Towards Automated Refactoring of Code Clones

Simon Baars
simon.mailadres@gmail.com

30 June 2019, 29 pages

Research supervisor: Dr. Ana Oprescu, ana.oprescu@uva.nl
Host/Daily supervisor: Xander Schrijen, x.schrijen@sig.eu
Host organisation/Research group: Software Improvement Group (SIG), http://sig.eu/

University of Amsterdam
Faculty of Physics, Mathematics and Informatics
Master Software Engineering
http://www.software-engineering-amsterdam.nl

Abstract

This should be done when most of the rest of the document is finished. Be concise, introduce context, problem, known approaches, your solution, your findings.

Duplication in source code can have a major negative impact on the maintainability of source code. There are several techniques that can be used in order to merge clones, reduce duplication, improve the design of the code and potentially also reduce the total volume of a software system. In this study, we look into the opportunities to aid in the process of refactoring these duplication problems for object-oriented programming languages. We focus primarily on the Java programming language, as refactoring in general is very language-specific.

We first look into redefinitions for different types of clones that have been used in code duplication research for many years. These redefinitions are aimed towards flagging only clones that are useful for refactoring purposes. Our definition defines additional rules for type 1 clones to make sure two cloned fragments are actually equal. We also redefined type 2 clones to reduce the number of false positives resulting from it.

We have conducted measurements that have indicated that more than half of the duplication in code is related to each other through inheritance, making it easier to refactor these clones in a clean way. Approximately a fifth of the duplication can be refactored through method extraction, the other clones require other techniques to be applied.

Contents

Chapter 1

Introduction

Context: what is the bigger scope of the problem you are trying to solve? Try to connect to societal/economical challenges. Problem Analysis: Here you present your analysis of the problem situation that your research will address. How does this problem manifest itself at your host organisation? Also summarises existing scientific insight into the problem.

Refactoring is used to improve quality related attributes of a codebase (maintainability, performance, etc.) without changing the functionality. There are many methods that have been introduced to help with the process of refactoring [**fowler2018refactoring**, **wake2004refactoring**]. However, most of these methods still require manual assessment of where and when to apply them. Because of this, refactoring takes up a signification portion of the development process [**lientz1978characteristics**, **mens2004survey**], or does not happen at all [**mens2003refactoring**]. For a large part, refactoring requires domain knowledge to do it right. However, there are also refactoring opportunities that are rather trivial and repetitive to execute. In this thesis, we take a look at the challenges and opportunities in automatically refactoring duplicated code, also known as "code clones". The main goal is to improve maintainability of the refactored code.

Duplication in source code is often seen as one of the most harmful types of technical debt. In Martin Fowler's "Refactoring" book [**fowler2018refactoring**], he exclaims that *"Number one in the stink parade is duplicated code. If you see the same code structure in more than one place, you can be sure that your program will be better if you find a way to unify them."*.

In this research, we focus on formalizing the refactoring process of dealing with duplication in code. We will measure open source projects from the We will show the improvement of the metrics over various open source and industrial projects. Likewise, we will perform an estimation of the development costs that are saved by using the proposed solution. We will lay the main focus on the Java programming language as refactoring opportunities do feature paradigm and programming language dependent aspects [**choi2011extracting**]. However, most practises used in this thesis will also be applicable with other object-oriented languages, like C#.

1.1 Problem statement

1.1.1 Research questions

Code clones can appear anywhere in the code. Whether a code clone has to be refactored, and how it has to be refactored, is dependent on where it exists in the code (it's context). There are many different contexts in which code clones can occur (in a method, a complete class, in an enumeration, global variables, etc.). Because of this, we first must collect some information regarding in what contexts code clones exist. To do this, we will analyze a set of Java projects for their clones, and generalize their contexts. To come to this information, we have formulated the following research question:

Research Question 1:
How can we group and rank clones based on their harmfulness?

As a result from this research question, we expect a catalog of the different contexts in which clones occur, ordered on the amount of times they occur. On basis of this catalog, we have prioritized the further analysis of the clones. This analysis is to determine a suitable refactoring for the clone type that has been found at the design level. For this, we have formulated the following research question:

Research Question 2:
To what extend can we suggest refactorings of clones at the design level?

As a result, we expect to have proposed refactorings for the most harmful clone patterns. On basis of these design level refactorings we will build a model, which we will proof using Java, that applies the refactorings to corresponding methods. For building this model, we have formulated the following research question:

Research Question 3:
To what extend can we automatically refactor clones?

As a result from this research question, we expect to have a model to be able to refactor the highest priority clones.

1.1.2 Research method

1.2 Contributions

Our research makes the following contributions:

1. We deliver several novel measurements regarding code clones on a large corpus of Java projects.
2. We deliver a novel clone detection tool that finds refactorable clones in Java.
3. We deliver a novel clone refactoring tool that suggests refactorings to be applied, and applies these refactorings.
4. We give further recommendations in how refactoring can be automatically applied to improve maintainability in software projects.

1.3 Scope

In this research we will look into code clones from a refactoring viewpoint. There are several methods that detect code clones using a similarity score to match pieces of code. This similarity is often based on the amount of tokens that match between two pieces of code. The problem with similarity based clones is that it is hard to assess the impact of merging clones that have different tokens, but what exactly this token is is unknown. Because of this, we will not focus on similarity based clone detection techniques, but rather on exact matches and predefined differences.

It is very disputable whether unit tests apply to the same maintainability metrics that applies to the functional code. Because of that, for this research, unit tests are not taken into scope. The findings of this research may be applicable to those classes, but we will not argue the validity.

1.4 Outline

In Chapter 2 we describe the background of this thesis. Chapter **??** describes ... Results are shown in Chapter 5 and discussed in Chapter 7. Chapter 8, contains the work related to this thesis. Finally, we present our concluding remarks in Chapter 9 together with future work.

Chapter 2

Background

This chapter will present the necessary background information for this thesis. Here, we define some basic terminology that will be used throughout this thesis.

2.1 Clone Class

As code clones are seen as one of the most harmful types of technical debt, they have been studied very extensively. A survey by Roy et al [**roy2007survey**] states definitions for various important concepts in code clone research. In this survey, he mentions the concept "clone pair", which is *a set of two code portions/fragments which are identical or similar to each other*. Furthermore, he defined "clone class" as *the union of all clone pairs*. Apart from this, we use the definition "clone instance", which is a single code portion/fragment that is part of either a clone pair or clone class.

Figure 2.1 displays an example of a clone pair or clone class. In this case, both cloned fragments, are found in the same class. Each of the cloned fragments can be defined as a "clone instance".

Figure 2.1: Example of a clone pair, as found in the ardublock project.

Figure 2.2 displays two clone classes. These clone classes are separated by a single line that is different. The first clone class spans over both the constructor and a method of this class.

Figure 2.2: Example of two clone classes.

2.2 Clone Types

?? Duplication in code is found in many different forms. Most often duplicated code is the result of a programmer reusing previously written code [**haefliger2008code**, **baxter1998clone**]. Sometimes this code is then adapted to fit the new context. To reason about these modifications, several clone types have been proposed. These clone types are described in Roy et al [**roy2007survey**]:

> **Type I:** Identical code fragments except for variations in whitespace (may be also variations in layout) and comments.
> **Type II:** Structurally/syntactically identical fragments except for variations in identifiers, literals, types, layout and comments.
> **Type III:** Copied fragments with further modifications. Statements can be changed, added or removed in addition to variations in identifiers, literals, types, layout and comments.
> **Type IV:** Two or more code fragments that perform the same computation but implemented through different syntactic variants.

A higher type of clone means that it's harder to detect and refactor. There are many studies that adopt these clone types, analyzing them further and writing detection techniques for them [**sajnani2016sourcerercc**, **kodhai2010detection**, **van2019novel**].

2.2.1 Type 2 clones

Relating this to the clone class example of 2.2, the complete class would be flagged as a type 2 clone. This is displayed in figure 2.3. Considering such clones displays the opportunity to refactor the class as a whole.

Figure 2.3: The clone displayed in figure 2.2 as a type 2 clone.

2.2.2 Type 4 clones

For this thesis we have chosen not to consider type 4 clones for refactoring, because they are both hard to detect and hard to refactor (how to choose the best alternative for a certain computation could be a thesis in itself). A study by Kodhai et al [**kodhai2013method**] looks into the distribution of the

different types of clones in several open source systems (see table 6 of his study). It becomes apparent that type 4 clones exist way less in source code than all of the other types of clones. For instance, for the J2sdk-swing system he finds 8115 type 1 clones, 8205 type 2 clones, 11209 type 3 clones and only 30 type 4 clones. Because of that, we can conclude that type 4 clones are relatively less relevant to study.

2.3 Clone Contexts

Code clones can be found anywhere in the code. The most commonly studied type of clone is the method-level clone. Method-level clones are duplicated blocks of code in the body of a method. Many clone detection tools only focus on method-level clones (like CPD[1], Siamese[2], Sysiphus[3]). The reason for this is that with method-level clones it's most likely that the clones are harmful, and they are more straight-forward to refactor.

A paper by Lozano et al [**lozano2007evaluating**] discusses the harmfulness of cloning. In this paper the author argues that 98% are produced at method-level. However, the paper that is cited to support this claim [**bergman2004ethnographic**] does not conclude this same information. First of all, the study that is referenced uses a very small dataset (460 copy & paste instances by 11 participants). Secondly, the group of subject only consists of IBM researchers (selection bias). Thirdly, it only focuses on copy and paste instances, as opposed to other ways clones can creep into the code. Finally, the "98%" is not stated explicitly, but is vaguely derivable from one of the figures (figure 1) in this paper. Because of this, there is no reliable overview of how many clones there are in different contexts.

This thesis will focus on measuring how many clones there are per context. This way we can determine the impact of focusing our search on a specific context, like the analysis of only method-level clones. Our hypothesis is that the 98% claim is not true (we think this should be far less). We also hypothesize that clones in different contexts than method-level are less likely to be harmful and less straight forward to refactor.

2.3.1 Clone refactoring in relationship to its context

How to refactor clones is highly dependent on their context. Method-level clones can be extracted to a method [**kodhai2013method**] if all occurrences of the clone reside in the same class. If a method level clone is duplicated among classes in the same inheritance structure, we might need to pull-up a method in the inheritance structure. If instances of a method level clone are not in the same inheritance structure, we might need to either make a static method or create an inheritance structure ourselves. So not only a single instance of a clone has a context, but also the relationship between individual instances in a clone class. This is highly relevant to the way in which the clone has to be refactored.

2.4 Code clone harmfulness

There has been a lot of discussion whether code clones should be considered harmful.

Most papers view clones as harmful regarding program maintainability. *"Clones are problematic for the maintainability of a program, because if the clone is altered at one location to correct an erroneous behaviour, you cannot be sure that this correction is applied to all the cloned code as well. Additionally, the code base size increases unnecessarily and so increases the amount of code to be handled when conducting maintenance work."* [**ostberg2014automatically**]

However, the harmfulness of clones depends on a lot of factors. A paper by Kapser et al [**kapser2006cloning**] describes several patterns of cloning that may not be considered harmful. In this paper Kapser names examples where eliminating clones would compromise other important program qualities. Another study by Jarzabek et al [**jarzabek2010clones**] categorized "Essential clones": clones that are essential because of the solution that is being modelled by the program. Overall, many of the benefits of code clones do not apply to most modern object-oriented programming languages.

[1] CPD is part of PMD, a commonly used source code analyzer: `https://github.com/pmd/pmd`

[2] Siamese is an Elasticsearch based clone detector: `https://github.com/UCL-CREST/Siamese`

[3] Sisyphus crawls the Java library for existing implementations of parts of a codebase: `https://github.com/fruffy/Sisyphus`

2.5 Related work

There have been some papers that take some steps towards code clone refactoring. Most research towards refactoring code clones has been conducted by Y. Higo et al. In a 2008 study [**higo2008metric**] the authors look at the refactoring of class-level, method-level and constructor-level clones in Java.

Chapter 3

Defining refactoring-oriented clone types

In section **??** we introduced the four clone types as defined in literature. These simple definitions are suitable for analysis of a codebase. Their detection results in simple to understand numbers to argue about a codebase. However, these clone types have a few flaws which makes it hard to argue to what extend two fragements of code are functionally related. For each of type 1-3 clones [**roy2007survey**] we list our solutions to their shortcomings to increase the chance that we can refactor the clone while improving the design. Due to the serious challenges involved in their detection and refactoring, type 4 clones are not considered in this study.

We also look into clone detection tools for their suitability to support the proposed clone type definitions. We selected a few criteria Most clone detection tools support these definitions of clone types. However, many of these tools use a vastly different approach. A study by Saini et al [**saini2018towards**] outlines different clone detection tools and compares their results for each of type 1-3 clones. Even though they operate on the same type definitions, the tools used in this study yield different results.

3.1 Shortcomings of clone types

The clone definitions, as outlined in section **??**, allow reasoning about the duplication in a software system. Clones by these definitions can relatively easily and efficiently be detected. This has allowed for large scale analyses of duplication. However, these clone type definitions have shortcomings which makes the clones detected in correspondance with these definitions less valuable for (automated) refactoring purposes.

In this section we discuss the shortcomings of the different clone type definitions which make them less suitable for (automated) refactoring. Because of that, these clones require more judgement whether they should and can be refactored.

3.1.1 Type 1 clones

Type 1 clones are *identical clone fragments except for variations in whitespace and comments* [**roy2007survey**]. This allows for the detection of clones that are the result of copying and pasting existing code, along with other reasons why duplicates might get into a codebase.

Type 1 clones are in most cases implemented as textual equality between code fragments (except for whitespace and comments). Although textually equal, method calls can still refer to different methods, type declarations can still refer to different types and variables can be of a different type. In such cases refactoring opportunities could be invalidated. An example of such a case is displayed in figure 3.1.

GameWinningScene.java

```java
package com.simonbaars.fruitcollector;

import com.simonbaars.fruitcollector.fruits.winning.Apple;
import com.simonbaars.fruitcollector.images.winning.*;

import com.musiclibrary.music.highvolume.*;

public class GameWinningScene extends GameScene
{
    protected final Player player;

    public GameScene(Player player)
    {
        this.player = player;
    }

    public String checkAppleCollected(Apple collectedApple)
    {
        if(player.inventory.contains(collectedApple)){
            showImage(new PNGImage());
        }
        playSong(new MusicPlayer());
    }

    @Override
    public void playSong(MusicPlayer musicPlayer){
        musicPlayer.playSong("theme.mp3");
    }
}
```

GameOverScene.java

```java
package com.simonbaars.fruitcollector;

import com.simonbaars.fruitcollector.fruits.rotten.Apple;
import com.simonbaars.fruitcollector.images.losing.*;

import com.musiclibrary.music.lowvolume.*;

public class GameOverScene extends GameScene
{
    protected final Player player;

    public GameOverScene(Player player)
    {
        super(player);
    }

    public String checkAppleCollected(Apple collectedApple)
    {
        if(player.inventory.contains(collectedApple)){
            showImage(new PNGImage());
        }
        playSong(new MusicPlayer());
    }

    @Override
    public void playSong(MusicPlayer musicPlayer){
        musicPlayer.playSong("theme.mp3");
    }
}
```

Figure 3.1: Example of a type 1 clone.

In the example in figure 3.1, we see a type 1 clone consisting of two methods. However, these clones might still be very hard to refactor as we cannot see by this example whether they are functionally equal. Both code fragments use different imported types, some of which imported via a wildcard. Because of this, it is hard to verify which of the used types have the same underlying implementation. This can make type 1 clones less suitable for refactoring purposes, as they require additional judgement regarding the refactorability of such a clone. When aiming to automatically refactor clones, applying refactorings to clones as shown in figure 3.1, is bound to be error prone and result in a uncompilable project or a difference in functionality.

Because of this, type 1 clones may not all be subject to refactoring. In section we describe an alternate approach towards detecting type 1 clones, which results in only clones that can be refactored.

3.1.2 Type 2 clones

Type 2 clones are *structurally/syntactically identical fragments except for variations in identifiers, literals, types, layout and comments* [**roy2007survey**]. This definition allows for the reasoning about code fragments that were copied and pasted, and then slightly modified. The two methods displayed in figure 3.2 are type 2 clones of each other.

```java
public boolean containsOnlyRedCircles(List<Circle> listOfCircles){
    return listOfCircles.stream().allMatch(Shape::isRed);
}

public Apple getEdibleAppleFromBasket(FruitBasket<Apple> appleBasket){
    return appleBasket.getFruitContainer().getAppleByCriterium(Fruit::hasNotYetBeenEaten);
}
```

Figure 3.2: Example of a type 2 clone.

Looking at the example in listing 3.2, we see an example of a type 2 clone that poses no harm to the design of the system. Both methods are, except for their matching structure, completely different in functionality. They operate on different types, call different methods, return different things, etc. Having such a method flagged as a clone does not provide much useful information.

When looking at refactoring, type 2 clones can be very difficult to refactor. For instance if we have variability in types, the code can describe operations on two completely dissimilar types. Type 2 clones do not differentiate between primitives and objects, which makes the these clones often not so useful for refactoring purposes.

3.1.3 Type 3 clones

Type 3 clones are *copied fragments with further modifications. Statements can be changed, added or removed in addition to variations in identifiers, literals, types, layout and comments* [**roy2007survey**]. Detection of clones by this definition can be very hard, as it may be hard to detect whether a fragment was copied in the first place if it was severely changed. Because of this, most clone detection implementations of type 3 clones work on basis of a similarity threshold [**svajlenko2014evaluating**, **cordy2011nicad**]. This similarity threshold has been implemented in different ways: textual similarity (for instance using levenshtein distance) [**lavoie2011automated**], token-level similarity of statement-level similarity.

Having a definition that allows for any change in code poses serious challenges on refactoring. A levenshtein distance of one can already change the meaning of a code fragment significantly, for instance if the name of a type differs by a character (and thus referring to different types).

3.2 Refactoring-oriented clone types

To resolve the shortcomings of clone types as outlined in the previous section, we propose alternative definitions for clone types to be directed at detecting clones that can and should be refactored. We have named these clones type 1R, 2R and 3R clones. These definitions share similarities with the literature definitions, the number of each type corresponds with the clone type it is modeled after. The "R" stands for refactoring-oriented (and may be less suitable for other analyses).

3.2.1 Type 1R clones

We propose an alternative definition of type 1 clones. This definition requires cloned fragments to be not just textually equal, but also functionally equal. Although requiring fragements to be functionally equal, type 1R clones do not allow for change in implementation (like type 4 clones). We check functional equality of two fragments by validating the equality of the fully qualityfied identifier for referenced types, methods and variables. Type 1R clones are always a subset of type 1 clones.

Referenced Types

Many object-oriented programming languages (like Java, Python, C#) require the programmer to import a type (or the class in which it is declared) before it can be used. Based on what is imported, the meaning of the name of a type can differ. For instance, if we import `java.util.List`, we get the interface which is implemented by all list datastructures in Java. However, importing `java.awt.List`, we get a listbox GUI component for the Java Abstract Window Toolkit (AWT). Because of this, for type 1R clones, we compare the fully qualified identifier for all referenced types.

Called methods

A codebase can have several methods with the same name. The implementation of these methods might differ. When we call two methods with an identical name, we can in fact call different methods. This is another reason that textually identical code fragements can differ functionally.

Because of this, for type 1R clones, we compare the fully qualified method signature for all method references. A fully qualified method signature consists of the fully qualified name of the method, the fully qualified type of the method plus the fully qualified type of each of its arguments. For instance, an `eat` method could become `java.lang.Boolean com.simonbaars.fruitgame.Apple.eat(java.util.List<com.simonbaars`

Variables

In typed programming languages, each variable declaration should declare a name and a type. When we reference a variable, we only use its name. If, in different code fragments, we use variables with the same name but different types, the code can be functionally unequal but still textually equal. As an example, see the code in figure **??**.

```java
public boolean containsOnlyRedCircles(List<Circle> listOfCircles){
    return listOfCircles.stream().allMatch(Shape::isRed);
}

public Apple getEdibleAppleFromBasket(FruitBasket<Apple> appleBasket){
    return appleBasket.getFruitContainer().getAppleByCriterium(Fruit::hasNotYetBeenEaten);
}
```

Figure 3.3: Variables with different types but the same name.

The body of both methods in figure **??** is equal. However, their functionality is not. The first method adds two numbers together and the other concatenates an integer to a String.

For type 1R clones variable references should be compared by both type and name.

3.3 Type 2R clones

Type 2R clones are modelled after type 2 clones, which allow any change in identifiers, literals, types, layout, and comments. For refactoring purposes, this definition is unsuitable; if we allow any change in identifiers, literals, and types, we cannot distinguish between different variables, different types and different method calls anymore. This could render two methods that have an entirely different functionality as clones (as shown in figure 3.2 previously). Refactoring such clones can be harmful instead of helpful.

We tackle these problems with type 2R clones to be able to detect such clones that can and should be refactored. Type 1R clones are a subset of type 2R clones. All rules that apply to type 1R clones also apply to type 2R clones. Additionally, type 2R clones allow variability in literals, variables and method calls. This variability however is constrained by a threshold. Type 2R do not allow any variability in types, as opposed to type 2 clones which do allow variability in types.

A threshold for variability in literals, variables and method calls

Type 2 clones allow any variability in literals, variables and method identifiers. However, this information tells a lot about the meaning of the code fragment. Most clone detection tools do not differentiate between a type 2 clone that differs by a single literal/identifier and one that differs by many. However, this does have a big impact on the meaning of the code fragment and thus the harmfulness of the duplication being there.

For type 2R clones we define a threshold for variability in literals, variables and method calls. We calculate the variability in literals, variables and method calls using the following formula:

T2R Variability = (Diff(l) + Diff(v) + Diff(m))/Total(t) Diff = Amount that differ from other clone instances in the clone class Total = Total number in the clone instance l = Number of literals v = Number of variables m = Number of method calls t = Number of tokens

Literal variability We allow only variability in the value of literals, but not in the type of literals. This is because a difference in literal type may have a big impact on the refactorability of the cloned fragment. When we refactor literals that have both the same type, in case of an "Extract Method" refactoring, we create a method argument for this literal and pass the corresponding literal from cloned locations. However, if two literals have different types, this might not be possible (or will have a negative effect on the design of the system).

Allow any change in some identifiers

- Identifier of method declarations - Identifier of locally declared variables - Identifier of class/interface/enum declaration - etc?

- **Considering types:** Type 2 clones do not consider types. However, this can make a code fragment very hard to refactor, as different types can describe different functional concepts. Because of this, we propose that type 2R clones should consider types like type 1R clones do.
- **Having a distinction between different variables:** For type 2 clones, no identifiers would be taken into account. We agree that a difference in identifiers may still result in a harmful clone, but we should still consider the distinction between different variables. For instance, if we call a

method like this: `myMethod(var1, var2)`, or call this method like this: `myMethod(var1, var1)`. Even if the variables have the same type, the distinction between the variables is important to ensure the functionality is the same after refactoring.

- **Defining a threshold for variability in literals:** For type 2 clones no literals would be taken into account. We agree, as when refactoring the clone (for example by extracting a method), we can turn the literal into a method parameter. However, we would argue that thresholds matter here. How many literals may differ for the segment still to be considered a clone with another segment? We need to define a threshold to be sure that, by refactoring, we are not replacing a code fragment by a worse maintainable design.
- **Consider method call signatures and define a threshold for variability in method calls:** As type-2 clones allow changes in identifiers, also the names of called methods may vary. However, because of this, completely different methods can be called in cloned fragments as a result. This poses serious challenges on refactoring and makes it more disputable whether such a clone is harmful for the maintainability of the code. This is because different method identifiers can describe a completely different functionality. Therefore, we suggest considering the call signatures of cloned methods when they are compared. We can allow variability in the rest of method identifiers by passing the function as a parameter. To limit the amount of parameters required we also recommend defining a threshold for variability in method call expressions, so only a limited number of method calls can vary.

3.4 Type 3R clones

Type 3 clones are even more permissive than type 2 clones, allowing added and removed statements. For these clones, thresholds matter a lot to make sure that not the whole project is detected as a clone of itself. The main question for this study regarding type 3 clones is: *"how can we refactor type 3 clones while improving the design?"*.

Clone instances in type 3 clones are almost always different in functionality. As we have to ensure equal functionality after refactoring the clone, we have to wrap the difference in statements between the clone instances in conditional blocks. We can then pass a variable to indicate which path should be taken through the code (either a boolean or an enumeration). Such a refactoring would make added statements that are contiguous less harmful for the design than added statements that are scattered throughout the cloned fragment.

We also argue that statements that are not common between two clone instances, should not count towards the size of the clone (and thus towards the threshold which determines whether the clone will be taken into account). As for the detection of type 3 clones, we think the easiest opportunity to detect these clones is to consider it as a postprocessing step after clone detection. By trying to find short gaps between clones, we can find opportunities to refactor clone classes into a single type 3 clone class. The amount of statements that this "short gap" can maximally span should be dependent on a threshold value.

3.5 The challenge of detecting these clones

To detect each type of clone, we need to parse the fully qualified identifier of all types, method calls and variables. This comes with serious challenges, regarding both performance and implementation. Also, to be able to parse all fully qualified identifiers, and trace the declarations of variables, we might need to follow cross file references. The referenced types/variables/methods might even not be part of the project, but rather of an external library or the standard libraries of the programming language. All these factors need to be considered for the referenced entity to be found, on basis of which a fully qualified identifier can be created.

3.6 Unifying the types

In this chapter we have proposed refactoring-oriented definitions using the type 1, 2 and 3 clone definitions from literature as a baseline. In literature, these definitions are mainly aimed towards reasoning about duplication in source code. When considering these types for refactoring, the goal becomes slighly different. Because of this, having separate clone type definitions does not have any value. Rather, we

need a single clone type definition by which we can detect all clones that can and should be considered for refactoring.

Because of this, the ultimate goal would be not to consider type 1R, 2R and 3R separately, but together. However, this is dependent on good thresholds for the type 2R variability and type 3R gap size. Because of this, we have dedicated section 5.1 to performing measurements to find good thresholds. The ultimate goal is to have a single unified definition of clones that can and should be refactored. Although it will be next to impossible to define such a definition and its corresponding thresholds that does not detect false positives. However, we strive to find at least a near-optimal set of thresholds regarding the type definitions proposed in this chapter.

3.7 Suitability of existing Clone Detection Tools towards our refactoring purposes

We conducted a short survey on (recent) clone detection tools that we could use to analyze refactoring possibilities. The results of our survey are displayed in table 3.1. We chose a set of tools that are open source and can analyze a popular object-oriented programming language. Next, we formulate the following four criteria by which we analyze these tools:

1. **Should find clones in any context.** Some tools only find clones in specific contexts, such as only method-level clones. We want to perform an analysis on all clones in projects to get a complete overview.
2. **Should find all clones in created control projects.** We assembled a number of test projects to assess the validity of clone detection tools. On basis of this, we checked whether clone detection tools can correctly find clones in diverse contexts.
3. **Can analyse resolved symbols.** When detecting clones for refactoring purposes, it is important that clone instances can be refactored. Sometimes, textual equality between code fragments does not imply that these can be refactored (this is described more elaborately in section 3.1.1). Because of this, we want to use a clone detection tool that can analyze such structures.
4. **Extensive detection configuration.** We aim to exclude expressions/statements from matching (more about our rationale in section 3). To achieve this, the tool needs to be able to allow those threshold changes. This can be either through simple changes of the source code, or by using some configuration file.

Table 3.1: Our survey on clone detection tools.

Clone Detection Tool	**(1)**	**(2)**	**(3)**	**(4)**
Siamese [**ragkhitwetsagul2019siamese**]				✓
NiCAD [**roy2008nicad**, **cordy2011nicad**]	✓	✓		
CPD [**roy2009comparison**]	✓	✓		
CCFinder [**kamiya2002ccfinder**] D-CCFinder [**livieri2007very**]	✓	✓		
CCFinderSW [**semura2017ccfindersw**]	✓			✓
SourcererCC [**sajnani2016sourcerercc**] Oreo [**saini2018oreo**]	✓			✓
BigCloneEval [**svajlenko2016bigcloneeval**]	✓	✓		
Deckard [**jiang2007deckard**]	✓		✓	
Scorpio [**higo2013revisiting**, **kamalpriya2017enhancing**]	✓		✓	✓

None of the state-of-the-art tools we identified implement all our criteria, so we decided to implement our own clone detection tool: CloneRefactor[1].

[1]CloneRefactor (WIP) is available on GitHub: `https://github.com/SimonBaars/CloneRefactor`. This repository contains all scripts that were used to retrieve the data that is displayed in this paper.

Chapter 4

Evaluation setup

In this chapter we describe the setup we use for our experiments. Our most prominent contribution is the proposal of a tool called CloneRefactor. This tool allows us to map clones with all clone definitions as described in chapter 3.

All data for our experiments, as displayed in chapter 5, is measured over a corpus of Java projects. In this chapter we will explain how we prepared this corpus.

4.1 CloneRefactor

CloneRefactor is the name of our clone detection and refactoring tool. It features the following novel functions:

- Detection of clone classes rather than clone pairs.
- A novel detection method, aimed at extensibility.
- Detection of refactoring-oriented clone types, in addition to the literature clone types.
- Allows for automated refactoring of a subset of the detected duplication issues.

In this section we describe our approach and rationale for the design decisions regarding this tool.

4.1.1 JavaParser

A very important design decision for CloneRefactor is the usage of a library named JavaParser [**tomassetti2017javaparse**
JavaParser is a Java library which allows to parse Java source files to an abstract syntax tree (AST). JavaParser allows to modify this AST and write the result back to Java source code. This allows us to apply refactorings to the detected problems in the source code.

Integrated in JavaParser is a library named SymbolSolver. This library allows for the resolution of symbols using JavaParser. For instance, we can use it to trace references (methods, variables, types, etc) to their declarations (these referenced identifiers are also called "symbols"). This is very useful for the detection of our refactoring-oriented clone types, as they make use of the fully qualified identifiers of symbols.

In order to be able to trace referenced identifiers SymbolSolver requires access to not only the analyzed Java projects, but also all its dependencies. This requires us to include all dependencies with the project. Along with this, SymbolSolver solves symbols in the JRE System Library (the standard libraries coming with every installation of Java) using the active Java Virtual Machine (JVM). This has a big impact on performance efficiency.

Because of the requirement of symbol resolution, the refactoring-oriented clone types are less suitable for large scale clone analysis.

4.1.2 Clone Detection

To detect clones, CloneRefactor parses the AST aqcuired from JavaParser to an unweighted graph structure. On basis of this graph structure, clones are detection. Dependent on the type of clones being detected, transformations may be applied. The way in which CloneRefactor was designed does not allow for several clone types to be detected simultaneously, in accordance with our clone type philosophy as described in chapter **??**.

4.1.3 Type specific transformations

4.2 The corpus

For our measurements we use a large corpus of open source projects [**githubCorpus2013**]. This corpus has been assembled to contain relatively higher quality projects. Also, any duplicate projects were removed from this corpus. This results in a variety of Java projects that reflect the quality of average open source Java systems and are useful to perform measurements on.

As indicated in chapter 3.5 CloneRefactor requires all libraries of software projects we test. As these are not included in the used corpus [**githubCorpus2013**], we decided to filter the corpus to only include Maven projects. Maven is a build automation tool used primarily for Java, and works on basis of an `pom.xml` file to describe the projects' dependencies. As no `pom.xml` files are included in the corpus, we cloned the latest version of each project in the corpus. We then removed each project that has no `pom.xml` file. As a final step, we collected all dependencies for each project by using the `mvn dependency:copy-dependencies -DoutputDirectory=lib` Maven command, and removed each project for which not all dependencies were available (due to non-Maven dependencies being used or unsatisfiable dependencies being referenced in the `pom.xml` file).

Some general data regarding this corpus is displayed in Table 4.1.

Table 4.1: General results for GitHub Java projects corpus [githubCorpus2013].

Amount of projects	1,361
Amount of lines (excluding whitespace, comments and newlines.)	1,414,996
Amount of statements/declarations	1,212,189
Amount of tokens (excluding whitespace, comments and newlines.)	11,643,194

Chapter 5

Results

In this chapter, we present the results of our experiments.

5.1 Thresholds

5.2 Relation, Location and Content Analysis of Clones

To be able to refactor code clones, it is very important to consider the context of the clone. We define the following aspects of the clone as its context:

1. The relation of clone instances among each other through inheritance (for example: a clone instance resides in a superclass of another clone instance in the same clone class).
2. Where a clone instance occurs in the code (for example: a method-level clone is a clone instance that is in a method).
3. The contents of a clone instance (for example: the clone instance spans several methods).

Figure 5.1: Abstract representation of clone classes and clone instances.

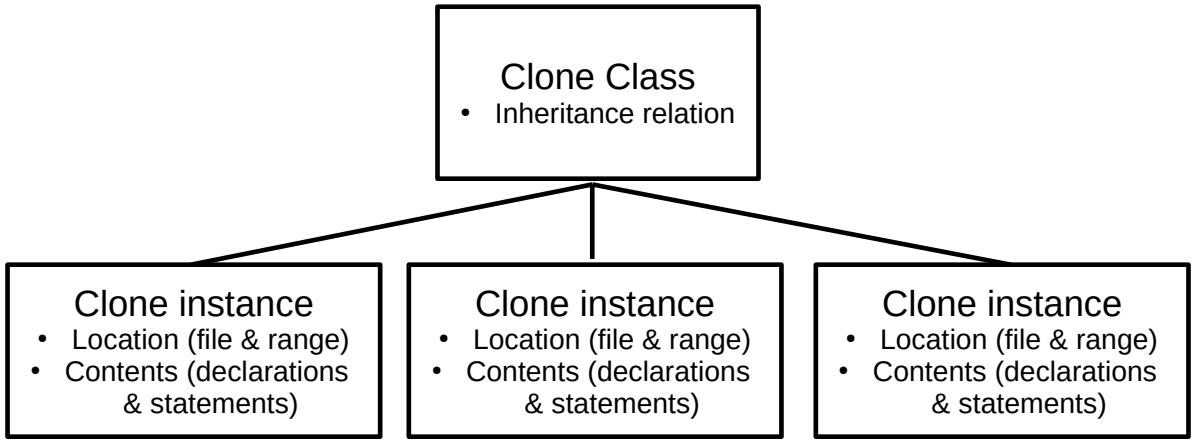


Figure 5.1 shows an abstract representation of clone classes and clone instances. The relation of clones through inheritance is measured on clone class level: it involves all child clone instances. The location and contents of clones is measured on clone instance level. A clone's location involves the file it resides in and the range it spans (for example: line 6 col 2 - line 7 col 50). A clone instance contents consists of a list of all statements and declarations it spans.

We analyzed the context of clones in a large corpus of open source projects. For these experiments, we used our CloneRefactor tool. These experiments follow the structure of the context: The relation between clone instances is explained, measured and discussed in chapter 5.4; the location of clone instances is explained, measured and discussed in chapter 5.5; the content of clone instances are explained, measured

and discussed in chapter 5.6.

5.3 Clone detection results

Currently, we have implemented two clone detection algorithms into CloneRefactor. The first one finds clones by comparing tokens (excluding whitespace, comments and newlines), equal to the definition of type 1 clones in literature [**roy2007survey**]. The second algorithm implements our type 1R, as explained in chapter 3.1.1. The differences between the clones found for these algorithms is displayed in Table 5.1.

Table 5.1: CloneRefactor clone detection results for the two different algorithms.

	Type 1	**Type 1R**
Amount of lines cloned	200,362	129,519
Amount of statements/ declarations cloned	182,466	118,980
Amount of tokens cloned	1,582,845	973,596

Looking at Table 5.1, it becomes apparent that the type 1R algorithm finds significantly less clones than the type 1 algorithm. This indicates that about a third of the clones have textual equality, but are not actually equal when considering the types of expressions. This makes these clones less suitable for automated refactoring.

5.4 Relations Between Clone Instances

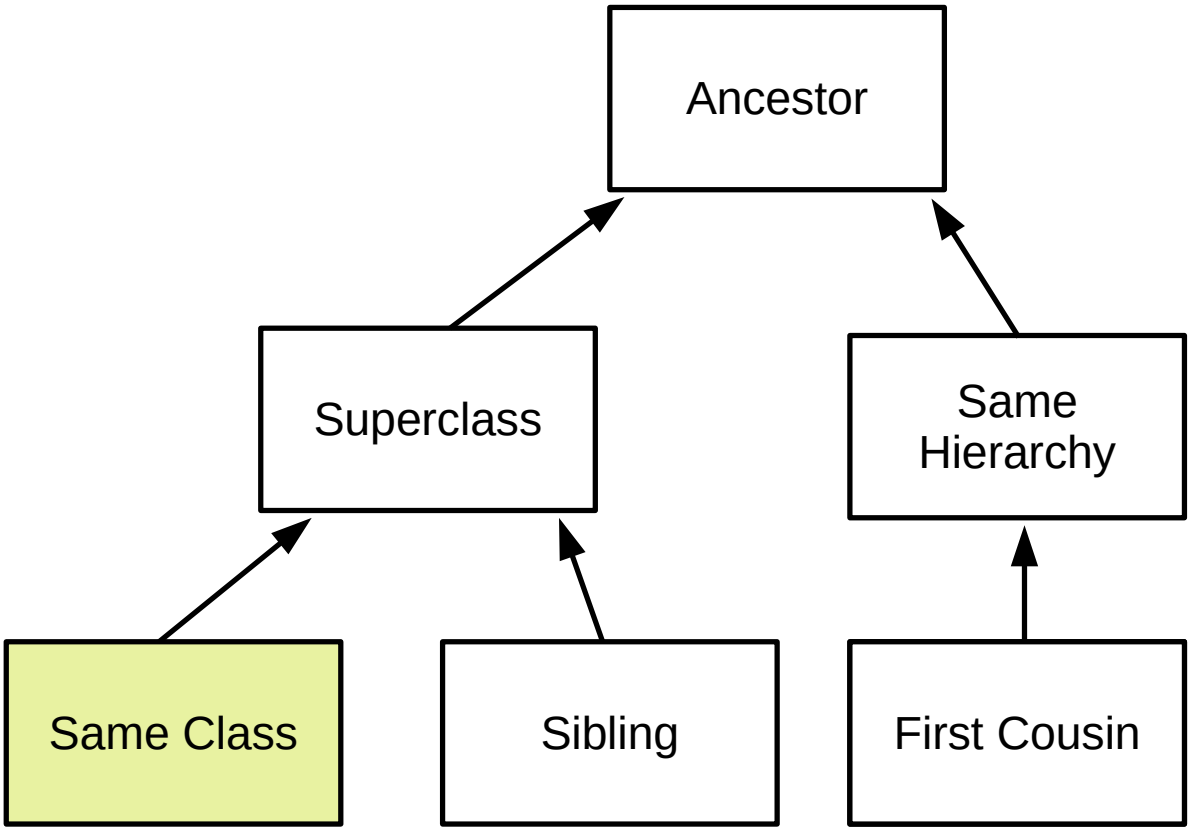
When merging code clones in object-oriented languages, it is very important to consider the relation between clone instances. This relation has a big impact on how a clone should be merged, in order to improve the software design in the process. In this chapter, we display measurements we conducted on the corpus introduced in chapter 4.2. These measurements are based on an experiment by Fontana et al. [**fontana2015duplicated**], which we will briefly introduce in chapter 5.4.1. We use a vastly different setup, which is explained in chapter 5.4.2. We then show our results in chapter 5.4.3.

5.4.1 Categorizing Clone Instance Relations

Fontana et al. [**fontana2015duplicated**] describe measurements on 50 open source projects on the relation of clone instances to each other. To do this, they first define several categories for the relation between clone instances in object-oriented languages. A few of these categories are shown in Figure 5.2. These categories are as follows:

1. **Same method**: All instances of the clone class are in the same method.
2. **Same class**: All instances of the clone class are in the same class.
3. **Superclass**: All instances of the clone class are children and parents of each other.
4. **Ancestor class**: All instances of the clone class are superclasses except for the direct superclass.
5. **Sibling class**: All instances of the clone class have the same parent class.
6. **First cousin class**: All instances of the clone class have the same grandparent class.
7. **Common hierarchy class**: All instances of the clone class belong to the same hierarchy, but do not belong to any of the other categories.
8. **Same external superclass**: All instances of the clone class have the same superclass, but this superclass is not included in the project but part of a library.
9. **Unrelated class**: There is at least one instance in the clone class that is not in the same hierarchy.

Figure 5.2: Abstract figure displaying some relations of clone classes. Arrows represent superclass relations.



Please note that none of these categories allow external classes (except for "same external superclass"). So if two clone instances are related through external classes but do not share a common external superclass, it will be flagged as "unrelated". The main reason for this is that it is (often) not possible to refactor to external classes.

5.4.2 Our setup

We use a similar setup to that used by Fontana et al. (Table 3 of Fontana et al. [**fontana2015duplicated**]). Fontana et al. measure clones using their own tool (DCRA). As explained in chapter 3.7, we chose to implement our own tool, CloneRefactor. Therefore, the setup for our measurements differs as follows from Fontana et al.:

- We consider clone classes rather than clone pairs. The rationale for this is given in chapter **??**.
- We use different thresholds regarding when a clone should be considered. Fontana et al. seek clones that span a minimum of 7 source lines of code (SLOC). We seek clones with a minimum size of 6 statements/declarations. This is explained detail in chapter **??**.
- We seek duplicates by statement/declaration rather than SLOC. This makes our analysis depend less on the coding style (in terms of newline usage) of the author of the software project.
- We test a broader range of projects. Fontana et al. use a set of 50 relatively large projects. We use the corpus as explained in 4.2, which contains a diverse set of projects (diverse both in volume and code quality).

5.4.3 Our results

Table 5.2 contains our results regarding the relations between clone instances. In this table, "T1" stands for the type 1 algorithm from literature and "1R" stands for our type 1R definition as explained in chapter 3.1.1.

Table 5.2: Clone relations

Relation	# T1	% T1	# 1R	% 1R
Unrelated	6,134	35.48	4,762	38.14
Same Class	4,772	27.60	3,131	25.07
Sibling	2,680	15.50	1,949	15.61
Same Method	2,247	13.00	1,685	13.49
External Superclass	794	4.59	558	4.47
First Cousin	269	1.56	197	1.58
Superclass	237	1.37	118	0.94
Common Hierarchy	123	0.71	73	0.58
Ancestor	35	0.20	14	0.11

The most notable difference when comparing it to the results of Fontana et al. [**fontana2015duplicated**] is that in our results most of the clones are unrelated (38.14% with type 1R), while for them it was only 15.70%. This might be due to the fact that we consider clone classes rather than clone pairs, and mark the clone class "Unrelated" even if just one of the clone instances is outside a hierarchy. It could also be that the corpus which we use, as it has generally smaller projects, uses more classes from outside the project (which are marked "Unrelated" if they do not have a common external superclass). About a fourth of all clone classes have all instances in the same class, which is generally easy to refactor. On the third place come the "Sibling" clones, which can often be refactored using a pull-up refactoring. There are no noteworthy differences between type 1 and type 1R clones.

5.5 Clone instance location

After mapping the relations between individual clones, we looked at the location of individual clone instances. A paper by Lozano et al. [**lozano2007evaluating**] discusses the harmfulness of cloning. The authors argue that 98% are produced at method-level. However, this claim is based on a small dataset and based on human copy-paste behavior rather than static code analysis. We validated this claim over our corpus. The results for the clone instance locations are shown in Table 5.3. We chose the following categories:

1. **Method/Constructor Level:** A clone instance that does not exceed the boundaries of a single method or constructor (optionally including the declaration of the method or constructor itself).
2. **Class Level:** A clone instance in a class, that exceeds the boundaries of a single method or contains something else in the class (like field declarations, other methods, etc.).
3. **Interface Level:** A clone that is (a part of) an interface.
4. **Enumeration Level:** A clone that is (a part of) an enumeration.

Please note that these results are measured over each clone instance rather than each clone class, hence the higher total amount in comparison to the results of chapter 5.4.3.

Table 5.3: Clone instance locations

Location	# T1	% T1	# 1R	% 1R
Method Level	32,861	66.02	19,075	58.23
Class Level	15,069	30.27	12,207	37.27
Constructor Level	1,391	2.79	1,080	3.30
Interface Level	282	0.57	247	0.75
Enum Level	171	0.34	147	0.45

Our results indicate that around 58% of the clones are produced at method-level. About 39% of clones either span several methods/constructors or contain something like a field declaration. Another 3% of the clones are found in constructors. The amount of clones found in interfaces and enumerations is very low. Regarding the differences between type 1 and type 1R, it seems that there are relatively less method level clones and more class level clones for type 1R. This is probably due to that the main reason for variability between type 1 and type 1R is variable references, which occur more at method level than class level.

5.6 Clone instance contents

Finally, we looked at the contents of individual clone instances: what kind of declarations and statements do they span. We selected the following categories to be relevant for refactoring:

1. **Full Method/Class/Interface/Enumeration:** A clone that spans a full class, method, constructor, interface or enumeration, including its declaration.
2. **Partial Method/Constructor:** A clone that spans a method partially, optionally including its declaration.
3. **Several Methods:** A clone that spans over two or more methods, either fully or partially, but does not span anything but methods (so not fields or anything in between).
4. **Only Fields:** A clone that spans only global variables.
5. **Includes Fields/Constructor:** A clone that spans a combination of fields and other things, like methods.
6. **Method/Class/Interface/Enumeration Declaration:** A clone that contains the declaration (usually the first line) of a class, method, interface or enumeration.
7. **Other:** Anything that does not match with above-stated categories.

The results for these categories are displayed in Table 5.4.

Table 5.4: Clone instance contents

Contents	**# T1**	**% T1**	**# 1R**	**% 1R**
Partial Method	32,214	64.72	18,791	57.37
Several Methods	10,542	21.18	8,514	25.99
Includes Constructor	1,772	3.56	1,213	3.70
Includes Field	1,681	3.38	1,487	4.54
Partial Constructor	1,389	2.79	1,078	3.29
Only Fields	962	1.93	888	2.71
Full Method	647	1.30	284	0.87
Includes Class Declaration	263	0.53	258	0.79
Other Categories	304	0.61	243	0.74

Unsurprisingly, most clones span a part of a method. More than a quarter of the clones (for type 1R) span over several methods, which either requires more advanced refactoring techniques or indicates a non-harmful clone.

Chapter 6

Merging duplicate code through refactoring

The most used technique to merge clones is method extraction (creating a new method on basis of the contents of clones). However, method extraction cannot be applied in all cases. Sometimes a clone spans a statement partially (like a for-loop of which only it's declaration and a part of the body is cloned). Merging the clones can be harder in such instances. Also, the cloned code can contain statements like `return`, `break`, `continue`. In these instances, more conditions may apply to be able to conduct a refactoring, if beneficial at all.

We measured the amount of clones that can be refactored through method extraction (without additional transformations being required). Our results are displayed in Table 6.1. In this table we use the following categories:

- **Can be extracted:** This clone is a fragment of code that can directly be extracted to a method. Then, based on the relation between the clone instances, further refactoring techniques can be used to merge the extracted methods (for instance "pull up method" for clones in sibling classes).
- **Complex control flow:** This clone contains `break`, `continue` or `return` statements.
- **Spans part of a block:** This clone spans a part of a statement.
- **Is not a partial method:** If the clone does not fall in the "Partial method" category of Table 5.4, the "extract method" refactoring technique cannot be applied.

Table 6.1: Refactorability through method extraction

	# T1	% T1	# 1R	% 1R
Is not a partial method	5,917	34.22	4,806	38.49
Complex control flow	5,511	31.87	3,158	25.29
Spans part of a block	3,989	23.07	3,152	25.24
Can be extracted	1,874	10.84	1,371	10.98

From Table 6.1, we can see that approximately ten percent of the clones can directly be refactored through method extraction (and possibly other refactoring techniques based on the relation of the clone instances). For the other clones, other techniques or transformations will be required. Looking into these techniques and transformations will be one of our next steps.

Chapter 7

Discussion

In this chapter, we discuss the results of our experiment(s) on ...

Finding 1: Highlight like this an important finding of your analysis of the results.

Refer to Finding 1.

Chapter 8

Related work

We divide the related work into ... categories: ...

Chapter 9

Conclusion

In the research we have conducted so far we have made three novel contributions:

- We proposed a method with which we can detect clones that can/should be refactored.
- We mapped the context of clones in a large corpus of open source systems.
- We mapped the opportunities to perform method extraction on clones this corpus.

We have looked into existing definitions for different types of clones [**roy2007survey**] and proposed solutions for problems that these types have with regards to automated refactoring. We propose that fully qualified identifiers of method call signatures and type references should be considered instead of their plain text representation, to ensure refactorability. Furthermore, we propose that one should define thresholds for variability in variables, literals and method calls, in order to limit the number of parameters that the merged unit shall have.

The research that we have conducted so far analyzes the context of different kinds of clones and prioritizes their refactoring. Firstly, we looked at the inheritance relation of clone instances in a clone class. We have found that more than a third of all clone classes are flagged unrelated, which means that they have at least one instance that has no relation through inheritance with the other instances. For about a fourth of the clone classes all of its instances are in the same class. About a sixth of the clone classes have clone instances that are siblings of each other (share the same superclass).

Secondly, we looked at the location of clone instances. Most clone instances (58 percent) are found at method level. About 37 percent of clone instances were found at class level. We defined "class level clones" as clones that exceed the boundaries of a single method or contain something else in the class (like field declarations, other methods, etc.). Thirdly, we looked at the contents of clone instances. Most clones span a part of a method (57 percent). About 26 percent of clones span over several methods.

We also looked into the refactorability of clones that span a part of a method. Over 10 percent of the clones can directly be refactored by extracting them to a new method (and calling the method at all usages using their relation). The main reason that most clones that span a part of a method cannot directly be refactored by method extraction, is that they contain `return`, `break` or `continue` statements.

9.1 Threats to validity

We noticed that, when doing measurements on a corpus of this size, the thresholds that we use for the clone detection have a big impact on the results. There does not seem to be one golden set of thresholds, some thresholds work in some situations but fail in others. We have chosen thresholds that, according to our manual assessment, seemed optimal. However, by using these, we definitely miss some harmful clones.

9.2 Future work

Acknowledgements

Thanks to you, for reading this :)

Appendix A

Non-crucial information