

Evaluating Performance of Clone Detection Tools in Detecting Cloned Co-change Candidates

Md Nadim, Manishankar Mondal, Chanchal K. Roy

Department of Computer Science, University of Saskatchewan, Saskatoon, Canada

Abstract

Code reuse by copying and pasting from one place to another place in a codebase is a very common scenario in software development which is also one of the most typical reasons for introducing code clones. There is a huge availability of tools to detect such cloned fragments and a lot of studies have already been done for efficient clone detection. There are also several studies for evaluating those tools considering their clone detection effectiveness. Unfortunately, we find no study which compares different clone detection tools in the perspective of detecting cloned co-change candidates during software evolution. Detecting cloned co-change candidates is essential for clone tracking. In this study, we wanted to explore this dimension of code clone research. We used 12 different configurations of nine promising clone detection tools to identify cloned and non-cloned co-change candidates from eight C and Java-based subject systems and evaluated the performance of those clone detection tools in detecting the cloned co-change fragments. Our findings show that a good clone detector may not perform well in detecting cloned co-change candidates. The amount of unique lines covered by clone fragments, the number of detected clone fragments, source file processing mechanism, types of detected clones plays an important role in its performance in detecting cloned co-change candidates.

Keywords: Clone Detection, Cloned Co-change Candidates, Commit operation, Software Maintenance.

Email address: {mdn769, mshankar.mondal, chanchal.roy}@usask.ca (Md Nadim, Manishankar Mondal, Chanchal K. Roy)

1. Introduction

A large number of software tools have already been introduced for detecting cloned code fragments. Two surveys, that were done in 2009 by Roy et al. [1] and in 2013 by Rattan et al. [2] reported 75% increase in the number of clone
5 detection tools in these four years. Roy and Cordy [3] reported the existence of about 200 tools for detecting cloned code fragments. Although a large number of clone detection tools currently exist, we found no study for comparing the performance of different tools based on their ability to be used in software main-
10 tenance activity such as predicting cloned co-change candidates during software evolution. In this study, we wanted to explore, whether a good clone detector also performs well in detecting cloned co-change fragments?

One of the common features of clone detection tools is to combine similar code fragments into a clone group or class. The code fragments in a particular clone class are expected to perform similar functionalities. If we want to make
15 changes to a particular clone fragment in a clone class, the other fragments in the class are likely to have similar changes to ensure consistency of the code-base. Considering this assumption, we can say that all the clone fragments in a clone class have the possibility of being a cloned co-change candidate with any change of that class members. We utilize the clone classes provided by the clone
20 detectors for these types of co-change prediction.

Finding the co-change candidates of a target code fragment is also known as change impact analysis [4] in the literature. Mondal et al. [5] investigated the use of clone detection tool can enhance the performance of a tool for finding change impact set or co-change candidates. They performed their investigation
25 using Nicad for detecting both the regular and micro-clones and found that use of detected clone results significantly enhanced the performance of Tarmaq Rolfsnes et al. [6]. As they only analysed the use of Nicad, in this study we wanted to compare some other good clone detection tools to find whether some other tools may perform better for detecting co-change candidates. We have

30 evaluated four different configurations of CloneWorks Svajlenko and Roy [7]
and eight other clone detectors in our investigation. Therefore we have a total
of 12 separate implementations of clone detection tools (we will consider them
as 12 separate tools in the rest of this paper) on eight open-source software
systems. Configuration of clone detection tools are given in Table 3 and the
35 software systems used in this study are reported in Table 1.

During software evolution, a developer makes changes in the codebase to
fulfil some change requests. Those change requests could be related to each
other or independent [8, 9]. Therefore, all the changes done in a single commit
need not be related to each other. Some changes in a single commit may be
40 dependent on each other and some may be independent. The related code
fragments are known as the co-change candidates in literature [10]. Some of
those co-change candidates may contain similar code-fragments i.e. they are
clones of one another, on the other than, other types of co-change candidates
may not be cloned fragments but they have a functional dependency or coupling
45 with each other. If a developer makes changes to a target code fragment, those
changes might also need to be reflected other similar fragments in the codebase
to ensure consistent evolution of the software system [5, 11]. Failing to change
a co-change candidate of a target fragment can introduce bugs in the software
system [12, 13].

50 We have analyzed thousands of commit operations from the evolutionary
histories of eight subject systems listed in Table 1. While analyzing a commit
operation, we identify which code fragments changed together (i.e., co-changed)
in that commit. Considering each fragment as the target fragment, we try to
predict the other actually co-changed fragments using each of our clone detec-
55 tors. We found some change fragment which is not detected by any of the
clone detectors. We excluded those change fragments from consideration during
calculating the performance measures of clone detectors. An example of our
detection process is demonstrated in Fig. 1. Let us assume that 21 changes,
C1 to C21, occurred in the codebase of a subject system in a particular commit

operation. We detect these changes using the UNIX diff operation. If we consider C1 as the target change, the other 20 changes, C2 to C21, are the actual co-change candidates (i.e., co-changed candidates) of C1. We apply different clone detectors to detect these co-change candidates for the target change C1. Let using Deckard we can detect five change fragments (C2, C6, C8, C15, C21) from those 20 fragments, similarly using Nicad we can detect four fragments (C5, C10, C16, C18). We will continue to detect co-change fragments using all the other clone detectors. After getting the results from all the clone detectors, we find 10 unique change fragments (C2, C5, C8, C15, C21, C5, C8, C10, C15, C21) out of 20 fragments by taking a union of the results of all the clone detectors. We will take those 10 unique change fragments as cloned co-change candidates and calculate the precision and recall of each of the clone detectors based on their number of detection among those cloned co-change candidates. For each subject system, we finally calculated average recall, average precision, and F1 Score for each of the clone detector and then compare the clone detectors based on their weighted average F1 Score considering all the subject systems in this study. Figure 2 shows the bar chart of Average Recall and Average Precision drawn from our experimental results and in Table 5 we have given the calculated weighted average of F1 Score. According to our preliminary findings and ranking of the clone detectors (Table 7), we can conclude that Deckard outperforms all the other 11 tools. Other two configuration of CloneWorks, CLW(3Pattern)-splitting the code fragment considering lines, and CLW(3Token)-splitting the code fragments considering token working very close to Deckard. CCFinder, CLW(2Blind), ConQAT, and iClones are the name of other good clone detection tools whose F1 Score is at least 10% in our investigation. From the final rank list, we also see that the clone detection tools which detecting only Type-1 clones (such as Duplo, CLW Type-1) are performing worst in finding co-change candidates. We also calculated the average number of distinct lines detected as cloned lines by each of the clone detectors in all the revisions of all the subject systems (Figure 3) and found that the clone detector which detects more distinct lines as a cloned line in the codebase also performs well in detecting cloned

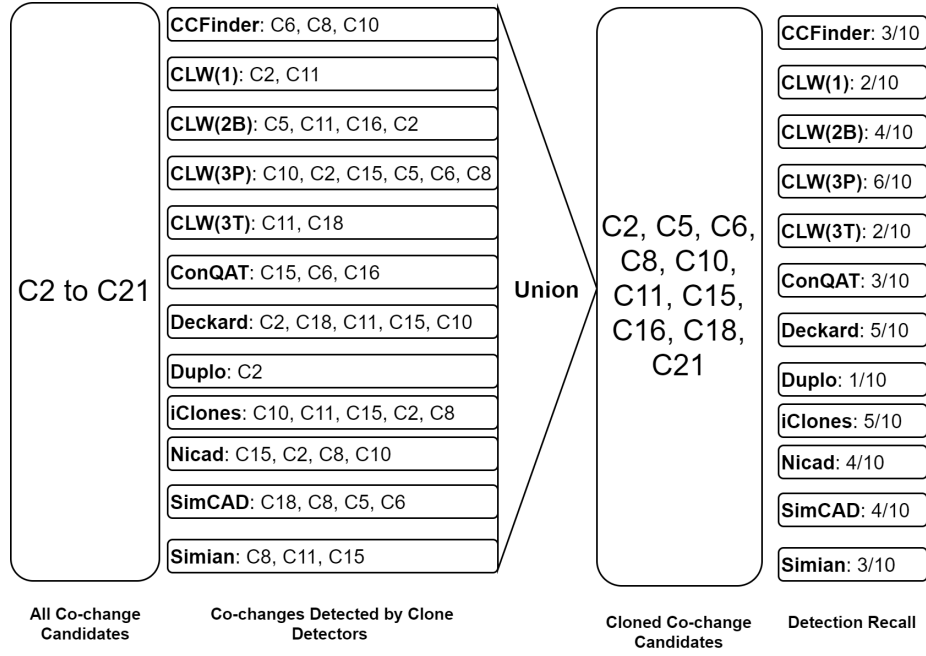


Figure 1: Demonstrating cloned co-change detection process

co-change candidates.

Based on this preliminary study, we tried to answer the following research questions:

RQ1: What is the comparison scenario of the clone detectors in predicting cloned co-change candidates?

RQ2: Why do different clone detectors perform differently in detecting cloned co-change candidates?

RQ3: Do clone detection tools source code processing technique (Pattern/Token/ Text based processing) have impact on the performance of detecting co-change candidates?

RQ4: Do clone detection tools designed for detecting different types of clones (Type 1, 2, 3) works differently in detecting cloned co-change candidates?

To the best of our knowledge, our study is the first one to compare clone de-

tection tools considering a particular maintenance perspective (e.g., considering
105 their capabilities in successfully suggesting cloned co-change candidates during
software evolution). From an initial assumption, it is obvious that a clone de-
tector which is good in detecting clone should also be good in detecting cloned
co-change candidates. In this exploratory study, we wanted to practically verify
this assumption. We selected 12 implementations of clone detectors detecting
110 different types of clones in our investigation to verify whether they are also good
at detecting co-change candidates.

We organized this paper in the following sections: Some related works are
described in Section 2, our methodology is in Section 3, we described the ex-
perimental result in Section 4, the discussion is in section 5, Section 6 explains
115 some possible threats to validity, and we conclude our paper by mentioning
future work in Section 7.

2. Related Work

There are several studies [14, 1, 15, 16] that have been focused on ranking
different clone detection tools based on their performance in detecting differ-
120 ent types of clone fragments and accuracy of those detection tools. Burd and
Bailey [17] did a study for comparing the performance of three clones and two
plagiarism detecting tools based on their precision and recall of the ability to
detect duplicated codes in a single file or across different files. Bellon et al. [15]
evaluated six clone detection tools based on eight large C and Java programs of
125 almost 850 KLOC and made a framework for comparing different clone detec-
tion tools with the data validated by one of the authors of this. Rysselberghe
and Demeyer [18] evaluated three representative clone detection techniques from
a refactoring perspective where they provided comparative results in terms of
portability, kinds of clone reported, scalability, number of false positive, and
130 number of useless clone detection. Svajlenko and Roy [14] evaluated eleven
modern clone detection tools using four benchmark frameworks and noted Con-
QAT, iClones, NiCad and SimCAD as very good tools for detecting clones of

all the three types (Type-1, Type-2, Type-3). Roy et al. [1] did a qualitative comparison and evaluation of the latest clone detection approaches and tools, and made a benchmark called BigCloneBench [3] which contains eight million manually validated clone pairs in a large inter-project source dataset of more than 25,000 projects and 365 million lines of code. They categorize, relate and assess different clone detection tools based on two different points of view such as classification based on the overlapping set of attributes in the different code fragments and the scenarios how Type-1, Type-2, Type-3, and Type-4 clones created. They also elaborated the procedure of using the result of their study to select the most suitable clone detection tool or technique in the context of a specific set of areas and limitations.

There are some studies which not only proposed a clone detection mechanism but also did a comparison of their proposed technique with some existing techniques. Koschke et al. [19] provided a technique to detect clone using suffix trees in abstract syntax trees and they also made a comparison to other techniques using the Bellon benchmark for clone detectors. Ducasse et al. [20] and Selim et al. [21] also utilized Bellon’s framework for measuring the performance of their proposed clone detection tools based on string comparison and intermediate source transformation respectively. Selim et al. [21] showed that their tool is capable of detecting Type-3 clones and their technique is better than the source-based clone detectors based on the value of recall through a slight drop in the precision using Bellon’s corpus where clone group is not complete. Compared to the standalone string and token-based clone detectors, their technique showed a little higher precision.

All the studies which compared different clone detectors have been focused on the precision, recall, computational complexity, and memory used or detecting a specific type of clone fragments such as Type-1, Type-2, Type-3, or Type-4 during the detection approach of duplicated code in a codebase. Our study to compare clone detectors is completely different from the previous comparisons. We do not want to compare clone detection tools based on the capability to detect clones. Our point of interest is to detect co-change candidates during the

software commit operations. Mondal et al. [10] did a study to predict and rank
165 the co-change candidates by analyzing evolutionary coupling from previously
done change history using generated clone fragments by NiCad but they did
not consider the result of other clone detection tools and also did not show any
comparative study among different clone detection tools in doing such prediction
and rank of co-change candidates. We found no study which compared different
170 clone detectors in this perspective of software maintenance. In this research,
we have analyzed the performance of nine clone detection tools in 12 different
configurations based on their capabilities in finding co-change candidates during
software evolution using their generated clone results. We have taken four clone
detection tools (ConQAT, iClones, NiCad, and SimCAD) suggested as good
175 tools in the study of Svajlenko and Roy [14] and two other tools, one of them is
text similarity-based (Simian) and the other is tree similarity-based (Deckard)
for evaluating their performance in our study. According to our knowledge, this
is the first such investigation of performance with clone detection tools.

3. Methodology

180 We have used eight subject systems listed in Table 1 and 12 clone detec-
tion tools (Table 3) for our analysis. Our analysis aims to rank these clone
detection tools based on their performance in successfully suggesting actual co-
change candidates (ACC) during the software evolution. Before starting our
main analysis, we have to resolve some issues and we have taken the following
185 considerations in this regard.

Selection of subject systems: To select subject systems for this study, we
considered both the popularity of programming language and availability of a
considerable amount of revisions. According to the TIOBE Programming Com-
munity index [22] (an indicator of the popularity of programming languages),
190 Java is dominating the list of popular programming languages for more than the
last ten years and C is the second most popular programming language within
this period. Considering this fact, we wanted to select subject systems written

in these two programming languages. Our other consideration was the availability of a considerable amount of revisions of each of the systems. Based on
195 both of the considerations, we have chosen the subject systems listed in Table 1.

Table 1: SUBJECT SYSTEMS

Systems	Language	Domains	Revisions
Brlcad	C	Computer Aided Design	2115
Camellia	C	Batch Job Server	301
Carol	Java	Game	1700
Ctags	C	Code Def. Generator	774
Freecol	Java	Game	1950
Jabref	Java	Reference Manager	1545
jEdit	Java	Text Editor	4000
Qmailadmin (QMA)	C	Mail System Manager	317

Selection of clone detectors: In this research, we wanted to examine those clone detection tools which are good in detecting all types of clones. To select such tools, we considered some related studies. We have taken CloneWorks
200 [7] as it is considered as a fast and flexible clone detector for large-scale near-miss clone detection experiments. CloneWorks tool provides the ability to change its processing mechanism by changing its configuration files. We applied four different configurations of CloneWorks to detect Type-3 Pattern, Type-3 To-

Table 2: SUMMARY OF DATA PROCESSED

Revisions/ Subject Systems	Brlcad	Camellia	Carol	Ctags	Freecol	Jabref	jEdit	QMA
Processed	2113	301	1700	774	1001	1540	215	317
Experiencing change	660	163	454	447	836	860	145	35
Experiencing more than one change	553	155	430	330	833	755	145	25

ken, Type-2 Blind, and Type-1 clones for investigating the impact of the types
 205 of clones in detecting co-change candidates. We included Duplo [23] as another
 type-1 clone detector for making the comparison with type-1 clones of
 CloneWorks in this investigation. ConQAT [24], iClones [25], NiCad [26], and
 SimCAD [27] have been reported as very good tools for detecting all type of
 clones in the study of Svajlenko and Roy [14]. Besides these, CCFinder [28],
 210 Deckard [29], iClones and NiCad are often considered as common examples of
 modern clone detectors that support Type-3 clone detection. The reason of
 taking Simian [30] in our analysis was its ability to find duplicated code by
 line-by-line textual comparison supporting identifier renaming with a fast de-
 tection speed on the large repository and extensive use in several clone studies
 215 [31, 32, 33, 34, 35]. NiCad, SimCAD, and Simian are textual similarity-based
 clone detection tools. Deckard works using tree comparison technique, on the
 other hand, CCFinder, ConQAT and iClones are token-based clone detection
 tools.

Determining if the extracted co-changes are related to each other
 220 **or not:** Even though we have extracted all the changes between two adjacent
 revisions (i.e., revision n and $n+1$), it is not possible to fully guarantee that all
 the changes are actually co-change candidates of each other. There might be
 some changes which do not depend on any other changes i.e. they may change
 independently. The inclusion of such dissimilar changes into our calculation
 225 can drop the detection accuracy of clone detectors. To minimize such drops,
 we excluded those co-changes which are not detected by any of the 12 clone
 detection techniques in our study. As none of the clone detectors considers
 them as co-change candidates, we considered those changes as dissimilar or
 independent changes.

Ensuring if the configuration parameters of all the clone detection
 230 **tools identical with each other or not:** As we wanted to compare different
 clone detectors based on their capability of successfully suggesting co-change
 candidates, it was important to configure them identically during detecting

clones from our subject systems. Wang et al. [32] introduced confounding configuration choice problem where the configuration of different tools during clone
235 detection may play a vital role and the result may be best or worst depending on the configuration. Our configuration of different tools is shown in Table 3. We have used similar configurations for each of the tools for obtaining a consistent result. We have taken configuration values similar to Svajlenko and Roy
240 [14] which they conducted to compare different clone detectors based on their efficiency in detecting cloned fragments. We provided 70% similarity threshold for all the clone detection tools (except Deckard) which takes similarity dissimilarity value as parameter. We have used 85% as the similarity threshold for Deckard 85% because we found a lot of unwanted clones in the result if we use
245 the similarity threshold 70%. These results include a lot of duplicated clone fragments and showing a lot of fragments as a clone to itself several times. We also tried some other percentage values such as 75%, and 80% but the detected result of Deckard becomes much desirable when we set it to 85%. Svajlenko and Roy [14] also used 85% similarity while running Deckard for Mutation Framework.
250 We have also selected identical parameter values such as the minimum number of tokens, minimum number of lines for different clone detection tools. As we wanted to compare different clone detectors based on their capability of successfully suggesting co-change candidates, it was very important to configure them identically during detecting clones from our subject systems.

The overall approach: Our overall processing is performed in some distinct
255 steps. Initially, we downloaded all the source files of all the revisions of all the subject systems from their respective SVN repositories. We then applied **diff** operation between each file of a revision with the respective file in the next revision and extracted the change information such as Name of the File
260 which is changed, the Line where the respective change begins, the Line where the change is ended from the output of **diff**. We did the change extraction for each of the revision (excluding the last one) of all the subject systems. After detecting all the changes, we started the clone detection on all the revisions of all the subject systems using all the clone detection tools. We started our main

265 analysis to find the accuracy of each of the clone detection tools after having the
result of all the clone detectors and change information from all the revisions.

The mechanism of calculating accuracy is demonstrated in our introduction
using Fig. 1. Suppose, we are examining a particular commit operation. The
number of fragments that were changed in this commit operation is n . Now,
270 let us consider one of these n fragments as the target fragment. Then the other
 $n - 1$ fragments are the actually co-changed candidates for the target fragment.
We excluded the non-cloned co-change candidates using the approach described
in the introduction. After this exclusion, we get the **Actually Cloned Co-
change** (ACC) for each of the target fragments.

275 Let us assume that the target change fragment intersects a particular clone
fragment from a particular clone class. The other fragments in that clone class
are considered as the **Predicted Cloned Co-change** (PCC) candidates. We
now determine how many of these PCC intersect with the ACC to obtain the
number of detected cloned co-change candidates by the clone detector.

280 These counts of predicted and actually co-changed candidates are considered
as the **true positives** to calculate Recall, Precision, and F1 Score. We calculate
these using the following equations (Eq. 1, 2, and 3).

$$Recall = \frac{|PCC \cap ACC|}{|ACC|} \quad (1)$$

$$Precision = \frac{|PCC \cap ACC|}{|PCC|} \quad (2)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

We repeat the calculating process of Recall and Precision for all the changes
in each of the subjects systems with the detected clone fragments generated by
285 all the clone detection tools. We then calculate F1 Score of the clone detectors
for each of the subject systems by taking the average values of Recall and
Precision which is reported in Table 5. We reported both the ranking of the
tools in each of the subject systems in Table 6 and overall ranking considering all

Table 3: CONFIGURATION OF PARTICIPATING CLONE DETECTION TOOLS

Tools	Configuration for Clone Detection
CCFinder	min. size: 50 tokens, min. token types: 12
CLW(T1)	termsplit=token, termproc=Joiner
CLW(T2Blind)	cfproc=rename-blind, cfproc=abstract literal, termsplit=token, termproc=Joiner
CLW(T3Pattern)	cfproc=rename-blind, cfproc=abstract literal, termsplit=line
CLW(T3Token)	termsplit=token, termproc=FilterOperators, termproc=FilterSeperators
ConQAT	block clones, clone min-length=5, gap ratio=0.3
Deckard	min. size: 30 tokens, 5 token stride, min. 85% similarity
Duplo	min. size: 10 lines, min. characters/line:1
iClones	minimum block: 30, minimum clone: 50, All Transformation
Nicad	block clones, blind renaming, max. threshold=0.3, minimum lines=5, maximum lines=2500
SimCAD	block clones, Source Transformation= generous
Simian	min. size: 5 lines, normalize literals/identifiers

CLW: *Clone Works*, **T1/2/3:** *Type-1/2/3*

the subject systems in Table 7. To calculate the overall ranking of the tools we
290 took weighted average (Total number of Changes is the corresponding weighting
factor in each subject system) of the performance measures (Precision, Recall,
F1 Score) in each subject systems.

4. Experimental Result

In this section, we will answer the research questions based on our overall
295 analysis and obtained results by the processing of each of the eight subject

Table 4: SUMMARY OF ACTUAL TARGET AND CO-CHANGE CANDIDATES

SS	# ATC	# ACC	% ATC	% ACC
Brlcad	2909	33578	7.45	1.89
Camellia	8052	346140	20.61	19.46
Carol	4582	254311	11.73	14.29
Ctags	718	3648	1.84	0.21
Freecol	6865	265213	17.57	14.91
Jabref	8313	455469	21.28	25.60
jEdit	5122	323277	13.11	18.17
QMA	2508	97396	6.42	5.47
Total	39069	1779032	100	100

* *SS: Subject Systems*

* *# ATC: Number of Actual Target Changes*

* *# ACC: Number of Actual Co-changes*

systems using all the 12 clone detection tool executions.

4.1. Answer to the *RQ 1*

What is the comparison scenario of the clone detectors in predicting cloned co-change candidates?

300

The key experimental results are in Figure 2, Table 4, Table 5, Table 6, and Table 7 where Fig. 2 shows the average Recall and average Precision of each of the clone detection tools. Table 4 shows the summary of target change candidates and detected co-change candidates for those target changes in each of the subject systems. We found the highest and lowest percentage of target change and its cloned co-change candidates from Jabref and Ctags respectively. Table 5 shows the F1 Score of each of the clone detectors in each of the subject systems. The F1 Score is calculated using Equation (3). Our experimental results concluded in Table 6 which shows that CLW(T3Pattern), CLW(T3Token), and Deckard shows top performance (Rank 1 or 2) in most of

310

the subject systems compared to all the other tools. The summary of the results in the Table 6 shows that among the subject systems, CLW(T3Pattern) is the best in all the subject systems except Camellia and Freecol where Deckard is showing the best performance. CLW(T3Token) shows the second-best performance in most of the subject systems. In an overall observation on individual rankings of different clone detection techniques reveals that CLW(3Pattern), Deckard, CLW(3Token), CCFinder show better performance in most of the subject systems compared to the other clone detectors. On the other hand, Duplo, CLW(Type1) shows worst performance in most of the subject systems. Other tools shows average performance considering individual ranking in different subject systems. We also calculated overall ranking considering weighted average of F1 Scores in individual subject systems which is reported in Table 7. Though, in individual ranking in Table 6 CLW(3Pattern) showed best performance compared to Deckard, weighted average of Deckard become higher in Table 7 as Deckard obtained much better F1 Score in Camellia than CLW(3Pattern). Camellia has very high number of target change candidates (#ATC in Table 4) which makes weighted average of F1 Score higher. CLW(3Pattern) obtained second-highest position in the rank list by obtaining very close weighed average F1 Score to Deckard. CLW(Type1) and Duplo obtained the bottom position in the final rank list.

As our analysis was based on the clone grouping into class or pair provided by the clone detection tools, we found that the efficiency of clone detection tools in suggesting cloned co-change candidates is mostly dependent on its effectiveness in making clone class/ pair. The tool which groups functionally similar clone fragments into a clone class/ pair effectively can perform well in successfully suggesting cloned co-change candidate(s). Different values of the accuracy of different clone detectors indicate the difference in their efficiency in this research domain.

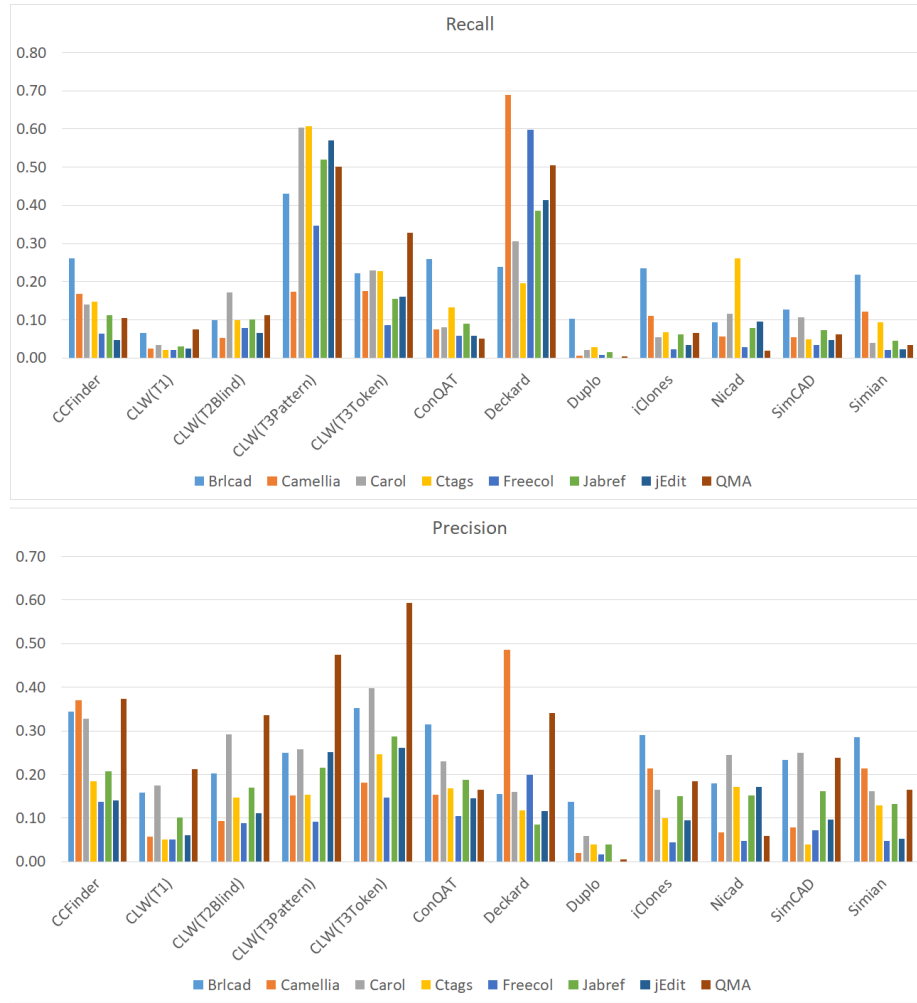


Figure 2: Average recall of different tools

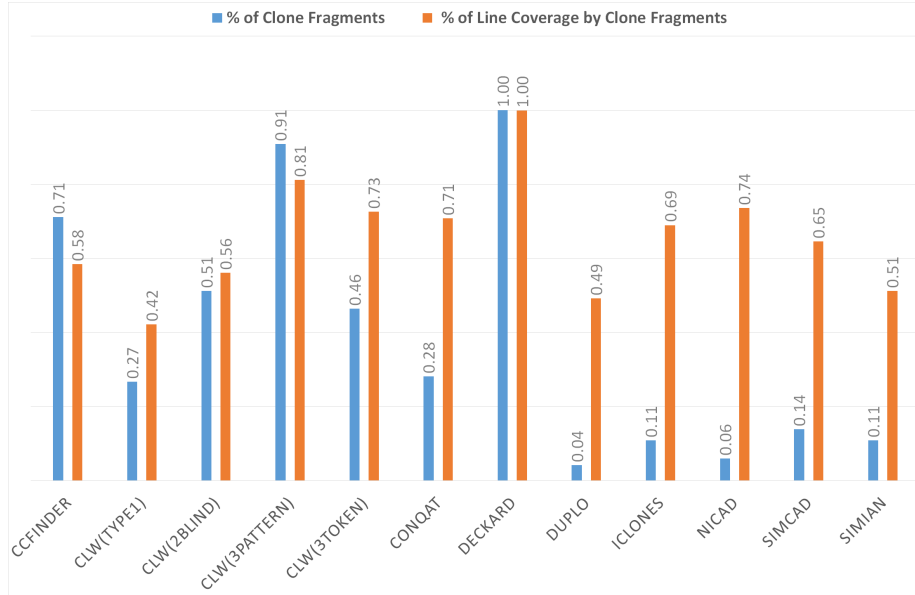


Figure 3: Comparing unique line coverage by clone fragments and number of clone fragments from different clone detectors.

4.2. Answer to the **RQ 2**

340 **Why do different clone detectors perform differently in detecting cloned co-change candidates?**

From the answer of our **RQ 1**, we found a difference in performance for different clone detection tools in suggesting cloned co-change candidates. We
 345 found a clone detection tool which is good in detecting clone fragments may not be good at detecting cloned co-change candidates. This motivates us to find out the reason to answer this research question.

We investigated the number of clone fragments and the number of unique lines covered by those clone fragments by all the 12 clone detectors from all the
 350 revisions of all the subject systems. Figure 3 shows the comparison scenario of Number of Clone Fragments and Line Covered by those Clone Fragments from different clone detectors. For better comparison, we bring the values in a single scale (between 0 and 1) where 0 and 1 represents the lowest and highest values

respectively compared to all the clone detectors under comparison. Considering
 355 both, the the number of clone fragments and the number of lines covered by
 those clone fragments from all the revisions of all the subject systems, if we
 order the clone detectors from the highest to the lowest, we find Deckard and
 CLW(3Pattern) in the top of the list. CLW(3Token) and CCFinder fall in the
 respective next position in providing highest number of clone fragments and
 360 covering highest number of unique lines in the source files. This scenario shows
 that, a good clone detector can perform bad in detecting cloned co-change can-
 didates if it does not detect enough clone fragments and does not cover enough
 unique lines by those clone fragments in the source file. Though, earlier study
 [10] suggests that NiCad is a very good clone detector, in both of these cases, it
 365 falls at the bottom of the list. Despite, NiCad performs very well in detecting
 clone fragments, it provides a lower number of clone fragment and also the lower
 number of line coverage by those clone fragments in the software systems. For
 that reason, while detecting the cloned co-change candidates, NiCad is show-
 ing lower F1 Score. The number of clone fragments and line coverage by those
 370 fragments seems to be an underlying factor behind the obtained comparison sce-
 nario of the clone detectors in predicting cloned co-change candidates, there can
 be several other factors such as overlapping of code clones and code similarity
 detection mechanism. We plan to investigate these factors in future.

4.3. Answer to the **RQ 3**

375 **Do clone detection tools source code processing technique (Pat-
 tern/ Token/ Text based processing) have impact on the performance
 of detecting co-change candidates?**

We can answer this research question by analysing our final ranking of the
 clone detector in Table 7. Top two clone detectors (Rank 1 and 2) works by
 380 extracting source code patterns from the source code. Deckard first generates
 vectors from the source file and then extracts tree like source pattern to match
 similarity among different source code fragments. CLW(3Pattern) also processes
 the source code terms by splitting in lines and then extracts code patterns.

The other five tools (Rank 3 to 7) in the rank list perform token-based source
code processing and the remaining five tools perform text based source code
processing for detecting clones from the source file. From this result we can
say that, text based clone detection tools are not good to be used in detecting
cloned co-change candidates during software evolution. The tools which can
detect more generalized clone fragments specially pattern based clone detectors
are very good for detecting co-change candidates.

4.4. Answer to the **RQ 4**

Do clone detection tools designed for detecting different types of clones (Type 1, 2, 3) works differently in detecting cloned co-change candidates?

From the final rank list of our clone detectors we also find the relation
of detected clone types with its ability to detect clone co-change candidates.
Rank list of clone detectors in Table 7 shows that, clone detecting tools such
as CLW(Type1), Duplo, which detects only Type 1 clone will not perform
good in detecting co-chagne candidates. On the other hand, tools such as
Deckard, CLW(3Pattern), CLW(3Token), CCFinder perform very good in de-
tecting cloned co-change candidates.

5. Discussion

There are two primary perspectives of managing code clones: (1) clone track-
ing and (2) clone refactoring. Our research essentially focuses on the clone
tracking perspective. The main task of a clone tracker is to suggest similar co-
change candidates when a programmer attempts to change a code fragment. For
suggesting co-change candidates, a clone tracker depends on a clone detector.
Our research compares 12 promising clone detectors based on their capabili-
ties in suggesting cloned co-change candidates. According to our investigation,
Deckard, CloneWorks (Type-3), and CCFinder are the most promising tools for
suggesting such co-change candidates. CloneWorks (Type-2) and ConQAT are

Table 5: F1 SCORE OF DIFFERENT TOOLS IN DETECTING CLONED CO-CHANGE

Tools/SS	Brlcad	Camellia	Carol	Ctags	Freecol	Jabref	jEdit	QMA
CCFinder	0.30	0.23	0.20	0.16	0.09	0.15	0.07	0.16
CLW(T1)	0.09	0.04	0.06	0.03	0.03	0.05	0.04	0.11
CLW(T2Blind)	0.13	0.07	0.22	0.12	0.08	0.13	0.08	0.17
CLW(T3Pattern)	0.32	0.16	0.36	0.25	0.15	0.30	0.35	0.49
CLW(T3Token)	0.27	0.18	0.29	0.24	0.11	0.20	0.20	0.42
ConQAT	0.28	0.10	0.12	0.15	0.08	0.12	0.08	0.08
Deckard	0.19	0.57	0.21	0.15	0.30	0.14	0.18	0.41
Duplo	0.12	0.01	0.03	0.03	0.01	0.02	0.00	0.00
iClones	0.26	0.15	0.08	0.08	0.03	0.09	0.05	0.10
Nicad	0.12	0.06	0.16	0.21	0.04	0.10	0.12	0.03
SimCAD	0.17	0.07	0.15	0.04	0.05	0.10	0.06	0.10
Simian	0.25	0.16	0.06	0.11	0.03	0.07	0.03	0.06

Table 6: RANKS OF CLONE DETECTORS BY F1 SCORE IN EACH OF THE SUBJECT SYSTEMS

Tools/SS	Brlcad	Camellia	Carol	Ctags	Freecol	Jabref	jEdit	QMA
CCFinder	2	2	5	4	4	3	7	5
CLW(T1)	12	11	10	11	9	11	10	6
CLW(T2Blind)	9	8	3	7	5	5	5	4
CLW(T3Pattern)	1	4	1	1	2	1	1	1
CLW(T3Token)	4	3	2	2	3	2	2	2
ConQAT	3	7	8	5	6	6	6	9
Deckard	7	1	4	6	1	4	3	3
Duplo	10	12	12	12	12	12	12	12
iClones	11	10	6	3	8	7	4	11
Nicad	8	9	7	10	7	8	8	8
SimCAD	6	5	11	8	11	10	11	10
Simian	5	6	9	9	10	9	9	7

* Tools are listed in alphabetic order in the left-most column.

* The numbers under each subject system represent the ranks of the tools for that system.

Table 7: FINAL RANK OF CLONE DETECTORS CONSIDERING ALL THE SUBJECT SYSTEMS

Clone Detectors	WA of Precision (p)	WA of Recall (r)	F1 Score $2pr/(p+r)$	Final Rank
Deckard	0.22	0.47	0.30	1
CLW(3Pattern)	0.21	0.43	0.28	2
CLW(3Token)	0.27	0.17	0.21	3
CCFinder	0.25	0.12	0.16	4
CLW(2Blind)	0.16	0.09	0.12	5
ConQAT	0.17	0.09	0.12	6
iClones	0.15	0.07	0.10	7
Nicad	0.13	0.07	0.09	8
Simian	0.14	0.07	0.09	9
SimCAD	0.14	0.07	0.09	10
CLW(Type1)	0.10	0.03	0.05	11
Duplo	0.03	0.02	0.02	12

also very good options according to our final ranking demonstrated in Table 7. Based on our overall observation, we can say that the performance of Deckard and CloneWorks (Type-3 Pattern) are much better compared to the other clone
415 detection tools in detecting co-change candidates during software evolution. As the clone classes/ pairs generated by different clone detectors played an important role in our analysis, we can say that the clone detectors which can group similar clone fragments into a class efficiently will perform better in detecting co-change candidates during the commit operation. From our findings we can
420 also say that the clone detectors which detect all the Type 1, 2, 3 clones are also perform good in detecting co-change candidates.

When a particular code fragment is changed, we apply the clone detectors to predict which other similar code fragments might also need to be co-changed. However, some dissimilar fragments might also be changed together with the
425 particular fragment. As we are applying only clone detectors, we cannot consider

those dissimilar co-change candidates in our research.

In our research, we do not compare the clone detectors considering their clone detection efficiency. We rather compare the clone detection tools based on their ability in suggesting cloned co-change candidates. Such a comparison of clone detectors focusing on a particular maintenance perspective was not done previously. Suggesting co-change candidates for a target program entity is an important impact analysis [4] task during software evolution. Thus, through our research, we investigate which of the clone detectors can be useful in change impact analysis to what extent. Findings from our research can identify which clone detector(s) can be promising for change impact analysis.

6. Threats to Validity

We have investigated eight subject systems in our study. While more subject systems could generalize our findings, we selected our systems focusing on their diversity, popularity of used programming language, and availability of a considerable number of revisions. For example, our systems are of different application domains, sizes, and revision history lengths. Thus, our findings are not biased by our choice of subject systems. We believe that our findings are important from the perspectives of software maintenance.

We have investigated 12 different configuration of nine clone detectors in our study. Detection parameter settings of the clone detectors can have an impact on their comparison. However, the parameters of different clone detectors were selected considering their equivalence. Thus, we believe that we have a fair comparison among the clone detectors.

Several code fragments might change together in a commit operation. While some of these fragments can be similar to one another, and some might be dissimilar. Similar code fragments co-change (i.e., change together) for ensuring consistency of the codebase. However, dissimilar code fragments can co-change because of their underlying dependencies which could have some impact on the generalization of this research outcome. As we aim to compare the clone

455 detection tools, we wanted to discard the dissimilar co-change candidates from our consideration. If a co-change candidate was not detected as a true positive by any of the clone detectors, we discarded the candidate. We believe that such a consideration is reasonable in our experiment aiming towards comparing clone detectors and our findings may inspire more similar research.

460 **7. Conclusion and Future Works**

In this research, we make a comparison among different clone detection tools from the perspective of software maintenance. In particular, we investigate their performances in successfully suggesting (i.e., predicting) cloned co-change candidates during evolution. We used eight open source subject systems written in C and Java for our analysis. According to our findings (Table 6 & 7) on thousands of revisions of these systems, Deckard and CloneWorks (Type-3) show the most promising results in most of the eight subject systems compared to the other tools. CCFinder and CloneWorks (Type-2) are also shows better performance in some of the subject systems. Although we have figured some reasons of the better performance of Deckard, CloneWorks, and CCFinder in the Discussion section of our study, we planned to extend this research by analyzing the clone detection mechanism of the clone detectors to find out some other reasons for their performance. We also want to investigate the impact of different similarity score of different clone detectors in finding co-change candidates in our future studies. Besides this, we want to include some other clone detection tools of different detection mechanism (i.e., tree/ token/ text-based) and subject systems written in some different programming languages (i.e. C/ C++, C#, Python) for extending our research.

Acknowledgment

480 This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and by a Canada First Research Excellence Fund (CFREF) grant coordinated by the Global Institute for Food Security (GIFS).

References

- [1] C. K. Roy, J. R. Cordy, R. Koschke, Comparison and evaluation of code
485 clone detection techniques and tools: A qualitative approach, *SCIENCE
OF COMPUTER PROGRAMMING* (2009).
- [2] D. Rattan, R. K. Bhatia, M. Singh, Software clone detection: A systematic
review., *Information Software Technology* 55 (2013) 1165–1199.
- [3] C. K. Roy, J. R. Cordy, Benchmarks for software clone detection: A ten-
490 year retrospective, in: *Proc. SANER*, 2018, pp. 26–37. doi:10.1109/SANER.
2018.8330194.
- [4] R. S. Arnold, *Software Change Impact Analysis*, IEEE Computer Society
Press, Los Alamitos, CA, USA, 1996.
- [5] M. Mondal, B. Roy, C. K. Roy, K. A. Schneider, Associating code clones
495 with association rules for change impact analysis, in: *Proc. SANER*, 2020,
p. 11pp.
- [6] T. Rølfesnes, S. Di Alesio, R. Behjati, L. Moonen, D. W. Binkley, Gen-
eralizing the analysis of evolutionary coupling for software change impact
analysis, in: *2016 IEEE 23rd International Conference on Software Analy-
500 sis, Evolution, and Reengineering (SANER)*, volume 1, 2016, pp. 201–212.
- [7] J. Svajlenko, C. K. Roy, Cloneworks: A fast and flexible large-scale near-
miss clone detection tool, in: *2017 IEEE/ACM 39th International Confer-
ence on Software Engineering Companion (ICSE-C)*, 2017, pp. 177–179.
- [8] M. Mondal, C. K. Roy, K. A. Schneider, An empirical study on change
505 recommendation, in: *Proc. CASCON, CASCON '15*, IBM Corp., Riverton,
NJ, USA, 2015, pp. 141–150.
- [9] M. Mondal, C. Roy, K. Schneider, Connectivity of co-changed method
groups: a case study on open source systems, 2012, pp. 205–219.

- [10] M. Mondal, C. K. Roy, K. A. Schneider, Prediction and ranking of co-change candidates for clones, in: Proc. MSR 2014, ACM, New York, NY, USA, 2014, pp. 32–41. doi:10.1145/2597073.2597104.
- [11] M. Mondal, B. Roy, C. K. Roy, K. A. Schneider, Investigating context adaptation bugs in code clones, in: Proc. ICSME, 2019, pp. 157–168. doi:10.1109/ICSME.2019.00026.
- [12] J. F. Islam, M. Mondal, C. K. Roy, K. A. Schneider, Comparing bug replication in regular and micro code clones, in: Proc. ICPC, 2019, pp. 81–92. doi:10.1109/ICPC.2019.00022.
- [13] J. F. Islam, M. Mondal, C. K. Roy, A comparative study of software bugs in micro-clones and regular code clones, in: Proc. SANER, 2019, pp. 73–83.
- [14] J. Svajlenko, C. K. Roy, Evaluating modern clone detection tools, in: Proc. ICSME, 2014, pp. 321–330. doi:10.1109/ICSME.2014.54.
- [15] S. Bellon, R. Koschke, G. Antoniol, J. Krinke, E. Merlo, Comparison and evaluation of clone detection tools, IEEE Transactions on Software Engineering 33 (2007) 577–591. doi:10.1109/TSE.2007.70725.
- [16] C. Roy, J. Cordy, Scenario-based comparison of clone detection techniques, 2008, pp. 153–162. doi:10.1109/ICPC.2008.42.
- [17] E. Burd, J. Bailey, Evaluating clone detection tools for use during preventative maintenance, in: Proc. SCAM, 2002, pp. 36–43. doi:10.1109/SCAM.2002.1134103.
- [18] F. V. Rysselberghe, S. Demeyer, Evaluating clone detection techniques from a refactoring perspective, in: Proc. ASE, 2004, pp. 336–339.
- [19] R. Koschke, R. Falke, P. Frenzel, Clone detection using abstract syntax suffix trees, in: 2006 13th Working Conference on Reverse Engineering, 2006, pp. 253–262. doi:10.1109/WCRE.2006.18.

- 535 [20] S. Ducasse, O. Nierstrasz, M. Rieger, On the effectiveness of clone detection
by string matching: Research articles, *J. Softw. Maint. Evol.* 18 (2006) 37–
58. doi:10.1002/smr.v18:1.
- [21] G. M. K. Selim, K. C. Foo, Y. Zou, Enhancing source-based clone detection
using intermediate representation, in: 2010 17th Working Conference on
540 Reverse Engineering, 2010, pp. 227–236.
- [22] T. Software, Tiobe index — tiobe - the software quality company, 2020
(accessed July 6, 2020). URL: <https://www.tiobe.com/tiobe-index/>.
- [23] S. Ducasse, M. Rieger, S. Demeyer, A language independent approach
for detecting duplicated code, in: Proceedings IEEE International Confer-
545 ence on Software Maintenance - 1999 (ICSM'99). 'Software Maintenance
for Business Change' (Cat. No.99CB36360), 1999, pp. 109–118.
- [24] E. Juergens, F. Deissenboeck, B. Hummel, Clonedetective - a workbench
for clone detection research, in: Proc. ICSE, 2009, pp. 603–606. doi:10.
1109/ICSE.2009.5070566.
- 550 [25] N. Göde, R. Koschke, Incremental clone detection, in: Proc. CSMR, 2009,
pp. 219–228. doi:10.1109/CSMR.2009.20.
- [26] J. R. Cordy, C. K. Roy, The nicad clone detector, in: Proc. ICPC, 2011,
pp. 219–220. doi:10.1109/ICPC.2011.26.
- [27] M. S. Uddin, C. K. Roy, K. A. Schneider, Simcad: An extensible and faster
555 clone detection tool for large scale software systems, in: Proc. ICPC, 2013,
pp. 236–238. doi:10.1109/ICPC.2013.6613857.
- [28] T. Kamiya, S. Kusumoto, K. Inoue, Ccfinder: a multilinguistic token-based
code clone detection system for large scale source code, *IEEE Transactions
on Software Engineering* 28 (2002) 654–670.
- 560 [29] L. Jiang, G. Mishherghi, Z. Su, S. Glondu, Deckard: Scalable and accurate
tree-based detection of code clones, in: Proc. ICSE, 2007, pp. 96–105.
doi:10.1109/ICSE.2007.30.

- [30] S. Harris, Simian - Similarity Analyser — Duplicate Code Detection for the Enterprise—Overview, 2003 (accessed July 6, 2020). URL: <http://www.harukizaemon.com/simian/>.
565
- [31] C. Ragkhitwetsagul, J. Krinke, D. Clark, Similarity of source code in the presence of pervasive modifications, in: Proc. SCAM, 2016, pp. 117–126. doi:10.1109/SCAM.2016.13.
- [32] T. Wang, M. Harman, Y. Jia, J. Krinke, Searching for better configurations: A rigorous approach to clone evaluation, in: Proc. ESEC/FSE, ACM, New York, NY, USA, 2013, pp. 455–465.
570
- [33] M. Mondal, M. S. Rahman, R. K. Saha, C. K. Roy, J. Krinke, K. A. Schneider, An empirical study of the impacts of clones in software maintenance, in: Proc. ICPC, 2011, pp. 242–245. doi:10.1109/ICPC.2011.14.
- [34] W. T. Cheung, S. Ryu, S. Kim, Development nature matters: An empirical study of code clones in javascript applications, Empirical Softw. Engg. 21 (2016) 517–564.
575
- [35] J. Krinke, N. Gold, Y. Jia, D. Binkley, Cloning and copying between gnome projects, in: Proc. MSR, 2010, pp. 98–101. doi:10.1109/MSR.2010.5463290.
580