

Bachelor's Proposal

COMMIT FEATURE INTERACTIONS

SIMON STEUER
(2579492)

July 16, 2023

Advisor:

Sebastian Böhm Chair of Software Engineering

Examiners:

Prof. Dr. Sven Apel Chair of Software Engineering

Chair of Software Engineering
Saarland Informatics Campus
Saarland University



PRESENTATION ABSTRACT

Short summary of the contents in English...a great guide by Kent Beck how to write good abstracts can be found here:

<https://plg.uwaterloo.ca/~migod/research/beck00PSLA.html>

INTRODUCTION

Goal of this Thesis

The primary focus of this thesis is to gain an overview of how commits interact with features in software projects. Our goal is to lay basic groundwork regarding this subject, while leaving more detailed questions to future research. As previously mentioned we investigate two types, structural and dataflow-based commit-feature Interactions. While using both types separately can already answer many research questions, we will also show applications utilizing a combination of both. We aim to reveal insights about the development process of features and usage of commits therein with the help of structural commit-feature interactions and high-level repository information. We will also investigate to what extent there exist commit-feature interactions through dataflow that cannot be discovered with a purely syntactical analysis.

Overview

RELATED WORK

Interactions between Features and Interactions between Commits have already been used to answer many research questions surrounding software projects. However investigating Feature Interactions has been around for a long time whereas examining Commit Interactions is a more recent phenomenon.

In an article published in 2023, Sattler et al. [4] analysed several open-source projects with their novel approach, SEAL. SEAL merges low-level data-flow with high-level repository information in the form of Commit Interactions. The paper shows the importance of a combination of low-level Program Analysis and high-level Repository Mining techniques by discussing research problems that neither analysis can answer on its own. For example SEAL is able to detect commits that are central in the dependency structure of a program. This was used to identify small commits affecting central code that would normally not be considered impactful to a program. Furthermore they investigated author interactions at a dataflow level with the help of commit interactions. Thus they can identify interactions between developers that cannot be detected by a purely syntactical approach. They found that, especially in smaller projects, there often exists one main developer authoring the majority of commits and thus, logically, accounting for most author interactions. It was also explained how SEAL makes it possible to relate occurrences of bad programming practices to developers. This is accomplished by SEAL enriching program analyses with computed repository information. Lillack et al. [2] first implemented functionality to automatically track load-time configuration options along program flow. Said configuration options can be viewed analogously to feature variables in our research. Their analysis tool Lotrack can detect which features, here configuration options, must be activated in order for certain code segments to be executed. They evaluated Lotrack on numerous real-world Android and Java applications and observed a high accuracy for the predicted code execution constraints.

Referencing this paper Kolesnikov et al. [1] published a case study on the relation of external and internal feature interactions. Internal feature interactions are control-flow feature

interactions that can be detected through static program analysis as mentioned above. They concluded that considering internal feature interactions could potentially help predict external, performance feature interactions.

BACKGROUND

In this section, we summarize previous research on the topic of code regions and interaction analysis. Thereby we give definitions of terms and discuss concepts that are fundamental to our research. In the next chapter, we will use the introduced definitions and concepts to explain and investigate structural as well as dataflow-based commit-feature interactions.

Code Regions

Software Programs consist of source code lines that are translated into an intermediate representation (IR) upon compilation. IR accurately represents the source-code information and is utilized in many code improvement and transformation techniques such as code-optimization. Furthermore, static program analyses are conducted on top of IR or some sort of representation using IR. Interaction analysis, which is an essential part of our research, is one such type of static program analyses. Inside the interaction analysis employed by us, code regions are used to represent abstract entities of a software project. Code regions are comprised of IR instructions that are directly consecutive in their control-flow. Each code region carries some kind of variability information, detailing its meaning. It should be noted that there can be several code regions with the same meaning scattered across the program. We discuss two kinds of variability information, namely commit and feature variability information allowing us to define two separate code regions.

Commits are used within a version control system to introduce the latest source code changes to its respective repository. Inside a repository revision, a commit encompasses all source code lines that were last changed by it.

Definition 1. The sum of all consecutive instructions, that stem from source code lines belonging to a commit, is called a *commit region*.

As the source code lines changed by a commit are not necessarily contiguous, a commit can have many commit regions inside a program. To properly address the entire set of these commit regions within a program, we introduce the regions of a commit. We say that the *regions of a commit* are comprised of the set of all commit regions that stem from source code lines of the commit.

In general, features are program parts implementing specific functionality. In this work we focus on features that are modelled with the help of configuration variables that specify whether the functionality of a feature should be active or not. Inside a program, configuration variables decide whether instructions, performing a feature's intended functionality, get executed. Detecting these control-flow dependencies is achieved by deploying extended static taint analyses, such as Lotrack.

Definition 2. The sum of all consecutive instructions, whose execution depends on a configuration variable belonging to a feature, is called a *feature region*.

Similarly to commits, features can have many feature regions inside a program, as the instructions implementing their functionalities can be scattered across the program. Here, we also say that the *regions of a feature* are comprised of the set of its feature regions.

Interaction Analysis

In the previous chapter, we have already shown how different abstract entities of a software project, such as features and commits, can be assigned concrete representations within a program. We can use interaction analyses to infer interactions between them by computing interactions between their concrete representations. This can improve our understanding of them and give insights on how they are used inside software projects. An important component of computing interactions between concrete representations is the concept of interaction relations between code regions introduced by Sattler et al. As we investigate structural and dataflow-based interactions, we make use of structural \odot and dataflow \rightsquigarrow interaction relations in this work.

The structural interaction relation between two code regions r_1 and r_2 is defined as follows:

Definition 3. The interaction relation $r_1 \odot r_2$ evaluates to true if at least one instruction that is part of r_1 is also part of r_2 .

The dataflow interaction relation between two code regions r_1 and r_2 is defined as:

Definition 4. The interaction relation $r_1 \rightsquigarrow r_2$ evaluates to true if the data produced by at least one instruction i that is part of r_1 flows as input to an instruction i' that is part of r_2 .

Essential to the research conducted in this paper is the interaction analysis introduced by Sattler et al. [4]. Their interaction analysis tool SEAL is implemented on top of LLVM and PhASAR. SEAL is used to determine dataflow interactions between commits by computing interactions between their respective commit regions. This computation is carried out using the dataflow interaction relation \rightsquigarrow . Their approach for this will be shortly discussed here, however we advice their paper for a more thorough explanation.

The first step of their approach is to annotate code by mapping information about its regions to the compiler's intermediate representation. This information is added to the LLVM-IR instructions during the construction of the IR. SEAL focuses on commit regions, which contain information about the commit's hash and its respective repository. The commit region of an instruction is extracted from the commit that last changed the source code line the instruction stems from. Determining said commit is accomplished by accessing repository meta-data.

The second analysis step involves the actual computation of the interactions. For this, Sattler et al implemented a special, inter-procedural taint analysis. It's able to track data flows between the commit regions of a given target program. For this, information about which commit region affected it, is mapped onto data and tracked along program flow. In this context, we would like to introduce the concept of taints, specifically commit taints. Taints are used to to give information about which code regions have affected an instruction through dataflow. For example, an instruction is tainted by a commit if its taint stems from a commit region. It follows, that two commits, with their respective commit regions r_1 and r_2 , interact through dataflow, when an instruction is part of one's commit region, r_2 , while being tainted by

r_1 . Consequently data produced by the commit region r_1 flows as input to an instruction within the commit region r_2 , matching the dataflow-based interaction relation as defined in **Definition 4**.

COMMIT FEATURE INTERACTIONS

In this section, we define structural and dataflow-based commit-feature interactions as well as properties related to them. Furthermore their meaning and relationship inside a software project is explained here.

In the background chapter we discuss the concept of code regions, especially commit and feature regions. Logically, we speak of structural interactions between features and commits when their respective regions structurally interact. This structural interaction between code regions occurs when at least one instruction is part of both regions. This is the case when code regions structurally interact through the interaction relation \odot .

Definition 5. A commit C with its commit regions r_{1C}, r_{2C}, \dots and a feature F with its feature regions r_{1F}, r_{2F}, \dots structurally interact, if at least one commit region r_{iC} and one feature region r_{jF} structurally interact with each other, i.e. $r_{iC} \odot r_{jF}$.

Structural commit-feature interactions carry an important meaning, namely that the commit of the interaction was used to implement or change the functionality of the feature. This can be seen when looking at an instruction that is both part of a commit as well as a feature region. From **Definition 1**, it follows that the instruction stems from a source code line that was last changed by the commit region's respective commit. From **Definition 2**, we also know that the source line implements the functionality of the feature region's respective feature. This means that the commit contributed to the code space implementing the feature. Following this we can say that the commits a feature structurally interacts with, implement the entire functionality of the feature. That's because each source code line of a git-repository was introduced by a commit. Thus every instruction that is part of a feature region is part of a commit region as well.

Knowing the commits used to implement a feature allows us to determine the authors that developed it. This is made possible by simply linking the commits, a feature structurally interacts with, to their respective authors.

Commit interactions based on dataflow were explained in the **Interaction Analysis** chapter. Now, we define dataflow interactions between commits and features. We find that investigating features affecting commits through dataflow doesn't offer enough interesting use-cases. Instead, we focus solely on commits affecting features through dataflow. Similarly to commits interacting with other commits through dataflow, commits interact with features through dataflow, when there exists dataflow from a commit to a feature region. This means that data produced within a commit region flows as input to an instruction located inside a feature region. This pattern can also be matched with the dataflow interaction relation \rightsquigarrow when defining dataflow-based commit-feature interactions.

Definition 6. A commit C with its commit regions r_{1C}, r_{2C}, \dots and a feature F with its feature regions r_{1F}, r_{2F}, \dots interact through dataflow, if at least one commit region r_{iC} interacts with a feature region r_{jF} through dataflow, i.e. $r_{iC} \rightsquigarrow r_{jF}$.

1. <code>int calc(int val) {</code>	▷ d93df4a	
2. <code>int ret = val + 5;</code>	▷ 7edb283	
3. <code>if (FeatureDouble) {</code>	▷ fc3a17d	▷ FeatureDouble
4. <code>ret = ret * 2;</code>	▷ fc3a17d	▷ FeatureDouble
5. <code>}</code>	▷ fc3a17d	▷ FeatureDouble
6. <code>return ret;</code>	▷ d93df4a	
7. <code>}</code>	▷ d93df4a	

Listing 1: Commit Feature Interactions

The code example contains both structural as well as dataflow-based commit feature interactions. It's obvious that commit fc3a17d implements the functionality of FeatureDouble for this function. It follows that a structural CFI can be found between them, as their respective commit and feature region overlap. Commit 7edb283 introduces a new variable that is later used inside the feature region of the "Double"-Feature. This accounts for a CFI through dataflow, as data that was produced within a commit region is used as input inside an instruction of a feature region later on in the program.

When investigating dataflow-based commit-feature interactions, it is important to factor in that structural interactions heavily coincide with dataflow-based interactions. This means that whenever commits and features structurally interact, they are likely to interact through dataflow as well. This becomes clear when looking at an instruction accounting for a structural commit-feature interaction. From **Definition 6.**, we know that the instruction belongs to a commit region of the interaction's respective commit. It follows that data changed inside the instruction produces commit taints for instructions that use the data as input. Now, if instructions that use the data as input are also part of a feature region of the interaction's respective feature, the commit and feature of the structural interaction will also interact through dataflow. However, such dataflow is very likely to occur, as features are functional units, whose instructions build and depend upon each other. Knowing this we can differentiate between dataflow-based commit-feature interactions that occur within the regions of a feature and those where data flows from outside the regions of a feature into it. From prior explanations, it follows that this differentiation can be accomplished by simply checking whether a commit that influences a feature through dataflow, also structurally interacts with it.

When examining commit feature interactions in a project, it is helpful to have a measure that can estimate the scope of a feature. We can use such a measure to compare features with each other and thus put the number of their interactions into perspective. Considering our implementation, it makes most sense to define the scope of a feature as the amount of instructions implementing its functionality inside a program. As feature regions implement a feature's functionality, we can define the scope of a feature as follows:

Definition 7. The sum of all instructions belonging to the regions of a feature is called the scope of said feature.

We do not miss any instructions implementing a feature's functionality with this definition, as each such instruction is part of a respective feature region. It's possible to calculate the defined scope of a feature by calculating the amount of instructions in which structural

commit-feature interactions occur. **That's** because every instruction that is part of a feature region accounts for a structural commit-feature interaction, as every instruction is part of a commit region as **well as** shown in beginning of this section. This implies that the instructions belonging to the regions of a feature are the same instructions accounting for the structural commit feature interactions between the feature and other commits.

Implementation

The detection of structural as well as dataflow-based commit-feature interactions is implemented in VaRA [3]. Additionally to commit regions, VaRA maps information about its feature regions onto the compiler's IR during its construction. Commit regions contain the hash and repository of their respective commits, whereas feature regions contain the name of the feature they originated from. VaRA also gives us access to every llvm-IR instruction of a program and its attached information. Thus, structural commit-feature interactions of a program can be collected by iterating over its compiled instructions. According to **Definition 5**, we can store a structural interaction between a commit and a feature, if an instruction is part of a respective commit and feature region. For each such interaction we also save the amount of instructions it occurs in. This is accomplished by incrementing its instruction counter if we happen to encounter a duplicate.

With this, it's possible to calculate the scope of a feature by iterating over the found structural commit-feature interactions. Thereby we can increase the scope of a feature when we encounter an interaction of the feature by the interaction's instruction counter.

In the *Interaction Analysis* section we have discussed the taint analysis deployed by SEAL. **VaRA uses the same analysis**, which computes information about which code regions have affected an instruction through dataflow. Focusing on commit regions allows us to extract information about which commits have tainted an instruction. Thus, dataflow-based commit-feature interactions can also be **iteratively** collected on instruction level. According to **Definition 6**, we can store a dataflow-based interaction between a commit and a feature, if an instruction has a respective commit taint while belonging to a respective feature region. Consequently data, that was changed by a commit region earlier in the program, flows as input to an instruction belonging to a feature region.

For our research we examine numerous software projects to get a wide range of reference data, as commit-feature interactions could potentially vary greatly between different code spaces. Accordingly, the VaRA-Tool-Suite was extended making it possible to generate a report comprising all found CFIs of an according type in a software project. This aids us in examining several software projects to gain sufficient and sensible data about commit-feature interactions. The created reports are also evaluated in the VaRA-Tool-Suite, which offers support to process and display statistics of the generated data.

METHODOLOGY


Research Questions



RQ1: How do commits interact with features structurally?

We intend to research two main properties which already provide a lot of insight into the development process of features and usage of commits therein. Firstly, we examine the amount of commits, features interact with structurally. This gives us a direct estimate on how many commits were used in the development of a feature. Our analysis also allows us to measure the scope of feature, which can put the amount of commits used to implement a feature into perspective. Secondly, we want to examine how many features a commit interacts with structurally, e.g. how many features a commit usually changes. This is especially interesting when considering best practices surrounding the usage of commits. **It is preferred to keep commits granular** meaning they should only fix a single bug or, in our case, change a single feature. Acquiring data on this issue might show how strictly this policy is enforced in the development of features.

RQ2: How do commits interact with features through dataflow?

Investigating dataflow can unveil interactions between parts of a program that were previously hidden from programmers. This can help a programmer understand the extend to which new changes affect other parts of a program. Deploying the introduced analysis in an **ad-hoc, functional manner** could even aid a programmer when fixing bugs. Bugs occurring in certain parts of a program could be traced back to their cause by factoring in recent changes affecting said parts through dataflow. 

Previous research has laid the groundwork for researching dataflow interactions between different parts of a program. However, it has focused solely on dataflow interactions between commits. That's why we want to provide first insights on the properties of dataflow-based commit-feature interactions. Specifically, we investigate how connected commits and features are by analyzing the amount of features a commit usually affects through dataflow. Knowing what fraction of all commits contributing code to a project are part of dataflow-based interactions can show how often new commits affect the data of a feature. Regarding this, it is worth considering that commits constituting code of a feature are very likely to influence said feature through dataflow. Since dataflow interactions coinciding with structural interactions are so obvious, programmers are also much more aware of them. Depending on the prevalence of feature-regions in a project's code space, this could heavily skew the data in one direction, as a large portion of all dataflow interactions would stem from these obvious interactions. Therefore we want to especially focus on commits that aren't part of a feature giving us more valuable and insightful information.

RQ3: How do authors implement features?

Usually there are many programmers working on the same software project, implementing different features, sometimes alone, sometimes with the help of colleagues. We want to shine some light on the exact statistics of this by combining structural commit-feature interactions with high-level repository information. One major question we want to answer is how many authors implement a feature on average, where considering feature-scope could help put this

data into perspective. The collected results could serve as advice for software companies on how to allocate programmers on to-be implemented features.

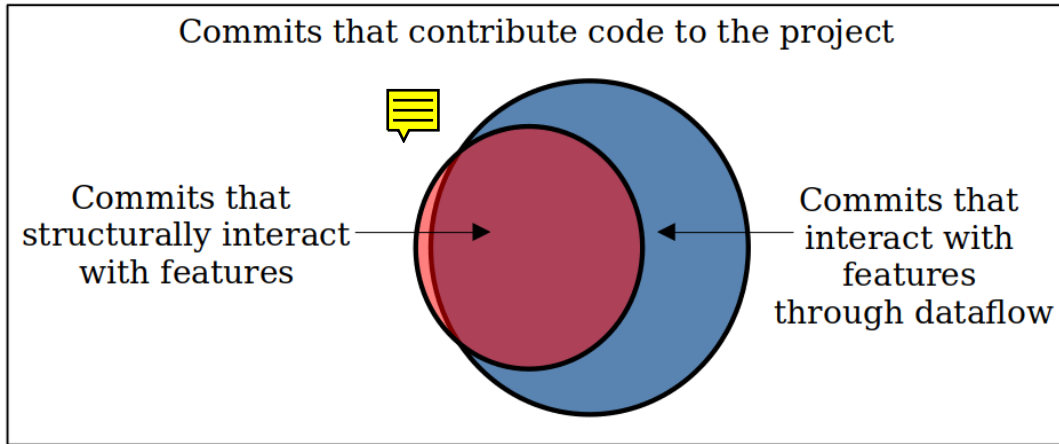


Figure 1: Kinds of Commits in a Software Project investigated in this work

In the first two **RQs** we have discussed different kinds of commits and the ways in which they interact with features. Figure 1 showcases them in a venn diagram and illustrates the dependencies and divisions between them.

Operationalization

RQ1: What are the characteristics of structural commit-feature interactions?

The needed data **will** be collected by creating reports comprising all structural commit-feature interactions of a chosen software project. The collected data is processed into statistics of a desired form which are then displayed graphically to facilitate a faster and better understanding of them. Concerning the first property of structural commit-feature interactions mentioned by us, displaying the amount of commits used to implement a feature in a normal distribution gives us a comprehensive overview of it. The discussed correlation with the scope of a feature is best depicted in a regression plot, which allows the reader to quickly estimate the strength of their correlation.

RQ2: How do commits interact with features through dataflow?

The projects investigated for dataflow-based commit-feature interactions will be the same projects investigated for structural commit-feature interactions. This choice will allow more insight into a single project and allow us to combine both analysis results as will be discussed below. In **RQ2** we consider all commits that currently contribute code to the project, which we can extract from high level repository information of the project. For each commit we will save whether and if true which features they interact with through dataflow. Similarly to **RQ1**, this is carried out by iterating over the dataflow-based commit-feature interactions in the created reports. The acquired information makes it possible to calculate what fraction of commits interact with features through dataflow. For commits that do have dataflow-based interactions with features, we will examine how many features they interact with on average. The discussed separation for said commits into those that are part of a feature and those that

aren't is accomplished with the usage of already created structural reports. We can find out whether a commit is part of a feature by checking if it is part of a structural commit-feature interaction in the according report.

RQ3: How do authors implement features?

Here, we will examine the same projects as the previous **RQs**. That way we can reuse data produced in **RQ1** to map each feature to the authors that implemented it. In **RQ1** we have already mapped each feature to the commits it interacts with, e.g. that contribute code to it. It's possible to retract the authors of these commits by searching through high-level repository information with their hashes. This will directly give us the authors that implemented a feature. The amount of instructions that stem from code belonging to a feature has also been calculated for **RQ1**. With this information we can correlate the size of a feature with the amount of developers that implemented it. Furthermore we want to estimate the amount of code a developer contributes to a feature. To accomplish this, we adapt the analysis used to map each feature to the authors that implement them. When iterating over the commit-feature interactions we not only save the commit, but additionally save the amount of instructions of that interaction. When extracting the authors from the commits that were mapped to a feature, we also add up the amount of instructions for each commit. Now we can estimate the amount of code an author contributes to a feature with the amount of instructions stemming from said code.

Expectations

How and to which extend **features are used** in the projects to be examined is not known and could potentially vary from project to project. Thus, some results of the discussed research topics are difficult to predict. For example, nesting of feature regions inside each other could lead to an increase in the amount of features a commit usually changes. Due to the discussed best practices of commits, we expect commits to change at most one feature on average if there happens to be little nesting. Because of the unknown size of features, it's not sensible to give an estimate about the amount of commits needed to implement a feature. We expect a rather strong positive correlation between the scope of a feature and its commits however. It was mentioned that we examine small projects meaning that the pool of developers is limited in size. Normally, features only encompass a **tiny share of a project's overall code**. Besides that they implement specific functionality that some programmer's might have a better understanding of than others. This leads us to the expectation that a feature is implemented by a small share of all developers contributing to a project. Because of the small pool of developers and prior research findings, we expect the existence of a main developer that contributes most if not all of the functionality of a feature. We know that commits structurally interacting with features most likely are part of dataflow interactions with them as well. Excluding such commits, the extend to which commits interact with features through dataflow depends heavily on what fraction of the code space is made up of feature regions. The purpose of features is to implement additional and sometimes necessary functionality separate from the main program. For this they access and change specific data according to their intended functionality. Provided that feature regions only make up a small portion of the program,

we do expect relatively few, albeit important dataflow interactions between commits and features.

Threats to Validity

There are some potential threats to the internal validity of our gathered data, which stem from our implementation in VaRA.

From **Definition 2** of feature regions, it follows that we implement feature regions in such a way that any instruction whose execution depends on a configuration variable, is part of a feature region. However, **not every such instruction also implements the functionality of a feature**, meaning that feature regions overapproximate the amount of instructions responsible for a feature's functionality. Since feature regions are used for computing both structural and dataflow-based commit-feature interactions, they are overapproximated to some **extend** as well. Thus, it's possible that commits of structural interactions don't actually implement functionality of a feature and commits of dataflow interactions don't actually affect instructions implementing a feature's functionality through dataflow. **Furthermore our deployed taint analysis does not necessarily detect all dataflows occurring in a program**. This results in taints being underapproximated, meaning that some instructions are not tainted when they correctly should be. Thus, some dataflow interactions could be missed by our deployed commit-feature interaction detection.

Concerning the external validity of our findings, most dangers come from the selection of which projects we investigate. Our pool of investigated projects is limited and it's likely that the way commits and features are used in them is different to other projects to some extent. In previous chapters we have discussed that the chosen projects are rather small and **are of a compression domain**. This could mean that our findings might not be applicable for projects of larger size or of different domains. As we already factor in the scope of a feature in the analysis of our data, we are able mitigate some doubts about the applicability of our results onto larger projects, as we can scale them accordingly.

CONCLUSION

In this work we ~~want to~~ research the main properties of structural and dataflow-based commit-feature interactions. Using **high-repository** information and a combination of **both types** allows us to gain additional knowledge on their properties. Following this, we aim to interpret the findings in a sensible way. Structural interactions and the injection of author information within them can be utilized to provide insights into feature development and usage of commits therein. Dataflow interactions can unveil interactions between features and commits that cannot be discovered through a purely structural analysis. Seeing how common they really are could encourage programmers to be more aware of them. Furthermore, they can improve our understanding on which impact new commits have on features.

Research involving interactions inbetween features and commits has shown that investigating interactions between program entities is a topic worthy of study. For example, **Sattler et al** used dataflow commit interactions to allow for a more detailed understanding of author interactions in software projects. Commit interactions also make it possible to identify seemingly insignificant changes that have a central impact on the program. Kolesnikov et al provides

further indication for the wide range of subjects interactions can be used for. Particularly, they argued that control-flow interactions between features can help predict performance interactions between them.

We extend VaRA to implement the detection of structural and dataflow-based commit-feature interactions. As we want to investigate several projects, we create reports containing all found interactions for them. These reports are created inside the VaRA-Tool-Suite and contain either all structural or dataflow-based interactions of a software project. The VaRA-Tool-Suite also allows use to work with the collected data and enrich it with high-repository information, such as information on which author a commit belongs to.



BIBLIOGRAPHY

- [1] Sergiy Kolesnikov, Norbert Siegmund, Christian Kästner, and Sven Apel. “On the relation of external and internal feature interactions: A case study.” In: *arXiv preprint arXiv:1712.07440* (2017).
- [2] Max Lillack, Christian Kästner, and Eric Bodden. “Tracking load-time configuration options.” In: *Proceedings of the 29th ACM/IEEE international conference on Automated software engineering*. 2014, pp. 445–456.
- [3] Florian Sattler. *VaRA is an analysis framework that enables users to build static and dynamic analyses for analyzing high-level concepts using advanced compiler and analysis technology in the background*. <https://vara.readthedocs.io/en/vara-dev/> [Accessed: (24.05.2023)]. 2023.
- [4] Florian Sattler, Sebastian Böhm, Philipp Dominik Schubert, Norbert Siegmund, and Sven Apel. *SEAL: Integrating Program Analysis and Repository Mining*. 2023.