout: analyzing the android permission specification," in *Proceedings of the 2012 ACM conference on Computer and communications security*, ser. CCS '12.

New York, NY, USA: ACM, 2012, pp. 217–228. [Online]. Available: http://doi.acm.org/10.1145/2382196.2382222

[23] A. Bartel, J. Klein, Y. Le Traon, and M. Monperrus, "Automatically securing permission-based software by reducing the attack surface: an application to android," in *Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE 2012. New York, NY, USA: ACM, 2012, pp. 274–277. [Online]. Available: http://doi.acm.org/10.1145/2351676.2351722

[24] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," in *Proceedings of the 18th ACM conference on Computer and communications security*, ser. CCS '11. New York, NY, USA: ACM, 2011, pp. 627–638. [Online]. Available: http://doi.acm.org/10.1145/2046707.2046779

[25] J. C. Platt, "Advances in kernel methods," B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, ch. Fast training of support vector machines using sequential minimal optimization, pp. 185–208. [Online]. Available: http://dl.acm.org/citation.cfm?id=299094.299105

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] J. R. Quinlan, *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[28] H. Zhang, "The Optimality of Naive Bayes." in *FLAIRS Conference*, V. Barr and Z. Markov, Eds. AAAI Press, 2004. [Online]. Available: http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf

[29] P. Lam, E. Bodden, O. Lhoták, and L. Hendren, "The soot framework for java program analysis: a retrospective," in *Cetus Users and Compiler Infastructure Workshop (CETUS 2011)*, 2011.

[30] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection." Morgan Kaufmann, 1995, pp. 1137–1143.

[31] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, ser. SP '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 95–109. [Online]. Available: http://dx.doi.org/10.1109/SP.2012.16

[32] M. C. Grace, W. Zhou, X. Jiang, and A.-R. Sadeghi, "Unsafe exposure analysis of mobile in-app advertisements," in *Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks*, ser. WISEC '12. New York, NY, USA: ACM, 2012, pp. 101–112. [Online]. Available: http://doi.acm.org/10.1145/2185448.2185464

[33] "The google maps geolocation api," aug 2013, https://developers.google.com/maps/documentation/business/geolocation/.

[34] "Android api differences report," aug 2013, https://developer.android.com/sdk/api_diff/9/changes.html.

[35] "Android api differences report," aug 2013, https://developer.android.com/sdk/api_diff/17/changes.html.

[36] "Android api differences report," aug 2013, https://developer.android.com/sdk/api_diff/18/changes.html.

[37] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner, "Android permissions demystified," in *Proceedings of the 18th ACM conference on Computer and communications security*, ser. CCS '11. New York, NY, USA: ACM, 2011, pp. 627–638. [Online]. Available: http://doi.acm.org/10.1145/2046707.2046779

[38] "Scandroid," apr 2013, https://github.com/SCanDroid.

[39] "A study of android application security - fortify rules," Apr. 2013, http://www.enck.org/tools/fsca_rules-final.xml.

[40] "cyanogenmod," Apr. 2013, http://www.cyanogenmod.org/.

[41] C. Fritz, S. Arzt, S. Rasthofer, and E. Bodden, "Flowdroid: Precise context-, flow-, object-sensitive and lifecycle-aware taint analysis for android apps," in *Submitted to ACM CCS 2013*.

[42] B. Livshits, A. V. Nori, S. K. Rajamani, and A. Banerjee, "Merlin: Specification inference for explicit information flow problems," *SIGPLAN Not.*, vol. 44, no. 6, pp. 75–86, Jun. 2009. [Online]. Available: http://doi.acm.org/10.1145/1543135.1542485

[43] K.-M. Schneider, "A comparison of event models for naive bayes anti-spam e-mail filtering," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, ser. EACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 307–314. [Online]. Available: http://dx.doi.org/10.3115/1067807.1067848

[44] A. A. Sebyala, T. Olukemi, L. Sacks, and D. L. Sacks, "Active platform security through intrusion detection using naive bayesian network for anomaly detection," in *In: Proceedings of London communications symposium*, 2002.

[45] B. P. Sarma, N. Li, C. Gates, R. Potharaju, C. Nita-Rotaru, and I. Molloy, "Android permissions: a perspective combining risks and benefits," in *Proceedings of the 17th ACM symposium on Access Control Models and Technologies*, ser. SACMAT '12. New York, NY, USA: ACM, 2012, pp. 13–22. [Online]. Available: http://doi.acm.org/10.1145/2295136.2295141

[46] H. Peng, C. Gates, B. Sarma, N. Li, Y. Qi, R. Potharaju, C. Nita-Rotaru, and I. Molloy, "Using probabilistic generative models for ranking risks of android apps," in *Proceedings of the 2012 ACM conference on Computer and communications security*, ser. CCS '12. New York, NY, USA: ACM, 2012, pp. 241–252. [Online]. Available: http://doi.acm.org/10.1145/2382196.2382224

[47] S. Chakradeo, B. Reaves, P. Traynor, and W. Enck, "Mast: triage for market-scale mobile malware analysis," in *Proceedings of the sixth ACM conference on Security and privacy in wireless and mobile networks*, ser. WiSec '13. New York, NY, USA: ACM, 2013, pp. 13–24. [Online]. Available: http://doi.acm.org/10.1145/2462096.2462100