Equivalent Mutants in Configurable Systems: An Empirical Study

Luiz Carvalho, Marcio Augusto Guimarães, Márcio Ribeiro
UFAL
Maceió, Brazil
{luiz,masg,marcio}@ic.ufal.br

Leonardo Fernandes
UFPE
Recife, Brazil
lfmo@cin.ufpe.br

Mustafa Al-Hajjaji
pure-systems GmbH
University of Magdeburg
Magdeburg, Germany
mustafa.alhajjaji@pure-sytems.com

Rohit Gheyi
UFCG
Campina Grande, Brazil
rohit@dsc.ufcg.edu.br

Thomas Thüm
TU Braunschweig
Braunschweig, Germany
t.thuem@tu-braunschweig.de

ABSTRACT

Mutation testing is a program-transformation technique that evaluates the quality of test cases by assessing their capability to detect injected artificial faults. The costs of using mutation testing are usually high, hindering its use in industry. Previous research has reported that roughly one-third of the mutants generated in single systems are equivalents. In configurable systems, a set of mutation operators that focus on preprocessor directives (e.g., `#ifdef`) has been proposed. However, there is a lack of existing studies that investigate whether equivalent mutants do occur with these operators. To perform this investigation, we provide a tool that implements the aforementioned mutation operators and we conduct an empirical study using 20 C files of four industrial-scale systems. In particular, we provide examples of equivalent mutants and detailed information, such as which mutation operators generate these mutants and how often they occur. Our preliminary results show that nearly 40% of the generated mutants are equivalent. Hence, testing costs can be drastically reduced if the community comes up with efficient techniques to avoid these equivalent mutants.

CCS CONCEPTS

• **General and reference → Verification**; • **Software and its engineering → Software testing and debugging**; *Software verification*; *Software development process management*;

KEYWORDS

Mutation Testing, Configurable Systems, Equivalent Mutants

ACM Reference Format:
Luiz Carvalho, Marcio Augusto Guimarães, Márcio Ribeiro, Leonardo Fernandes, Mustafa Al-Hajjaji, Rohit Gheyi, and Thomas Thüm. 2018. Equivalent Mutants in Configurable Systems: An Empirical Study. In *VAMOS 2018: 12th International Workshop on Variability Modelling of Software-Intensive Systems, February 7–9, 2018, Madrid, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3168365.3168379

VAMOS 2018, February 7–9, 2018, Madrid, Spain

ACM ISBN 978-1-4503-5398-4/18/02…$15.00
https://doi.org/10.1145/3168365.3168379

1 INTRODUCTION

Mutation testing [12] is a program-transformation technique that injects artificial faults to check whether the existing test cases can detect them. The quality of mutation testing largely depends on the used mutation operators. Ideally, mutation operators represent realistic faults, i.e., they mimic faults that programmers usually make. Recently, researchers have proposed to apply mutation testing for highly-configurable systems [2, 7, 16, 22, 26, 29]. Most of the existing approaches mutate feature models with aim of generating an effective set of products or to assessing the quality of the generated ones [7, 16, 22, 26, 29]. Since a large percentage of the faults occur in the domain artifacts (i.e., source code) [1], mutating the source code of configurable systems has a huge potential for fault detection. While conventional mutation operators can be used, special mutation operators are needed to mimic variability-related faults. Thus, mutation operators that focus on preprocessor-based directives (e.g., `#ifdef`) have been proposed [2]. Evaluating these operators showed that they can simulate variability-related faults [2].

Mutation testing suffers from high costs [18]. One kind of mutant increases such costs: the equivalent mutant [15]. An equivalent mutant is useless because it shows the same behavior as the original program [10, 18, 24]. Recent researches have found that the rate of equivalent mutants in single systems might lie between 4% and 39% [24]. In this context, it is unknown how often equivalent mutants occur for industrial-scale configurable systems and which mutation operators produce most equivalent mutants.

To evaluate whether equivalent mutants are indeed a problem for configurable systems, we have implemented a tool, called #MUTAF, to mutate `#ifdefs` in C code. In particular, we have implemented 10 out of 13 mutation operators proposed in prior work [2]. The implemented operators make changes in the variability models, in the domain artifact (i.e., source code) and in the mapping between models and domain artifact. Then, we use the tool to generate mutants in 20 files of four preprocessor-based industrial-scale systems: *OpenSSL*, *Vim*, *lighttpd*, and *nginx*. To identify equivalent mutants, we take the original files and mutate them. Then, we

preprocess both files with all macros (i.e., `#define` directives) combinations to yield all configurations. Afterwards, to check whether the corresponding configurations are equivalent, we use compiler optimization and diff techniques [8, 21, 25, 27]. In case all original configurations are equivalent to the corresponding mutant configurations, we define this mutant as *totally equivalent*. In case only a strict, but non-empty subset of configurations is equivalent, we define the mutant as *partially equivalent*.

In particular, we present how often the total and partial equivalent mutants occur for configurable systems. As a result, we investigate whether equivalent mutants in configurable systems are a problem, i.e. whether researchers need to avoid them. In addition, we investigate which mutation operators lead to more equivalent mutants than other operators. This information can be helpful for testers, because they can avoid applying these operators or use them carefully.

Our results reveal that 6.5% and 31.7% of the generated mutants are partially and totally equivalent, respectively. Regarding the mutation operators, we find that the ones that add and remove `#ifdef` conditions generate more totally equivalent mutants.

In summary, this paper provides the following contributions:

- A mutant generator tool that considers mutation operators for configurable systems; and
- An empirical study to present numbers of partially and totally equivalent mutants in C preprocessor-based configurable systems, i.e., *OpenSSL*, *Vim*, *lighttpd*, and *nginx*. In addition, we provide numbers about the operators that lead to more equivalent mutants than others.

2 EQUIVALENT MUTANTS IN CONFIGURABLE SYSTEMS

In the context of configurable systems, several approaches exploited mutation testing to generate an effective set of products [7, 16] or to assess the quality of a set of generated products [29]. For this purpose, several mutation operators have been proposed to mutate configurable systems. However, most of the existing ones focus mainly on feature models [7, 16, 22, 29]. In general, the existing conventional mutation operators in single-system engineering can also be used to mutate configurable systems by applying these operators to the generated products or to the original source code. Nevertheless, these mutation operators may not able to mimic faults that are caused due to variability [1]. Thus, mutation operators that take the variability into account may cause faults in configurable systems that cannot be triggered using the conventional operators. Recently, mutation operators for preprocessor-based configurable systems have been proposed [2]. In this work, we consider the aforementioned operators to generate mutants in configurable systems.

The costs of mutation testing are usually high due to the large amount of possible mutants [18]. Kintis et al. [21] show that, for some cases, more than a third of these mutants for single systems are equivalents. The equivalent mutant problem has been a barrier that prevents mutation testing from being more widely used. To detect if a program and one of its mutants are equivalent is undecidable [10]. As a result, the detection of equivalent mutants alternatively may have to be carried out by humans, but manually checking mutant equivalence is error-prone (people judged equivalence correctly in about 80% of the cases [6]) and time consuming (approximately 15 minutes per equivalent mutant [31]).

In mutation testing for configurable systems, mutating a configurable program S yields a configurable mutant S'. To analyze whether the mutant is equivalent, we need to compare each configuration yielded by S against each corresponding configuration produced by S' for each valid configuration.[1]

Given this context, we now explain the problem of equivalent mutants in configurable systems and introduce the concepts of *totally* and *partially* equivalent mutants.

We assume a semantics function $[\![S]\!]_c$ that takes a configurable system S and a configuration c to generate the respective product. Two products $[\![S]\!]_c$ and $[\![S']\!]_c$ are equivalent, denoted by $[\![S']\!]_c = [\![S]\!]_c$, if $[\![S']\!]_c$ preserves the observable behavior of $[\![S]\!]_c$. Otherwise, they are not equivalent denoted by $[\![S']\!]_c \neq [\![S]\!]_c$. Furthermore, we define function wf that takes a product $[\![S]\!]_c$ as input and indicates whether $[\![S]\!]_c$ is well-formed or not:

$$wf([\![S]\!]_c) = \begin{cases} true & \text{if } [\![S]\!]_c \text{ has no compilation errors} \\ false & \text{if } [\![S]\!]_c \text{ has compilation errors} \end{cases}$$

Given a configurable mutant M derived by applying a mutation operator o to a configurable system S (i.e., $M = o(S)$) and the set of configurations of S as C_S, we define that the configurable mutant M is *totally equivalent* if and only if

$$(\forall c \in C_S : wf([\![M]\!]_c) \rightarrow [\![M]\!]_c = [\![S]\!]_c) \wedge (\exists c' \in C_S : wf([\![M]\!]_{c'})).$$

Similarly, a configurable mutant M is *partially equivalent* if and only if

$$(\exists c \in C_S : wf([\![M]\!]_c) \wedge [\![M]\!]_c \neq [\![S]\!]_c) \wedge (\exists c' \in C_S : wf([\![M]\!]_{c'}) \wedge [\![M]\!]_{c'} = [\![S]\!]_{c'}).$$

Notice the following two examples that illustrate the definition presented. Figure 1 shows a code snippet from the *vim* configurable system.[2] At the top, we illustrate the original version of the code and a mutant yielded by using the AICC (Adding `#ifdef` Condition around Code) mutation operator [2] (see the added lines, i.e., 138 and 157). In this example, the mutant is *totally equivalent*, as all valid configurations generated by using this mutant are equivalent to all configurations generated by the original. Observing from the tester's perspective this mutant is useless, since no configuration produced a mutated program with a behavior different from the original program. Then all configurations can be discarded.

Figure 2 illustrates a mutant generated by using the RFIC (Removing Feature of `#ifdef` Condition) mutation operator [2]. The mutant removed part of the `#ifdef` condition (see line 15 at the top-left corner of Figure 2). In this second example, we have a *partially equivalent* mutant, as only three out of four valid configurations are equivalent, i.e., `[]`, `[FEAT_CRYPT]`, and `[PROTO, FEAT_CRYPT]`. From the point of view of the tester, this mutant may be interesting to be tested, but only for the configuration that produces behavior different from the original program. In this case, only the equivalent configuration need to be discarded.

[1] Valid configuration in this context means compilable or specified in artifacts such as Feature Models and Configuration Knowledge

[2] https://github.com/vim/vim/blob/edf3f97ae2af024708ebb4ac614227327033ca47/src/crypt_zip.c

Figure 1: Example of totally equivalent mutant. All configurations generated by the mutant (AICC) are the same when compared to all configurations generated by the original code.

Analogously to single systems, equivalent mutants increase costs in configurable systems. This means that executing the test suite against the configurations of the original configurable system will lead to the same results as executing the test suite against the corresponding configurations of the mutant. This way, the equivalent mutant problem is even more challenging for configurable systems, as we need to deal with it in many different configurations.

In this paper, we present an empirical study to raise the awareness of equivalent mutants in configurable systems. In particular, we focus on totally and partially equivalent mutants, as the ones we present in this section.

3 EMPIRICAL STUDY

To better understand the problem of equivalent mutants in configurable systems, we perform an empirical study. In what follows, we present the settings of our study (Section 3.1), the results and discussion (Section 3.2), and threats to validity (Section 3.3).

3.1 Settings

In our study, we intend to answer the following research questions:

Table 1: Open-source configurable systems we use in our study.

System	Domain	LOC
OpenSSL	TSL and SSL protocol	269,621
Vim	Text editor	312,201
lighttpd	Web server	48,043
nginx	Http server	120,459

- **RQ1:** How often do totally and partially equivalent mutants occur for configurable systems?
- **RQ2:** Which mutation operators lead to more equivalent mutants than other operators?

Answering **RQ1** is important to understand the equivalent mutant problem for configurable systems. Answering **RQ2** will help (i) testers to be aware of costly operators and thus avoid them; and (ii) researchers to know which mutation operators need to be improved. To answer these questions, we select 20 C files of four industrial-scale configurable systems that contain preprocessor directives. We list the systems we use in Table 1. To make our analysis feasible, we randomly select files with at most six different feature macros.

Figure 2: Example of partially equivalent mutant. One corresponding configuration is different, i.e., [PROTO].

To generate the mutants, we rely on the mutation operators for configurable systems proposed in prior work [2]. For our evaluation, we have implemented 10 out of 13 operators and packaged them in a tool called #MUTAF (Mutation of `#ifdef`).[3] In Table 2, we present the operators that we have implemented and provide a description for each of them. We did not consider three mutation operators proposed previously: RFDM (Remove Feature from Model), MFDM (Modify Feature Dependency in Model), and CACO (Conditionally Applying Conventional Operator). The reasons for excluding these operators are the following. The operators RFDM and MFDM require feature models, which are not available for the configurable systems used in our evaluation, whereas operator CACO leads to conventional mutants, which can bias the experiment to results similar of previous studies in single systems [21, 27].

We execute #MUTAF on each file to generate the mutants. We then generate all configurations available in both the original files as well as the mutated ones, and perform a comparison as illustrated in Figures 1 and 2. In case a certain configuration does not compile in the original code, we do not analyze such configuration in any mutant. In addition, if the mutant does not compile for a certain configuration, we disregard this configuration from our numbers. For example, suppose a mutant with four configurations.

[3]https://github.com/lzmths/mutaf

Table 2: Mutation operators implemented in #MUTAF.

Mutation Operator	Acronym	Description
Remove Conditional Feature Definition	RCFD	Removes a feature definition.
Add Condition to Feature Definition	ACFD	Adds an #ifdef condition around an existing define statement.
Adding #ifdef Condition Around Code	AICC	Adds an #ifdef or #ifndef condition around a code fragment.
Adding Feature to#ifdef Condition	AFIC	Manipulates an #ifdef condition by inserting an additional logical dependency to the expression.
Remove #ifdef Condition	RIDC	Deletes an #ifdef condition.
Removing Feature of #ifdef Condition	RFIC	Removes the occurrence of a feature from an #ifdef expression.
Replacing #ifdef Directive with #ifndef Directive	RIND	Changes an existing #ifdef directive to an #ifndef directive.
Replacing #ifndef Directive with #ifdef Directive	RNID	Changes an #ifndef directive to an #ifdef directive.
Removing Complete #ifdef Block	RCIB	Deletes an entire #ifdef block.
Moving Code Around #ifdef Blocks	MCIB	Moves a certain code fragment around an #ifdef block.

If three are equivalents and one does not compile, we count this mutant as totally equivalent. This definition may seem strange to some researchers because by having a configuration different from the original, this mutant could not be called totally equivalent. However, we are observing from the tester's perspective. In this case, the mutant is completely useless and is not suitable for use in

mutation testing. Not counting mutants that do not compile is in accordance to mutation testing previous research, as they produce invalid mutations [20].

To detect equivalent mutants, we rely on compiler optimizations for checking equivalence [8, 21, 25]. So, a mutant is equivalent to the original program if, after the transformations made by compile optimization, they end up with the same object code. Notice that this approach is sound in the sense that two equal binaries mean that the programs have the same behavior. However, we do not detect equivalent mutants with the same behavior, but with different object codes. This leads us to have no false positives. But, this approach is definitely not free to have false negatives.

We compile the configurations with the `gcc`[4] compiler. This compiler has many levels of optimization. In this study, we decided to setup the compiler to use the option *-O3*. This option has been used by previous studies to detect equivalent mutants [21, 27]. Then, we check whether two binaries are equivalent by using the `diff`[5] utility with the flag "--binary."

We execute our study on a Linux kernel 4.9.0 , `gcc` 6.3.0-11, and `diff` 3.5-3.

3.2 Results and Discussion

In this section, we present the results and answer our research questions. All results of our study are available online: https://lzmths.github.io/vamos2018/. Figure 3 shows the distribution of percentages of the partially and the totally equivalent mutants for each analyzed system. We observe that large percentages of the mutants are totally equivalent, especially for configurable systems *Vim* and *nginx*, with median values 61.6% and 76.5%, respectively. Table 3 presents the results in more detail and answer **RQ1**. For each file, we illustrate the number of preprocessor macros, the number of valid mutants (the ones that compile), and the number of partially equivalent and totally equivalent mutants.

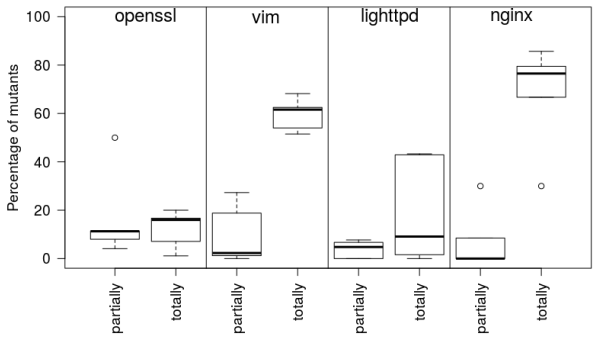

Figure 3: Partially and totally equivalent mutants.

For example, when considering the `crypt_zip.c` file (*Vim*), #MUTAF generated 48 valid mutants. This file has 3 preprocessor macros which give us 8 possible configurations, without considering information from potential constrains (e.g., feature model). Notice that

[4]https://gcc.gnu.org/
[5]https://www.gnu.org/software/diffutils/

Table 3: Number of partially and totally equivalents mutants for configurable systems.

Project	File	Macros	Valid Mutants	Partially Equivalent		Totally Equivalent	
OpenSSL	crypto/rand/rand_lib.c	4	98	4	4.1%	11	11.2%
	ssl/s3_enc.c	3	35	4	11.4%	7	20.0%
	ssl/ssl_sess.c	4	168	19	11.3%	12	7.1%
	ssl/bio_ssl.c	1	6	3	50.0%	1	16.7%
	ssl/ssl_cert.c	3	25	2	8.0%	4	16.0%
Vim	src/hashtab.c	1	86	1	1.2%	53	61.6%
	src/crypt_zip.c	3	48	9	18.8%	30	62.5%
	src/json.c	6	66	0	0.0%	34	51.5%
	src/nbdebug.c	4	22	6	27.3%	15	68.2%
	src/popupmnu.c	4	87	2	2.3%	47	54.0%
lighttpd	src/buffer.c	3	220	17	7.7%	20	9.1%
	src/chunk.c	1	125	6	4.8%	2	1.6%
	src/fdevent_libev.c	1	28	0	0.0%	0	0.0%
	src/fdevent_poll.c	2	30	2	6.7%	13	43.3%
	src/keyvalue.c	1	7	0	0.0%	3	42.9%
nginx	core/ngx_rwlock.c	3	24	2	8.3%	16	66.7%
	core/ngx_string.c	3	10	3	30.0%	3	30.0%
	mail/ngx_mail.c	3	17	0	0.0%	13	76.5%
	http/ngx_http_ups_rr.c	4	56	0	0.0%	48	85.7%
	core/ngx_inet.c	5	73	0	0.0%	58	79.5%
Total		59	1,231	80	6.5%	390	31.7%

18.8% and 62.5% of the mutants are partially and totally equivalent, respectively. The examples we present in Figures 1 and 2 are from the `crypt_zip.c` file.

In general, the number of totally equivalent mutants is higher than the number of partially equivalent mutants. In particular, 6.5% and 31.7% of the mutants are partially and totally equivalent, respectively. In this sense, these results answer our **RQ1** and evidence the importance of creating technical solutions to avoid equivalent mutants in configurable systems and thus reduce costs.

To answer **RQ2**, we refer to Table 4. The mutation operators that contribute most with totally equivalent mutants are AFIC, RFIC, and RIDC. On the other hand, RCIB is the one that most contributes to partially equivalent mutants. To better understand the results, we refer to the code snippets presented in Figure 4. We now proceed by explaining and discussing the results from the operators that caused the highest and lowest numbers.

AFIC adds a feature to `#ifdef` condition. RFIC, on the other hand, removes. In case the original code contains, for example, `#if defined(WIN32)`, AFIC may yield a mutant with the following: `#if defined(WIN32) && defined(BUFFER_SIZE)` (Figure 4(A)). Here, when comparing the original code with the mutant, only one configuration will be different, i.e., when we enable `WIN32` and disable `BUFFER_SIZE`. In our evaluation, this only one "different" configuration (such as enabling `A` and disabling `B`) did not compile in the majority of the cases. Because we do not count the configurations that do not compile, we ended up with many totally equivalent mutants. The same situation happens for the RFIC mutation operator. AFIC and RFIC generated 92.0% and 85.7% of totally equivalent mutants, respectively.

RIDC removes the `#ifdef`. In case the `#ifdef` is encompassing declarations of variables and functions (see Figure 4(B)), they will be unused after applying RIDC and preprocessing the code with the macro disabled (e.g., `HAVE_PCRE_H`). Unused variables and functions are detected and properly removed during the compiler optimization, yielding equivalent mutants. However, when applying RIDC, we may also have compilation errors, i.e., undeclared variables and functions (Figure 4(C)). As configurations that do not compile are

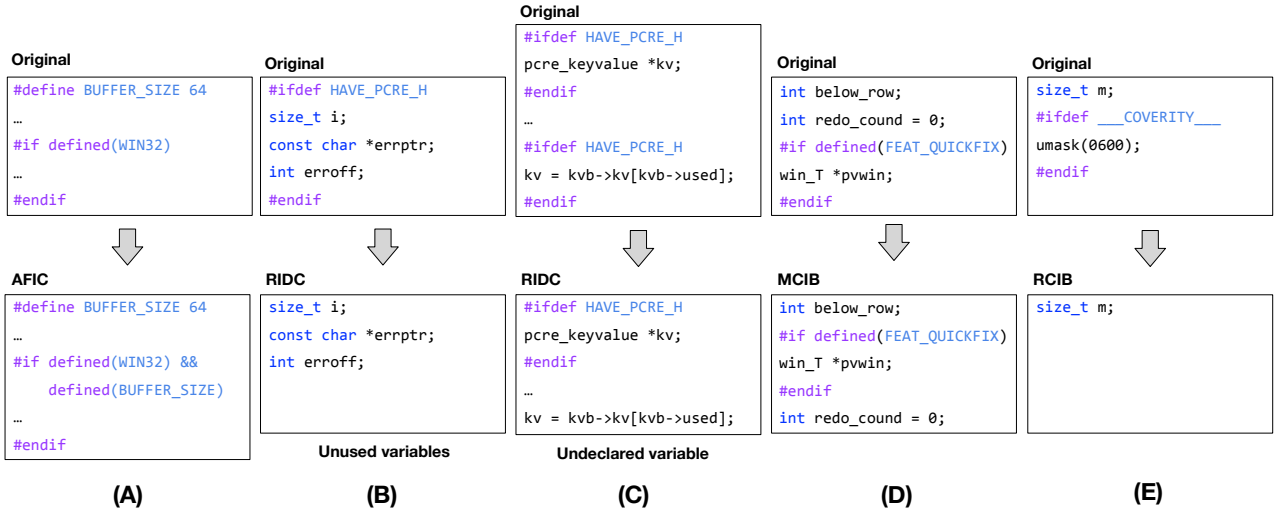

Figure 4: Examples of applying some mutation operators.

not present in our numbers, we increase the odds of having totally equivalent mutants (e.g., for four configurations, three are equivalent and one does not compile). RIDC generated 81.6% of totally equivalent mutants.

MCIB generated a moderate number of totally equivalent mutants (36.6%), especially in cases where the operator moved variable declarations around `#ifdef` blocks (see Figure 4(D)). AICC generated a few number of totally equivalent mutants. In general, the totally equivalent cases happened when the operator adds the `#ifdef` with a macro that has already been defined by using the `#define` directive (see the example presented in Figure 1).

RIND and RNID generated a few number of totally equivalent mutants (7.7% and 10.3%, respectively). The main reason is that the configurations tend to be different when compared to the original code due to the different blocks that are added and removed after preprocessing the code.

RCIB generated a moderate number of partially equivalent mutants, i.e., 39.3%. For example, when removing the entire block encompassed by `#ifdef ___COVERITY___`, the configuration where `___COVERITY___` is disabled is equivalent (Figure 4(E)). On the other hand, in case such macro is enabled, we have a non-equivalent configuration. This balance between equivalent and non-equivalent configurations increases the odds of having partially equivalent mutants.

3.3 Threats to Validity

The low number of configurable systems we used represents a threat to external validity. To alleviate this threat, we selected projects of different sizes and domains. We intend to conduct further experiments in future using different sizes of configurable systems in order to generalize the outcomes of this paper.

Our implementation of #MUTAF is a threat to internal validity. Because there is no formal specification of the mutation operators, the implementation may vary according to the tool. For example,

Table 4: Number of partially and totally equivalents mutants per mutation operator.

File	Valid Mutants	Partially Equivalent		Totally Equivalent	
AFIC	50	4	8.0%	46	92.0%
RFIC	7	1	14.3%	6	85.7%
RIDC	141	10	7.1%	115	81.6%
RCFD	3	0	0.0%	2	66.7%
ACFD	63	5	7.9%	41	65.1%
MCIB	164	2	1.2%	60	36.6%
AICC	679	34	5.0%	106	15.6%
RCIB	56	22	39.3%	8	14.3%
RNID	29	0	0.0%	3	10.3%
RIND	39	2	5.1%	3	7.7%
Total	1,231	80	6.5%	390	31.7%

for some mutation operators we decided to avoid an exponential growth of the number of mutants: AICC (Adding `#ifdef` Condition Around Code) only applies `#ifdef` and `#ifndef` around functions. AFIC (Adding Feature to `#ifdef` Condition) only adds a logical dependency to the expression in case the added feature is defined in the same file. However, for other mutation operators, we have expanded the scope: RIDC and RCIB do not work only with `#ifdef`. They also consider preprocessors with `#ifndef`, `#if`, and `#else` condition. To check whether the tool is creating the mutants as expected, we sampled five mutants of each mutation operator and manually analyzed them. Moreover, the `gcc` compiler and the `diff` utility may also have faults. Nevertheless, we minimize this threat by relying on those systems that are heavily tested, deployed, and used in practice.

In this paper, we use the brute-force (all possible configurations) in files with at most six macros. It represents a threat because many configurable systems have a dedicated building tool that filters the possible configurations before performing compilation (eg.: `kconfig`). Our intention was to assess whether equivalent mutants represent a problem for configurable systems. Notice that

this approach would not scale in case we decide to analyze files with many more macros.

The aforementioned manual analysis also represents a threat. We alleviate such a threat by double checking the questionable cases with a second researcher.

In our evaluation, we rely on compiler optimizations to check equivalence. Despite sound (no false positives), this technique may have false negatives. In this context, our results are conservative and represent a lower bound.

4 RELATED WORK

Recently, several mutation testing approaches have been proposed for highly configurable systems [2, 4, 7, 13, 14, 16, 29]. Al-Hajjaji et al. [2] propose mutation operators for preprocessor-based configurable systems and show that applying these operators can cause real faults. They aim with these operators to make changes that cause variability-related faults. In our work, we implemented the majority of those mutation operators and use them to generate mutants. We considered these generated mutants in our evaluation.

Arcaini et al. [7] propose a fault-based approach that considers finding faults in feature models. For this purpose, they propose a set of mutation operators that mutate feature models. These mutated feature models are tested against configurations that are generated from the original ones. Reuling et al. [29] present a fault-based approach that generates an effective set of products with respect to the ability of finding faults. To achieve their goal, they propose a set of atomic and complex mutation operators that mutates feature diagrams. Similarly, Henard et al. [16] propose two mutation operators to alter the propositional formula of feature models. Using these operators, they generate a set of configurations that has the ability to detect the mutated feature models. Papadakis et al. [26] report that considering fault-based approaches (i.e., mutation testing) to generate a set of products is more effective than generating products with combinatorial interaction testing with respect to the ability of detecting faults. Lackner et al. [22] assess the testing quality of configurable systems by exploiting mutation testing to measure the capability of detecting faults. With their approach, they consider model-based mutation operators. The aforementioned approaches [7, 16, 22, 26, 29] consider only mutation operators on the feature model level.

However, Abal et al. [1] report that most of the reported faults are located in the domain artifacts (i.e., source code). In addition, they do not consider any techniques that may reduce the mutation testing costs. In our work, we mutate the source code of configurable systems. Furthermore, we investigate how often the equivalent mutants occur in configurable systems and how many equivalent mutants are generated by each mutation operator.

Al-Hajjaji et al. [4] propose to reduce the cost by decreasing the possibility of generating equivalent mutants. For this purpose, they propose to combine static analysis and T-wise testing to indicate in which code segments the mutation operators should be applied. However, further experiments are required to evaluate the effectiveness of their approach. In addition, Reuling et al. [29] propose two strategies to reduce the number of mutants, namely mutation selection and higher-order mutation. With mutation selection, they select a set of operators randomly. Furthermore, they also exploit the similarity notion in operators selection, where they consider dissimilar operators to be selected. However, these strategies can be applied to the implemented mutation operators in future to reduce the mutation testing cost.

For highly configurable systems, several approaches have been proposed to select a set of products to be tested [3, 17, 19, 28]. For example, Johansen et al. [19], Al-Hajjaji et al. [3], Perrioun et al. [28] propose T-wise testing approaches to sample configurable systems. Henard et al. [17] suggest an alternative approach to T-wise testing by proposing a search-based approach that selects a set of products. Our approach can be used to assess the quality of the generated products with respect to the capability to find faults. Furthermore, numerous product prioritization approaches have been proposed to increase the fault detection rate by ordering products under test [5, 23, 30]. These prioritization approaches can be applied to prioritize the generated mutants, which may increase the testing effectiveness.

In previous work, Braz et al. [9] proposes a change-centric approach that aims to compile only the configurations that affected by the changes. This previous approach can be used to avoid the equivalent mutants by considering only configurations that affected by changes as a result of applying the mutation operators.

In single systems, as surveyed by Jia and Harman [18], efforts have been made to reduce the mutation testing costs. For example, some approaches propose that selecting only small percentages of mutants are sufficient to achieve a high accuracy of mutation score [11, 32]. However, applying these techniques of the single-system engineering may achieve promising results in the context of configurable systems. In a recent work [15], we propose an approach to avoid generating equivalent mutants. In particular, we propose rules that are required to be applied during mutants generation. Therefore, the aforementioned approach may be also considered in future in the configurable systems to reduce the mutation testing cost.

5 CONCLUSIONS

We empirically assess whether the equivalent mutant in configurable systems is a problem. In particular, we distinguish between partially and totally equivalent mutants in configurable systems. In our evaluation, we have implemented a tool, called #MUTAF, to mutate `#ifdefs` in C code. Then we execute our tool in *OpenSSL*, *Vim*, *lighttpd*, *nginx*. To check whether the mutants are indeed equivalents, we rely on compiler optimizations and diff techniques. Our results revealed that 31.7% of the mutants are totally equivalent, i.e., no configuration generated by these mutants are useful for the mutation testing, and 6.5% of the mutants are partially equivalents, i.e., only a few configurations would be useful for testing. We also found that mutation operators that add or remove `#ifdef` conditions generate totally equivalent mutants more often. Our results bring evidence that equivalent mutants is a non-negligible problem for configurable systems. In addition, our evaluation is important to make testers aware of costly mutation operators. Last but not least, our findings are important to improve the mutation operators we explored in this paper.

As future work we intend to investigate *duplicated mutants*. Duplicated mutants happen when two mutants are equivalent to

each other. In this case, only one of them is useful. In addition, we intend to improve #MUTAF to: (i) automatically execute the test suite, (ii) compute the mutation score, and (iii) integrate techniques to avoid those kinds of equivalent and duplicated mutants.

ACKNOWLEDGMENTS

We would like to thank the Federal Institute of Alagoas (IFAL) for partially supporting this work. In addition, this work has been partially supported by CAPES/PROCAD grant 175956, CAPES/PGCI grant 117875, CNPq (grants 306610/2013-2, 460883/2014-3, 307190/2015-3, 308380/2016-9, 409335/2016-9), FAPEAL PPGs 14/2016 (60030 000435/2017), INES - National Institute of Science and Technology for Software Engineering, grants CNPq 465614/2014-0, Federal Ministry of Education and Research in project CrESt (funding id: 01|S16043N) , and by the German Research Foundation within the project IMoTEP (grant agreement LO 2198/2-1). The responsibility for the content rests with the authors.

REFERENCES

[1] Iago Abal, Claus Brabrand, and Andrzej Wasowski. 2014. 42 Variability Bugs in the Linux Kernel: A Qualitative Analysis. In *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*. 421–432.

[2] Mustafa Al-Hajjaji, Fabian Benduhn, Thomas Thüm, Thomas Leich, and Gunter Saake. 2016. Mutation Operators for Preprocessor-Based Variability. In *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems*. 81–88.

[3] Mustafa Al-Hajjaji, Sebastian Krieter, Thomas Thüm, Malte Lochau, and Gunter Saake. 2016. IncLing: Efficient Product-line Testing Using Incremental Pairwise Sampling. In *Proceedings of the International Conference on Generative Programming: Concepts Experience*. 144–155.

[4] Mustafa Al-Hajjaji, Jacob Krüger, Fabian Benduhn, Thomas Leich, and Gunter Saake. 2017. Efficient Mutation Testing in Configurable Systems. In *Proceedings of the International Workshop on Variability and Complexity in Software Design (VACE '17)*. 2–8.

[5] Mustafa Al-Hajjaji, Thomas Thüm, Malte Lochau, Jens Meinicke, and Gunter Saake. 2017. Effective Product-Line Testing Using Similarity-Based Product Prioritization. *Software & Systems Modeling* (2017). To appear.

[6] Jr. Allen Troy Acree. 1980. *On Mutation*. Ph.D. Dissertation. Georgia Institute of Technology.

[7] Paolo Arcaini, Angelo Gargantini, and Paolo Vavassori. 2015. Generating tests for detecting faults in feature models. In *Software Testing, Verification and Validation*. 1–10.

[8] Douglas Baldwin and Frederick Sayward. 1979. *Heuristics for Determining Equivalence of Program Mutations*. Technical Report. DTIC Document.

[9] Larissa Braz, Rohit Gheyi, Melina Mongiovi, Márcio Ribeiro, Flávio Medeiros, and Leopoldo Teixeira. 2016. A Change-centric Approach to Compile Configurable Systems with #ifdefs. In *Proceedings of the International Conference on Generative Programming: Concepts and Experiences*. 109–119.

[10] Timothy Budd and Dana Angluin. 1982. Two notions of correctness and their relation to testing. *Acta Informatica* 18, 1 (1982), 31–45.

[11] Timothy Alan Budd. 1980. *Mutation Analysis of Program Test Data*. Ph.D. Dissertation. Yale University, New Haven, CT, USA.

[12] Richard DeMillo, Richard Lipton, and Frederick Sayward. 1978. Hints on Test Data Selection: Help for the Practicing Programmer. *Computer* 11, 4 (1978), 34–41.

[13] Xavier Devroey, Gilles Perrouin, Maxime Cordy, Mike Papadakis, Axel Legay, and Pierre-Yves Schobbens. 2014. A Variability Perspective of Mutation Analysis. In *Proceedings of the International Symposium on Foundations of Software Engineering*. 841–844.

[14] Xavier Devroey, Gilles Perrouin, Mike Papadakis, Axel Legay, Pierre-Yves Schobbens, and Patrick Heymans. 2016. Featured Model-based Mutation Analysis. In *Proceedings of the International Conference on Software Engineering*. 655–666.

[15] Leonardo Fernandes, Márcio Ribeiro, Luiz Carvalho, Rohit Gheyi, Melina Mongiovi, André Santos, Ana Cavalcanti, Fabiano Ferrari, and José Carlos Maldonado. 2017. Avoiding Useless Mutants. In *Proceedings of the 16th International Conference on Generative Programming: Concepts Experience*. 187–198.

[16] Christopher Henard, Mike Papadakis, and Yves Le Traon. 2014. Mutation-based generation of software product line test configurations. In *International Symposium on Search Based Software Engineering*. 92–106.

[17] Christopher Henard, Mike Papadakis, Gilles Perrouin, Jacques Klein, Patrick Heymans, and Yves Le Traon. 2014. Bypassing the Combinatorial Explosion: Using Similarity to Generate and Prioritize T-Wise Test Configurations for Software Product Lines. *IEEE Transaction Software Engineering* 40, 7 (2014), 650–670.

[18] Yue Jia and Mark Harman. 2011. An Analysis and Survey of the Development of Mutation Testing. *IEEE Transactions on Software Engineering* 37, 5 (2011), 649–678.

[19] Martin Fagereng Johansen, Øystein Haugen, and Franck Fleurey. 2012. An Algorithm for Generating T-Wise Covering Arrays from Large Feature Models. In *Proceedings of the International Software Product Line Conference*. ACM, NY, USA, 46–55.

[20] René Just. 2014. The Major mutation framework: Efficient and scalable mutation analysis for Java. In *Proceedings of the International Symposium on Software Testing and Analysis*. 433–436.

[21] Marinos Kintis, Mike Papadakis, Yue Jia, Nicos Malevris, Yves Le Traon, and Mark Harman. 2017. Detecting Trivial Mutant Equivalences via Compiler Optimisations. *IEEE Transactions on Software Engineering* PP, 99 (2017), 1–1.

[22] Hartmut Lackner and Martin Schmidt. 2014. Towards the assessment of software product line tests: a mutation system for variable systems. In *Proceedings of the 18th International Software Product Line Conference: Companion Volume for Workshops, Demonstrations and Tools-Volume 2*. 62–69.

[23] Sascha Lity, Mustafa Al-Hajjaji, Thomas Thüm, and Ina Schaefer. 2017. Optimizing Product Orders Using Graph Algorithms for Improving Incremental Product-line Analysis. In *Proceedings of the International Workshop on Variability Modelling of Software-intensive Systems*. 60–67.

[24] Lech Madeyski, Wojciech Orzeszyna, Richard Torkar, and Mariusz Jozala. 2014. Overcoming the Equivalent Mutant Problem: A Systematic Literature Review and a Comparative Experiment of Second Order Mutation. *IEEE Transactions on Software Engineering* 40, 1 (2014), 23–42.

[25] A. Jefferson Offutt and W. Michael Craft. 1994. Using compiler optimization techniques to detect equivalent mutants. *Software Testing, Verification and Reliability* 4, 3 (1994), 131–154.

[26] Mike Papadakis, Christopher Henard, and Yves Le Traon. 2014. Sampling program inputs with mutation analysis: Going beyond combinatorial interaction testing. In *Software Testing, Verification and Validation*. 1–10.

[27] Mike Papadakis, Yue Jia, Mark Harman, and Yves Le Traon. 2015. Trivial Compiler Equivalence: A Large Scale Empirical Study of a Simple, Fast and Effective Equivalent Mutant Detection Technique. In *IEEE International Conference on Software Engineering*. 936–946.

[28] Gilles Perrouin, Sagar Sen, Jacques Klein, Benoit Baudry, and Yves Le Traon. 2010. Automated and Scalable T-Wise Test Case Generation Strategies for Software Product Lines. In *Proceedings of the International Conference on Software Testing, Verification and Validation*. IEEE, Washington, 459–468.

[29] Dennis Reuling, Johannes Bürdek, Serge Rotärmel, Malte Lochau, and Udo Kelter. 2015. Fault-Based Product-Line Testing: Effective Sample Generation Based on Feature-Diagram Mutation. In *Proceedings of the the International Conference on Software Product Line*. 131–140.

[30] Ana B. Sánchez, Sergio Segura, and Antonio Ruiz-Cortés. 2014. A Comparison of Test Case Prioritization Criteria for Software Product Lines. In *Proceedings of the International Conference on Software Testing, Verification and Validation*. 41–50.

[31] David Schuler and Andreas Zeller. 2013. Covering and Uncovering Equivalent Mutants. *Software Testing, Verification and Reliability* 23, 5 (2013), 353–374.

[32] Lingming Zhang, Milos Gligoric, Darko Marinov, and Sarfraz Khurshid. 2013. Operator-Based and Random Mutant Selection: Better Together. In *Proceedings of the International Conference on Automated Software Engineering*. IEEE, 92–102.