Tracking Null Checks in Open-Source Java Systems

Haidar Osman
SCG, University of Bern
Switzerland
osman@inf.unibe.ch

Manuel Leuenberger
University of Bern
Switzerland
manuel.leu@students.unibe.ch

Mircea Lungu
University of Groningen
The Netherlands
m.flungu@rug.nl

Oscar Nierstrasz
SCG, University of Bern
Switzerland
oscar@inf.unibe.ch

Abstract—It is widely acknowledged that null values should be avoided if possible or carefully used when necessary in Java code. The careless use of null has negative effects on maintainability, code readability, and software performance. However, a study on understanding null usage is still missing.

In this paper we analyze null checks in 810 open-source Java systems and manually inspect 100 code samples to understand when and why developers use null. We find that 35% of all conditional statements contain null checks. A deeper investigation reveals many questionable practices with respect to using null. Uninitialized member variables, returning null in methods, and passing null as a method parameter are among the most recurrent reasons for introducing null checks. Developers often return null in methods to signal errors instead of throwing a proper exception. As a result, 71% of the values checked for null are returned from method calls.

Our study provides a novel evidence of an overuse of null checks and of the null value itself in Java, and at the same time, reveals actionable recommendations to reduce this null usage.

Keywords—Null Checks; Null Usage; Static Analysis

I. INTRODUCTION

Tony Hoare\(^1\) considers null as his “billion-dollar mistake” [1]. Besides the bugs it introduces in running systems, null usage hinders performance [2], increases maintenance costs [3], and decreases code readability.

Recent studies show that a considerable number of bug fixes are recurrent [4][5][6][7]. Interestingly, the \textit{if}-related bug category is dominant [4], and more particularly, \textit{missing null check} is the most frequent pattern of bugs in Java systems [7]. These results suggest that null dereferencing is a major source of bugs in Java programs, forcing developers to add guards (null checks) on objects before using them.

Many tools and techniques have been introduced to solve the null dereferencing problem. However, a study is still missing on how null is used, why null checks are introduced, and where the checked-for-null objects come from in the source code.

In this paper, we aim at understanding when, why, and how often developers introduce null checks. More concretely, we pose the following three research questions:

\textbf{RQ1: How common are null checks?}

We answer this question by measuring the ratio between the overall number of conditional statements and those containing null checks.

\textbf{RQ2: What kind of objects are checked against null?}

We consider whether the checked object is a parameter, a member variable, or a local variable.

\textbf{RQ3: How are the checked-for-null objects initialized?}

We analyze the kind of expression that was assigned to the checked-for-null object.

\textbf{RQ4: How is null used in Java programs?}

We manually investigate 100 samples from our corpus to understand what null represents in Java programs (e.g., errors, default values, or other special values).

To answer the posed research questions, we developed a tool, \textit{NullTracker}, that statically analyzes Java source code files and approximates the statistics about null check usage. Applying our analysis to a large Java software corpus, we find that 35\% of the conditional statements in our corpus are null checks.

We also find that 24\% of the checked objects are member variables, 23\% are parameters, and 50\% are local variables. Unsurprisingly, 71\% of the checked objects come from method calls. In other words, developers insert null checks mainly when they use methods that may return null. Passing null values to methods and uninitialized member variables are recurrent reasons for introducing null checks.

We manually review 100 code samples from our corpus to understand the contexts and discover patterns of null check usage. In 76 samples null is used to represent errors, in 15 samples it is a representation of absence of a value, and in 9 samples null represents a meaningful value. Most of these instances of null usage can (or should) be replaced by proper exceptions or \textit{special case objects} such as a \texttt{Null Object} [8][9].

The rest of the paper is organized as follows: in Section II, we explain why using null leads to problems. We explain the procedure we followed to extract the null checks and the devised heuristics to analyze the checked-for-null objects in Section III. Then, we explain the experimental setup,
the terminology, and the procedure we followed to analyze null checks and null usage in Section III. In Section IV, we demonstrate the results and answer the posed research questions. In Section V, we discuss the research questions and the implications of the results. We then explain in Section VI the possible threats to validity in our study and how we tried to mitigate them. In Section VII, we discuss the related work and how the null-related problems are approached. Finally, we conclude this paper in Section VIII.

II. Motivation

Pominville et al. achieved 2% to 10% performance gain in Java bytecode when they annotated Java class files with assumptions about the “nullness” and array bounds [2]. With respect to null, their framework, SOOT [10], performs intra-procedural null analysis to make assumptions about variables being null or not to be able to remove unnecessary null checks in the bytecode level. This means that null checks impose a non-negligible performance overhead on Java programs.

In a managed language like Java, null is an unusable value. Its interpretation depends on the context and when it causes a problem, its value yields “no useful debugging information” [11]. For instance, the listFiles() method in the File class returns an array of the files, File[], in the specified directory. However, if the directory does not exist, it returns null. This returned null value might mean that the File object does not exist, that it is not a directory, or that an I/O error has occurred. This inherent ambiguity of null leads to increased difficulties in code comprehension.

Missing null checks are the most recurrent bug pattern in Java systems [7]. This bug manifests itself as the Java NullPointerException. Debugging this kind of exception is particularly hard because the stack traces do not provide enough information about the source of the null. Acknowledging this problem, Bond et al. introduced Origin Tracking [11], which records code locations where unused values (such as null) are assigned. Origin Tracking gathers the necessary information dynamically at run time so they can be reported at failure time to aid the debugging activities. This indirectly means that the overuse of null in program increases maintenance efforts.

To this end, we establish that the use of null in Java code often leads to performance degradation, code that is harder to read, more defective software, and increased project maintenance efforts. In the following sections we explore how often null is used in Java code and in what contexts. This knowledge can help software engineers to build better static code checkers and develop better practices for writing and reviewing Java code.

III. Null Check Analysis

A. Experimental Corpus

For our experiment, we used the same software corpus from a previous study [7]. This corpus was built using a crawler that queries Github\(^2\) for Java projects that have more than 5 stars (popular) and are more than 100KB in size (relatively large). This corpus contains 810 Java projects, 371,745 Java source files, and 34,894,844 lines of code. We are making the corpus available for download through the Pangea infrastructure\(^3\) [12].

B. Terminology

Before we explain the analysis, we define the terms used in this paper as follows (depicted in Figure 1):

- A **Conditional**: is a binary comparison expression that evaluates to a boolean value, such as:
  - \( y > 0 \)
  - \( x \neq \text{null} \)

- A **Conditional Statement**: is a Java statement that contains a conditional, such as:
  - if\((y > 0)\) ...
  - isNull = \((x \neq \text{null});\)

- A **Null Check**: is a conditional that contains the null literal as a left hand side or a right hand side operand. In other words, It is a comparison to null. For instance:
  - Person != null
  - iterator.next() != null

- An **N-Comparand**: is the expression that is compared to null in the null check (Usually an assignable l-value.)

\(^2\)http://www.github.com/
\(^3\)http://scg.unibe.ch/research/pangea
A. How Common Are Null Checks?

To our knowledge, only Kimura et al. have measured the density of null checks in source code [3]. They measured the ratio between the number of null checks and the number of lines of code. They found this ratio to be from one to four per 100 lines of code, depending on the project.

We, on the other hand, go one step further and measure the ratio of the conditional statements containing null checks with respect to all conditional statements. This will enable us in answering the first research question: **RQ1: How Common are Null Checks?**

We call the ratio between the null checks and the overall number of conditional statements the null check ratio. Analyzing our dataset, we found 2,329,808 conditional, 818,335 of which contain null checks. This means that a staggering 35% of the conditional statements contain null checks.

As detailed in Table I and Figure 3, the null check ratio exhibits a bell-shaped distribution around the value of 32%. In other words, more than half of the projects have a null check ratio of more than 32%. Our results show evidence of an existing overuse (or abuse) of null checks by Java programmers, which indicates the over-use of the “null” value itself. As discussed in Section II, this practice affects the readability of code, the maintainability of the project, and the performance of the running system.

C. Analysis

We implemented a tool, NullTracker\(^4\), to extract null checks and analyze the kinds and definitions of the N-Comparands. NullTracker is designed as a pipeline, following a pipes and filters architecture. NullTracker analyzes each Java source file as follows:

1) Parse the Java source file and extract the null checks.
2) For each null check, extract the N-Comparand.
3) Parse the N-Comparand and determine its kind (e.g., name, method call, field access, etc.).
4) When the N-Comparand is a name expression, determine its kind (member variable, local variable, or parameter) by looking for its NC-Declaration within the current method for local variables and parameters, and within the current type declaration for member variables.
5) When the N-Comparand is assignable (name, array access, field access), extract all the NC-Def-expressions that appear lexically (textually) before the null check and within the same method as the null check itself. Then, parse them and extract the kind of the NC-Def-values (method call, null literal, object creation, or any expression that can evaluate to a reference value).
6) Finally, the resulting data, which conforms to the model illustrated in Figure 2, is saved in the database for further analysis.

\(^4\)https://github.com/haidaros/NullTracker

![Figure 2. The data model of NullTracker analysis.](image)

D. Manual Inspection

After the analysis phase, we manually inspect multiple instances of the null checks belonging to different categories and types. More concretely, we inspect 10 random samples of each of the following categories to gain more insight:

1) Method call N-Comparands.
2) Field access N-Comparands.
3) local variable name N-Comparands.
4) Parameter name N-Comparands.
5) Member variable name N-Comparands.
6) Method invocation NC-Def-value.
7) Null literal NC-Def-value.
8) Cast NC-Def-value.
9) Object creation NC-Def-value.
10) Name NC-Def-value.

In this phase we aim at understanding how and why developers use null values and null checks. We look specifically at the following:

1) The intended semantics of the null value.
2) Potentially missing null checks (e.g., a member variable that is sometimes checked against null and sometimes not before dereferencing it).
3) The type of the checked-for-null object (String, List, Tree, Number, etc.).
4) The source of the null value. (e.g., uninitialized local variables, a return null statement in a method body, etc.)

We do not derive any statistics from this phase, as we only want to gain deeper insights into how null and null checks are used in the code and for what reasons.

IV. RESULTS

We applied our analysis to the 810 Java projects in our dataset and we manually reviewed 100 code samples. In the following subsections which are organized around the research questions, we explain the results.

A. How Common Are Null Checks?

As detailed in Table I and Figure 3, the null check ratio exhibits a bell-shaped distribution around the value of 32%. In other words, more than half of the projects have a null check ratio of more than 32%. Our results show evidence of an existing overuse (or abuse) of null checks by Java programmers, which indicates the over-use of the “null” value itself. As discussed in Section II, this practice affects the readability of code, the maintainability of the project, and the performance of the running system.
Figure 3. The distribution of the per-project null check ratio. The boxplot in (a) shows that more than half the projects have null check ratios of more than 32%. The histogram in (b) shows that the null check ratio distribution demonstrates a bell-shaped curve around the value of 32%.

### Table I
A SUMMARY OF THE PER-PROJECT NULL CHECK RATIO.

<table>
<thead>
<tr>
<th></th>
<th>Number of Conditionals</th>
<th>Number of Null Checks</th>
<th>Null Check Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>1.0</td>
<td>0.0</td>
<td>0%</td>
</tr>
<tr>
<td>1st quartile</td>
<td>117.5</td>
<td>33.0</td>
<td>23%</td>
</tr>
<tr>
<td>Median</td>
<td>531.5</td>
<td>138.5</td>
<td>32%</td>
</tr>
<tr>
<td>Mean</td>
<td>2,876.3</td>
<td>1,010.3</td>
<td>32%</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>2,171.8</td>
<td>671.5</td>
<td>40%</td>
</tr>
<tr>
<td>Max</td>
<td>196,812.0</td>
<td>76,609.0</td>
<td>100%</td>
</tr>
</tbody>
</table>

B. What Entities Do Developers Check For Null?

To answer the second research question (**RQ2**: What kind of objects are checked against null?), first, we analyze the kind of the N-Comparand itself. Second, we look for the NC-Declaration of the N-Comparand when it is an l-value (i.e., name expression, array, or field access expression).

We find that N-Comparands are mainly name expressions (78%) and method call expressions (15%), as Figure 4 shows. Figure 5 shows that 50% of the name N-Comparands are local variables, 24% are member variables, and 23% are parameters.

Inspecting 10 code samples where the null check is against a method call, the method calls are all for getter methods either from the same class or from another class. This puts field access N-Comparands, method call N-Comparands, and member variable name N-Comparands in the same category, which is member variable null check. Member variables are checked against null because they might not be initialized. This happens in the manually inspected code when one or more of the following is true:

- There exists a constructor that does not initialize it.
- There exists a constructor or a setter method that can accepts null as a parameter and sets it to null.
- There exists a method that explicitly sets to null.
- The member variable is public or is returned by address in the getter method.

The code in all inspected 10 samples where the N-Comparand is a member variable can be improved to avoid having to check for null every time the member variable needs to be used. In our samples, there is no obvious reason why member variables are not explicitly initialized in every constructor. In fact in 5 of the inspected samples, the member variable should even be final. This suggests a failure in applying well-established object-oriented design principles, in particular that of establishing class invariants [13].

As we see in Figure 5, a considerable percentage of null checks are guards on method parameters. In the inspected code samples, we observe that developers check parameters
against null mainly to validate them. However, we differentiate between two recurring patterns of null checks on parameters. In the first pattern, a method throws an exception if a certain parameter is null. Listing 1 shows a real example of this pattern. The second, and more questionable, pattern is shown in Listing 2. The method skeletons in Listing 2 are the most recurrent usage scenarios of a parameter null check. In these scenarios, the method does not accept null as a parameter. However, instead of throwing a proper exception, the method does nothing and returns silently without informing the caller of any problem.

```java
public File writeToFile(final HttpEntity entity) throws ClientProtocolException, IOException {
    if (entity == null) {
        throw new LibRuntimeException(LibResultCode.E_PARAM_ERROR);
    }
}
```

Listing 1. Throwing a proper exception when the parameter is null.

```java
SOME_TYPE method1(Param p){
    if (p==null){
        return null;
    }
    METHOD_BODY
}

void method2(Param p){
    if (p==null){
        return;
    }
    METHOD_BODY
}

void method3(Param p){
    if (p!=null){
        METHOD_BODY
    }
}
```

Listing 2. The most recurrent usage scenarios of a parameter null check.

More often than not, this indicates a poor API design. One can use, for instance, the specification pattern [14][15] to extract the validation code, which throws a proper exception, in a dedicated method or class. In any case, it is widely acknowledged that passing null as an argument to methods is a bad practice and “the rational approach is to forbid passing null by default” [16]. Nevertheless, developers often add null checks on parameters because they expect null to be passed as a parameter. Our results show a clear gap between what is considered a good practice and how software is implemented in reality.

C. Where Does The Null Come From?

When the N-Comparand is a name expression, we analyze the NC-Def-values assigned to it in all the NC-Def-expressions preceding the null check within the same method. Figure 6 shows the kinds of the assigned NC-Def-values. As an answer to the third research question (RQ3: How are the checked-for-null objects initialized?) we find that 71% of the time the NC-Def-value is a method call expression, which means that null checks are mostly applied to values returned from method invocations. In other words, when methods possibly return null, they tend to cause NullPointerExceptions in the invoking methods forcing developers to add null checks.

There is a long debate about whether methods should return null or not. In a previous study [7], we found that missing null checks represent the most frequent bug in Java programs. In this study, we show that null checks are applied to the results of method invocations. Both studies combined provide evidence that returning null in methods is a major cause of bugs. Hence, we side with the opinion that developers should avoid returning null in their method implementations and either throw an exception or return a special case object [16] such as a Null Object [8][9].

Surprisingly, some Java standard libraries exhibit this questionable design [16]. In our manually inspected code samples, we find 5 null checks because of methods from the Java standard API e.g., Map.get(...), List.get(...), Iterator.next().

Another less frequent reason for checking a local variable for null is when it is initialized within a method call or a constructor that might throw an exception before completing. We observe this pattern in our manually-inspected code samples, as the code skeleton shows in Listing 3.
Figure 6. This diagram shows that the checked-for-null objects are mainly set or initialized using method invocations.

Figure 5. The type of name expression comparands. *Undefined* indicates that 3.36% cannot be determined, as explained in Section VI.

variable is set to null first, initialized in a try-catch block, then checked for null to make sure that initialization completed and no exception was thrown.

...
null values, adds a null check before using the local variable obj.

When such scenarios accumulate in a codebase, the code starts to get harder to maintain and reason about, leading to the problems discussed in Section II.

However, our study also reveals some actionable recommendations to reduce null checks and null usage:

1) A method should not return null in case of errors. A method should always throw a proper exception that explains the exact reason and even possible solutions in case of errors.

2) Null should not be passed to public methods and public methods should not accept null as a parameter. In other words, public method arguments should be non-null by default.

3) Member variables should be initialized either in all constructors or through the use of the Builder pattern. The point here is that objects should be fully constructed before being created and class invariants should be explicitly established.

4) String instances should be initialized to empty strings " ".

5) List instances should be initialized to empty lists.

Following the aforementioned practices can prevent or at least mitigate the problems coming from null usage (as discussed in Section II). These practices can be ensured manually during code review or automatically using static code analyzers and annotations. An even more radical approach is to forbid the usage of null altogether in the language and observe the effects on the code quality, but this is a topic for further study.

VI. Threats to Validity

The internal threats to validity come from the known limitations of static analysis itself on one hand, and the limitations of our heuristics on the other hand.

- As can be seen in Figure 5, NullTracker cannot trace 3.36% of the variables back to their declarations, as can be seen in Figure 5, because we parse and analyze one Java source file at a time. For instance, inherited member variables cannot be discovered.

- NullTracker extracts all the NC-Def-expressions that appear lexically before the null check regardless of the actual data flow taking all the possible NC-Def-expressions into account.

- NullTracker cannot detect whether an N-Comparand is changed by passing it as a parameter to other methods. Only NC-Def-expressions are considered.

- When an N-Comparand appears within the same method in multiple null checks, every time it is considered a different N-Comparand leading to possible duplicates in the analysis of the N-Comparand kinds.

The external threats to validity come from the fact that we only analyzed 810 Java open-source projects. The results might not generalize to all open source projects or to industrial closed-source projects.
VII. RELATED WORK

A closely related work is a study on the maintainability burden imposed by the “return null” statement [3]. Kimura et al. found that return null statements are modified more frequently than other return statements but null conditionals are not. This indicates that the presence of return null is costly to maintain. They also found that the density of null checks are from one to four per 100 lines of code [3]. Our definition of null check ratio expresses null check usage better than null checks per LOC because it quantifies the added complexity null checks have on the code.

Null-related bugs attract the attention of many researchers and practitioners who propose various approaches to detect them as early as possible.

The first family of solutions incorporates data flow analysis techniques to detect possible null values. Some techniques are simple, fast, and intra-procedural [18][19][20][21][22] and some are more complex, thorough, and inter-procedural [23][24][25][26]. For instance, Hovemeyer and Pugh [22] perform intra-procedural forward data flow analysis to approximate the static single assignment for the values of variables. Then they analyze the dereferences for the backward data flow over the SSA approximation. This algorithm replaces the previous basic forward data flow analysis approach [27] and is now part of FindBugs [28], a static analysis tool for finding bugs in Java.

An interesting study by Ayewah and Pugh [29] compares several null dereference analysis tools and observes that, besides the reported false positives, many of the reported null dereferences (true positives) do not manifest themselves as bugs at run time. The authors claim that when the null dereference passes the initial software testing, it rarely causes bugs and “reviewing all potential null dereferences is often not as important as many other undone software quality tasks” [29]. We argue that Ayewah and Pugh underestimate the frequency of the bugs caused by dereferencing null as we show in our previous study [7].

The second family of solutions proposes to annotate the “nullness” in code. Fähndrich and Leino propose to distinguish the non-null references from the possibly-null ones at the type level (using annotations) to detect null-related bugs [30]. Papi and Ernst introduced the @NonNull annotation on types [31]. Loginov et al. [23], beside their inter-procedural null dereference analysis, propose null-related annotations to ensure the soundness and safety of the analysis. The idea of annotations made it to widely-used Java libraries like Checker Framework5 and Guava6. Also some programming languages introduce the idea of reference declarations that are not null. For instance the Spec# programming system extends the C# programming language with the support of contracts (like non-null types), allowing the Spec# compiler to statically enforce these contracts [32].

The third family of solutions tries to solve the problem by introducing language constructs. Haskell [33] and Scala [34] have the “Maybe” and the “Option” types, which are object containers. In a similar fashion, Oracle introduced the “Optional” type in Java 8 recently [35]. Groovy and C# have the safe navigation “?.” to safely invoke a method on a possibly-null object.

None of the above solutions deals with null usage problem thoroughly. The first family of solutions does not reduce null usage but points out potential null dereferencing locations in the code, encouraging developers to add even more null checks. The second family of solutions can mainly ensure that method parameters are not null, but cannot, for instance, prevent methods from returning null. The third family of solutions just encapsulates the problem with syntactic sugar rather than solving it. We argue that more research is needed to solve the problems associated with null usage by dealing with the cause of the problems instead of dealing with the symptoms. A good solution should foster good programming practices and a disciplined usage of null.

VIII. CONCLUSIONS

Null-related problems are very common in Java. The overuse of the null value may introduce more bugs, hinder performance, and lead to maintenance difficulties.

In this paper, we study null checks as a proxy to understand null usage. We conduct a census of the null checks in Java systems showing that 35% of the conditionals are null checks. Our analysis reveals many bad practices in terms of null usage. Returning null in methods, passing null as arguments, and uninitialized member variables are the most frequent, and questionable, null usage patterns causing the high null check density. Finding out the root causes behind improper null usage, we provide actionable recommendations to avoid null-related problems. These recommendations can be checked using static code analyzers to ensure a disciplined use of null and increase the quality of the code.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding of the Swiss National Science Foundations for the project Agile Software Assessment (SNSF project No. 200020-144126/1, Jan 1, 2013 - Dec. 30, 2015)7.

The authors also acknowledge the financial support of the Swiss Object-Oriented Systems and Environments (CHOOSE)8 for the presentation of this research.

5http://types.cs.washington.edu/checker-framework/
6https://github.com/google/guava
7http://p3.snf.ch/Project-144126
8http://www.choose.s-i.ch


