

Dependent or Not: Detecting and Understanding Collections of Refactorings

Thiago Ferreira , James Ivers , Jeffrey J. Yackley , Marouane Kessentini , Ipek Ozkaya, *Senior Member, IEEE*, and Khoulood Gaaloul 

Abstract—Refactoring is a program transformation to improve the internal structure of a program while preserving its external behavior. Developers frequently apply multiple refactorings that depend on each other to achieve goals such as improving code reusability. Although manually applying a sequence of dependent refactorings is a common practice, existing refactoring recommendation tools treat refactorings in isolation without revealing the dependencies among them to developers. One reason is that these relationships among refactorings are poorly understood. Current approaches treat refactoring recommendations as a strictly ordered sequence limiting developers' ability to understand, validate, and apply recommended refactorings. To address this gap, this paper describes a theory for reasoning about collections of refactorings through defining an ordering dependency relation among refactorings and organizing collection of refactorings as a set of refactoring graphs. We propose an algorithm for identifying refactoring dependencies and illustrate these concepts with a tool for visualizing such refactoring dependencies and refactoring graphs. Our validation results demonstrate that 43% of the 1,457,873 recommended refactorings from 9,595 projects that we studied are part of dependent refactoring graphs. Furthermore, refactorings are not only commonly involved in dependent relations, but also when applied, dependent refactoring graphs improve all of the quality attribute metrics in our experiments more than individual refactorings.

Index Terms—Dependency, refactoring, search based software engineering.

I. INTRODUCTION

EVEN for the most competent organizations, building and maintaining high performing software applications with high quality-code is a challenging and expensive endeavor [1]. Working in a fast-paced environment that demands frequent releases across several products and deployment environments, developers are often forced to compromise high quality standards

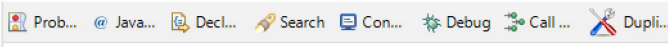
Manuscript received 23 December 2021; revised 7 December 2022; accepted 23 January 2023. Date of publication 17 February 2023; date of current version 14 June 2023. This work was supported by the Department of Defense under Grant FA8702-15-D-0002, with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. Recommended for acceptance by L. Tan. (Corresponding author: Marouane Kessentini.)

Thiago Ferreira and Jeffrey J. Yackley are with the College of Innovation & Technology, University of Michigan-Flint, Flint, MI 48502 USA (e-mail: thiagod@umich.edu; jyackley@umich.edu).

James Ivers and Ipek Ozkaya are with the Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: jivers@sei.cmu.edu; ozkaya@sei.cmu.edu).

Marouane Kessentini and Khoulood Gaaloul are with the Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309 USA (e-mail: marouane@umich.edu; kgaaloul@umich.edu).

Digital Object Identifier 10.1109/TSE.2023.3244123



Refactoring Type	Source Entity	Target Class
Move Method	SmellDetector.SourceModel.Resolver:...	SmellDetector.SourceMod...
Move Method	SmellDetector.SourceModel.Resolver:...	SmellDetector.SourceMod...
Move Method	SmellDetector.smells.designSmells.Ab...	SmellDetector.metrics.Typ...
Move Method	SmellDetector.smells.ThresholdsParse...	SmellDetector.smells.Thres...
Move Method	SmellDetector.smells.implementation...	SmellDetector.SourceMod...
Move Method	SmellDetector.SourceModel.Resolver:...	SmellDetector.SourceMod...

Fig. 1. Output Example of a refactoring recommendations tool: JDeodorant.

in favor of meeting deadlines [2]. Lack of robust automated tools results in buggy and poor quality software that causes financial losses, high maintenance costs, increased fault-proneness, and delayed or canceled projects [3]. Software refactoring is widely recognized as an effective approach for maintaining high quality software by restructuring existing code without changing its external behavior [4].

Resolving code smells and broader code quality improvements across a project often requires applying multiple refactorings, which has lead to the creation of tools that recommend collections of refactorings to developers [5]. Manually applying a sequence of refactorings is common practice [5], [6], [7], however these tools treat each refactoring in the sequence in isolation. For instance, Cinnéide et al. [8] investigate the impact of only individual refactorings on quality attribute metrics, such as using Move Method to reduce the coupling of a class, without studying the impact of a sequence of refactorings. Bibiano et al. [6] evaluates the relationships between refactoring types (e.g., Move Class, Extract Interface) and code smells. However, the study is based on the assumption that refactorings are only related if applied to the same code location, e.g., class. However, most refactoring types in fact modify multiple code fragments e.g., Move Method modifies two classes, the source and target of the move.

Current tools generate extensive lists of refactorings as recommendations to developers. However, these lists present refactorings as a sequence, with an implied strict ordering among refactorings. In practice, however, many of these orderings are not significant while others retain significant meaning. Developers lack a theory to tell the difference, which (among other things) makes refactoring recommendations harder to understand than is necessary. Fig. 1 shows an output example of the refactoring

recommendations of JDeodorant [9] where, similar to existing automated refactoring tools, the dependencies between the refactoring are not revealed, thus leaving the challenging task of interpreting the sequence of refactoring recommendations to the programmers.

The dependencies between refactorings is critical when developers select which refactorings to be applied as this helps them understand how a sequence of refactorings is dependent and what is the impact of making changes in some of the recommended refactorings. The integration of refactoring dependencies in recommender tools is still lacking in existing research.

To close this gap, in this paper we describe a theory for reasoning about collections of refactorings through a definition of ordering dependencies among refactorings and an algorithm for identifying these dependencies. We aim to improve the accuracy of refactoring recommendation tools by detecting refactoring dependencies, which allows the developers to efficiently interact with such refactoring recommendation tools. We propose defining refactoring recommendations as sets of refactoring graphs rather than as refactoring sequences. We illustrate these concepts with a tool for visualizing refactoring dependencies and refactoring graphs.

Refactorings, when formalized, have clear pre-conditions defining the circumstances in which they can be applied and post-conditions defining the effects of applying them. For instance, one of the pre-conditions of a Move Method refactoring is that the method exists in the class from which it will be moved and one of its post-conditions is that the method must exist in the target class afterward. Therefore, a refactoring dependency exists when any post-condition of one refactoring matches any pre-condition of another refactoring e.g., the method exists in the relevant class. These linked refactorings then can be organized into groups based upon their dependencies. We represent these groups as directed acyclic graphs, where the nodes are the refactorings and the edges are the dependencies. This approach offers three main benefits: 1) developers can quickly and intuitively understand the dependencies among refactorings that constrain recommendations e.g., which refactorings must be done together; 2) developers can more easily compare recommendations by focusing on essential, rather than accidental, differences; and 3) tool builders can easily integrate new features to detect invalid refactorings and improve their recommendation algorithms.

We validated our proposed theory based on 1,457,873 refactorings recommended for 9,595 Java projects publicly available on GitHub. We considered 14 types of refactorings that are most commonly used in practice [10], [11]. We also developed a web tool, DPRef, that implements the proposed ordering dependency detection algorithm. It transforms a refactoring sequence recommended by existing refactoring tools [9], [12] into refactoring graphs based on the detected dependencies. We conducted experiments to evaluate the correctness of detected dependencies, discover what portion of refactorings in recommendations are actually dependent rather than independent, and estimate the potential impact of dependent refactorings on several quality

attribute metrics. Finally, we conducted human validation with 27 developers to manually evaluate the correctness of the detected dependencies and their relevance.

Our implemented algorithm achieved 100% in correctly detecting all dependencies between refactorings and identifying invalid refactorings. Furthermore, our findings demonstrate that 43% of the 1,457,873 recommended refactorings are part of dependent refactoring graphs. This finding confirms that refactorings are commonly involved in dependent relations and cannot be applied truly independently. Furthermore, dependent refactorings improve all six QMOOD quality attribute metrics [13] in our experiments better than independent refactorings. The manual validation of the refactorings by 27 developers shows that all the identified dependencies are correct for a sample of 233 refactorings after applying them directly on the code of 61 open source projects based on the order proposed by DPRef. The post-study survey with the developers confirmed the relevance of detecting the dependencies to help them understand the sequence of recommended refactorings.

The authors also provide a *replication package*¹ that includes the refactoring dependency detection tool and necessary data for our large scale validation. The replication package will enable researchers and tool builders to integrate the refactoring dependency feature into existing refactoring recommendation and detection tools and further investigate the relationships among refactorings.

The rest of the paper is organized as follows: Section II discusses the related work and presents a motivating example; Section III provides our definitions of refactoring dependencies and an algorithm to detect them; Section IV presents and discusses the obtained results; Section V highlights the threats to validity; Section VI summarizes our research agenda; and Section VII concludes.

II. RELATED WORK AND MOTIVATING EXAMPLE

Catalogs like Fowler's [4] have identified many types of refactorings e.g., Move Method, Extract Class, Pull Up Field, each of which is a semantics preserving code transformation that improves code structure. Developers routinely apply such refactorings in their day-to-day work, with modern development environments providing limited support for applying selected refactorings as directed by a developer. In this section, we survey two categories of refactoring research: recommendation tools and refactoring dependencies. We then summarize the challenges addressed in this paper through a motivating example.

A. Related Work

1) *Refactoring Recommendations*: There has been both industry and research interest in developing automated and semi-automated refactoring tools to support developers [14]. One representative example is JDeodorant, the tool proposed by

¹<https://sites.google.com/umich.edu/refactoring-dependency/home>

TABLE I
REFACTORING TYPES AND THEIR PRE- AND POST-CONDITION RULES

Refactoring	Cond.	Rules
Move Method (c1, c2, m)	Pre	exist(c1, c2, m) && NOT(inheritanceHierarchy(c1, c2)) && defines(c1, m) && NOT(defines(c2, sig(m)))
	Post	NOT(defines(c1, m)) && defines(c2, m)
Move Field (c1, c2, f)	Pre	exist(c1, c2, f) && NOT(inheritanceHierarchy(c1, c2)) && defines(c1, f) && NOT(defines(c2, f))
	Post	NOT(defines(c1, f)) && defines(c2, f)
Pull Up Field (c1, c2, f)	Pre	exist(c1, c2, f) && isSuperClassOf(c2, c1) && defines(c1, f) && NOT(defines(c2, f))
	Post	defines(c2, f) && NOT(defines(c1, f))
Pull Up Method (c1, c2, m)	Pre	exist(c1, c2, m) && isSuperClassOf(c2, c1) && defines(c1, m) && NOT(defines(c2, sig(m)))
	Post	defines(c2, m) && NOT(defines(c1, m))
Push Down Field (c1, c2, f)	Pre	exist(c1, c2, f) && isSuperClassOf(c1, c2) && defines(c1, f) && NOT(defines(c2, f))
	Post	defines(c2, f) && NOT(defines(c1, f))
Push Down Method (c1, c2, m)	Pre	exist(c1, c2, m) && isSuperClassOf(c1, c2) && defines(c1, m) && NOT(defines(c2, sig(m)))
	Post	defines(c2, m) && NOT(defines(c1, m))
Extract Class (c1, c2, {elements})	Pre	exist(c1) && NOT(exist(c2)) && methods(c1) > 2 && $\forall e \in \text{elements: defines}(c1, e)$
	Post	exist(c2) && $\forall e \in \text{elements: defines}(c2, e)$ && NOT(defines(c1, e))
Extract Super Class (c1, c2, {elements})	Pre	exist(c1) && NOT(exist(c2)) && methods(c1) > 2 && $\forall e \in \text{elements: defines}(c1, e)$
	Post	exist(c2) && isSuperClassOf(c2, c1) && $\forall e \in \text{elements: defines}(c2, e)$ && NOT(defines(c1, e))
Extract Sub Class (c1, c2, {elements})	Pre	exist(c1) && NOT(exist(c2)) && methods(c1) > 2 && $\forall e \in \text{elements: defines}(c1, e)$
	Post	exist(c2) && isSuperClassOf(c1, c2) && $\forall e \in \text{elements: defines}(c2, e)$ && NOT(defines(c1, e))
Encapsulate Field (c1, f)	Pre	exist(c1) && defines(c1, f) && NOT(defines(c1, set_f)) && NOT(defines(c1, get_f))
	Post	defines(c1, set_f) && defines(c1, get_f)
Decrease Field Security (c1, f, newSecLvl)	Pre	exist(c1) && defines(c1, f) && ((newSecLvl == "public" $\implies$ (NOT(isPublic(f)) && isProtected(f))) && (newSecLvl == "protected" $\implies$ (NOT(isProtected(f)) && NOT(isPublic(f)) && isPrivate(f))))
	Post	(newSecLvl == "public" $\implies$ isPublic(f)) (newSecLvl == "protected" $\implies$ isProtected(f))
Decrease Method Security (c1, m, newSecLvl)	Pre	exist(c1) && defines(c1, m) && ((newSecLvl == "public" $\implies$ (NOT(isPublic(m)) && isProtected(m))) && (newSecLvl == "protected" $\implies$ (NOT(isProtected(m)) && NOT(isPublic(m)) && isPrivate(m))))
	Post	(newSecLvl == "public" $\implies$ isPublic(m)) && (newSecLvl == "protected" $\implies$ isProtected(m))
Increase Field Security (c1, f, newSecLvl)	Pre	exist(c1) && defines(c1, f) && ((newSecLvl == "private" $\implies$ (NOT(isPrivate(f)) && isProtected(f))) && (newSecLvl == "protected" $\implies$ (NOT(isProtected(f)) && NOT(isPrivate(f)) && isPublic(f))))
	Post	(newSecLvl == "private" $\implies$ isPrivate(f)) && (newSecLvl == "protected" $\implies$ isProtected(f))
Increase Method Security (c1, m, newSecLvl)	Pre	exist(c1) && defines(c1, m) && ((newSecLvl == "private" $\implies$ (NOT(isPrivate(m)) && isProtected(m))) && (newSecLvl == "protected" $\implies$ (NOT(isProtected(m)) && NOT(isPrivate(m)) && isPublic(m))))
	Post	(newSecLvl == "private" $\implies$ isPrivate(m)) && (newSecLvl == "protected" $\implies$ isProtected(m))

Tsantalis and Chatzigeorgiou [9]. JDeodorant and similar recommendation tools [11], [15], [16], [17], [18], [19] generate recommendations as sequences of refactoring instances. The experiments described in this paper take this form of refactoring recommendation as input. Thus, our discussion in this section focuses on this category of studies. We point the interested reader to the survey by Bavota et al. [20] for an overview of approaches supporting code refactoring recommendations. In another refactoring recommendation tool, O’Keeffe and Cinnéide [21] formulate refactoring tasks as a multi-objective search problem to generate alternative designs by applying a sequence of refactoring operations. Such a search is guided by a quality evaluation function based on eleven object-oriented design metrics that reflect refactoring goals. Harman and Tratt [17] were the first to introduce the concept of Pareto optimality to search-based refactoring. They used it to combine two metrics, namely CBO (Coupling Between Objects) and SDMP (Deviation of Methods Per Class), into a fitness function and showed its superior performance as compared to a mono-objective technique [17]. The two aforementioned studies [17], [21] paved the way for several search-based approaches aimed at recommending refactorings [12], [15], [22], [23], [24], [25].

A representative example of these search-based refactoring techniques is the work by Ouni et al. [12], who propose

a multi-criteria code refactoring approach aimed at optimizing contrasting objectives: (i) minimizing the number of code smells; (ii) minimizing the refactoring cost (i.e., the number of recommended refactorings); (iii) preserving the design semantics (meaning considering textual information embedded in code identifiers and comments in the refactoring recommendation); and (iv) maximizing the consistency with code changes performed over the system’s change history. In this study, we use the refactoring recommendations generated by this tool based on 1) its superior performance compared to the state of the art [12]; 2) the large number of supported refactoring types, and 3) its being publicly available. The contribution of this paper is not generating refactoring recommendations. Any refactoring recommendation approach can be used to generate the refactorings (input of our proposed approach) if they support some or all of the refactoring types summarized in Table I.

2) *Refactoring Dependencies*: Chavez et al. [26] investigated how refactoring types affect five quality attributes based on the version history of twenty-three open source projects. They found that 94% of refactorings are applied to code with at least one low quality attribute value, with 65% of refactorings improving attributes and 35% of all refactorings being neutral on the system. Similarly, Cinnéide et al. [8] studied the impact of individual refactorings on quality attributes, such as using

Move Method to reduce the coupling of a class. None of these studies considered the impact of a sequence of refactorings on the quality attributes.

Murphy-Hill et al. [7] investigated refactoring tool usage through both sampling developers' code and manually checking if their refactorings were performed with tool support and looked at 240,000 tool-assisted refactorings to find assumptions on how developers, in general, refactor code. They ultimately concluded that refactoring tools are rarely used by developers in practice with 90% of refactorings being performed manually, and that 40% of refactorings occur in batches.

Bibiano et al. [6] analyzed batch refactoring characteristics and their effects on code smells in open and closed source projects and concluded that 57% of batches/patterns are simple compositions of only two types of refactorings. They highlight lack of tool support to automatically detect refactoring dependencies as a barrier. However, this study is based on the assumption that refactorings are only related if applied to the same code location, which often is not the case for types of refactorings that modify multiple code fragments.

Mens et al. [27], [28] define and detect mutual exclusions, sequential dependencies and asymmetric conflicts between refactorings. These studies analyze dependencies at the model-level working with UML and they use graph transformation techniques to detect invalid refactorings. The detection of conflicts between refactorings at the model level (UML) is based on a set of rules (matrix where the lines and columns are model refactoring types) that are manually defined. The type of refactorings is different and simplified compared to code-level ones. Furthermore, the authors were looking for mutually exclusive UML refactorings rather than detecting dependencies.

Liu et al. [29], [30] propose a conflict-aware scheduling approach, which schedules refactorings according to the conflict matrix of refactorings and effects of each individual refactoring using a multi-objective optimisation model. In this work, the authors focused on identifying the best schedule to apply refactorings where the conflicts are defined based on which code smells to be fixed. Thus, the same refactorings fixing different code smells or applied in the same locations are grouped together. The notion of dependencies is defined in our work in a different way than Liu et al. where they are more about the conflicts between the refactoring themselves and not their goals.

Sousa et al. [31] identify and analyze composite refactorings within and across commits from the commit history of 48 GitHub software projects. The concept of defining dependencies in this work is different than our paper where the dependency is about grouping the refactorings applied within the same commits/locations.

Overall, existing studies do not provide a rigorous definition of ordering dependencies among refactorings. They mainly define what might be better considering similarity relations, such as a collection of refactorings that have similar effects (fixing a code smell) or similar context (applied by the same developer or to the same code location) [32], [33].

B. Motivating Example

The key to applying refactorings successfully is the decision of which refactorings to apply and where to apply them. In essence, this requires a developer to instantiate a refactoring by supplying the parameters that allow a type of refactoring to be unambiguously applied to code. For example, to instantiate a Move Method, a developer must supply parameters that indicate which method to move and where to move it. Throughout this paper, when we talk about refactorings and dependencies among refactorings, we are talking about refactoring instances.

While refactoring recommendations generated by tools to mimic this activity are typically represented as sequences, not all orderings in these sequences are significant. That is, the same code could be generated by two solutions that contain the same refactorings, but simply apply them in a different order. This is because while many refactorings are independent of one another, other refactorings are dependent on each other such that removing or reordering a refactoring from a solution could make other refactorings invalid. These refactoring recommendations tools typically do not intend to imply a strict ordering of dependencies among elements of the sequence. The sequence is simply a concise means to communicate multiple steps. In that sense, many tools that report a sequence of refactorings today may communicate an unintended meaning (strict sequential order) that this work can help clarify for users of such tools. With the current growth of interactive tools to support refactoring [11], [18], developers are offered solutions that contain dozens to hundreds of refactorings and the option to selectively apply elements of a solution. Without a theory for reasoning about refactoring dependencies, developers can inadvertently make decisions (e.g., ignoring part of a solution) that result in failure (code cannot be successfully refactored) and enter a tedious trial and error loop. Making refactoring dependencies visible improves developers' understanding of how refactorings work together and allows them to make sound inferences regarding their application.

Fig. 2 shows a simplified example of a solution composed of 5 refactorings to be applied to the JFreeChart project. Three of the refactorings (#3, #4, #5) depend on another refactoring (#2) because the Extract Super Class refactoring (#2) creates a new class (Class_7), on which refactorings #3, #4, and #5 operate. If the new class is not created first, then refactorings #3, #4, and #5 will fail. Thus, there exists an ordering dependency from each of #3, #4, #5 to #2. Refactoring #1, however, has distinct parameters, indicating that it operates on different code elements, thus it has no ordering dependencies on any others in this solution. Presenting these dependencies to a developer clarifies the options that the developer has. For example, the developer could choose not to apply any refactoring except for refactoring #2 without consequences; if the developer chooses not to apply refactoring #2, then refactorings #3, #4, and #5 cannot be applied either. Detecting ordering dependency relationships among refactorings is essential to more effectively applying refactorings.

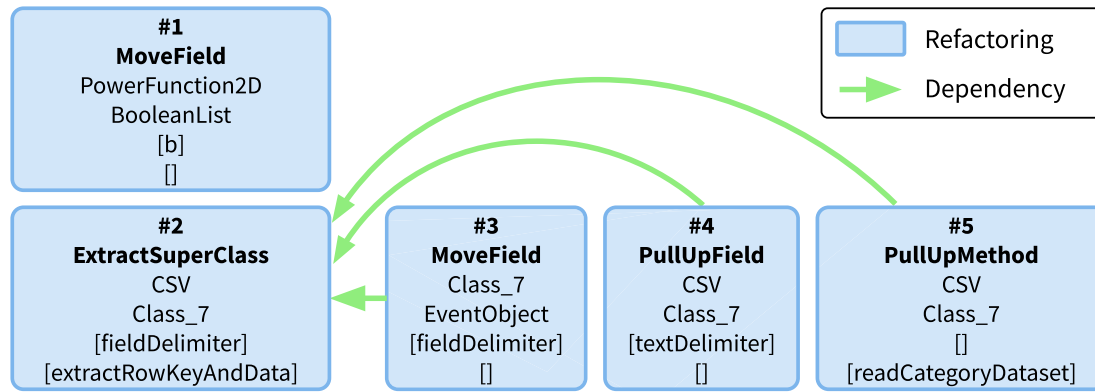


Fig. 2. A simplified solution of 6 refactorings for the JFreeChart project.

III. REFACTORING DEPENDENCY THEORY

The refactoring dependency theory for reasoning about collections of refactorings is built upon two concepts. The first is the definition of an ordering dependency relation among refactorings in a collection of refactorings. Pre- and post-conditions for refactoring types are used to detect refactoring dependencies, i.e., a set of predicates associated with each refactoring that reflect changes. The second is the organization of a collection of refactorings as a set of refactoring graphs. Together, these concepts improve our ability to understand the meaning of collections of refactorings, allowable operations on them, and their composition in practice.

In this section we describe the elements of our proposed theory, the algorithm for detecting refactoring dependencies and an associated web tool that implements this detection algorithm.

A. Definitions

Our proposed dependency relation captures an ordering dependency between pairs of refactoring instances. Specifically, an *ordering dependency* ($r f_2 \mapsto r f_1$) between two refactoring instances ($r f_1$ and $r f_2$) exists when $r f_2$ can only be successfully applied after $r f_1$ has been applied. That is, $r f_1$ makes a change to code that is necessary in order to apply $r f_2$. This condition can be evaluated based on the combination of pre- and post-conditions of the types of refactorings involved and the parameters of each refactoring instance. For example, to apply Move Method (a type of refactoring) to move method m_1 from class c_1 to class c_2 (m_1 , c_1 , and c_2 being the parameters of the refactoring instance), several preconditions must hold (e.g., m_1 , c_1 , and c_2 must all exist and m_1 must be defined on c_1). The pre- and post-conditions of each type of refactoring will be described in the next sub-section.

Building on this ordering dependency definition, we organize collections of refactorings as sets of refactoring graphs rather than as sequences of refactorings. A *refactoring graph* is a weakly connected directed acyclic graph composed of refactoring instance vertices and ordering dependency edges. Using the ordering dependencies as the basis for forming refactoring

graphs (Algorithm 1) results in a set of graphs with the following traits:

- Each refactoring instance is an element of exactly one refactoring graph.
- Some graphs contain a single refactoring instance because that refactoring is truly independent of all others. We call these *trivial graphs* comprised of a single node of a refactoring instance.
- The remaining graphs contain multiple refactoring instances, each of which is part of one or more dependencies. We call these *non-trivial graphs*.
- Each refactoring graph is independent of every other graph in the solution.

Refactoring recommendations typically comprises a collection of compatible refactorings, and as such positive dependencies are more relevant to common use cases. The idea of negative refactorings would be more applicable if a recommendation contained mutually exclusive advice (e.g., three Move Method refactorings that move the same method to three different locations). This is not the common use case, but this work would be easily adapted. The essence of identifying a refactoring that precluded (or invalidated) another could be performed using the same pre- and post-conditions, but with a modification to check for differences rather than commonalities (e.g., refactoring #1's post-condition moves the location of a method to class A and refactoring #2's pre-condition requires that same method to reside in class B). It may require additional work to consider the initial state of a program, but the same principles would likely apply. existing tools that do not use refactoring dependencies.

In this paper, we are planning to use refactoring dependencies to present recommendations in terms of refactoring graphs that clearly convey when an order is required and when it is not. Most recommendations based on this work will not imply a strict sequential order for all refactorings. For instance, let us consider the example of the non-trivial graph shown in Fig. 2 as an example. In an approach based on refactoring dependencies, #4 and #5 can only be applied after #2, but no ordering is implied between #4 and #5. The sequential order is critical when the refactorings are dependent to each other.

1) *Refactoring Pre- and Post-Conditions*: As our approach for detecting ordering dependencies relies on the pre- and post-conditions of types of refactorings, we began with validated conditions in the current literature [4], [12], [21], [32] related to 14 types of refactorings. We selected the refactoring types summarized in Table I since they were those most frequently used in practice based on existing studies [7], [10], [26] and since our work focuses more on complex/composite refactoring operations, which have more complex/sophisticated pre/post conditions. The pre- and post-conditions sets published in current literature were extensively validated for correctness and completeness [4], [12], [21], [32]. The complete list of updated pre- and post-conditions for the 14 supported types of refactorings are organized in Table I. In this table, the post-conditions that are presented are only those that represent a change. This is important for the dependency detection algorithm, allowing it to efficiently identify changes that enable pre-conditions of dependent refactorings. Also, the functions used for describing the pre- and post-conditions and their meanings are defined in [12].

B. Algorithm for Detecting Refactoring Dependencies

Algorithm 1 describes the process for detecting refactoring ordering dependencies. These dependencies are detected based on comparisons between pre- and post-conditions of refactoring instances. The proposed algorithm takes a list of refactoring instances as input and generates a set of refactoring graphs as output.

Lines 1 and 2 initialize the lists of refactoring instances (nodes of the graph, V) and refactoring dependencies (edges of the graph, E). Then, the post-conditions of each refactoring instance of the solution C (collection of refactorings) are evaluated for matching with the remaining refactoring instances in C (Lines 3–13). Specifically, the algorithm looks for any match between predicates of pre- and post-conditions from Table I. That is, if any predicate of the post-condition of one refactoring (any element of P) matches any predicate of the pre-condition of another refactoring (any element of Q), then a dependency has been detected and an edge is added to the graph between those refactorings (Lines 5–10). We repeat this process until all the refactorings have been visited. Then, Lines 14 identifies the different trivial and non-trivial graphs that are formed based on the detected dependencies.

To illustrate Algorithm 1, consider the motivating example described in Fig. 2. This example contains six refactoring instances:

```
#1 ExtractClass(OverwriteDataSet;Class_6;[x]; [add-
ChangeListener])
#2 MoveField(PowerFunction2D;BooleanList;[b];[])
#3 ExtractSuperClass(CSV;Class_7;[fieldDelimiter];
[extractRowKeyAndData])
#4 MoveField(Class_7;EventObject;[fieldDelimiter];[])
#5 PullUpField(CSV;Class_7;[textDelimiter];[])
#6 PullUpMethod(CSV;Class_7;[];[readCategory
Dataset])
```

Algorithm 1: Dependency Detection Algorithm.

Input: refactoring solution $C = \{r_1, r_2, r_3, \dots, r_n\}$
Output: forest of refactoring graphs
 $F = \{f_1, f_2, f_3, \dots, f_m\}$

```
1  $V \leftarrow \emptyset, E \leftarrow \emptyset;$ 
2 foreach  $r_i \in C$  do
3    $V \leftarrow V \cup r_i;$ 
4    $P \leftarrow \text{post\_conditions}(r_i);$ 
5   foreach  $r_j \in C \mid j > i$  do
6      $Q \leftarrow \text{pre\_conditions}(r_j);$ 
7      $M \leftarrow P \cap Q;$ 
8     if  $|M| \neq 0$  then
9        $E \leftarrow E \cup \{r_j, r_i\};$ 
10    end
11  end
12 end
13  $G \leftarrow (V, E);$ 
14  $F \leftarrow \text{partition}(G);$ 
15 return  $F;$ 
```

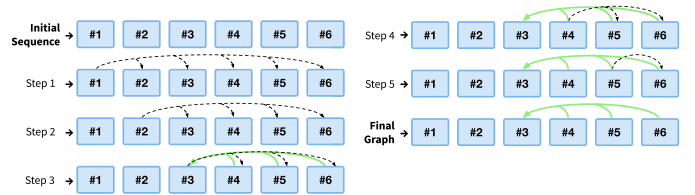


Fig. 3. Execution of Algorithm 1 on the example of Fig. 2.

Fig. 3 illustrates the application of the algorithm to this example. In Step 1, the post-conditions of refactoring #1 are compared to the pre-conditions of all remaining refactorings. Since there is no match among pre- and post-conditions for any of them, no dependency is added. Next, the post-conditions of refactoring #2 are compared in Step 2 to the pre-conditions of the next refactorings in the sequence (#3 - #6). Again, no match is found. In Step 3, the post-conditions of refactoring #3 are compared to pre-conditions of refactorings #4, #5, and #6. For each, a match *exist(Class_7)* is found. Thus, three dependencies are added, from each of #4, #5, and #6 to #3. The algorithm continues, but no additional matches are found. Thus, Algorithm 1 transforms this sequence into three refactoring graphs (two trivial graphs and one non-trivial graph that includes four refactorings).

When the developers interact with the tool by modifying or rejecting some of the refactorings then the dependencies detection algorithm is re-executed to check the impact of those changes on the graphs. Thus, all the graphs will be updated instantly during those interactions.

C. DPREf

To validate our definitions and algorithm and to make our refactoring dependency detection approach available to the community, we implemented DPREf [34], a free and open-source web platform that allows users to provide a sequence of refactorings to be applied as input and generates a set of refactoring graphs as output.

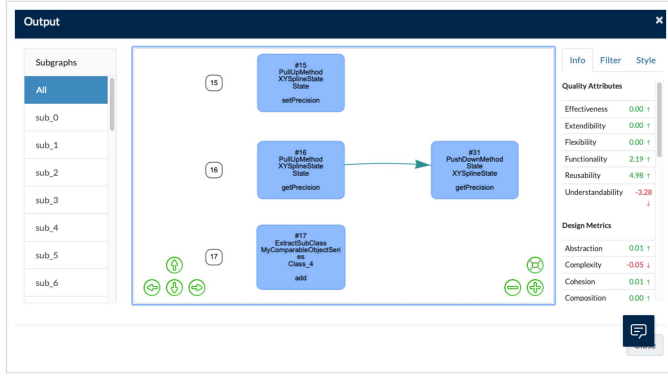


Fig. 4. DPREf, a tool for detecting refactoring dependencies.

Fig. 4 shows DPREf’s output, which includes the generated set of refactoring graphs and associated information like the impact on quality attributes for each graph or individual refactoring. Our tool automatically applies the refactoring graphs on the source code then the quality metrics are calculated on the code after applying those refactorings. Then, developers are able to change the generated refactoring graphs and apply filters to show, for instance, just the graphs that significantly improve specific quality attributes. DPREf also includes an Eclipse plug-in to execute any refactorings selected by the user on the actual source code. The tool is available as part of our replication package.

IV. EMPIRICAL STUDY

In this section, we present our research questions, validation methodology, experimental setup, and discuss our findings. Our implementation, data and results are publicly available [34].

A. Research Questions

The following research questions guide the evaluation of our refactoring dependency theory:

- RQ1.** (Precision) To what extent are the detected refactoring ordering dependencies correct?
- RQ2.** (Relation) To what extent are refactorings dependent?
- RQ3.** (Improvement) To what extent do non-trivial refactoring graphs improve quality attributes compared to trivial refactoring graphs (independent refactorings)?

We collected data from 9,595 open source repositories to evaluate the correctness of the detected ordering dependency relationships among refactorings. For each project, we executed existing refactoring recommendation tools [9], [12] to find relevant refactorings to be applied on the code of those projects. In this study, we use the refactoring recommendations generated by those tools based on their superior performance compared to the state of the art with over 90% of precision, recall and manual correctness based on large open source and industry projects; the large number of supported refactoring types; and its being publicly available. We describe the parameter settings used by such refactoring recommendation tools in the next section. Finally, we detected dependencies among refactorings and generated refactoring graphs based on Algorithm 1 for

TABLE II
QMOOD QUALITY METRICS

QMOOD Metrics	Definition / Computation
Reusability	$-0.25 * Coupling + 0.25 * Cohesion + 0.5 * Messaging + 0.5 * DesignSize$
Flexibility	$0.25 * Encapsulation - 0.25 * Coupling + 0.5 * Composition + 0.5 * Polymorphism$
Understandability	$-0.33 * Abstraction + 0.33 * Encapsulation - 0.33 * Coupling + 0.33 * Cohesion - 0.33 * Polymorphism - 0.33 * Complexity - 0.33 * DesignSize$
Functionality	$0.12 * Cohesion + 0.22 * Polymorphism + 0.22 * Messaging + 0.22 * DesignSize + 0.22 * Hierarchies$
Extensibility	$0.5 * Abstraction - 0.5 * Coupling + 0.5 * Inheritance + 0.5 * Polymorphism$
Effectiveness	$0.2 * Abstraction + 0.2 * Encapsulation + 0.2 * Composition + 0.2 * Inheritance + 0.2 * Polymorphism$

all generated refactoring solutions to understand if and how refactorings are applied together rather than in isolation. Finally, we studied and compared the impact of refactoring graphs on different well-defined quality attributes, based on the QMOOD model [13] detailed in Table II.

To answer **RQ1**, we use two methods: automated and manual correctness. We used these complementary methods because the manual evaluation can be error-prone, time-consuming and not scalable while the automated evaluation may lack insights and feedback from developers.

For the automated correctness, we removed refactorings from *valid* non-trivial graphs and determined whether the removal invalidated the graph. We define a valid graph as a refactoring graph for which all pre-conditions of all refactorings hold (Table I). As defined earlier, if an ordering dependency exists between two refactorings, for instance, rf_2 is the head refactoring (the refactoring depending on another) and rf_1 is the tail refactoring (the refactoring on which another depends) for a refactoring dependency ($rf_2 \mapsto rf_1$), then the head refactoring can only be successfully applied after the tail; removing any tail refactoring should then invalidate at least one pre-condition of a head refactoring. As such, our specific test was to remove one tail refactoring from each valid non-trivial graph with more than two refactorings and then count the number of valid and invalid non-trivial graphs. To answer **RQ1**, we calculated the Rate of Correctness (RC) after removing one tail refactoring from all non-trivial graphs as follows:

$$RC = \frac{\#of Invalid Non - trivial Graphs}{\#of Non - trivial Graphs}. \quad (1)$$

Our assertion is that the refactoring dependency detection algorithm is correct if all non-trivial graphs become invalid if at least one tail refactoring is removed from each, as above.

For the manual evaluation, we asked 27 full-time developers to manually check the correctness of 50 valid non-trivial graphs

totaling 233 refactorings for 5 open source projects^{2,3,4,5,6} that contained at least 5 K LOC and involved significant refactorings in the last 2 years. The graphs are a sequence of refactorings and some of them are large in size. We made sure during the sampling process to use the following criteria to avoid any bias in the selected refactorings for the manual validation: refactoring types, projects size, projects domain, and locations of the refactorings (files).

The participants were asked to use our tool to identify refactoring dependencies, assess the correctness of those dependencies, and apply and compile the refactorings. First, developers checked the applicability and ability to properly compile the refactorings without producing any conflicts. Each refactoring that caused the code to fail to compile was deemed invalid and was discarded. Then, they identified the refactorings for which the set of dependencies was both correct and complete among all the applicable and compilable refactorings. When checking correctness, developers were asked to evaluate each refactoring independently to determine whether each identified dependency was necessary and whether there were any missing dependencies to evaluate the accuracy of both tail-related dependencies and those not connected to the tail. The participants looked to the generated dependencies graph of the refactorings using our visual representation in the DPRef tool. Then, they reviewed the code before and after applying any selected refactorings. In case of any doubt, the participants can select the refactorings from the graph and they will be automatically executed on the code. Conflicting refactorings may generate errors in the code which may confirm the missing dependencies. In this case, the refactoring is considered as invalid in this exercise. Otherwise, a refactoring for which the set of dependencies was both correct and complete is considered as a valid refactoring. We define a manual correctness score (MC) as

$$MC = \frac{\# \text{ of Valid Refactorings}}{\# \text{ of Evaluated Refactorings}}. \quad (2)$$

To answer **RQ2**, we calculated the number of dependencies (edges) and graphs (trivial and non-trivial) for all projects. We also counted the number of refactorings in non-trivial graphs and the most frequently occurring refactoring types in them, as well as the Non-Trivial Rate (NTR) defined as follows:

$$NTR = \frac{\# \text{ of Refactorings in Non-trivial Graphs}}{\# \text{ of Refactorings}}. \quad (3)$$

These evaluation metrics allow us understand the extent of refactoring dependencies. Furthermore, we can evaluate the refactoring types that are less commonly applied in isolation and also understand the complexity of the non-trivial graphs based on their sizes. To answer **RQ3**, we consider all the trivial and non-trivial graphs to evaluate their impact on the quality attributes

TABLE III
STATISTICS OF THE SUBJECT PROJECTS

Metric	Min	Average	Max
Collaborators	2	2.6	65
Code Size	5.0k	25.9k	4,997.7k
# of Classes	1	931.3	23802
# of Refactorings	150	151.9	200

and the design metrics from QMOOD. We compared the number of graphs that improve the quality attributes and design metrics from QMOOD [13] for both trivial and non-trivial graphs. We also considered the rates of improvement, in percentage, for all graphs taking into consideration the reusability, flexibility, understandability, functionality, extendibility, and effectiveness quality attributes captured by QMOOD metrics and available in Table II, as well as, basic metrics such as coupling, cohesion, etc. We also calculated a Total Quality Index (TQI), aggregating all the metrics, after normalization, with equal weights into one metric.

These evaluation metrics are useful to understand the impact of collections versus individual refactorings on improving the quality and which quality attributes are more likely to be significantly improved using non-trivial graphs or independent refactorings. We want to also highlight that the comparison of non-trivial graphs and trivial graphs could be the result of the number of refactoring instances, instead of the result of dependency.

B. Experimental Settings

We considered a total of 9,595 open-source Java projects provided by [35] to address the above research questions. The selection process limited consideration to projects with ≥ 5 k LOC and at least 2 collaborators. We also eliminated any duplicate (cloned) projects from consideration. We applied these criteria on the list of one million GitHub projects. We performed this selection process in an attempt to eliminate small projects, such as student projects and small hobby/learner programs, that were not likely to be good candidates for refactorings.

Table III shows the min, average, and max for the number of the collaborators, code size (in LOC), # of classes, and # of recommended refactorings generated. The list of subject projects is also available in the replication package, along with all results (e.g., refactorings, quality metrics, refactoring graphs, etc.).

The total number of refactorings collected from recommendations for these 9,595 projects is almost 1.5 million (1,457,873 refactorings). We used the parameter settings recommended by the authors of the refactoring recommendations tool [12]: Single Point crossover with probability = 0.7, Bit Flip mutation with probability = 0.4, and stopping criterion was set to 100,000 evaluations. We also set the initial population size to 100 and utilized a tournament selection operator with $n=2$. The minimum and maximum number of refactorings per solution are limited to 150 and 200, respectively.

For the manual validation of the refactoring dependencies, we recruited 27 full-time developers from our networks, each of

²<https://github.com/phunware/maas-ads-android-sdk>

³<https://github.com/solita/query-utils>

⁴<https://github.com/forgo/roaster>

⁵<https://github.com/goobi/goobi-ugh>

⁶<https://github.com/kongchen/swagger-maven-example>

TABLE IV
SELECTED PARTICIPANTS

System	# of Subjects	Prog. Exp. (Years) [Min-Avg-Max]	Avg. Ref. Exp.
Dataset #1	5	[6.0 - 5.5 - 7.5]	High
Dataset #2	5	[5.5 - 7.0 - 8.0]	High
Dataset #3	7	[6.5 - 7.5 - 10.5]	High
Dataset #4	5	[6.0 - 6.5 - 8.5]	High
Dataset #5	5	[6.5 - 7.5 - 9.0]	High

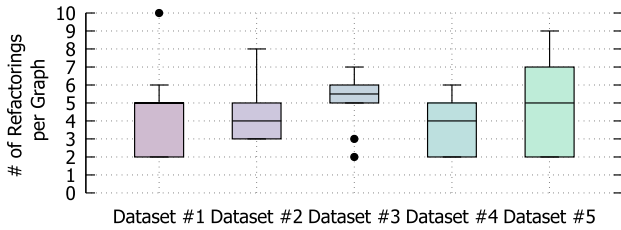


Fig. 5. Number of refactorings per non-trivial graph in each dataset.

whom was unaware of our algorithm. These participants were first asked to fill out a pre-study questionnaire containing six questions. The questionnaire helped to collect background information such as their role within the company, their programming experience, and their familiarity with software refactoring. The list of the pre- and post-study questions of all the questionnaires, the validation data and the obtained results can be found in the online appendix, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TSE.2023.3244123>. Although the vast majority of participants were already familiar with refactoring as part of their jobs and graduate studies, all the participants attended a two-hour lecture on refactoring by two the organizers of the experiments. The details of the selected participants can be found in Table IV, including their years of programming experience, familiarity with refactoring, etc.

All participants had a minimum of 6 years experience and work as active programmers with strong backgrounds in refactoring, Java, and software quality metrics. We divided the participants into 5 datasets where each dataset contains 10 samples of valid non-trivial graphs. We selected these samples based on the distribution of the refactoring types and number of refactorings in each graph as described in Fig. 5.

Each participant was asked to assess the correctness of the refactoring dependencies and to identify missing dependencies between the refactorings using our DPRef tool. They were asked to execute the sequence of refactoring using our Eclipse plug-in and compile the code after applying the refactorings of each valid non-trivial graph. In addition to evaluating the refactorings, the participants were asked to configure, run, and interact with the tool on the different systems. We assigned tasks to the participants according to the datasets and developers' experience. Each participant was given a post-study survey. This second survey was more general as it collected the practitioners' opinions and their perception of the importance and relevance of

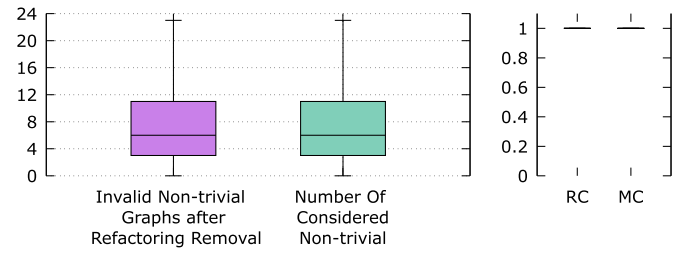


Fig. 6. Boxplots of refactoring dependency correctness for the 9,595 projects.

detecting refactoring dependencies along with the usability of DPRef. The samples used in the manual validation study as well as the full details of our pre-study survey results can be found in the replication package.

C. Results and Discussions

1) *Results for RQ1*: Fig. 6 summarizes the distribution of the results of removing one tail refactoring from all non-trivial graphs with more than two refactorings. One important outcome is that all non-trivial graphs become invalid after removing a tail refactoring (RC), which confirms that the proposed algorithm accurately identifies refactoring dependencies. The RC evaluation metric has a value of 1.0 (or 100%) across all the 9,595 projects, i.e., for all non-trivial graphs. The total number of non-trivial graphs is 257,725 with an average of 26.8 per project.

Another interesting result was that our algorithm detected many invalid refactorings among the solutions generated by the tool of Ouni et al. [12]. An average of 6.2 invalid non-trivial graphs were identified per project. The main reason is that the crossover operator used to exchange refactorings between solutions did not ensure that refactoring pre-conditions would remain satisfied after the exchange. As discussed in the future work section, the theory proposed in this paper can be integrated into existing refactoring recommendation tools to improve their correctness and contribute to the definition of intelligent change operators (including crossover) for search-based refactoring. Non-trivial graphs that were initially invalid were excluded from the calculation of RC and the removal of refactorings shown in Fig. 6.

For the manual evaluation, all the non-trivial graphs were correctly executed by the participants on the open source projects and they agreed that the dependencies were correctly and completely identified for each refactoring, thus the MC scores were 100% on all the selected datasets as shown in Fig. 6. Thus, the manual correctness results confirm the automated method.

The fact that a refactoring was recommended by Ouni et al. [12] and was not applied before by developers does not mean that the recommendation is not correct but simply that probably the developers may not think about that refactoring and may not have time to do it. The manual validation scores of [12] are more than 90% on large scale systems which means that the manual check by developers confirmed that the vast majority of the recommendations are correct and useful. We clarified these observations in the validation section when explaining our choices.

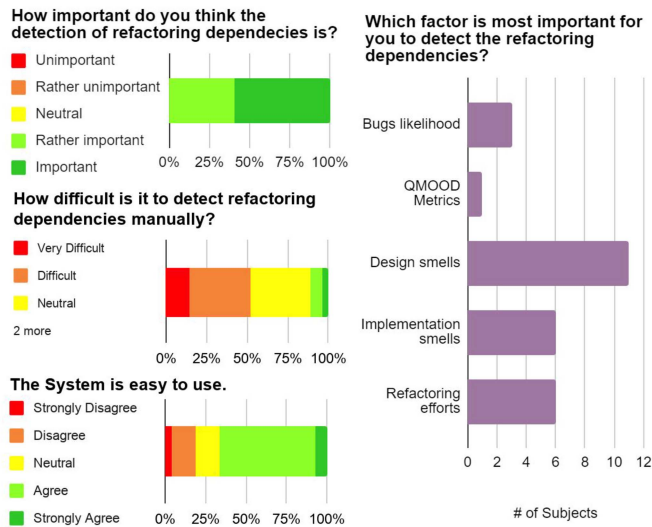


Fig. 7. Participant survey.

Maturation of automated tools to assist developers with complex tasks such as refactoring is an important gap to fill. The developers that we surveyed also agree with this. All the participants rated the importance of detecting refactoring dependencies as important or rather important, which confirms the need for tools to detect dependencies in a refactoring sequence. The vast majority of participants (24 out of 27) also rated the task of manually identifying these dependencies as difficult or very difficult based on their experiences in using existing refactoring tools. These developers described that the most important motivations of automatically detecting dependencies is to understand how they can fix code smells (typically fixed using a sequence of refactorings) and also to reduce the refactoring effort by applying refactorings incrementally.

We asked also the participants to rate their agreement on a Likert scale from 1 (complete disagreement) to 5 (complete agreement) with the following question on how important it is to detect refactoring dependencies automatically.

As shown in Fig. 7, all the participants rated this feature as important or rather important which confirms the need for tools to detect dependencies in a refactoring sequence. The majority of participants (14 out of 27) also rated the task of manually identifying these dependencies as difficult/very difficult based on their experiences in using existing refactoring tools. Fig. 7 also describes that developers think that the most important motivations of automatically detecting dependencies is to understand how they can fix code smells (typically fixed using a sequence of refactorings) and also to reduce the refactoring effort by applying refactorings in an incremental way. The figure also shows that DPRef was easy to use by at least 18 out of the 27 participants and they were able to understand the dependencies without complications.

Key findings: To answer RQ1, the proposed algorithm for detecting refactoring dependencies achieved 100% correctness on all projects.

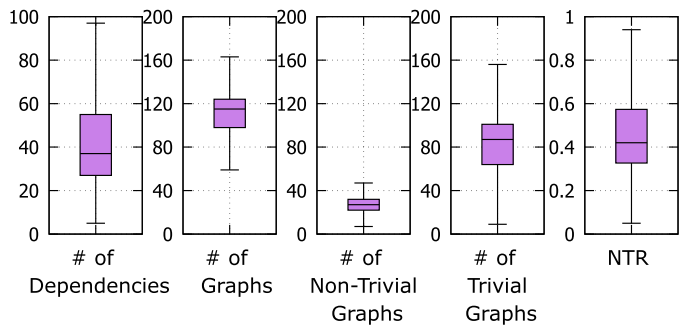


Fig. 8. Distribution of refactorings in trivial versus non-trivial graphs based on the 9,595 projects.

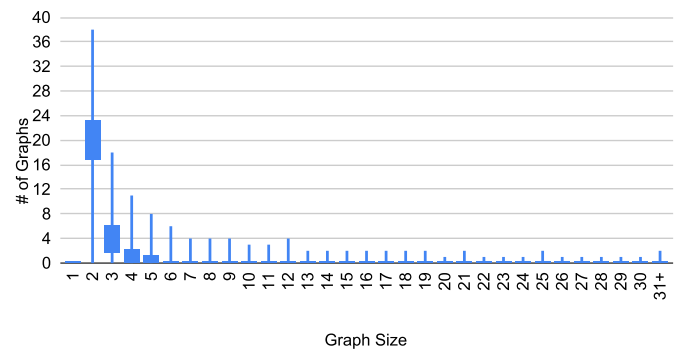


Fig. 9. Size of non-trivial refactoring graphs in the 9,595 projects.

2) *Results for RQ2:* Fig. 8 shows that while truly independent refactorings are more common, the mean NTR shows that more than 40% of all recommended refactorings are part of non-trivial graphs and for some projects, all refactorings are part of a single non-trivial graph (NTR=1.0). The portion of refactorings that are part of refactoring dependencies is significant.

Across all projects, non-trivial graphs have a mean number of dependencies of almost 40 (recall that this is based on solutions with a mean of roughly 150 refactorings), with nearly 100 dependencies observed for some projects. This indicates a range of connectedness in non-trivial graphs, with the more highly dependent/coupled likely being difficult for developers to understand and having a high risk of generating large numbers of invalid refactorings if they are not applied together. We noticed that the most connected refactoring graphs are more likely to be associated with small projects in which many refactorings are applied to common code locations.

Comparing the range and distribution of the number of graphs, most projects have a mean of more than 100 graphs per solution, in which the number of trivial graphs is greater than non-trivial. However, as shown by NTR, the number of refactorings in non-trivial graphs is very similar or exceeds the number of refactorings in trivial graphs.

Fig. 9 shows the distribution of the size of the non-trivial graphs; notice that there is a small number of non-trivial graphs including 31+ refactorings. However, the vast majority of non-trivial graphs include 2-5 refactorings. This may confirm that collections of dependent refactorings tend to be small, which

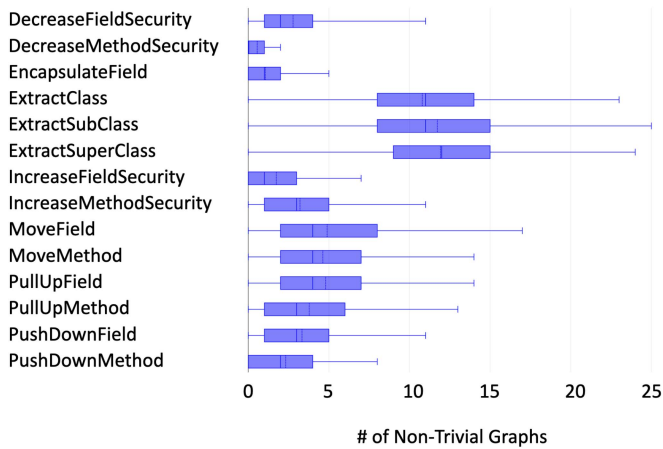


Fig. 10. Distribution of the refactoring types among non-trivial graphs for the 9,595 projects.

can offer flexibility to developers if they want to modify these collections. Projects vary, however, with several including non-trivial graphs with 10-50 refactorings and with one including a graph of 138 refactorings.

Regarding refactoring types, Fig. 10 shows that the most common refactoring types found in non-trivial graphs are *Extract Class*, *Extract Super Class* and *Extract Sub Class*. This finding confirms that modifying the hierarchy of code requires a combination of several refactorings and cannot be done with one isolated refactoring. Furthermore, the figure also shows that *Decrease Method Security*, *Encapsulate Field*, and *Increase Field Security* are the least common refactoring types in non-trivial graphs. It most likely indicates that these refactorings can be applied independently without requiring major restructuring effort.

Key findings: To answer RQ2, while more refactorings appear in trivial graphs than non-trivial graphs, the difference is not large. The mean value of NTR is 43%, indicating nearly half of all refactorings participate in refactoring dependencies and as such cannot be applied without consideration of other refactorings. Some refactoring types are more likely to be applied with other dependent refactorings than others.

3) *Results for RQ3:* Fig. 11 shows how trivial and non-trivial refactoring graphs improve quality attribute and design metrics. The number of trivial graphs that improves each metric is greater than the number of non-trivial graphs that do so. However, the impact of each kind of graph on the improvement of each quality metric can vary considerably. We also analysed the improvements caused by the two kinds of refactoring graphs for a subset of the quality attribute metrics. To statistically compare the distributions of the number of graphs that improved the quality metrics across the quality metrics for trivial and non-trivial graphs, we used the Wilcoxon rank sum test [36] with the level of significance (α) set to 0.05. In all cases, the test rejected the null hypothesis (p -values < 0.005).

Fig. 12 shows the rate of improvement in % for Effectiveness, Extendibility, Flexibility, Functionality, Reusability and Understandability. This data shows that, for most projects, the improvement caused by the non-trivial graphs is greater than for trivial graphs. The implication is that non-trivial graphs may be more useful in practice for developers than marginal improvements obtained by individual refactorings. Making refactoring dependencies explicit allows users to see, for example, that disregarding a particular refactoring may invalidate several others in the same refactoring graphs. As non-trivial graphs often result in greater improvements, that decision may be more significant to the user. Furthermore, some metrics, like Extendibility (note the different scale) and Reusability, are more significantly improved using dependent refactorings. Even small changes are significant when considered as aggregate measures across an entire code base. This result is consistent with the fact that Extract [Super/Sub] Class refactorings are more likely to occur within a non-trivial graph and that these refactorings are natural choices for improving the Extendibility and Reusability of software.

The results are also consistent with the test which returned p -values lower than 0.05 for Functionality and Reusability, while the p -value is greater than 0.05 for the remaining metrics. However, quality metrics may be conflicting, hence the execution of a refactoring could increase some metrics and deteriorate others. The cumulative impact will depend on the types of the refactorings in the non-trivial graph. Indeed, several studies examined relationships and correlations between quality metrics and confirmed the possibility of a conflict between two quality metrics, which means that improving one quality may seriously degrade another such as coupling and cohesion. For instance, it has been found that maintainability and efficiency are negatively correlated, as are functionality and understandability, performance and evolvability, or performance and reliability. Negative relationships between extensibility and understandability, for example, result from the fact that systems with extensional features are more challenging to learn and maintain. Most of the time, a solution entails prioritization and a compromise when applying refactoring, which could increase some metrics and deteriorate others.

Key findings: To answer RQ3, non-trivial refactoring graphs improve all six quality attribute metrics in our experiments better than independent refactorings. In particular, the improvement from the application of non-trivial graphs over trivial graphs is particularly significant for Functionality and Reusability.

V. THREATS TO VALIDITY

Conclusion Validity. We used Design of Experiments (DoE) [37] to mitigate the internal threat related to parameter tuning used in our experiments. DoE is a methodology for systematically applying statistics to experimentation and is one of the most efficient techniques for parameter settings of evolutionary algorithms for the used refactoring recommendation tool. Each parameter has been uniformly discretized in some

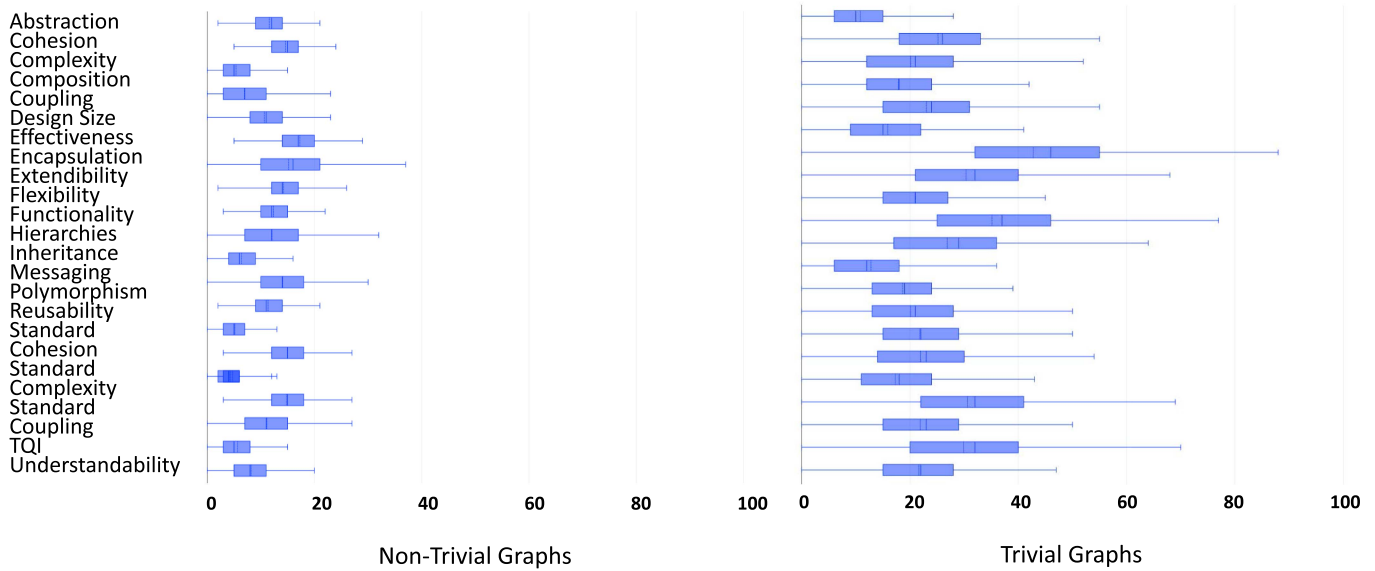


Fig. 11. The number of graphs that improved the quality metrics.

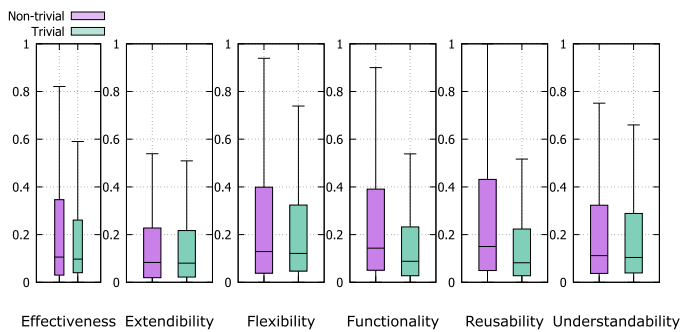


Fig. 12. Rate of quality improvement (%) for the refactoring graphs per metric.

intervals. Values from each interval have been tested for our application and we chose the best values.

Internal Validity. We collected data from a very large number of repositories resulting in about 1.5 million refactorings for 9,595 projects. A possible internal threat is the diversity of these projects in terms of domains, size, etc. To mitigate this threat, we used several criteria (e.g., more than one contributor per project, over 5 K lines of code, etc.) to select the projects and eliminate redundant ones or those with small size to avoid considering student projects on GitHub and so on. Also, different tools, with different recommendation strategies might be more or less dominated by non-trivial graphs. Historical refactorings from commit histories could likewise have a different composition. When we conclude that dependencies are common, the source of the source data matters. We make a replication package available including all the collected data that can be used and improved by the community.

Another possible internal threat is the technique used for the manual validation. We did not inform the participants that the non-trivial graphs are all valid based on our tool and we wanted

them to confirm that by validating the code after refactoring and executing it. The invalid graphs can be easily validated as the code will not even compile after refactoring. As described in the pre-study questionnaire, we asked the participants about their experience in interacting with the tool and measured the time that they spent as well. Thus, our questions were not just to ask them about the importance of detecting dependencies but more on how they could be useful for them in understanding the refactoring recommendations and so on.

Construct Validity. The refactorings used in our experiments are generated using an existing refactoring tool [12]. Thus, it is possible that some of them are not relevant (e.g., small impact on quality). However, our goal is to evaluate the dependencies among the refactorings independently from their relevance. In addition, our approach can take as input any sequence of refactorings.

External Validity. The types of refactorings considered in our experiments may threaten the generalizability of our results. Besides, our study was limited to the use of specific quality attributes to measure the impact of the application of refactoring graphs. Future replications of this study are necessary to further confirm our findings. Also, the number of participants can be extended in our future work to validate more refactoring dependencies. Moreover, There is no consensus in refactoring studies about the most representative types. Several existing works [38], [39], [40], [41] show that the used refactoring types in this study are the most frequently used by developers.

VI. IMPLICATIONS AND FUTURE WORK

Our proposed theory of organizing collections of refactoring instances as a set of refactoring graphs offers several advantages that address the challenges confirmed by developers:

- *explainability*: each refactoring graph is smaller and more coherent than a long sequence of refactorings. As each

can be explained independently, the cognitive burden on a developer is much lower (e.g., contrast with determining which refactorings scattered across a sequence of dozens or hundreds of refactorings are related).

- *comparability*: search-based refactoring recommendation tools typically generate multiple recommendations on a Pareto front, leaving developers to choose one. Identifying common elements of different recommendations is simplified by comparing sets of graphs that do not contain the spurious orderings found in sequence representations.
- *search efficiency*: search-based refactoring recommendation tools that use genetic algorithms gain new options. Specifically, crossover operations can be more reliable (reducing failures) when using dependency analysis; graphs may also be better genomes for crossover than individual refactorings.

Consequently, there are several directions for future work:

Refactoring Pattern Extraction. One important implication of the proposed refactoring dependency theory is the ability to extract common refactoring patterns by mining software repositories using tools such as RefMiner [16]. These patterns are the common non-trivial graphs that can be extracted on different commits/pull-requests of the same project or multiple projects. Such patterns of non-trivial graphs can be linked to refactoring opportunities such as resolving different types of code smells repeatedly. In the future, we plan to use the refactoring dependencies to understand the common refactoring patterns from the history of commits and pull requests of software repositories using existing refactoring detection tools such as RefMiner.

Refactoring Collaborations Between Developers. Studying the collaborations among multiple developers when refactoring code is a promising next step. Refactoring graphs extracted from commit histories can be linked to the authors of those commits. Then, a graph of collaborations among developers can be generated based on the dependencies among the applied refactorings. This can lead to new insights into why and when developers collaborate for refactoring.

Change Operator in Search-Based Refactoring. Random selection and application of crossover and mutation when evolving a population of solutions is a challenge in search-based refactoring. Refactoring dependency analysis can be used to avoid destroying good patterns in refactoring solutions and make change operators more intelligent, which can lead to better solutions and faster convergence.

Interactive Refactoring Tool Support. Developers can more easily understand the implications of selecting which refactorings from a recommendation to apply, improving the interactive process and increasing their confidence in the recommendation tool. The only restriction in applying non-trivial refactoring graphs is that a refactoring can only be applied if every other refactoring that it depends on (transitively) is also applied. Thus, invalid refactorings can be detected and highlighted on the fly.

VII. CONCLUSION

Although manually applying a collection of refactorings is common practice, existing empirical studies and refactoring

recommendation/detection tools treat refactorings in isolation. In this paper, we proposed a definition for ordering dependencies among refactorings and an algorithm for detecting these dependencies. We also proposed defining refactoring recommendations as sets of refactoring graphs rather than as refactoring sequences and illustrated these concepts with a tool for visualizing refactoring dependencies and sets of refactoring graphs. We elaborated our research agenda for future work in Section VI.

We validated the proposed approach on 1,457,873 refactorings recommended for 9,595 projects. Our results show that the proposed approach achieved 100% in correctly detecting all dependencies among refactorings. Furthermore, we found that 43% of the 1,457,873 recommended refactorings are part of dependent refactoring graphs, which confirms that refactorings are commonly involved in dependent relations and cannot be applied truly independently. These concepts advance a theory for reasoning about refactorings collectively, rather than individually, and offer clear benefits to developers applying refactoring recommendations (explainability and comparability) and authors of tools for recommending refactorings (search efficiency and improving correctness of recommendations).

ACKNOWLEDGMENTS

Copyright 2021 IEEE. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute. DM21-0546

REFERENCES

- [1] E. Tom, A. Aurum, and R. Vidgen, "An exploration of technical debt," *J. Syst. Softw.*, vol. 86, no. 6, pp. 1498–1516, 2013.
- [2] M. Kuuttila, M. Mäntylä, U. Farooq, and M. Claes, "Time pressure in software engineering: A systematic review," *Inf. Softw. Technol.*, vol. 121, 2020, Art. no. 106257.
- [3] S. A. Slaughter, D. E. Harter, and M. S. Krishnan, "Evaluating the cost of software quality," *Commun. ACM*, vol. 41, no. 8, pp. 67–73, 1998.
- [4] M. Fowler, *Refactoring: Improving the Design of Existing Code*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.
- [5] G. Bavota, A. D. Lucia, M. D. Penta, R. Oliveto, and F. Palomba, "An experimental investigation on the innate relationship between quality and refactoring," *J. Syst. Softw.*, vol. 107, pp. 1–14, 2015.
- [6] A. C. Bibiano et al., "A quantitative study on characteristics and effect of batch refactoring on code smells," in *Proc. IEEE/ACM 13th Int. Symp. Empir. Softw. Eng. Meas.*, Brazil, 2019, pp. 1–11.
- [7] E. Murphy-Hill, C. Parnin, and A. P. Black, "How we refactor, and how we know it," *IEEE Trans. Softw. Eng.*, vol. 38, no. 1, pp. 5–18, Jan./Feb. 2012.
- [8] M. Ó. Cinnéide, L. Tratt, M. Harman, S. Counsell, and I. H. Moghadam, "Experimental assessment of software metrics using automated refactoring," in *Proc. IEEE-ACM Int. Symp. Empir. Softw. Eng. Meas.*, Lund, Sweden, 2012, pp. 49–58.
- [9] N. Tsantalis and A. Chatzigeorgiou, "Identification of move method refactoring opportunities," *IEEE Trans. Softw. Eng.*, vol. 35, no. 3, pp. 347–367, May/Jun. 2009.
- [10] M. Kim, T. Zimmermann, and N. Nagappan, "An empirical study of refactoring challenges and benefits at Microsoft," *IEEE Trans. Softw. Eng.*, vol. 40, no. 7, pp. 633–649, Jul. 2014.
- [11] V. Alizadeh, M. Kessentini, W. Mkaouer, M. O. Cinnéide, A. Ouni, and Y. Cai, "An interactive and dynamic search-based approach to software refactoring recommendations," *IEEE Trans. Softw. Eng.*, vol. 46, no. 9, pp. 932–961, Sep. 2020.

- [12] A. Ouni, M. Kessentini, H. Sahraoui, K. Inoue, and K. Deb, "Multi-criteria code refactoring using search-based software engineering: An industrial case study," *ACM Trans. Softw. Eng. Methodol.*, vol. 25, no. 3, 2016, Art. no. 23.
- [13] J. Bansiya and C. G. Davis, "A hierarchical model for object-oriented design quality assessment," *IEEE Trans. Softw. Eng.*, vol. 28, no. 1, pp. 4–17, Jan. 2002.
- [14] C. Abid, V. Alizadeh, M. Kessentini, T. D. N. F. Ferreira, and D. Dig, "30 years of software refactoring research: A systematic literature review," 2020, *arXiv:2007.02194*.
- [15] M. W. Mkaouer, M. Kessentini, S. Bechikh, K. Deb, and M. O. Cinnéide, "Recommendation system for software refactoring using innovization and interactive dynamic optimization," in *Proc. IEEE-ACM 29th Int. Conf. Autom. Softw. Eng.*, Vasteras, Sweden, 2014, pp. 331–336.
- [16] N. Tsantalis, M. Mansouri, L. M. Eshkeviri, D. Mazinanian, and D. Dig, "Accurate and efficient refactoring detection in commit history," in *Proc. ACM 40th Int. Conf. Softw. Eng.*, Gothenburg, Sweden, 2018, pp. 483–494.
- [17] M. Harman and L. Tratt, "Pareto optimal search based refactoring at the design level," in *Proc. 9th ACM Annu. Conf. Genet. Evol. Comput.*, London, England, 2007, pp. 1106–1113.
- [18] Y. Lin, X. Peng, Y. Cai, D. Dig, D. Zheng, and W. Zhao, "Interactive and guided architectural refactoring with search-based recommendation," in *Proc. ACM SIGSOFT Int. Symp. Found. Softw. Eng.*, Seattle, USA, 2016, pp. 535–546.
- [19] T. Sharma and D. Spinellis, "A survey on software smells," *J. Syst. Softw.*, vol. 138, pp. 158–173, 2018.
- [20] G. Bavota, A. D. Lucia, A. Marcus, and R. Oliveto, "Recommending refactoring operations in large software systems," in *Recommendation Systems in Software Engineering*, Berlin, Germany: Springer, 2014, pp. 387–419.
- [21] M. O’Keeffe and M. O. Cinnéide, "A stochastic approach to automated design improvement," in *Proc. ACM 2nd Int. Conf. Princ. Pract. Program. Java*, Kilkenny City, Ireland, 2003, pp. 59–62.
- [22] O. Seng, J. Stammel, and D. Burkhart, "Search-based determination of refactorings for improving the class structure of object-oriented systems," in *Proc. ACM 8th Annu. Conf. Genet. Evol. Comput.*, Seattle, USA, 2006, pp. 1909–1916.
- [23] M. Kessentini, W. Kessentini, H. Sahraoui, M. Boukadoum, and A. Ouni, "Design defects detection and correction by example," in *Proc. IEEE 19th Int. Conf. Prog. Comprehension*, Kingston, Canada, 2011, pp. 81–90.
- [24] A. Ouni, M. Kessentini, and H. Sahraoui, "Search-based refactoring using recorded code changes," in *Proc. IEEE 17th Eur. Conf. Softw. Maintenance Reengineering*, Genova, Italy, 2013, pp. 221–230.
- [25] W. Mkaouer et al., "Many-objective software remodularization using NSGA-III," *ACM Trans. Softw. Eng. Methodol.*, vol. 24, no. 3, pp. 17:1–17:45, 2015.
- [26] A. Chávez, I. Ferreira, E. Fernandes, D. Cedrim, and A. Garcia, "How does refactoring affect internal quality attributes? A multi-project study," in *Proc. ACM 31st Braz. Symp. Softw. Eng.*, Fortaleza, Brazil, 2017, pp. 74–83.
- [27] T. Mens, G. Taentzer, and O. Runge, "Analysing refactoring dependencies using graph transformation," *Softw. Syst. Model.*, vol. 6, no. 3, pp. 269–285, 2007.
- [28] T. Mens, G. Taentzer, and O. Runge, "Detecting structural refactoring conflicts using critical pair analysis," *Electron. Notes Theor. Comput. Sci.*, vol. 127, no. 3, pp. 113–128, 2005.
- [29] H. Liu, G. Li, Z. Ma, and W. Shao, "Conflict-aware schedule of software refactorings," *IET Softw.*, vol. 2, no. 5, pp. 446–460, 2008.
- [30] H. Liu, Z. Ma, W. Shao, and Z. Niu, "Schedule of bad smell detection and resolution: A new way to save effort," *IEEE Trans. Softw. Eng.*, vol. 38, no. 1, pp. 220–235, Jan./Feb. 2012.
- [31] L. Sousa et al., "Characterizing and identifying composite refactorings: Concepts, heuristics and patterns," in *Proc. 17th Int. Conf. Mining Softw. Repositories*, 2020, pp. 186–197.
- [32] H. Melton and E. Tempero, "Identifying refactoring opportunities by identifying dependency cycles," in *Proc. ACM 29th Australas. Comput. Sci. Conf.*, Australia, 2006, pp. 35–41.
- [33] N. Yoshida, Y. Higo, T. Kamiya, S. Kusumoto, and K. Inoue, "On refactoring support based on code clone dependency relation," in *Proc. IEEE 11th Int. Softw. Metrics Symp.*, Como, Italy, 2005, pp. 10–pp.
- [34] DPREf, 2022. [Online]. Available: <https://github.com/iselab-dearborn/dpref-refactoring-dependencies>
- [35] N. Munaiah, S. Kroh, C. Cabrey, and M. Nagappan, "Curating github for engineered software projects," *Empir. Softw. Eng.*, vol. 22, no. 6, pp. 3219–3253, 2017.
- [36] J. H. McDonald, *Handbook of Biological Statistics*, vol. 2. Baltimore, MD, USA: Sparky House Publishing, 2009.
- [37] J. R. Koehler and A. B. Owen, "Computer experiments," in *Handbook of Statistics*, Amsterdam, Netherlands: Elsevier Science, 1996, pp. 261–308.
- [38] M. Alshayeb and Mohammad, "Empirical investigation of refactoring effect on software quality," *Inf. Softw. Technol.*, vol. 51, pp. 1319–1326, Sep. 2009.
- [39] F. Palomba, A. Zaidman, R. Oliveto, and A. De Lucia, "An exploratory study on the relationship between changes and refactoring," in *Proc. IEEE/ACM 25th Int. Conf. Prog. Comprehension*, 2017, pp. 176–185.
- [40] C. Vassallo, G. Grano, F. Palomba, H. C. Gall, and A. Bacchelli, "A large-scale empirical exploration on refactoring activities in open source software projects," *Sci. Comput. Program.*, vol. 180, pp. 1–15, 2019.
- [41] G. Szke, G. Antal, C. Nagy, R. Ferenc, and T. Gyimthy, "Empirical study on refactoring large-scale industrial systems and its effects on maintainability," *J. Syst. Softw.*, vol. 129, no. C., pp. 107–126, Jul. 2017, doi: [10.1016/j.jss.2016.08.071](https://doi.org/10.1016/j.jss.2016.08.071).



Thiago Ferreira received the PhD degree in computer science from the Federal University of Parana, in 2019. He is an assistant professor with the College of Innovation & Technology (CIT), University of Michigan-Flint. His research interests focus on the use of user preferences, optimization algorithms, and artificial intelligence techniques to address several software engineering problems such as software requirements, software testing, and software refactoring. For more information, see thiagod@umich.edu.



James Ivers is the lead of the Carnegie Mellon University Software Engineering Institute's software architecture group, which develops and matures tools and practices to support software architects. He is also the co-author of the Documenting Software Architectures book. For more information, see jivers@sei.cmu.edu.



Jeffrey J. Yackley is currently working toward the PhD degree with the University of Michigan - Dearborn. He is co-advised by Dr. Marouane Kessentini in the ISE Lab and Dr. Bruce R. Maxim in the GAME Lab. His research focuses on search based software engineering and machine learning in order to address problems with software architecture, refactoring, and testing in addition to his research on computer science education where he focuses on applying active learning techniques in the classroom. For more information, see jjackley@umich.edu.



Marouane Kessentini received the PhD degree from the University of Montreal, Canada, in 2012. He is a full professor, chair of the CSE Department, Oakland University, and director of the NSF IUCRC Center on Pervasive AI. He is a recipient of the prestigious 2018 President of Tunisia distinguished research award, the University of Michigan-Dearborn distinguished teaching award, the University of Michigan-Dearborn distinguished digital education award, the University of Michigan-Dearborn/College of Engineering and Computer Science distinguished research award, 4 best paper awards including and the prestigious IEEE 10 Year Most Influential Paper award (2011–2021), and his AI-based software refactoring invention, licensed and deployed by Fortune 500 companies, and selected as one of the Top 8 inventions with the University of Michigan for 2018 (including the three campuses), among more than 500 inventions, by the UM Technology Transfer Office. He received various multi-million grants from both industry and federal agencies and published more than 180 papers in top journals and conferences. He has extensive collaborations with the industry in different areas related to Edge AI, AI/MLOps, AI and cyber-physical systems, intelligent software bots, etc. He is the co-founder of many workshops, general chair of SSBSE16 and ASE22, and PC chair of MODELS19, SANER 2021, GECCO, etc. He served as a keynote speaker with various venues including ICSR, SSBSE, GECCO, WCCI, etc. He graduated more than 18 PhD students and served as associate editor in 7 journals and PC member of more than 200 conferences.



Khoulood Gaaloul received the PhD degree from The University of Luxembourg. She is a postdoctoral researcher with the University of Oakland University under the supervision of Dr. Marouane Kessentini in the ISELab. She held the position of a post-doctoral researcher with The SnT Centre for Security, Reliability, and Trust, University of Luxembourg and the position of a post-doctoral researcher with the University of Michigan-Dearborn. Her research interests include model-based software development and analysis of Cyber-Physical Systems, search-based testing and machine learning. She has been conducting her research in close collaboration with industry partners in the aerospace sector. For more information, see kgaaloul@oakland.edu



Ipek Ozkaya (Senior Member, IEEE) received the PhD degree in computational design from CMU. She is a technical director with the Carnegie Mellon University Software Engineering Institute, where she develops methods and practices for software architectures, agile development, and managing technical debt in complex systems. She coauthored a book on *Managing Technical Debt: Reducing Friction in Software Development* (2019). She is 2019–2021 editor-in-chief of *IEEE Software Magazine*. For more information, see ozkaya@sei.cmu.edu.