JOURNAL OF SOFTWARE: EVOLUTION AND PROCESS
J. Softw. Evol. and Proc. 0000; **00**:1–16
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/smr

There and Back Again: Can you Compile that Snapshot?

Michele Tufano[1], Fabio Palomba[2], Gabriele Bavota[3], Massimiliano Di Penta[4]
Rocco Oliveto[5], Andrea De Lucia[2], Denys Poshyvanyk[1]

[1]*The College of William and Mary, USA* — [2]*University of Salerno, Italy* — [3]*Università della Svizzera italiana (USI), Switzerland* — [4]*University of Sannio, Italy* — [5]*University of Molise, Italy*

SUMMARY

A broken snapshot represents a snapshot from a project's change history that cannot be compiled. Broken snapshots can have significant implications for researchers, as they could hinder any analysis of the past project history that requires code to be compiled. Noticeably, while some broken snapshots may be observable in change history repositories (*e.g.,* no longer available dependencies), some of them may not necessarily happen during the actual development. In this paper we systematically study the compilability of broken snapshots in 219,395 snapshots belonging to 100 Java projects from the Apache Software Foundation, all relying on Maven as an automated build tool. We investigated broken snapshots from two different perspectives: (i) how frequently they happen and (ii) likely causes behind them. The empirical results indicate that broken snapshots occur in most (96%) of the projects we studied and that they are mainly due to problems related to the resolution of dependencies. On average, only 38% of the change history of project systems is *currently* successfully compilable.
Copyright © 0000 John Wiley & Sons, Ltd.

Received ...

KEY WORDS: Broken Snapshots, Mining Software Repositories, Software Quality, Empirical Studies

1. INTRODUCTION

A broken snapshot represents a snapshot from a project's change history that cannot be compiled[a]. The problem of having incompilable snapshots can have significant implications for researchers preventing the use of tools working on byte code as well as any sort of dynamic analysis. Other than representing problems for developers, as previous work has shown [14, 32, 38, 19, 29], broken snapshots represent a strong limitation for researchers interested in running Mining Software Repositories (MSR) studies. To better illustrate the problem, let us consider the following scenarios.

Gabriele is a researcher studying the impact of refactoring activities on code smells. To this aim, *Gabriele* is using a smell detector, JDeodorant [36, 11] and a refactoring detector, RefFinder [27]. All these tools work on byte code, hence require source code to be compiled. When trying to compile all 1,000 revision snapshots of a project *A*, *Gabriele* realizes that there are 400 snapshots that failed to compile, hence, creating gaps in the observed project history, because JDeodorant and RefFinder cannot work on these snapshots. *Gabriele* feels that these gaps would seriously impact his analysis, especially because some of them occur in periods where release notes announced major projects' restructuring. Certainly, *Gabriele* could decide to conduct the whole study at release level; however, refactoring actions can be precisely identified only at commit level granularity. Indeed, differencing

[a]In this paper we focus on broken snapshots based on compiling software, while other broken snapshots could be the result of test failures, deployment failures, *etc.*

Copyright © 0000 John Wiley & Sons, Ltd.

two releases would introduce noise because multiple changes, not only refactorings, have likely occurred on the same artifacts. This represents a significant threat to validity for such analysis [4].

A similar situation occurs to *Alex*, a researcher studying how the quality of test suites changes over time. *Alex* is interested in measuring the code coverage of the test suites in a project after each commit, to analyze the phenomenon at a fine-grained level (*e.g.*, to understand the characteristics of commits that lead to a strong decrease in code coverage). To measure the code coverage *Alex* is using Cobertura[b]. Unfortunately, this tool also works on byte code, requiring the compilation of each project's snapshot that *Alex* is interested in analyzing. *Alex* is unable to compile about 450 out of the 900 projects' snapshots, i.e., half of them. He performs a deep investigation to identify the likely causes of lack of compilability and, for 100 snapshots, he realizes that the build script is unable to solve library dependencies, either because they refer to local files, or because the libraries are no longer available in the *Maven* central repository. He is able to solve the problem for 80 of them, by manually searching for the needed library version. However, (i) the whole investigation took two days of *Alex*'s effort, and still the achieved compilability rate is not enough for the analysis to be performed. As a consequence, *Alex* decides to change his plan, shifting the focus of the study to a coarser-grained level, only studying the quality of the test suite in each issued release (already compiled) of the analyzed project.

Clearly, the higher the rate of broken snapshots in the project, the higher the chances that *Gabriele* and *Alex* will not be able to derive useful (actionable) insights for the projects they are studying, since they will not be able to run the needed tools on those broken commits.

Paper contribution. While broken snapshots is a significant research problem, there has been little effort in empirically exploring and understanding the root causes and implications of these problems. The goal of this paper is to fill this gap by systematically studying the compilability of Java project snapshots downloaded from Version Control Systems (VCS), with the purpose of understanding:

1. *how frequently* compile broken snapshots (in the following referred to as simply "broken snapshots") occur, and how long they last before the snapshot becomes compilable again in post-mortem analysis (noteworthy, we focus on compilability "here and now" for mining purposes, as opposed to compilability in the past for development purposes).

2. *why* broken snapshots occur, *i.e.*, what are the typical reasons behind their introduction.

The study has been conducted in the *context* of 100 Java open source projects by relying on *Maven* as an underlying build system. Overall, we downloaded, attempted to compile and analyze 219,395 snapshots of these 100 projects.

Motivations and Implications. The results of our study can be used from multiple *perspectives*. One straightforward perspective is the one of a researcher—or somebody interested in gaining actionable analytical insights for an evolving software project—mining change histories for various purposes [6]. Broken snapshots can hinder any analysis performed at commit level, because many such tools require byte code to work [9, 2, 10, 35], while specific analyses, *e.g.*, related to dynamic analysis or test suite assessment, require the system to be executed. From this perspective, the higher rates of broken snapshots would either trigger tedious and costly repair activities (assuming they are even possible), or force the researcher to perform the analysis at a coarse-grained level, *i.e.*, limit the analyses to compilable releases (and potentially loose useful data at commit level granularity). However, there are numerous examples in the research literature showing that the choice of granularity (commit *vs.* release) is an important design decision, since analysis at different granularity may produce different or even contradictory results. For instance, the evolutionary study of code clone genealogies at commit level by Kim *et al.* showed that a significant number of clones in software was short-lived [16]. However, another study of code clone genealogies at release level by Saha *et al.* demonstrated contradictory results, finding that many code genealogies persisted in follow-up releases [31, 5].

[b]http://cobertura.github.io/cobertura/

Copyright © 0000 John Wiley & Sons, Ltd.
Prepared using smrauth.cls

More studies, related to fix-inducing refactorings [4] or to changes occurring on design patterns [1] have been performed by detecting refactorings and design patterns at release level. However, in the first case [4] it would have been desirable to perform a more precise detection of refactorings at commit level, and, in the second case [1], to study how design patterns were introduced by developers in specific commits. Yet, the refactoring detector [27][c] and the design pattern detector [37] work on byte code. Similarly, a study on the evolution of test cases performed by Zaidman *et al.* [40] could have been performed at fine-grained level aiming at observing specific activities triggering the need to augment test suites, as well as changes leading to better test suites achieving higher coverage. Of course, there are also cases (see the clone studies above) for which analysis at release level [31] can be easier and probably more meaningful [5]. All these considerations outline the need for the study presented in this paper: An empirical investigation into the phenomenon of broken snapshots is needed aiming at shedding some light into the reasons on *when* and *why* such changes happen.

Main results. The results of our study indicate that broken intervals (i.e., time periods including one or several consecutive broken snapshots) occur in 96% of the projects we analyzed with a median number of four in each system. Broken intervals are mostly small, with a median size of three subsequent broken snapshots. However, they can severely impact the compilability of software systems. As a matter of fact, on average, only 38% of the change history of a project system is successfully compilable today. The substantial number of broken snapshots shows their potential to hinder commit-based analysis of the project's change history when this requires compiling a snapshot of a system (*e.g.*, in the context of MSR-driven research). Moreover, broken snapshots are in most of the cases related to dependency resolution problems. Finally, snapshots' age highly impacts the likelihood of a successful compilation. Indeed, *Recent Snapshots* (*i.e.,* snapshots referring to commits performed recently in the change history) are almost 3.76 times more likely to compile with respect to *Early Snapshots* (*i.e.,* early snapshots in the change history of the project).

Paper structure. Section 2 describes the study definition, research questions and planning. Results are reported and discussed in Section 3, while the threats to its validity are discussed in Section 4. Section 5 discusses the related literature on the analysis of build systems, while Section 6 concludes the paper and outlines directions for future work.

2. EMPIRICAL STUDY DEFINITION AND DESIGN

The *goal* of the study is to analyze the change history of open source software projects, with the *purpose* of investigating the extent to which broken snapshots occur and which are the typical errors that a miner can observe when automatically compiling change history code snapshots. The *quality focus* of the study is on compiling projects' snapshots downloaded at commit-level granularity. The *perspective* is of researchers interested in performing mining studies over project histories using approaches and tools requiring to compile the project. The *context* consists of 100 Java open source projects using a Maven-based build infrastructure.

More specifically, the study aims at addressing the following **R**esearch **Q**uestions (RQs):

- **RQ$_1$:** *How many snapshots are currently compilable in the change history?* This research question aims at quantifying the relevance and magnitude of the observed phenomenon by investigating the frequency with which broken snapshots occur over time, and to what extent automatically compiling the complete change history of software project can be successful. We also investigate whether the size of the change history and the snapshot's age impact the likelihood of a successful compilation. We have chosen to study the effect of these two factors because it is possible that recent snapshots could be less affected by the problem, e.g., because dependencies are still available.

[c]This tool is reported as an example, as refactoring detection tools working on source code also exist.

Copyright © 0000 John Wiley & Sons, Ltd.

Prepared using smrauth.cls
DOI: 10.1002/smr

Table I. 100 projects considered in the study.

#Classes	KLOC	#Commits	Mean history length	Min-max history length
4-6,027	0-887	219,395	5.21	0-13

The output of RQ_1 will provide a quantification of the possible problems researchers could experience in mining and compiling the change histories of open source systems.

- **RQ_2:** *Which types of errors can be observed by automatically compiling code snapshots?* The second research question sheds the light on the possible causes behind broken snapshots, by categorizing the compilation errors we identified in our study. Knowing *why* a broken snapshot happens can help in understanding if it could be automatically/manually fixed with a reasonable effort while mining the software change history.

2.1. Context Selection

The *context* of our study consists of 100 open source projects from the Apache ecosystem[d]. Our choice of focusing on these software projects is not random but aimed at considering systems that are (i) widely used in mining studies and, more generally, in studies carried out in the software engineering community [12, 22, 3, 30], (ii) using state-of-the-art build systems, and (iii) having a considerable change histories to analyze. Overall, to date the Apache ecosystem is composed of 249 software projects. Among those, we only selected the ones fully written in Java (187 projects) and relying on Maven[e] as an automated build tool. As we will explain later, the use of a fully automated build tool is a requirement for considering a system in our study, while the focus on systems written in Java is dictated by the fact that Maven is mainly used in the context of Java-based projects. As of today, Maven is one of the most popular build systems for Java projects. According to a survey (performed in mid 2013)[f] Maven is being used by 71% of the interviewed developers, while 35% use Gradle[g] and 17% Ant[h]. Note that the sum is beyond 100% as there are developers using multiple tools. In summary, we decided to focus on Maven not only because of its popularity (especially in the case of the Apache ecosystem [20]), but also because of its capability to resolve dependencies (*e.g.*, a broken snapshot for an Ant-based project could just be due to the need for manually downloading a library and putting it in a specific location). Moreover, as it will be explained in Section 2.2, Maven generates error codes for different kinds of compilation errors, which allows us to address **RQ_2** without the risk of introducing a subjective or error-prone classification.

To identify which of the 187 Java projects make use of Maven we cloned their *Git* repositories and verified the presence of `pom.xml` in the last systems' snapshot. This selection resulted in identifying 136 projects (72% of the total) using Maven. Due to the very computationally expensive analysis that we needed to perform to answer our two research questions, we decided to limit our study to 100 randomly selected projects. Table I reports the characteristics of the subject systems, and in particular (i) the size ranges in terms of the number of classes and KLOC, (ii) the overall number of commits analyzed, and (iii) the average, minimum, and maximum length of the projects' history (in years) analyzed. All the considered projects are hosted in *Git* repositories. The list of 100 projects considered in our study can be found in our replication package (see Section 2.4).

In order to assess the extent to which the chosen set of projects represents a large body of open source Java projects (in terms of code size, team size, and project activity), we used the diversity metric proposed by Nagappan *et al.* [23]. We matched the list of the systems in our study against the projects available in Boa [8], and ended up with only ten out of 100 projects that were matched by

[d] http://www.apache.org
[e] http://maven.apache.org
[f] http://arhipov.blogspot.it/2013/08/java-build-tools-survey-results.html
[g] https://gradle.org
[h] https://ant.apache.org

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/smr

name. The diversity metric calculated on this subset was 0.4984, which means that only 10% of our dataset covers almost half of the open-source projects according to four dimensions: developers, project age, number of committers, and number of revisions. The scores for each dimensions are 0.89, 0.99, 0.92, 0.91 respectively, indicating that our subset covers relevant dimensions for our analysis.

2.2. *Data Extraction Process*

The first step needed to answer the two research questions is to identify, for each analyzed project P_i, the time interval T_b during which it used the Maven building tool. To identify T_b for each P_i we mined P_i's complete change history looking for the commit C_{start} introducing the Maven build file `pom.xml`. Then, we verified that there was no commit in the P_i's history deleting the build file from the versioning system until the last commit considered in our study (we cloned the repositories during September 2014). Thus, T_b for each of the considered projects begins at the date in which C_{start} was performed and ends at the date in which the last commit was performed in the cloned repository. This process resulted in the exclusion, on average, of 0.79 years of change history across the 100 systems (*i.e.*, $\approx$ 3 years). Overall we analyzed 437 cumulative years of change history for all the projects.

Once the portion T_b of change history to analyze was defined for each project, the steps detailed in the following subsections have been performed to extract the data needed to answer our RQs. The overall data extraction process took eleven weeks despite being distributed across three Linux servers of which two were equipped with 8 ten-core 3.2 GHz CPU (80 cores) and 192 Gb of RAM and one with 40 quad-core 2.13GHz CPU (160 cores) and 160 Gb of RAM.

RQ$_1$ and RQ$_2$: Compiling Snapshots for each Commit. The basic data needed for answering the two research questions is the information about the compilability of each snapshot present in the change history (T_b) of each project P_i. To extract such information we built a mining tool using Maven to automatically compile the system at each snapshot in T_b. During this phase each snapshot S is classified as *compilable* or *not compilable*.

If a snapshot S is *not compilable* it is classified as a *broken snapshot*. We define a *broken interval* as a series of continuous broken snapshots $S_i, S_{i+1}, \ldots, S_{j-1}, S_j$, such that S_{i-1} and S_{j+1} are either *compilable* snapshots or not part of the change history (*i.e.*, S_{i-1} and/or S_{j+1} do not exist). The snapshot S_i is the *broken interval starting point* represented with BI_{start}, while the snapshot S_j is the *broken interval ending point* represented with BI_{end}. We limited our analysis to the *master* branch (including merges of other branches) for several reasons. First of all, according to the guidelines provided by *Git*[i], the master branch *always reflects a production-ready state*. Thus, it can be considered as the most stable one and, therefore, it is worth investigating the phenomenon of broken snapshots on such a branch. We also verified this conjecture applies to the analyzed projects by analyzing the documentation page of 10 randomly selected Apache projects. Secondly, most of the mining studies are conducted through the analysis of the master branch (*e.g.*, [24, 40, 37]). Since our goal is to analyze how the phenomenon of broken snapshots can impact the activities of researchers interested in compiling code snapshots, we decided to focus our attention only on this branch.

Also, note that for our perspective of researcher interested to perform fine-grained historical analysis of compilable commits, we are interested in the compilability of each single commit. This might not be the case when looking at the same phenomenon from a developer's perspective, for which what matters is the compilability before performing a push or pull request.

In order to compile all the 219,395 snapshots in our study, the tool that we built relies on Maven 3.2.3 (the latest version available as of September 2014, which ensures backward compatibility with previous versions), by programmatically invoking it through the Java API provided by Apache[j]. More specifically, our tool creates an `Invoker` of type `InvocationRequest` setting `compile`

[i] https://git-scm.com/doc
[j] org.apache.maven

Copyright © 0000 John Wiley & Sons, Ltd.
Prepared using smrauth.cls

DOI: 10.1002/smr

as the building goal. The `InvocationResult` returned by the `Invoker` reports the compiling's results, including its success (*i.e.*, the snapshot was compilable) or failure (*i.e.*, the snapshot was not compilable).

$\mathbf{RQ}_2$: Classifying Compile Errors. While the classification of snapshots as *compilable* or *not compilable* and the identification of *breaking intervals* is sufficient to answer $\mathbf{RQ}_1$, in the context of $\mathbf{RQ}_2$ we need to analyze the issues observed when compiling *broken snapshots*. To this aim, we mapped the errors captured while compiling *not compilable* snapshots into four possible categories related to different *building actions* that may fails into one of these categories:

- *Resolution*: errors related to the resolution of artifacts (*e.g.*, download of dependencies);
- *Parsing*: errors due to the parsing operation (*e.g.*, malformed build files);
- *Compilation*: errors occurring during the compilation phase (*e.g.*, syntactic errors in the source code);
- *Other*: generic errors that cannot be mapped to any of the above listed actions (*e.g.*, Maven itself crashes during the run).

These four categories have been iteratively defined by two of the authors that manually analyzed the errors stored by our mining tool for each snapshot tagged as *not compilable*. The errors thrown by Maven during the snapshot compilation have been manually and independently analyzed by each of the two involved authors, who classified them into categories of errors related to the different compiling actions. Specifically, the two authors manually mapped the exceptions thrown by Maven into the corresponding *action* categories. To accomplish this task, they read the Maven exceptions' documentation available online[k]. After having independently mapped the errors into categories, the two authors met to resolve the conflicts and to refine their classification, leading to the four categories listed above.

2.3. Data Analysis

To answer $\mathbf{RQ}_1$ we report descriptive statistics about the number of *broken intervals* found while attempting to automatically compile the change history of the analyzed software systems and the length of such *broken intervals*. Also, we measure the length of *broken intervals* BI in terms of the number of *broken snapshots* between BI_{start} and BI_{end} (considering BI_{start} and BI_{end} too).

Since we want to report absolute values as they are easier to be interpreted than percentages of commits over the total history length, we need to take into account the different size of the change histories of the subject systems. To this aim, we classified projects in short, medium and long history length, on the basis of the available, observed history. This also allows us to verify the impact of the history length on the likelihood of observing broken snapshots. To classify the projects into the above categories, we compute the first (Q_1) and the third (Q_3) quartile of the distribution representing the change history length of the subject systems. We classify systems into the following categories: (i) *short history* length if they have a number of commits C_n lower than Q_1; (ii) *medium history* length if $Q_1 \leq C_n < Q_3$, and (iii) *long history* length for which $C_n \geq Q_3$. We report and discuss results on the number and size of broken intervals for the three categories separately as well as for all categories together. In order to provide an overall view of the extent to which *broken snapshots* are found in the change history, we show the distribution of the *compilability* (that is, the percentage of snapshots successfully compilable in the change history of each of the analyzed systems). This analysis complements the previous one performed with absolute numbers.

Moreover, we determine whether the number of broken intervals and project's compilability is correlated with the history's length. We show scatter plots, compute Kendall's rank correlation, and also check, using a Kruskal-Wallis test, whether the number of broken intervals differs across the three categories.

[k] `http://tinyurl.com/ja4ncal`

Copyright © 0000 John Wiley & Sons, Ltd.
Prepared using smrauth.cls

Finally, we analyzed whether the *age* of the snapshot in a software project is an important factor characterizing its likelihood of successful compilability. We classified the snapshots of each project in accordance to their *position* in the change history. Similarly to what previously was done for the change history length, we use quartiles to define the categories of snapshots. In particular, given a software project's change history $CH = \{S_1, S_2, \ldots, S_n\}$ we compute the first (Q_1) and third (Q_3) quartile of the distribution of the snapshots over time. We classify the snapshot of each project CH in the following categories:

- *Early Snapshots*: snapshots belonging to the early stage of the change history ($S_i | i < Q_1$). Representing the first 25% of the snapshots in the CH.
- *Intermediate Snapshots*: snapshots belonging to the middle stage of the change history ($S_i | Q_1 \leq i \leq Q_3$). Representing the middle 50% of the snapshots in the CH.
- *Recent Snapshots*: snapshots belonging to the recent stage of the change history ($S_i | i > Q_3$). Representing the last 25% of the snapshots in the CH.

We report and discuss the results on the distribution of compilability for such categories of snapshots. The intuition behind this analysis is that we expect that snapshots belonging to the early stage of a project's change history are less likely to be successfully compilable with respect to more recent snapshots, given the fact that references to dependencies might have been changed, missing or deprecated. To verify the statistical significance of such a hypothesis, also in this case we apply the Kruskal-Wallis test.

For **RQ**$_2$ we report the results of classifying non-compilable snapshots according to the possible causes behind those errors (see Section 2.2).

2.4. Replication Package

The data set used in our study is publicly available at http://www.cs.wm.edu/semeru/data/breaking-changes/. Specifically, we provide: (i) the list of analyzed systems with links to their versioning system and GitHub page; (ii) the raw data collected for the two research questions; and (iii) the `R` scripts and working data sets used to run the statistical tests and produce the figures and tables presented.

3. EMPIRICAL RESULTS

This section reports the analysis of the results aimed at answering two research questions formulated in Section 2.

3.1. RQ$_1$: Occurrence of Broken Snapshots

We divided the projects in three categories based on the length of the change history using the quartiles of such distribution as described in Section 2.3. The quartile values are: $Q_1 = 235$ and $Q_3 = 1,968$ commits. Note that these quartiles refer to the portion of the change history in which we identified a build file. Since we are using Q_1 and Q_3 to define our sets, the number of projects for each category is 25, 50 and 25 respectively for projects in the *short, medium* and *long* history category.

Table II shows the number of *broken intervals* observed during the change history of the studied systems. We found that 96% of projects experienced at least one *broken interval* in their history. Also, the value for the first quartile (2.00) for all the projects indicates that at least 75% of the analyzed projects experienced at least two *broken intervals*. Also the median (4.00) and the mean (8.99) values confirm that, generally speaking, the phenomenon of broken snapshots cannot be ignored by researchers interested in mining (and compiling) the change history of a software system. When looking at the data for projects belonging to the three groups, on the one hand, we can notice that the median and mean number of broken intervals slightly increases in case of projects with

Copyright © 0000 John Wiley & Sons, Ltd.

Prepared using smrauth.cls
DOI: 10.1002/smr

Table II. Number of broken intervals for projects having different length of change history.

Projects	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Short history	1.00	2.00	3.00	3.12	4.00	7.00	1.56
Medium history	0.00	1.00	5.00	7.84	9.00	32.00	8.35
Long history	0.00	1.00	7.00	16.50	26.00	62.00	18.96
All	0.00	2.00	4.00	8.99	9.00	62.00	12.26

Table III. Size of broken intervals.

Projects	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Short history	1.00	1.00	5.00	18.96	16.00	139.00	33.20
Medium history	1.00	1.00	4.00	53.09	30.00	1212.00	150.39
Long history	1.00	1.00	2.00	184.00	26.00	6955.00	676.01
All	1.00	1.00	3.00	110.50	25.75	6955.00	478.03

Table IV. Percentage of Compilable Snapshots

Projects	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Short history	0.00	24.44	44.00	48.91	79.61	97.01	33.63
Medium history	0.00	5.44	29.60	39.46	72.50	100.00	34.69
Long history	0.00	0.00	18.54	23.78	33.26	100.00	28.38
All	0.00	5.43	29.60	38.13	66.00	100.00	33.79

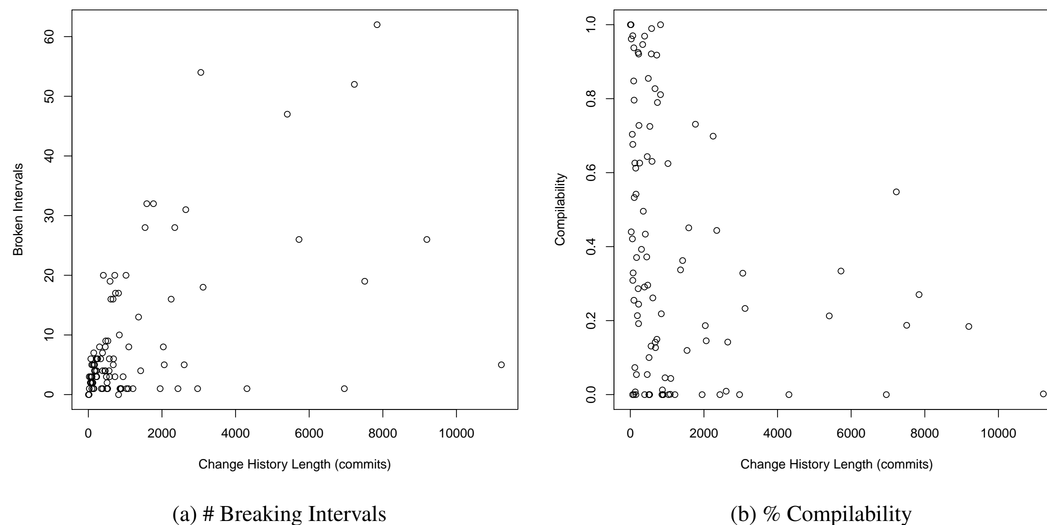
medium or long history. On the other hand, the scatter plot in Figure 1a shows that there is no clear correlation between the size of the change history and the number of broken intervals found. This result is confirmed by the correlation value ($\tau = 0.29$), which indicates a low correlation between the phenomenons. Note that even though we obtained a low correlation value, this is definitely not negligible, therefore worth to further investigate using different dataset. Finally a statistical comparison among the three groups performed using the Kruskal-Wallis test indicates a lack of statistically significant difference (p-value=0.06).

To get a better idea of what these numbers mean in practical terms (*i.e.*, to what extent these broken intervals affect the compilability of the subject systems), Table III shows the descriptive statistics summarizing the length of the broken intervals (*i.e.*, the number of commits performed between a BI_{start} and its corresponding BI_{end} —see Section 2.3). The median length of broken intervals is 3.00, showing that at least half of the broken intervals only affect few commits (less than four). However, relatively high values of the mean (110.50) and standard deviation (478.03) tell another part of the story, highlighting the presence of very long broken intervals. For instance, we identified in our dataset a very long breaking interval in the `Apache Oozie` system. The BI_{start} snapshot is dated Sep 2, 2011 and its corresponding BI_{end} is found on Jul 17, 2012 for a total of 408 broken snapshots. In order to understand the reason behind this long breaking interval we manually analyzed what was the change that actually broke the compilation, finding that it was related to the update of a dependency, i.e., `org.slf4j`, that cannot be resolved. As for the end of the broken interval, we found that it corresponds to a commit in which a developer updated most of the dependencies of the project, likely resolving the `org.slf4j` dependency.

The effect of broken intervals on compilability of the analyzed systems' snapshots is described in Table IV, which reports the percentage of compilable snapshots. The process adopted to verify the snapshots' compilability is the one described in Section 2.2. Surprisingly, the percentage of compilable snapshots is quite low (mean=38%, median=70%). This clearly poses strong limitations on the kind of analysis one can carry out on these non-compilable snapshots, for instance, the kinds of analyses outlined in the introduction. On 18 projects we were not able to compile any of the snapshots available. We manually investigated such cases in order to understand the possible reasons behind this result. We found that all of them are actually sub-projects belonging to parent

Copyright © 0000 John Wiley & Sons, Ltd.

Prepared using smrauth.cls
DOI: 10.1002/smr



(a) # Breaking Intervals

(b) % Compilability

Figure 1. Scatter plots of (a) size of broken intervals per change history length; and (b) compilability rate per change history length.

projects. Even if they have their own `pom.xml` file, their compilation depends on the compilation of the parent projects. As an example, `Log4j-receivers` is part of the `Log4j` project. In one of the first commits, a developer moved the build plugin configuration to the parent project[l], making it impossible to compile any of the snaphots of the system. Note that we do not look at the relationships among the software projects since our goal is to systematically investigate the compilability of Apache projects, actually simulating the behavior of a researcher interested in performing a large scale mining study.

On the other hand, we also observed several exceptions to the general low compilability trend, like the `Apache Quid-proton` project that reported 100% compilability for all the snapshots despite the relatively long change history (1,759 mined commits). Interestingly, during its change history this project only established dependencies toward two libraries, *i.e.*, `JUnit` and `org.mockito`. However, it is worth noting that developers always declared dependencies with major releases of such libraries.

Moreover, we investigated whether the length of the change history (negatively) correlates with the compilability of the project. The scatter plot in Figure 1b clearly shows a very low correlation, which is confirmed by the Kendall's $\tau = -0.25$.

Finally, we analyzed the impact of the snapshots' age on the compilability. Table V shows the distribution of compilability for the three categories of snapshots based on their age in the software project. The trend is clear - the age of the snapshots is an important factor characterizing the likelihood of a successfully compilation. Indeed, looking at the median values, *Recent Snapshots* are almost two times more likely to compile than *Intermediate Snapshots* and 3.76 times more likely to compile with respect to *Early Snapshots*. The result is statistically significant (p-value$<$0.01) as highlighed by the Kruskal-Wallis test.

[l]http://tinyurl.com/hodp7fu

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/smr

Table V. Percentage of compilable snapshots based on snapshots' age.

Snapshots' Age	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Early Snapshots	0.00	0.00	12.39	36.20	83.93	100.00	41.06
Intermediate Snapshots	0.00	0.00	23.60	36.42	71.01	100.00	38.62
Recent Snapshots	0.00	0.46	46.56	43.89	75.69	100.00	38.51

Table VI. Likely causes behind broken snapshots.

Action	%
Resolution	58%
Parsing	10%
Compilation	4%
Other	28%

Summary for RQ$_1$. Among the analyzed systems, 96% exhibit at least one broken interval in their history, and this phenomenon is not necessarily more severe for projects with a longer change history. Most of the broken intervals only result in a few uncompilable commits, while the others persist in the system for a long time resulting in a high number of snapshots (median 30%) that can not be properly compiled or built. Snapshots' age in a software system represents an important factor characterizing the likelihood of a successful compilation, with recent snapshots being more compilable than early snapshots.

3.2. *RQ$_2$: Likely Causes for Broken Snapshots*

Table VI reports classification results of the causes behind the compiling errors that we identified in the subject projects (see Section 2.2). A large number of the errors (58%) raised during the *Resolution* of an artifact, usually representing a dependency to be solved (*e.g.*, a library linked by the project under analysis). Indeed, the most frequently raised exceptions we found during the Maven building were `DependencyResolutionException` (32%) and `ArtifactResolution-Exception` (26%). The former is raised, for example, during the compilation of an `Apache Oltu`'s snapshot S_B, due to the problem of downloading the `OAuth` library, responsible for the implementation of a secure authorization protocol via HTTP. This error has been fixed after 193 commits in a commit S_F by modifying the reference pointing to the `OAuth` library in the `pom.xml` file. It should be noted that the compilation error we identified during these 193 commits could be the result of two scenarios having different implications:

1. *The compilation error was actually present at the time T in which S_B was committed.* This means that the developers actually put a wrong reference to the `OAuth` library inside the build file. Thus, starting from T, it was not possible to compile the system for the subsequent 193 commits, when S_F fixed the issue. If this is the scenario behind S_B, this change clearly resulted in a long broken interval for `Apache Oltu`, with all possible consequent problems.

2. *The compilation error was not present when S_B was committed.* Suppose that the reference to the `OAuth` library present in the build file after S_B was valid at that time, and has remained so until S_F, in which the developers updated the `OAuth`'s reference to a new one, given the unavailability of the previous reference. Clearly, by mining the `OAuth`'s versioning system today, the old reference is no more available since the time in which S_F has been committed results in compiling issues during the 193 commits between S_B and S_F. In this scenario, the S_B commit did not cause any issues for developers during the broken interval we observed, but it only represents a problem for people interested in post-mortem analysis — compiling snapshots of a system for each commit, which is a typical use case in the field of mining software repositories. It is worth noting that it is not obvious to determine to what extent

Copyright © 0000 John Wiley & Sons, Ltd.
Prepared using smrauth.cls

this phenomenon occurs, because we cannot determine whether the URL was valid when the `pom.xml` file was committed. However, it must be pointed out that on our interpretation this still constitutes a broken interval for researchers interested to analyze the past history of a project or use such data to build a recommender.

Parsing and *Compilation* errors account only for a small fraction of broken snapshots identified in the projects (10% and 4%, respectively), showing that the build and the source code files are generally correct from a syntactic point of view. One example of such broken interval is the one we observed in the `Apache Jackrabbit` system due to compilability problems in one of the libraries it exploited. Such a problem was fixed after three commits with a commit reporting the following message: "*Bundle plugin 2.3.5 requires Java 6. Downgrade to 2.3.4 that works with Java 5*". The problem was due to the different Java versions exploited by the client project (*i.e.*, `Apache Jackrabbit`, requiring Java 5) and the specific version of the exploited library (*i.e.*, `Bundle plugin 2.3.5`, requiring Java 6). The problem was solved by downgrading the dependency to a previous version of the library requiring Java 5. Unlike the example discussed for the `Apache Oltu` system, in this case we can observe that the broken interval has been directly observed by the developers of the `Apache Jackrabbit` project.

The remaining 28% of errors fall in the *Other* category, containing errors that we were not able to map to a specific compile action.

Summary for $\mathbf{RQ}_2$. The causes behind broken snapshots are often related to problems encountered during the resolution of dependencies (58% of cases), while we rarely observed errors due to syntactic errors in the build and source code files.

4. THREATS TO VALIDITY

Threats to *construct validity* concern the relationship between theory and observation, and in this work, they are mainly related to the measurements we performed in our study. As discussed in Section 3.2 it could be the case that we observe missing dependencies because a library URL contained in a Maven `pom.xml` file is no longer valid. While this can be considered as a false positive while interpreting the results from the perspective of somebody compiling the system at that time (*e.g.*, a tester), as explained in Section 3.2, this is a true broken interval if one is interested to analyze the past history of a project, for various purposes. Imprecisions in our study can also be due, as explained, to the fact that we only observed the evolution history of a project by looking at the master branch. However, as explained in Section 2.2, this is intentional in order to identify a specific evolution timeline.

Threats to *internal validity* are related to factors, internal to our study, that can influence our results. The main threat to internal validity is related to cause-effect relationships we claim, especially in $\mathbf{RQ}_2$, where we analyze the likely causes of broken snapshots. As explained earlier, we mainly rely on error messages raised by Maven, which may not necessarily capture the true underlying causes of a broken snapshot (this is why we refer to them as "likely causes"). Nevertheless, we must point out that our focus is to observe *what breaks the compilability today, rather than what broke compilability at development time*.

As explained in Section 2.2 we limited our analysis to the projects' history master branch, under the assumption that it is the main and stable development trunk. However, we could not exclude that, for some projects, the main development trunk occurs in different branches.

Threats to *conclusion validity* concern the relationship between experimentation and outcome. As explained in Section 2.3, wherever possible we have used appropriate statistical procedures, including effect size measures, to address our research questions.

Threats to *external validity* concerns the generalization of our findings. Our results are clearly confined to a specific category of software projects, *i.e.*, Java projects using Maven. Moreover,

Copyright © 0000 John Wiley & Sons, Ltd.
Prepared using smrauth.cls

DOI: 10.1002/smr

McIntosh *et al.* [20] found that Maven adoption drops drastically in projects outside of the Apache foundation. We have explained in Section 2.1 why we decided to focus on such projects. Truly, nowadays other build tools such as Gradle[m] are becoming even more popular than Maven, however, our focus on Maven for a mining past history study like ours is motivated by its popularity in the past five years or so. Nevertheless, our study definitely needs to be extended on projects using other build tools and, above all, to projects developed in different programming languages, including C/C++ programs using Make. Last, but not least, although we have shown that the selected pool of projects achieve a good level of diversity [23], future studies should look at projects deployed on different forges or belonging to different open source communities other than the Apache Software Foundation, because different communities may impose different promotion policies.

5. RELATED WORK

Despite the significant research and practical importance of a problem of broken snapshots, there have been only few studies aimed at understanding this phenomena and its implications. While there is some work on studying continuous integration practices [7, 38] and build systems [26, 21], there are no general recipes on how to handle broken snapshots.

Seo *et al.* [32] presented an empirical study of 26.6 million builds from Google's cloud-based build system used internally for Google projects. The authors analyzed the log of the build system in order to understand (i) how often compiling actions fail; (ii) why they fail; (iii) how long it takes to fix them. Although our study appears to be similar to the one conducted by Seo *et al.*, there is a fundamental difference between the contexts of these two studies. Indeed, while Seo *et al.* investigated the *developers' building activity* (*i.e.*, every time a developer attempts to build the project she is working on), our study analyzed the *repository's history* (*i.e.*, the committed source code). Clearly, compiling activities do not match the activity performed on the repository. Indeed, *a developer could obtain several compile failures using the building tool, before performing a commit that does not contain any broken snapshot*. Also, while the reasons behind a compiling failure and a broken snapshot can be shared between these two studies, the results regarding the frequency of compile errors and how long does it takes to fix them cannot be compared since they are related to different contexts (building vs. committing). Finally, it is worth noting that in our study we perform our analysis relying on Maven as building tool, rather than using an internal Google's building system. This different choice allows us to directly analyze projects typically used in MSR studies. Thus, our study can be seen as strongly complementary to the one proposed by Seo *et al.*.

The closest work to our is the study by Kerzazi *et al.* [14], who conducted an empirical study considering 3,214 builds of an industrial system during a period of six months. They analyzed (i) the frequency of broken snapshots on one software project, (ii) under which circumstances builds usually breaks, and (iii) which are the factors impacting broken snapshots. Firstly, the results indicate that broken snapshots are frequent, *i.e.*, they found a median of 8.3% of broken compilations per day. Secondly, they conducted semi-structured interviews involving 28 developers with at least three years of experience and authors of most of the breaking and fixing compiles they found. The outcome of the interviews showed that developers usually break the compilation when they forget to commit some files or update dependencies. This happens especially because they were active in several branches of a system at the same time. Finally, Kerzazi *et al.*, explored the factors that impact the rate of broken snapshots, demonstrating that the development timeline is the most important factor causing the breaks. While our study shares similar goals with the one by Kerzazi *et al.* [14], the context of our study is much more extensive as we conduct the study on 219,395 snapshots of 100 Java open source projects. In addition to quantitatively studying frequencies of broken snapshots, the goal of our study is to obtain and in-depth and systematic understanding of the phenomena into *how* and *why* broken snapshots happen. Moreover, it is important to remark that our study is performed at the *commit level*, which has important implications for the practical tools and research analyses

[m] https://gradle.org

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/smr

that need to be performed at commit level granularity [34, 42, 43, 41, 33, 22, 28, 39, 15, 17]. At the same time, our paper reinforces some of the findings provided by Kerzazi *et al.*, in terms of broken snapshots' diffusion and time windows in which breaks are located.

McIntosh *et al.* [19] studied, on the one hand, how frequently source code changes require changes to the build file, and, on the other hand, the proportion of developers responsible for build maintenance. The study involved ten open source systems and showed that a good percentage of source code items (ranging between 4% and 27%) requires an accompanying build change. Moreover, about 79% of source code developers and 89% of test code developers are involved in build maintenance activities. Cataldo and Herbsleb [7] studied major factors behind integration failures in distributed software projects, whereas Hassan and Zhang [13] defined a model for predicting the outcome of the compilation process. Unlike these works, in this paper we do not provide methods that a researcher can use for assessing the output of the compilation process, but we have the main goal of systematically investigating to what extent the phenomenon of broken snapshots can impact the ability to conduct a mining software repository study.

Phillips *et al.* [25] designed and executed the study aimed at studying how release engineers and managers make integration decisions. The work by Kwan *et al.* [18] investigated the effect of socio-technical congruence (*i.e.*, the alignment between work and social relationships between team members) on the outcome of the compilation process on a large industrial software project. Their findings suggest that congruence affects build success probabilities for different types of builds. Our work does not take into account social factors, but it is focused on the analysis of the likely causes behind broken snapshots.

Broken snapshots have been studied from a completely different perspective by Raemaekers *et al.* [29]. Specifically, they investigated seven years of library releases on the Maven central repositories, to understand the phenomenon of incompatibility caused by new versions of such libraries. They found that the "semantic versioning", *i.e.*, the strict usage of major and minor release numbers do not provide developers with enough information about the compatibility of the library interfaces. Differently from our study, they looked at a very specific problem more related to the evolution of library interfaces, while we are interested in investigating various kinds of compilability problems, from dependencies to parsing errors.

6. CONCLUSION

This paper presents a study aimed at investigating compile broken snapshots in 100 Java projects from the Apache Software Foundation, all of them relying on Maven as an underlying build tool. More specifically, the study aims to investigate (i) the magnitude of the phenomenon, *i.e.*, frequency of broken intervals and their length in terms of the number of broken snapshots, (ii) the likely causes of broken snapshots, as identified from the Maven's error messages.

In summary, the main results of the study are that the size of broken intervals is relatively short, with a median of three commits, although the long-tailed distribution highlights some cases in which the broken intervals can be very long. Looking at the percentage of snapshots that could not be compiled (median 30%), we can easily understand that compilability breaks can seriously affect any kind of research activity studying the evolution of software projects and needing compiled snapshots. Also, similar problems arise for recommender systems requiring code to be compiled and/or executed and at the same time exploiting projects' past history.

The causes behind the majority of the compile broken snapshots can be mostly attributed to the problems with dependencies. While Maven is considered to be one of the most advanced build tools in terms of handling dependencies, the problems still arise, especially when observing the past change history of a project, *e.g.*, libraries are no longer available or URLs are outdated. Again, while this may not be a problem during actual development, this phenomenon can seriously affect any kind of post-mortem analysis either in research or practice.

Results of our study clearly highlight—at least within the investigated set of projects, language (Java), and build tool (Maven)—the major challenges of performing mining studies at commit level should such studies require source code to be compiled. One of the main reasons for that is to

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/smr

be found in dependency resolution problems. In some cases, dependencies are resolved through resources available online whose URLs are outdated at present time. For this reason, a possible way to circumvent this problem would be to resolve the dependency with an updated link to the missing resource. Therefore, researchers interested in performing software evolution analysis could tweak such references in order to successfully compile the project under analysis.

Finally, we found that the snapshots' age highly impacts the likelihood of a successful compilation. Indeed, *Recent Snapshots* are almost twice more likely to compile than *Intermediate Snapshots* and 3.76 times more likely to compile with respect to *Early Snapshots*.

In summary, if we observe one project's past change history today, it is likely that many of the snapshots at the commit level are likely to be incompilable. If compilation is required for our analyses, then we would have to seek the first compilable snapshot. Clearly, this could make our studies less precise because we are no longer observing all the changes (although, fortunately, systems are often not compilable just for few consecutive snapshots). Alternatively, compile broken snapshots (and outdated/missing dependencies in particular) have to be manually fixed, which makes the analysis effort-prone.

Researchers could maximize the number of snapshots successfully analyzed while minimizing the operational costs by focusing only on the recent snapshots of each system under analysis (if the desired analysis allows to discard part of the change history).

In our future work we are planning on enlarging the study, especially with the aim of considering the impact of different types of build tools and, consequently, the projects developed in different programming languages (or even mixes of programming languages), *e.g.*, C/C++ projects using Make as a build tool. Moreover, it is interesting to analyze whether the phenomenon of broken snapshots is different when analyzing branches different from the master branch. In addition, we are also planning on empirically assessing the impact of broken snapshots on reliability of observations when studying past change history of software projects. Last, but not least, we will investigate how the lack of a detailed, compilable history analysis can affect the accuracy of some previously conducted mining studies.

REFERENCES

1. L. Aversano, G. Canfora, L. Cerulo, C. Del Grosso, and M. Di Penta. An empirical study on the evolution of design patterns. In *Proceedings of the 6th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT International Symposium on Foundations of Software Engineering, 2007, Dubrovnik, Croatia, September 3-7, 2007*, pages 385–394, 2007.
2. T. Bakota, A. Beszédes, R. Ferenć, and T. Gyimóthy. Continuous software quality supervision using sourceinventory and columbus. In *Companion of the 30th International Conference on Software Engineering*, ICSE Companion '08, pages 931–932, New York, NY, USA, 2008. ACM.
3. G. Bavota, G. Canfora, M. Di Penta, R. Oliveto, and S. Panichella. How the apache community upgrades dependencies: an evolutionary study. *Empirical Software Engineering*, pages 1–43, 2014.
4. G. Bavota, B. D. Carluccio, A. De Lucia, M. Di Penta, R. Oliveto, and O. Strollo. When does a refactoring induce bugs? an empirical study. In *12th IEEE International Working Conference on Source Code Analysis and Manipulation, SCAM 2012, Riva del Garda, Italy, September 23-24, 2012*, pages 104–113, 2012.
5. N. Bettenburg, W. Shang, W. M. Ibrahim, B. Adams, Y. Zou, and A. E. Hassan. An empirical study on inconsistent changes to code clones at the release level. *Sci. Comput. Program.*, 77(6):760–776, 2012.
6. R. P. L. Buse and T. Zimmermann. Information needs for software development analytics. In *Proceedings of the 34th International Conference on Software Engineering*, ICSE '12, pages 987–996, Piscataway, NJ, USA, 2012. IEEE Press.
7. M. Cataldo and J. D. Herbsleb. Factors leading to integration failures in global feature-oriented development: An empirical analysis. In *Proceedings of the 33rd International Conference on Software Engineering*, ICSE '11, pages 161–170, New York, NY, USA, 2011. ACM.
8. R. Dyer, H. A. Nguyen, H. Rajan, and T. N. Nguyen. Boa: A language and infrastructure for analyzing ultra-large-scale software repositories. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, pages 422–431, Piscataway, NJ, USA, 2013. IEEE Press.
9. R. Ferenc, A. Beszedes, and T. Gyimothy. Fact extraction and code auditing with columbus and sourceaudit. In *Software Maintenance, 2004. Proceedings. 20th IEEE International Conference on*, pages 513–, Sept 2004.
10. R. Ferenc, L. Lango, I. Siket, T. Gyimothy, and T. Bakota. Source meter sonar qube plug-in. In *Source Code Analysis and Manipulation (SCAM), 2014 IEEE 14th International Working Conference on*, pages 77–82, Sept 2014.
11. M. Fokaefs, N. Tsantalis, E. Stroulia, and A. Chatzigeorgiou. Jdeodorant: identification and application of extract class refactorings. In *Proceedings of the 33rd International Conference on Software Engineering, ICSE 2011, Waikiki, Honolulu , HI, USA, May 21-28, 2011*, pages 1037–1039, 2011.

Copyright © 0000 John Wiley & Sons, Ltd.

Prepared using smrauth.cls
DOI: 10.1002/smr

12. N. Franke and E. von Hippel. Satisfying heterogeneous user needs via innovation toolkits: the case of apache security software. *Research Policy*, 32(7):1199 – 1215, 2003. Open Source Software Development.
13. A. Hassan and K. Zhang. Using decision trees to predict the certification result of a build. In *Automated Software Engineering, 2006. ASE '06. 21st IEEE/ACM International Conference on*, pages 189–198, Sept 2006.
14. N. Kerzazi, F. Khomh, and B. Adams. Why do automated builds break? an empirical study. In *Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on*, pages 41–50, Sept 2014.
15. M. Kim, D. Cai, and S. Kim. An empirical investigation into the role of api-level refactorings during software evolution. In *Proceedings of the 33rd International Conference on Software Engineering*, ICSE '11, pages 151–160, New York, NY, USA, 2011. ACM.
16. M. Kim, V. Sazawal, D. Notkin, and G. Murphy. An empirical study of code clone genealogies. *SIGSOFT Softw. Eng. Notes*, 30(5):187–196, Sept. 2005.
17. S. Kim, K. Pan, and E. J. Whitehead. E.j.: Micro pattern evolution. In *In: MSR (2006*.
18. I. Kwan, A. Schroter, and D. Damian. Does socio-technical congruence have an effect on software build success? a study of coordination in a software project. *Software Engineering, IEEE Transactions on*, 37(3):307–324, May 2011.
19. S. McIntosh, B. Adams, T. Nguyen, Y. Kamei, and A. Hassan. An empirical study of build maintenance effort. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 141–150, May 2011.
20. S. McIntosh, M. Nagappan, B. Adams, A. Mockus, and A. E. Hassan. A large-scale empirical study of the relationship between build technology and build maintenance. *Empirical Software Engineering*, 20(6):1587–1633, 2015.
21. S. McIntosh, M. Poehlmann, E. Juergens, A. Mockus, B. Adams, A. E. Hassan, B. Haupt, and C. Wagner. Collecting and leveraging a benchmark of build system clones to aid in quality assessments. In *Companion Proceedings of the 36th International Conference on Software Engineering*, ICSE Companion 2014, pages 145–154, New York, NY, USA, 2014. ACM.
22. A. Mockus, R. T. Fielding, and J. D. Herbsleb. Two case studies of open source software development: Apache and mozilla. *ACM Trans. Softw. Eng. Methodol.*, 11(3):309–346, July 2002.
23. M. Nagappan, T. Zimmermann, and C. Bird. Diversity in software engineering research. In *Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, ESEC/FSE'13, Saint Petersburg, Russian Federation, August 18-26, 2013*, pages 466–476, 2013.
24. F. Palomba, G. Bavota, M. Di Penta, R. Oliveto, A. De Lucia, and D. Poshyvanyk. Detecting bad smells in source code using change history information. In *Automated Software Engineering (ASE), 2013 IEEE/ACM 28th International Conference on*, pages 268–278, Nov 2013.
25. S. Phillips, G. Ruhe, and J. Sillito. Information needs for integration decisions in the release process of large-scale parallel development. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1371–1380, New York, NY, USA, 2012. ACM.
26. S. Phillips, T. Zimmermann, and C. Bird. Understanding and improving software build teams. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE 2014, pages 735–744, New York, NY, USA, 2014. ACM.
27. K. Prete, N. Rachatasumrit, N. Sudan, and M. Kim. Template-based reconstruction of complex refactorings. In *26th IEEE International Conference on Software Maintenance (ICSM 2010), September 12-18, 2010, Timisoara, Romania*, pages 1–10, 2010.
28. D. Qiu, B. Li, and Z. Su. An empirical analysis of the co-evolution of schema and code in database applications. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2013, pages 125–135, New York, NY, USA, 2013. ACM.
29. S. Raemaekers, A. van Deursen, and J. Visser. Semantic versioning versus breaking changes: A study of the maven repository. In *14th IEEE International Working Conference on Source Code Analysis and Manipulation, SCAM 2014, Victoria, BC, Canada, September 28-29, 2014*, pages 215–224, 2014.
30. P. C. Rigby, D. M. German, and M.-A. Storey. Open source software peer review practices: A case study of the apache server. In *Proceedings of the 30th International Conference on Software Engineering*, ICSE '08, pages 541–550, 2008.
31. R. Saha, M. Asaduzzaman, M. Zibran, C. Roy, and K. Schneider. Evaluating code clone genealogies at release level: An empirical study. In *Source Code Analysis and Manipulation (SCAM), 2010 10th IEEE Working Conference on*, pages 87–96, Sept 2010.
32. H. Seo, C. Sadowski, S. Elbaum, E. Aftandilian, and R. Bowdidge. Programmers' build errors: A case study (at google). In *Proceedings of the 36th International Conference on Software Engineering*, ICSE 2014, pages 724–734, 2014.
33. J. Sheoran, K. Blincoe, E. Kalliamvakou, D. Damian, and J. Ell. Understanding "watchers" on github. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, pages 336–339, New York, NY, USA, 2014. ACM.
34. J. Śliwerski, T. Zimmermann, and A. Zeller. When do changes induce fixes? *SIGSOFT Softw. Eng. Notes*, 30(4):1–5, May 2005.
35. D. Spinellis. Tool writing: A forgotten art? *IEEE Software*, 22(4):9–11, July/August 2005.
36. N. Tsantalis and A. Chatzigeorgiou. Identification of move method refactoring opportunities. *IEEE Trans. Software Eng.*, 35(3):347–367, 2009.
37. N. Tsantalis, A. Chatzigeorgiou, G. Stephanides, and S. T. Halkidis. Design pattern detection using similarity scoring. *IEEE Trans. Software Eng.*, 32(11):896–909, 2006.
38. T. Wolf, A. Schroter, D. Damian, and T. Nguyen. Predicting build failures using social network analysis on developer communication. In *Proceedings of the 31st International Conference on Software Engineering*, ICSE '09, pages 1–11, Washington, DC, USA, 2009. IEEE Computer Society.
39. R. Wu, H. Zhang, S. Kim, and S.-C. Cheung. Relink: Recovering links between bugs and changes. In *Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering*,

Copyright © 0000 John Wiley & Sons, Ltd.

DOI: 10.1002/smr

ESEC/FSE '11, pages 15–25, New York, NY, USA, 2011. ACM.
40. A. Zaidman, B. V. Rompaey, S. Demeyer, and A. van Deursen. Mining software repositories to study co-evolution of production & test code. In *First International Conference on Software Testing, Verification, and Validation, ICST 2008, Lillehammer, Norway, April 9-11, 2008*, pages 220–229, 2008.
41. T. Zimmermann, S. Diehl, and A. Zeller. How history justifies system architecture (or not). In *Proceedings of the 6th International Workshop on Principles of Software Evolution*, IWPSE '03, pages 73–, Washington, DC, USA, 2003. IEEE Computer Society.
42. T. Zimmermann, S. Kim, A. Zeller, and E. J. Whitehead, Jr. Mining version archives for co-changed lines. In *Proceedings of the 2006 International Workshop on Mining Software Repositories*, MSR '06, pages 72–75, New York, NY, USA, 2006. ACM.
43. T. Zimmermann, P. Weisgerber, S. Diehl, and A. Zeller. Mining version histories to guide software changes. In *Proceedings of the 26th International Conference on Software Engineering*, ICSE '04, pages 563–572, Washington, DC, USA, 2004. IEEE Computer Society.

Copyright © 0000 John Wiley & Sons, Ltd.
Prepared using smrauth.cls

DOI: 10.1002/smr