Formal Characterization of Illegal Control Flow in Android System
Mariem Graa, Nora Cuppens-Boulahia, Frédéric Cuppens, Ana Cavalli

To cite this version:

HAL Id: hal-00924480
https://hal.archives-ouvertes.fr/hal-00924480
Submitted on 6 Jan 2014
Formal Characterization of Illegal Control Flow in Android System

Mariem Graa†, Nora Cuppens-Boulahia*, Frédéric Cuppens*, and Ana Cavalli†
*Telecom-Bretagne, 2 Rue de la Châtaigneraie, 35576 Cesson Sévigné - France
Emails: {mariem.benabdallah, nora.cuppens, frederic.cuppens}@telecom-bretagne.eu
†Telecom-SudParis, 9 Rue Charles Fourier, 91000 Évry - France
Emails: {mariem.graa, ana.cavalli}@it-sudparis.eu

Abstract—The dynamic taint analysis mechanism is used to protect sensitive data in the Android system. But this technique does not control flows which can cause an under-tainting problem. This means that some values should be marked as tainted, but are not. The under-tainting problem can be the cause of a failure to detect a leak of sensitive information. To solve this problem, we use a set of formally defined rules that describes the taint propagation. We prove the completeness of these rules. Also, we provide a correct and complete algorithm based on these rules to solve the under-tainting problem.

Keywords—dynamic taint analysis; android system; control flows; under-tainting; formal rules; complete algorithm

I. INTRODUCTION

Recent years have witnessed an increase in the use of embedded systems such as smartphones. According to a recent Gartner report [1], 417 million of worldwide mobile phones were sold in the third quarter of 2010, which corresponds to 35 percent increase from the third quarter of 2009. To make mobile phones more fun and useful, users usually download third-party applications. For example, we can show an increase in third-party apps of Android Market from about 15,000 third-party apps in November 2009 to about 150,000 in November 2010. These applications are used to capture, store, manipulate, and access to data of a sensitive nature in mobile phone. An attacker can launch flow control attacks to compromise confidentiality and integrity of the Android system and can leak private information without user authorization. In the study presented in Black Hat conference, Daswani [2] analyzed the live behavior of 10,000 Android applications and show that more than 800 were found to be leaking personal data to an unauthorized server. Therefore, there is a need to provide adequate security mechanisms to control the manipulation of private data by third-party apps. The dynamic taint analysis mechanism is used to protect sensitive data in the Android system against attacks [3]. But this technique does not detect control flows which can cause an under-tainting problem i.e. that some values should be marked as tainted, but are not. Let us consider the attack shown in Figure 1 that presents an under-tainting problem which can cause a failure to detect a leak of sensitive information. The variables $x$ and $y$ are both initialized to false. On Line 4, the attacker tests the user’s input for a specific value. Let us assume that the attacker was lucky and the test was positive. In this case, Line 5 is executed, setting $x$ to true and $y$ is tainted. Variable $y$ keeps its false value, since the assignment on Line 7 is not executed and $y$ is not tainted because dynamic tainting occurs only along the branch that is actually executed. As $y$ is not tainted, it is leaked to the network (Line 8) without being detected. Since $y$ has not been modified, it informs the attacker about the value of the user private contact. Thus, an attacker can circumvent an android system through the control flows.

In a previous work [4], we have proposed an approach that combines static and dynamic taint analysis to propagate taint along control dependencies and to track implicit flows in embedded systems such as the Google Android operating system. In this paper, we formally specify the under-tainting problem and we provide an algorithm to solve it based on a set of formally defined rules describing the taint propagation. We prove the completeness of those rules and the correctness and completeness of the algorithm.

The rest of this paper is organized as follows. Section 2 gives a technical overview of our approach. Section 3 presents some definitions and theorems that are used in other sections. Section 4 describes our formal specification of the under-tainting problem. In section 5, we specify an algorithm based on a set of formally defined rules describing the taint propagation policy that we use to solve the under-tainting problem. Related work about existing solutions to solve the under-tainting problem is analyzed in section 6. Finally, section 7 concludes with an outline of future work.

II. APPROACH OVERVIEW

TaintDroid [3], an extension of the Android mobile-phone platform, implements dynamic taint analysis to track the information flow in real-time and control the handling of private data. It addresses different challenges specific to mobile phones like the resource limitations. TaintDroid is composed of four modules: (1) Explicit flow module that tracks variable at the virtual machine level, (2) IPC Binder module that tracks messages between applications, (3) File module that tracks files at the storage level and (4) Taint propagation module that is implemented in the native methods level. It only tracks explicit flows and does not track control flows. In a previous work [4], we have proposed a technical approach that enhances the TaintDroid approach by tracking control flow in the Android system to solve the under-tainting problem. To track implicit flow, we have added an implicit flow module in the Dalvik VM bytecode verifier which checks instructions of methods at load time. We have defined two additional rules to propagate taint in the control flow. At class load time, we have built an array of variables to which a value is

```plaintext
1. x = false;
2. y = false;
3. char c[256];
4. if ( gets(c) != user_contact )
5.   x = true;
6. else
7.   y = true;
8. NetworkTransfer (y);
```

Figure 1. Attack using indirect control dependency
assigned to handle the branch that is not executed. Figure 2 presents the modified architecture to handle implicit flow in TaintDroid system.

In this paper, we prove the completeness of the two additional rules and we provide a correct and complete taint algorithm based on these rules. Our algorithm is based on a hybrid approach that combines and benefits from the advantages of static and dynamic analyses. We use static analysis to detect control dependencies. This analysis is based on the control flow graphs [5], [6] which will be analyzed to determine branches in the conditional structure. A basic block is assigned to each control flow branch. Then, we detect the flow of the condition-dependencies from blocks in the graph. Also, we detect variable assignment in a basic block of the control flow graph to handle not executed branches. The dynamic analysis uses information provided by the static analysis and allows tainting variables to which a value is assigned in the conditional instruction. To taint these variables, we create an array of context taints that includes all condition taints. We use the context taints array and the condition-dependencies from block in the graph to set the context taints. Finally, we apply the propagation rules to taint variables to which a value is assigned whether the branch is taken or not.

III. NOTATIONS, DEFINITIONS AND THEOREMS

Definition 1. Direct graph
A directed graph \( G = (V, E) \) consists of a finite set \( V \) of vertices and a set \( E \) of ordered pairs \((v, w)\) of distinct vertices, called edges. If \((v, w)\) is an edge, \( w \) is a successor of \( v \) and \( v \) is a predecessor of \( w \).

Definition 2. Complete directed graph
A complete directed graph is a simple directed graph \( G = (V, E) \) such that every pair of distinct vertices in \( G \) are connected by exactly one edge. So, for each pair of distinct vertices, either \((x, y)\) or \((y, x)\) (but not both) is in \( E \).

Definition 3. Control flow graph
A control flow graph \( G = (V, E, r) \) is a directed graph \((V, E)\) with a distinguished \( \text{Exit} \) vertex and start vertex \( r \), such that for any vertex \( v \in V \) there is a path from \( r \) to \( v \). The nodes of the control flow graph represent basic blocks and the edges represent control flow paths.

The concept of post-dominator and dominator tree are used to determine dependencies of blocks in the control flow graph.

Definition 4. Dominator
A vertex \( v \) dominates another vertex \( w \neq v \) in \( G \) if every path from \( r \) to \( w \) contains \( v \).

Definition 5. Post-Dominator
A node \( v \) is post-dominated by a node \( w \) in \( G \) if every path from \( v \) to \( \text{Exit} \) (not including \( v \)) contains \( w \).

Theorem 1. Every vertex of a flow graph \( G = (V, E, r) \) except \( r \) has a unique immediate dominator. The edges \( \{(\text{idom}(w), w)\} \mid w \in V \setminus \{r\} \) form a directed tree rooted at \( r \), called the dominator tree of \( G \), such that \( v \) dominates \( w \) if and only if \( v \) is a proper ancestor of \( w \) in the dominator tree[7], [8].

Computing post-dominators in the control flow graph is equivalent to computing dominators [5] in the reverse control flow graph. Dominators in the reverse graph can be computed quickly by using the Fast Algorithm [9] or a linear-time dominators algorithm [10] to construct the dominator tree. Using these algorithms, we can determine the post-dominator tree of a graph.

Definition 6. Control Dependency
Let \( G \) be a control flow graph. Let \( X \) and \( Y \) be nodes in \( G \). \( Y \) is control dependent on \( X \) noted \( \text{Dependency}(X, Y) \) if:

1) There exists a directed path \( P \) from \( X \) to \( Y \) with any \( Z \) in \( P \) (excluding \( X \) and \( Y \)) post-dominated by \( Y \) and

2) \( X \) is not post-dominated by \( Y \).

Given the post-dominator tree, Ferrante et al. [11] determine control dependencies by examining certain control flow graph edges and annotating nodes on the corresponding tree paths.

Definition 7. Context Taint
Let \( G \) be a control flow graph. Let \( X \) and \( Y \) be basic blocks in \( G \). If \( Y \) is control dependent on \( X \) that contains \( \text{Condition} \) then we associate to \( Y \) a \( \text{Context Taint} \) with \( \text{Context Taint}(Y) = \text{Taint}(\text{Condition}) \).

We use the completeness theorem to prove the completeness of the taint propagation rules in section V-A. We use the soundness theorem to prove this completeness from left to right and the compactness theorem and theorem 2 to prove from right to left. These theorems [12] are given below.

Completeness Theorem. For any sentence \( G \) and set of sentences \( \mathcal{F}, \mathcal{F} \models G \) if and only if \( \mathcal{F} \vdash G \).

Soundness Theorem. For any formula \( G \) and set of formulas \( \mathcal{F} \), if \( \mathcal{F} \vdash G \), then \( \mathcal{F} \models G \).

Compactness Theorem. Let \( \mathcal{F} \) be a set of formulas. \( \mathcal{F} \) is unsatisfiable if and only if some finite subset of \( \mathcal{F} \) is unsatisfiable.

Definition 8. CNF formula
A formula\( \mathcal{F} \) is in conjunctive normal form (CNF) if it is a conjunction of disjunctions of literals. That is,
We have a situation of under-tainting when \( A \rightarrow B \) are two variables.

**Theorem 2.** Let \( F \) and \( G \) be formulas of the first-order logic. Let \( H \) be the CNF formula obtained by applying the CNF algorithm [12] to the formula \( F \land \neg G \). Let \( \text{Res}^*(H) \) be the set of all clauses that can be derived from \( H \) using resolvents. The following are equivalent:

1. \( F \models G \)
2. \( F \not\models G \)
3. \( \emptyset \in \text{Res}^*(H) \)

IV. THE UNDER-TAINING PROBLEM

In this section we formally specify the under-tainting problem based on Denning’s information flow model. Denning [13] defined an information flow model as:

\[
FM = \langle N, P, SC, \oplus, \rightarrow \rangle.
\]

\( N \) is a set of logical storage objects (files, program variables, ...). \( P \) is a set of processes that are executed by the active agents responsible for all information flow. \( SC \) is a set of security classes that are assigned to the objects in \( N \). \( SC \) is finite and has a lower bound \( L \) attached to objects in \( N \) by default. The class combining operator “\( \oplus \)” specifies the class result of any binary function on values from the operand classes. A flow relation “\( \rightarrow \)” between pairs of security classes \( A \) and \( B \) means that “information in class \( A \) is permitted to flow into class \( B \)”. A flow model \( FM \) is secure if and only if execution of a sequence of operations cannot produce a flow that violates the relation “\( \rightarrow \)”.

We draw our inspiration from the Denning information flow model to formally specify under-tainting. However, we assign taint to the objects instead of assigning security classes. Thus, the class combining operator “\( \oplus \)” is used in our formal specification to combine taints of objects.

**Syntactic definition of connectors \( \{\rightarrow, \oplus, \rightarrow, \oplus\} \):**

We use the following syntax to formally specify under-tainting: \( A \) and \( B \) are two logical formulas and \( x \) and \( y \) are two variables.

- \( A \rightarrow B \): If \( A \) then \( B \)
- \( x \rightarrow y \): Information flow from object \( x \) to object \( y \)
- \( x \leftarrow y \): The value of \( y \) is assigned to \( x \)
- \( \text{Taint}(x) \oplus \text{Taint}(y) \): specifies the taint result of combined taints.

**Semantic definition of connectors \( \{\rightarrow, \leftarrow, \oplus\} \):**

- The \( \rightarrow \) connector is reflexive: If \( x \) is a variable then \( x \rightarrow x \).
- The \( \rightarrow \) connector is transitive: \( x, y \) and \( z \) are three variables, if \( (x \rightarrow y) \land (y \rightarrow z) \) then \( x \rightarrow z \).
- The \( \leftarrow \) connector is reflexive: If \( x \) is a variable then \( x \leftarrow x \).
- The \( \leftarrow \) connector is transitive: \( x, y \) and \( z \) are three variables, if \( (x \leftarrow y) \land (y \leftarrow z) \) then \( x \leftarrow z \).
- The \( \rightarrow \) and \( \leftarrow \) connectors are not symmetric.
- The \( \oplus \) relation is commutative: \( \text{Taint}(x) \oplus \text{Taint}(y) = \text{Taint}(y) \oplus \text{Taint}(x) \)
- The \( \oplus \) relation is associative: \( \text{Taint}(x) \oplus (\text{Taint}(y) \oplus \text{Taint}(z)) = (\text{Taint}(x) \oplus \text{Taint}(y)) \oplus \text{Taint}(z) \)

**Definition 9. Under-Tainting**

We have a situation of under-tainting when \( x \) depends on a condition, the value of \( x \) is assigned in the conditional branch and condition is tainted but \( x \) is not tainted.

Formally, an under-tainting occurs when there is a variable \( x \) and a formula condition such that:

\[
\text{IsAssigned}(x, y) \land \text{Dependency}(x, \text{condition}) \\
\land \text{Tainted}(\text{condition}) \land \neg \text{Tainted}(x)
\]

where:

- \( \text{IsAssigned}(x, y) \) associates with \( x \) the value of \( y \).
- \( \text{Dependency}(x, \text{condition}) \) defines an information flow from condition to \( x \) when \( x \) depends on the condition.

**V. THE UNDER-TAINING SOLUTION**

In this section, we specify a set of formally defined rules that describe the taint propagation. We prove the completeness of these rules. Then, we provide an algorithm to solve the under-tainting problem based on these rules. Afterwards, we analyse some important properties of our algorithm such as Correctness and Completeness.

**A. The taint propagation rules**

Let us consider the following axioms:

\[
(x \rightarrow y) \Rightarrow (\text{Taint}(y) \leftarrow \text{Taint}(x))
\]

\[
(x \leftarrow y) \Rightarrow (y \rightarrow x)
\]

\[
(\text{Taint}(x) \leftarrow \text{Taint}(y)) \land (\text{Taint}(x) \leftarrow \text{Taint}(z)) \Rightarrow (\text{Taint}(x) \leftarrow \text{Taint}(y \oplus \text{Taint}(z)))
\]

**Theorem 3.** We consider that Context_Taint is the taint of the condition. To solve the under-tainting problem, we use the two rules that specify the propagation taint policy:

- Rule 1: if the value of \( x \) is modified and \( x \) depends on the condition and the branch is taken, we will apply the following rule to taint \( x \):

  \[
  \text{IsModified}(x, \text{explicit flowstatement}) \\
  \land (\text{Dependency}(x, \text{condition})) \\
  \land (\text{BranchTaken}(\text{br}, \text{conditional statement})) \\
  \Rightarrow (\text{Taint}(x) \leftarrow \text{Context_Taint}) \\
  \oplus \text{Taint}(\text{explicit flowstatement})
  \]

  where:

  - The predicate \( \text{BranchTaken}(\text{br}, \text{conditional statement}) \) specifies that branch \( \text{br} \) in the conditional statement is executed.

So, an explicit flow which contains \( x \) is executed. \( \text{IsModified}(x, \text{explicit flowstatement}) \) associates with \( x \) the result of an explicit flow statement.

\[
\text{IsModified}(x, \text{explicit flowstatement}) \overset{\text{def}}{=} (x \leftarrow \text{explicit flowstatement})
\]

- Rule 2: if the value of \( y \) is assigned to \( x \) and \( x \) depends on the condition and the branch \( \text{br}' \) in the conditional statement is not taken (\( x \) depends only on implicit flow and does not depend on explicit flow), we will apply the following rule to taint \( x \):
Proof of taint propagation rules

To prove completeness of propagation taint rules, we use the basic rules cited in Table I for derivations.

<table>
<thead>
<tr>
<th>Premise</th>
<th>Conclusion</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$ is in $F$</td>
<td>$F \supset G$</td>
<td>Assumption</td>
</tr>
<tr>
<td>$F \supset G$ and $F \subseteq F'$</td>
<td>$F' \supset G$</td>
<td>Monotonicity</td>
</tr>
<tr>
<td>$F \supset F, F \supset G$</td>
<td>$F \supset (F \land G)$</td>
<td>$\land$-Introduction</td>
</tr>
<tr>
<td>$F \supset (F \land G)$</td>
<td>$F \supset (G \land F)$</td>
<td>$\land$-Symmetry</td>
</tr>
</tbody>
</table>

Table I

Basic rules for derivations

We start by proving completeness of the first rule. We suppose that $F = \{I\text{isModified}(x, \text{explicit}\text{flowstatement}), \text{Dependency}(x, \text{condition}), \text{BranchTaken}(br, \text{conditional\text{statement})}\}$ and $G = Taint(x) \leftarrow \text{Context\text{Taint}} \oplus \text{Taint}(\text{explicit}\text{flowstatement})$.

We prove soundness, left to right, by induction. If $F \vdash G$, then there is a formal proof concluding with $F \vdash G$ (see Table II). Let $M$ be an arbitrary model of $F$, we will demonstrate that $M \models G$. $G$ is deduced by Modus ponens of $G_1, G_j \rightarrow G$ then by induction, $M \models G_1$ and $M \models G_j \rightarrow G$ and it follows $M \models G$.

<table>
<thead>
<tr>
<th>Statement</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ($\text{condition} \rightarrow x) \vdash (\text{Taint}(x) \leftarrow \text{Taint(\text{condition}))}$</td>
<td>Axiom (2)</td>
</tr>
<tr>
<td>2. ($\text{condition} \rightarrow x) \vdash (\text{Taint}(x) \leftarrow \text{Context\text{Taint})}$</td>
<td>Taint(\text{condition}) $\rightarrow$ Context\text{Taint}</td>
</tr>
<tr>
<td>3. $F \vdash (\text{Taint}(x) \leftarrow \text{Context\text{Taint})}$</td>
<td>Monotonicity applied to 2</td>
</tr>
<tr>
<td>4. $x \vdash (x \leftarrow x)$</td>
<td>The relation $\leftarrow$ is reflexive</td>
</tr>
<tr>
<td>5. $F \vdash (x \leftarrow x)$</td>
<td>Monotonicity applied to 3</td>
</tr>
<tr>
<td>6. $(x \leftarrow x) \vdash (\text{Taint}(x) \leftarrow \text{Taint}(x))$</td>
<td>Axiom (2)</td>
</tr>
<tr>
<td>7. $F \vdash (\text{Taint}(x) \leftarrow \text{Taint}(x))$</td>
<td>Modus ponens applied to 5 and 6</td>
</tr>
<tr>
<td>8. $F \vdash (\text{Taint}(x) \leftarrow \text{Taint}(x)) \land (\text{Taint}(x) \leftarrow \text{Taint(\text{explicit}\text{flowstatement)}))$</td>
<td>$\land$-Introduction applied to 3 and 7</td>
</tr>
<tr>
<td>9. $F \vdash (\text{Taint}(x) \leftarrow \text{Taint}(x)) \land (\text{Taint}(x) \leftarrow \text{Context\text{Taint}})$</td>
<td>$\land$-Symmetry applied to 8</td>
</tr>
<tr>
<td>10. $F \vdash G$</td>
<td>Modus ponens</td>
</tr>
</tbody>
</table>

Table II

Formal proof of the first rule

Conversely, suppose that $F \vdash G$. Then $F \cup \neg G$ is unsatisfiable. By compactness, some finite subset of $F \cup \neg G$ is unsatisfiable. So there exists finite $F_0 \subseteq F$ such that $F_0 \cup \neg G$ is unsatisfiable and, equivalently, $F_0 \vdash G$. Since $F_0$ is finite, we can apply Theorem 2 to get $F_0 \vdash G$. Finally, $F \vdash G$ by Monotonicity.

B. The algorithm

The tainting algorithm that we propose, $\text{Taint\text{Algorithm}}$, allows solving the under-tainting problem. It takes as input a control flow graph of a binary program. In this graph, nodes represent a set of instructions that represent basic blocks. Firstly, it determines the control dependency of the different blocks in the graph using $\text{Dependency\text{Algorithm}}$ [11]. Afterwards, we parse the $\text{Dependency\text{List}}$ generated by $\text{Dependency\text{Algorithm}}$ and we set the context taint of blocks to include the taint of the condition that depends on whether the branch is taken or not. Finally, using the context taint and the two propagation rules, we taint all variables to which a value is assigned in the conditional branch.

Algorithm 1 Taint\text{Algorithm} (Control flow graph G)

Input: $G = (V, E, r)$ is a control flow graph of a binary program
Output: $\text{Tainted\text{Variables\text{List}}}$ is the list of variables that are tainted.

```
x \in V
y \in V
Dependency\text{List} \leftarrow Dependency\text{Algorithm}(G)
while (x, y) \in \text{Dependency\text{List}} do
    Set\text{Context\text{Taint}(y, List\text{Context\text{Taint})}
    Tainted\text{Variables\text{List} \leftarrow Taint\text{Assigned\text{Variable}(y)}
end while
```
C. Running example

We analyze the control flow graph $G = (V, E, r)$ (see Figure 5) of the bytecode given in Figure 4 to illustrate the operation of the $\text{Taint\_Algorithm}$. The source code is given in Figure 3. The $\text{Taint\_Algorithm}$ takes as input the control flow graph $G = (V, E, r)$ where:

- $V = \{ BB(1), BB(2), BB(3), BB(4) \}$
- $E = \{(BB(1), BB(2)), (BB(1), BB(3)), (BB(2), BB(4)), (BB(3), BB(4))\}$
- $r = \{BB(1)\}$

The $\text{Dependency\_Algorithm}$ checks the dependency of the blocks in the control flow graph. It generates a $\text{Dependency\_List} = \{(BB(1), BB(2)); (BB(1), BB(3))\}$. As, $BB(2)$ depends on $BB(1)$ and $BB(3)$ depends on $BB(1)$, the $\text{Taint\_Algorithm}$ sets context_taint of $BB(2)$ and context_taint of $BB(3)$ to condition taint in $BB(1)$. If $x = true$, the first branch is executed but the second is not. The first rule is used to taint modified variable $y$ in $BB(2)$ : $\text{Taint}(y) = \text{ContextTaint} \oplus \text{Taint}(\text{explicit}\_\text{flow}\_\text{statement})$. The second rule is used to taint the variable $y$ in $BB(3)$ : $\text{Taint}(y) = \text{ContextTaint} \oplus \text{Taint}(y)$. So, all variables that depend on the condition will be tainted and stored in $\text{Tainted\_Variables\_List}$ whether the branch is taken or not and we do not have an under-tainting problem.

D. Properties of the algorithm

First, we prove the correctness of the $\text{Taint\_Algorithm}$ and then we prove its completeness.

1) Correctness: We want to prove the correctness of the $\text{Taint\_Algorithm}$. Let us assume that the control flow graph is correct [14]. The proof consists of three steps: first prove that $\text{Dependency\_Algorithm}$ is correct, then prove that $\text{Set\_Context\_Taint}$ is correct, and finally prove that $\text{Taint\_Assigned\_Variable}$ is correct. Each step relies on the result from the previous step.

Correctness proof for Dependency\_Algorithm:

The $\text{Dependency\_Algorithm}$ is defined by Ferrante et al. [11] to determine dependency of blocks in the graph. This algorithm takes as input the post-dominator tree for an augmented control flow graph (ACFG). Ferrante et al. add to the control flow graph a special predicate node ENTRY that has one edge labeled ‘T’ going to START node and another edge labeled ‘F’ going to STOP node. ENTRY corresponds to whatever external condition causes the program to begin execution. The post-dominator tree of ACFG can be created using the algorithms defined in [9], [10]. These algorithms are proven to be correct.

Basic steps in the Dependency\_Algorithm:

Given the post-dominator tree, Ferrante et al. [11] determine control dependencies as follows:

- Find $S$, a set of all edges $(A, B)$ in the ACFG such that $B$ is not an ancestor of $A$ in the post-dominator tree (i.e., $B$ does not postdominate $A$).
- For each edge $(A, B)$ in $S$, find $L$, the least common ancestor of $A$ and $B$ in the post-dominator tree.

CLAIM: Either $L$ is $A$ or $L$ is the parent of $A$ in the post-dominator tree.

Ferrante et al. consider these two cases for $L$, and show that one method of marking the post-dominator tree with the appropriate control dependencies accommodates both cases.

- Case 1. $L = parent of A$. All nodes in the post-dominator tree on the path from $L$ to $B$, including $B$ but not $L$, should be made control dependent on $A$.
- Case 2. $L = A$. All nodes in the post-dominator tree on the path from $A$ to $B$, including $A$ and $B$, should be made control dependent on $A$.

Given $(A, B)$ in $S$ and its corresponding $L$, the algorithm given by Ferrante et al. traverses backwards from $B$ in the post-dominator tree until they reach $L$ and mark all nodes visited; mark $L$ only if $L = A$.

- Statements representing all marked nodes are control dependent on $A$ with the label that is on edge $(A, B)$.

They prove that the correctness of the construction follows directly from the definition of control dependency (see section III).

Referring back to this definition, for any node $M$ on the path in the post-dominator tree from (but not including) $L$ to $B$, (1) there is a path from $A$ to $M$ in the control flow graph that consists of nodes post-dominated by $M$, and (2) $A$ is not post-dominated by $M$. Condition (1) is true because the edge $(A, B)$ gives us a path to $B$, and $B$ is post-dominated by $M$. Condition (2) is true because $A$ is either $L$, in which case it post-dominates $M$, or $A$ is a child of $L$ not on the path from $L$ to $B$.

We can therefore conclude that $\text{Dependency\_Algorithm}$ is correct.
Correctness proof for Set_Context Taint

We include the taint of the condition in the context taint of the dependent blocks. As the condition taint is valid thus the inclusion operation is valid. We can conclude that Set_Context Taint is correct.

Correctness proof for Taint_Assigned Variable

We use the two propagation rules to taint variables to which a value is assigned. We proved the completeness of the two propagation rules in section V-A, thus we can conclude that Taint_Assigned Variable is complete. Therefore, we can conclude the completeness of the Taint Algorithm.

2) Completeness: Let us assume that the control flow graph is complete (see Definition 2). To prove the completeness of the Taint Algorithm, we will prove the completeness of Dependency Algorithm and Taint_Assigned Variable.

The Dependency Algorithm takes as input the post-dominator tree of the control flow graph. The post-dominator tree can be constructed using the complete algorithm defined in [10]. The Dependency Algorithm is based on the set of the least common ancestor (L) of A and B in the post-dominator tree for each edge (A, B) in S. According to the value of L, Ferrante et al. define two cases to determine the control dependency. To prove the completeness of the Dependency Algorithm, we show that Ferrante et al. prove that there does not exist another value of L (either A’s parent or A itself) to consider.

Proof: Let us assume that X is the parent of A in the post-dominator tree. So, X is not B because B is not an ancestor of A in the post-dominator tree (by construction of S). Ferrante et al. perform a proof reductio ad absurdum to demonstrate that X post-dominates B, and suppose it does not. Thus, there would be a path from B to STOP that does not contain X. But, by adding edge (A, B) to this path, a path from A to STOP does not pass through X (since, by construction, X is not B). This contradicts the fact that X post-dominates A. Thus, X post-dominates B and it must be an ancestor of B in the post-dominator tree. If X, immediate post-dominator of A, post-dominates B, then the least common ancestor of A and B in the post-dominator tree must be either X or A itself.

As only two values of L exist, there does not exist another case to compute the control dependency. The Case 2 captures loop dependency and all other dependencies are determined according to Case 1. Thus, Dependency Algorithm is complete.

We proved the completeness of the two propagation rules in Section V-A thus we can conclude that Taint_Assigned Variable is complete. Therefore, we can conclude the completeness of the Taint Algorithm.

E. Time complexity of the algorithm

The Dependency Algorithm performs with a time of at most $O(N^2)$ where N is the number of nodes in the control flow graph. Linear time algorithm to calculate control dependencies have been proposed in [15] but no proof of correctness of this algorithm was given. For each (X, Y) examined in the Dependency List, setting context taint and tainted variables can be done in constant time $O(N)$. Thus, the Taint Algorithm requires linear time using algorithm defined in [15] and at most $O(N^2)$ using Ferrante et al. algorithm.
completeness of these rules. We draw our inspiration from the Denning approach, but we define formally a set of taint propagation rules to solve the under-tainting problem and we improve the approach of Graa et al. by proving the correctness and completeness of these rules.

VII. DISCUSSION

Cavallaro et al. [35] describe evasion techniques that can easily defeat dynamic information flow analysis. These evasion attacks can use control dependencies. They demonstrate that a malware writer can propagate an arbitrarily large amount of information through control dependencies. Cavallaro et al. see that it is necessary to reason about assignments that take place on the unexecuted program branches. We implement the same idea in our taint propagation rules. Unfortunately, this will lead to an over-tainting problem (false positives). The problem has been addressed in [28] and [36] but not solved though. Kang et al. [28] used a diagnosis technique to select branches that could be responsible for under-tainting and propagated taint only along these branches in order to reduce over-tainting. However a smaller amount of over tainting occurs even with DTA++, as we can see by comparing the "Optimal" and "DTA++" results in the evaluation. Bao et al. [36] define the concept of strict control dependencies (SCDs) and introduce its semantics. They use a static analysis to identify predicate branches that give rise to SCDs. They do not consider all control dependencies to reduce the number of false positives. Their implementation gives similar results as DTA++ in many cases, but is based on the syntax of a comparison expression. Contrariwise, DTA++ uses a more general and precise semantic-level condition, implemented using symbolic execution.

In our approach, we taint all variables in the conditional branch. This causes an over-tainting. But it provides more security because all confidential data are tainted. So, the sensitive information cannot be leaked. We are interested in solving the under tainting because the false negatives are much more dangerous than the false positives since the false negatives can lead to a false sense of security. We can try to reduce the over-tainting problem by considering expert rules.

VIII. CONCLUSION

Smartphones are extensively used. However, the use of untrusted android applications can provoke the leakage of private data by exploiting the under-tainting problem. In this paper, we propose a formal approach to detect control flow and to solve the under-tainting problem in android system. We formally specify the under-tainting problem. As a solution, we provide an algorithm based on a set of formally defined rules that describe the taint propagation. We prove the completeness of those rules and the correctness and completeness of the algorithm. In the future works, we will evaluate our approach in terms of overhead and false alarms. We will also test our approach and show that it resists to code obfuscation attacks based on control dependencies.

REFERENCES


[23] Peter Hornyack, Seungyeop Han, Jaeyeon Jung, Stuart Schechter, and David Wetherall, “These aren’t the droids you’re looking for; retrofitting android to protect data from imperious applications,” in Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011, pp. 639–652.


