



PairSmell: A Novel Perspective Inspecting Software Modular Structure

Chenxing Zhong
State Key Laboratory of Novel
Software Technology
Nanjing University, China
zhongcx@smail.nju.edu.cn

Daniel Feitosa
Faculty of Science and Engineering
University of Groningen
the Netherlands
d.feitosa@rug.nl

Paris Avgeriou
Faculty of Science and Engineering
University of Groningen
the Netherlands
p.avgeriou@rug.nl

Huang Huang
State Grid Nanjing
Power Supply Company, China
sgcc.huang.huang@gmail.com

Yue Li
State Key Laboratory of Novel
Software Technology
Nanjing University, China
yueli.dom@outlook.com

He Zhang*
State Key Laboratory of Novel
Software Technology
Nanjing University, China
hezhang@nju.edu.cn

Abstract—Enhancing the modular structure of existing systems has attracted substantial research interest, focusing on two main methods: (1) software modularization and (2) identifying design issues (e.g., smells) as refactoring opportunities. However, re-modularization solutions often require extensive modifications to the original modules, and the design issues identified are generally too coarse to guide refactoring strategies. Combining the above two methods, this paper introduces a novel concept, *PairSmell*, which exploits modularization to pinpoint design issues necessitating refactoring. We concentrate on a granular but fundamental aspect of modularity principles—*modular relation (MR)*, i.e., whether a pair of entities are separated or collocated. The main assumption is that, if the actual MR of a pair violates its ‘apt MR’, i.e., an MR agreed on by multiple modularization tools (as raters), it can be deemed likely a flawed architectural decision that necessitates further examination.

To quantify and evaluate *PairSmell*, we conduct an empirical study on 20 C/C++ and Java projects, using 4 established modularization tools to identify two forms of *PairSmell*: inapt separated pairs *InSep* and inapt collocated pairs *InCol*. Our study on 260,003 instances reveals that their architectural impacts are substantial: (1) on average, 14.60% and 20.44% of software entities are involved in *InSep* and *InCol* MRs respectively; (2) *InSep* pairs are associated with 190% more co-changes than properly separated pairs, while *InCol* pairs are associated with 35% fewer co-changes than properly collocated pairs, both indicating a successful identification of modular structures detrimental to software quality; and (3) both forms of *PairSmell* persist across software evolution. This evidence strongly suggests that *PairSmell* can provide meaningful insights for inspecting modular structure, with the identified issues being both granular and fundamental, making the enhancement of modular design more efficient.

Index Terms—Modular Structure, Architectural Smell, Architecture Analysis

I. INTRODUCTION

Software Modularity is an essential quality attribute reflecting how a system is structured into different parts (*modules*) [1]. This attribute has demonstrated a substantial impact on software reuse [2], and has been considered in various mod-

ern design scenarios, e.g., microservices-based systems [3] and LLM-enabled systems [4]. Although the debate over “*what constitutes a single module*” has sparked broad academic interests, determining appropriate modules is still challenging in practice. The reason is that modules can evolve quickly [5], due to updated business features and infrastructure technologies. Modules that worked well in the past might not fit into the current system. Thus, a large part of the community’s effort was spent on providing methodological support to improve the modularity of existing systems [5]–[8], focusing primarily on two methods:

On one hand, *software modularization* has been extensively investigated for nearly 30 years, with at least 143 papers published in the past decade [9]. Relevant work [6], [7], [10], [11], treats modularization as an optimization problem, and searches for a (near-)optimal modular solution to replace the original modules. Such solutions often ask for expensive changes to original systems, which might prevent developers from adopting them. For example, even with refactoring effort as the optimization objective, a solution may introduce up to 170 move class [12] operations to a system [5].

On the other hand, some studies focus on identifying issues in modular structure, e.g., based on quality metrics [13]–[15], anti-patterns or smells [16]–[18]. The issues are regarded as opportunities for refactoring in subsequent development, aiming at improving the degraded modules. The problem is that most of the issues are coarse at the module level, making it difficult for developers to determine refactoring strategies [19]. A typical example is Cycle Dependency [20], where the chain of relations among several modules breaks the desirable acyclic nature of modules’ dependency structure. Although we know that cycle dependencies should be broken, it is difficult to decide which dependencies to break [21].

Our idea in this paper builds on both aforementioned methods. Rather than replacing the original modules, we propose that modularization tools can help identify issues that require

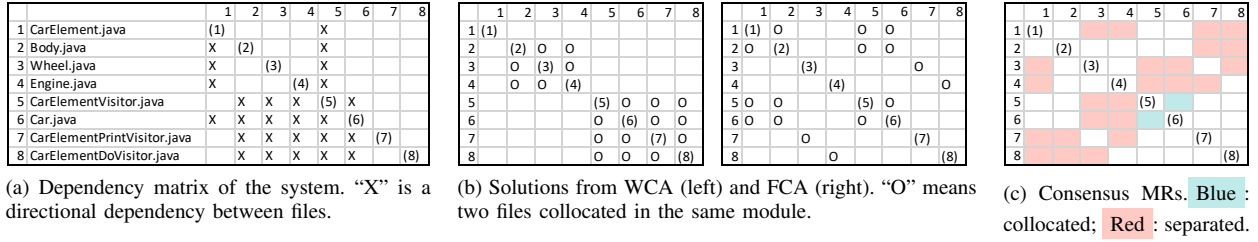


Fig. 1. MRs agreed upon by multiple modularization tools are more reliable, as they comply with multiple rule sets.

refactoring. We focus on issues with a specific granularity: *whether an entity (file) pair is collocated or separated within the same module*, termed **modular relation (MR)**. This characteristic is central to several fundamental modularity principles like Common Closure Principle [22] and Single Responsibility Principle [23]. For instance, the Common Closure Principle suggests grouping entities that often change together [22]. Moreover, entity pairs and their relationships are fundamental to many architecture analyses [24]–[26]. Our assumption is that *if multiple modularization tools consensually designate a MR as collocated or separated, it can be deemed a promisingly ‘apt MR’*, due to the consideration of diverse viewpoints. This assumption, inspired by *consensus clustering* [27], [28] where similar cluster assignments indicate strong grouping between a pair of entities, reflects the consensus-based decision making in software development (e.g., [29]–[31]). On the contrary, *if the MR of a pair violates the apt MR, this violation indicates an inappropriate architectural decision* [32], which we refer to as **Pairwise Modular Smell (PairSmell)**. In a nutshell, *PairSmell* offers granular and fundamental insights, helping developers inspect and improve software modules more effectively. It aims to identify issues necessitating refactoring based on multiple modularization tools, making development effort more targeted.

In this paper, we introduce, quantify, and evaluate *PairSmell* as a novel type of issue for inspecting modular structure. To assess the severity of this issue, we conducted an empirical study involving 20 C/C++ and Java projects from GitHub. To support this study, we developed a tool, integrating 4 established modularization tools, to automatically detect *PairSmell* from the modular structure of a development architectural view. We mined 22,528 code commits across 473 diverse snapshots, and inspected 146,668,710 separated and 3,866,940 collocated pairs of entities. Based on the dataset, our study identified 260,003 *PairSmells*, including 73,536 inapt separated pairs (a.k.a., *InSep*) and 186,467 inapt collocated pairs (*InCol*).

The empirical results reveal that: (1) *PairSmell* is prevalent among projects, with *InSep* and *InCol* instances covering 14.60% and 20.44% of the entities on average; (2) on average, entities in *InSep* MRs co-change 190% more than in other separated pairs, and entities in *InCol* MRs co-change 35% less than other collocated pairs, dramatically deviating from well-structured modules; (3) *PairSmells* persist in software projects if left unaddressed, where the percentages of *InSep* and *InCol* pairs remain stable as systems grow. In summary, our study

makes the following contributions:

- 1) A novel type of architectural smell and its identification approach are proposed, enabling the revelation of granular and fundamental modular issues.
- 2) An empirical study on the architectural smell is present, revealing that such smells are prevalent but detrimental to software maintenance and change, and could persist for long if left unaddressed.
- 3) The novel smell is discussed, and its implications for practice and future research are illustrated.
- 4) The benchmarks and replication package collected from 20 open source projects are publicly available [33] for continued research of the novel type of smell.

II. PAIRWISE MODULAR SMELL

Before proposing *PairSmell*, we first illustrate its assumption inspired by *consensus clustering* [27], [28], where similar cluster assignments for a pair of entities, indicate that these entities should be grouped together. Specifically, a simple Java system from [34] is modularized using two widely used modularization tools, WCA [35] and FCA [6] (detailed in Section II-C), as shown in Fig. 1 (a) and (b). Comparing these resulting solutions, we found promising insights in cases where both solutions consensually design the MR of a pair, as Fig. 1 (c). An instance of collocated MR is found between *CarElementVisitor* (row 5) and *Car* (6). By inspecting the architecture in Fig. 1 (a), we notice that these two files are structurally connected, featuring two direct dependencies and multiple indirect dependencies (i.e., via files in row 2, 3, and 4). Another example is a separated MR between *CarElement* (1) and *Wheel* (3), where we observe only a direct dependency. This separated MR appears justifiable, given the densely connected nature of this architecture.

To summarize, this example illustrates the rationale behind using MRs agreed by multiple modularization tools as promisingly apt MRs. It is crucial to recognize that these MRs are not infallible; they inherit potential biases from the individual tools involved. Nevertheless, this usage is justified as it diminishes the risk of biases that might be present when relying solely on a single modularization tool, thus offering a more reliable inspection regarding potential MR issues. In the remainder of this section, we first define *PairSmell* and then present an automated approach for its identification.

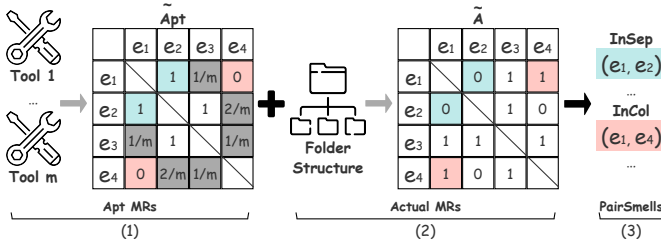


Fig. 2. Overview of identifying PairSmell.

A. PairSmell Definition

We define a *PairSmell* as a 3-tuple regarding a pair of entities e_i and e_j , where the actual MR violates its apt MR:

$$PairSmell = \langle (e_i, e_j), MR_{act}(e_i, e_j), MR_{apt}(e_i, e_j) \rangle \quad (1)$$

The first element (e_i, e_j) denotes a pair of entities in a target system, where $e_i \neq e_j$. In this study, we consider an entity as a single code file, following the common practice of architecture-level analyses (e.g., [13], [36]). Both the second and third elements ($MR_{act}(e_i, e_j)$ and $MR_{apt}(e_i, e_j)$) denote modular relations between the entities e_i and e_j . The MR of a pair in a specific design d is separated or collocated, formally:

$$MR_d(e_i, e_j) = \begin{cases} 0, & \text{if } mod_d(e_i) \neq mod_d(e_j) \\ 1, & \text{if } mod_d(e_i) = mod_d(e_j) \end{cases} \quad (2)$$

where $mod_d(e_i)$ is the module to which e_i belongs in design d . $MR_{act}(e_i, e_j)$ is the actual modular relation of the pair, which could be extracted from a system snapshot. Inspired by *consensus clustering* [27], [28], this work considers an MR apt if it is agreed upon by multiple modularization tools. In contrast, if modularization tools disagree, it suggests that the pair may be reasonably designed as either collocated or separated. Formally, an apt MR exists if:

$$MR_{d_1}(e_i, e_j) = \dots = MR_{d_m}(e_i, e_j) \quad (3)$$

where m is the number of modularization tools considered.

B. Identification Approach

For *PairSmell* identification, we first infer the apt MRs agreed upon by multiple modularization tools, and then utilize them as references to identify smell candidates. Fig. 2 illustrates our approach in three steps.

1) Inferring Apt MRs: This step infers apt MRs by comparing m solutions from distinct modularization tools.

Given n entities in a system, the solution from modularization tool t_i could be denoted as $Mod_i = \{mod_i(e_1), \dots, mod_i(e_n)\}$, where $mod_i(e_j)$ is the module to which e_j belongs in the solution. We could construct a Co-Association matrix $\tilde{Apt} \in \mathbb{R}^{n \times n}$, as defined in the consensus clustering field [27], [37], to denote the frequency that e_i and e_j occur in the same module across m solutions:

$$\tilde{Apt}_{ij} = \frac{1}{m} \sum_{k=1}^m MR_k(e_i, e_j) \quad (4)$$

where $MR_k(e_i, e_j)$ is the modular relation between e_i and e_j in the solution from modularization tool t_k .

In an \tilde{Apt} matrix, if most of the modularization tools separate e_i and e_j (i.e., \tilde{Apt}_{ij} near to 0), these two entities are very likely to belong to different modules. Similarly, if most of the tools group e_i and e_j into the same module (i.e., \tilde{Apt}_{ij} near to 1), the two entities are very likely to belong together. On the contrary, for the entity pairs with \tilde{Apt}_{ij} between 0 and 1 but near to neither, the modularization tools suggest relatively inconsistent MRs. That is, these pairs could be reasonably implemented as either *separated* or *collocated*, and there is no apt MR for them. We define two types of apt MRs that could be inferred from the matrix.

- *apt separated*: denotes a pair which should be separated according to the tools. This occurs when all tools consistently suggest separating the MR for a pair, i.e., $\tilde{Apt}_{ij} = 0$.
- *apt collocated*: indicates a pair which should be collocated. An *apt collocated* exists if the suggested MRs for a pair are collocated by all tools, i.e., $\tilde{Apt}_{ij} = 1$.

A cell in the matrix denotes an apt MR if all modularization tools agree with the MR, i.e., with \tilde{Apt}_{ij} as 0 or 1 (as Fig. 2). Note that we left out those pairs whose MRs are inconsistent (i.e., $0 < \tilde{Apt}_{ij} < 1$), denoted as gray in the figure. In this sense, we reduce the biases that could be introduced by individual modularization tools (e.g., due to specific rules), thereby enhancing the reliability of the apt MRs we derive.

2) Recovering Actual MRs: In this step, we collect the actual MRs from a system's existing modules.

A key question is, what are the existing 'modules' in a system? In this study, we consider the folder structure of a system as its existing modules and extract from it the actual MRs. This is because folders represent the actual code organization structure in the development environment, which is created by the developers of the systems [13]. In fact, folders display a development architectural view, dating back to Kruchten's seminal 4+1 view model [38].

Folder structure can be represented by a tree hierarchy of folders and sub-folders. Each leaf of the tree is an entity contained in a folder, which itself may belong to a higher-level folder (super-folder). To align with the prevailing notion of mutually exclusive modules in software engineering, e.g., in [9], [10], we do not consider all folders as 'existing' modules. Instead, we select only the lowest level folders to serve as the existing modules for specificity. Consequently, two entities are considered co-located in the same module only if they reside in exactly the same folder.

Based on the recovered modules, we define an actual MR matrix $\tilde{A} \in \mathbb{R}^{n \times n}$, as Fig. 2. Each cell of \tilde{A} represents the MR between two entities within the existing system. The possible value for each cell adheres to formula 2.

3) Identifying PairSmell Candidates: Next, we detect smell candidates, by comparing the apt MRs with the actual MRs.

For each pair of entities e_i and e_j , the apt MR and the actual MR constitute two binary expressions: $MR_{apt}(e_i, e_j)$ and $MR_{act}(e_i, e_j)$. The possible value for each expression is

	1	2	3	4	5	6	7	8
1 KStreamTransformValues.java	(1)	0.75	0.75	0.75	0.75	0.75	1	0.75
2 KTableFilter.java	0.75	(2)	1	0.75	0.75	0.75	0.75	1
3 KTableImpl.java	0.75	1	(3)	0.75	0.75	0.75	0.75	1
4 KTableKTableAbstractJoin.java	0.75	0.75	0.75	(4)	0.75	0.75	0.75	0.75
5 KTableKTableLeftJoin.java	0.75	0.75	0.75	0.75	(5)	0.75	0.75	0.75
6 KTableKTableRightJoin.java	0.75	0.75	0.75	0.75	0.75	(6)	0.75	0.75
7 KTableReduce.java	1	0.75	0.75	0.75	0.75	0.75	(7)	0.75
8 Processor.java	0.75	1	1	0.75	0.75	0.75	0.75	(8)

Fig. 3. An *InSep* example. Each number indicates the average frequency that two entities are grouped together by tools. Entities in a lined rectangle actually belong to one module.

	1	2	3	4	5	6	7	8
1 Consumer.java	(1)	0.5	0.5	0.5	0.5	0.5	0.75	0
2 ConsumerConfig.java	0.5	(2)	0.75	0.5	0.75	0.75	0.5	0
3 ConsumerInterceptor.java	0.5	0.75	(3)	0.5	0.75	0.75	0.5	0
4 ConsumerRecord.java	0.5	0.5	0.5	(4)	0.5	0.5	0.5	0
5 ConsumerRebalanceListener.java	0.5	0.75	0.75	0.5	(5)	0.75	0.2	0
6 MockConsumer.java	0.5	0.75	0.75	0.5	0.75	(6)	0.5	0
7 KafkaConsumer.java	0.75	0.75	0.5	0.5	0.2	0.5	(7)	0
8 OffsetResetStrategy.java	0	0	0	0	0	0	0	(8)

Fig. 4. An *InCol* example with the same annotations as Fig. 3.

1 or 0. We enumerated all 4 combinations of these two expressions. The 2 combinations with consistent $MR_{apt}(e_i, e_j)$ and $MR_{act}(e_i, e_j)$ indicate that the actual MR between e_i and e_j is appropriate, as it aligns with the promisingly apt design. For the other two combinations where $MR_{apt}(e_i, e_j)$ and $MR_{act}(e_i, e_j)$ are inconsistent, we define 2 specific forms of *PairSmell*—*InSep* and *InCol* as follows:

Inapt Separated (*InSep*)—two entities are separated into different modules in the actual system but the *apt* MR is collocated according to modularization tools. This smell means that the two separated entities are highly related, *e.g.*, they may depend on each other, thus all tools group them together. The inapt MR of these two entities may hamper the independence of corresponding modules, making changes of one module propagating to another module [39]. All entity pairs that match this form is denoted by *InSep* and identified by:

$$InSep = \{(e_i, e_j) | \neg MR_{act}(e_i, e_j) \wedge MR_{apt}(e_i, e_j)\} \quad (5)$$

Fig. 3 shows an instance of *InSep* detected in the project *Kafka*. *Processor* is located in a separate module from all other files. However, we can see from the cells annotated with 1 that all tools assigned the *Processor* file to be collocated with files *KTableFilter* (row 2) and *KTableImpl* (3).

Inapt Collocated (*InCol*)—two entities are actually implemented as collocated but the *apt* MR by all tools is to separate them. This smell indicates that the two entities combined in one module are to some extent irrelevant, *e.g.*, they address different (or even orthogonal) concerns. This inapt MR may impede the cohesion of the current module, violating the single responsibility principle [23]. All pairs that match this form are denoted by *InCol* and identified as:

$$InCol = \{(e_i, e_j) | MR_{act}(e_i, e_j) \wedge \neg MR_{apt}(e_i, e_j)\} \quad (6)$$

Fig. 4 depicts an instance of *InCol* in *Kafka*. *OffsetResetStrategy* is in the same module with other files. However, the cells annotated with a zero number in the DSM reveal that all modularization tools separate it from other files.

C. Tool Implementation

To implement the identification of *PairSmell*, we used four modularization tools as the basis to infer apt MRs. Our selection criteria are five-fold: (1) The tool is able to modularize C/C++ and Java projects, which are our focus in this study; (2) Its analysis unit is the code file, which is the entity considered in this paper; (3) Its approach should either be established with promising results in prior empirical studies (WCA, LIMBO, and ACDC) or advanced from the latest research published within the last five years (like FCA); (4) Its source code should be accessible; (5) The tool uses deterministic techniques, *i.e.*, each execution yields the same result, to avoid the effects of randomness in our results. The final tools are:

- **WCA** [35] is a hierarchical clustering algorithm using inter-cluster distance to extract software modules. The distance can be calculated by two similarity metrics: *UE* and *UENM*. We used *UENM* as it outperforms *UE* [40].
- **LIMBO** [41] is a hierarchical algorithm that clusters categorical data using a distance measure (called mutual information) to minimize information loss in the clusters.
- **ACDC** [42] clusters software entities based on specific patterns (*e.g.*, body-header) and uses orphan adoption technique to assign remaining entities to clusters.
- **FCA** [6] is a clustering algorithm that maximizes intra-connectivity and minimizes inter-connectivity in clusters.

For running these tools, we provide a dependency graph, as required by the tools. The nodes of the graph represent software entities, and the edges represent the structural relations between entities. This paper uses *Depends* [43] to recover structural relations, as it is capable of extracting 13 dependency types by analyzing the syntactic structures of C/C++/Java programs, such as *call*, *contain*, and *implement*. In this sense, the extensive data collected enables us to recover more accurate dependency graphs of software systems, which generally results in enhanced modularization solutions.

Tool Evaluation. *PairSmell* is defined based on the deviations of actual MRs from the apt MRs. Since there is no a set of *ground-truth* apt MRs, it is hard to construct a validation set. Thus, we decide to manually examine the detection results of our tool, similar to the methodology used by Kim et al. [44].

We start by executing the tool on 20 projects, listed in Section III-A, which results in 9,415 smell instances. Then, we derive a sample to be manually validated. We randomly select 370 out of 9,415 smells, based on a 95% confidence level and 5% confidence interval [45]. This includes 129 *InSep* and 241 *InCol* pairs. Each pair is then independently examined by two authors to decide the correctness. Both annotators produced identical results after completing the tagging, and our validation process achieved a precision of 100%. However, we cannot evaluate the recall due to the lack of oracles.

III. EMPIRICAL STUDY

The *goal* of this study is to provide empirical evidence for assessing the severity of *PairSmell* in practice, focusing particularly on its prevalence, impact, and evolution—three fun-

TABLE I
SUMMARY OF THE STUDIED SOFTWARE PROJECTS.

P_i	Project(l)	Domain	Version	#Entity	#Link	#Cmt
P_1	Arrow(c)	Memory analytics	0.15.0	568	3,003	5,159
P_2	Brpc(c)	RPC framework	1.5.0	385	345	2,032
P_3	Cassandra(j)	Row store	0.6.10	283	5,569	1,752
P_4	Druid(j)	Analytics database	0.7.0	1,045	7,651	4,980
P_5	Gobblin(j)	Data management	0.9.0	1,279	9,743	3,717
P_6	Hadoop(j)	Distributed framework	0.20.0	890	17,266	3,461
P_7	Hbase(j)	Storage system	1.0.2	1,456	34,968	10,061
P_8	Httpd(c)	Web server	2.0.46	229	3,349	11,539
P_9	Impala(c)	SQL framework	2.7.0	439	490	4,934
P_{10}	Iotdb(j)	Data management	0.11.0	836	19,273	4,209
P_{11}	Kafka(j)	Event streaming	0.10.2.1	747	11,593	3,247
P_{12}	Kudu(c)	Storage engine	0.7.0	514	78	4,022
P_{13}	Kvrocks(c)	NoSQL database	2.8.0	220	4,716	1,262
P_{14}	Lucene(j)	Search engine	2.9.2	1,006	21,377	4,042
P_{15}	Mahout(j)	DSL framework	0.6	1,052	12,939	2,269
P_{16}	Mesos(c)	Cluster manager	0.21.2	281	554	3,713
P_{17}	Ozone(j)	Object store	1.0.0	1,380	8,595	2,698
P_{18}	Pulsar(j)	Pub-sub messaging	2.3.0	1,142	20,519	2,892
P_{19}	Thrift(c)	RPC framework	0.12.0	202	483	5,384
P_{20}	Traffic(c)	Caching proxy server	4.2.0	963	34,589	4,301

damental aspects critical to investigating a phenomenon [46]–[49]. Specifically, the study addresses three research questions:

- RQ1. *To what extent does PairSmell appear in software projects?* This question aims to quantitatively assess the prevalence of *PairSmell* in software projects. If *PairSmell* is prevalent, *a.k.a.*, its amount is notable per project, it suggests that the proposed smell merits further attention.
- RQ2. *To what extent does PairSmell impact software maintenance?* This question assesses how *PairSmell* affects software maintenance by analyzing the co-change extent manifested in project revision history. If the co-change extent of smelly pairs significantly and detrimentally differs from non-smelly pairs, it indicates a deviation from the ideal modular structure of well-maintained systems.
- RQ3. *How does the amount of PairSmells evolve across time?* With this question, we aim to investigate the amount of *PairSmell* as systems evolve. If *PairSmell* proliferates, or at least does not diminish, across time, it would denote a significant motivation for its removal.

A. Data Collection

For empirically answering the research questions, we choose open-source software projects as study subjects by following three predefined criteria: (1) C/C++ and Java projects on GitHub because they are among the most popular programming languages; (2) projects with at least 2 years of change history and over 1,000 commits, so that they can provide sufficient evolution data for analyzing the impact of *PairSmell* in software evolution and maintenance; (3) non-trivial projects with at least 100 entities, because architecture smells turn to be significant especially for non-trivial projects [47]. The selected projects are shown in Table I, together with their number of entities (#Entity), relationships between the entities (#Link), and commits (#Cmt). These projects differ in their scale, business domains, and other characteristics. All data we used are publicly available [33].

TABLE II
INSEP AND INCOL INSTANCES IN THE CURRENT VERSION.

P_i	InSep			InCol		
	Pair(%)	Entity(%)	Density	Pair(%)	Entity(%)	Density
P_1	3(<0.01)	6(1.06)	1.00	143(1.98)	86(15.14)	3.33
P_2	0(0)	0(0)	0.00	13(0.29)	13(3.38)	2.00
P_3	27(0.07)	34(20.14)	1.59	365(16.60)	171(60.42)	4.27
P_4	85(0.02)	119(11.39)	1.43	90(1.01)	63(5.00)	2.86
P_5	80(0.01)	129(10.09)	1.24	61(0.96)	64(5.00)	1.91
P_6	334(0.09)	242(27.19)	2.76	162(1.08)	114(12.81)	2.84
P_7	960(0.09)	434(29.81)	4.42	176(0.73)	65(4.46)	5.42
P_8	1(<0.01)	2(0.87)	1.00	199(15.05)	95(41.49)	4.19
P_9	0(0)	0(0)	0.00	57(0.51)	43(9.80)	2.65
P_{10}	134(0.04)	167(19.98)	1.60	189(5.90)	172(20.57)	2.20
P_{11}	93(0.04)	125(16.73)	1.49	66(0.68)	80(10.71)	1.65
P_{12}	1(<0.01)	2(0.39)	1.00	36(0.27)	26(5.06)	2.77
P_{13}	0(0)	0(0)	0.00	335(14.56)	130(59.09)	5.15
P_{14}	398(0.08)	279(27.73)	2.85	536(2.70)	332(33.00)	3.23
P_{15}	987(0.18)	349(33.18)	5.66	36(0.73)	39(3.71)	1.85
P_{16}	0(0)	0(0)	0.00	75(2.19)	33(11.75)	4.55
P_{17}	119(0.01)	151(10.94)	1.58	44(0.65)	48(3.48)	1.83
P_{18}	66(0.01)	108(9.46)	1.22	113(1.33)	114(9.98)	1.98
P_{19}	0(0)	0(0)	0.00	67(5.64)	32(15.84)	4.19
P_{20}	0(0)	0(0)	0.00	3,364(16.10)	742(77.05)	9.07
Avg.	164(0.03)	107(14.60)	1.44	306(4.45)	123(20.44)	3.40

B. RQ1: Prevalence of PairSmell

1) *Setup*: To answer RQ1, we identify *PairSmell* on the current version of 20 projects (cf. Table I). We study how frequently *PairSmell* appears at both pair and entity levels. Please note that an entity affected by *PairSmell* is involved in at least one smell instance. To provide a comparative statistic, we calculate the percentages of *PairSmell* relative to the total number of corresponding program elements (pairs or entities). For example, we calculate at the pair level, the proportion of *InSep* among all separated pairs in a project, indicating the extent of inappropriate MR design for separated pairs. In addition, we calculate smell density to measure the ‘smelliness’ of a specific smell form x (*InSep* or *InCol*) among the affected entities. This metric quantifies the average number of smell instances concurrently affecting each entity, and is computed as follows:

$$Density(x) = \frac{Total\ instances\ of\ x \times 2}{Number\ of\ entities\ involved} \quad (7)$$

2) *Results*: Table II presents the prevalence of *InSep* and *InCol* in different projects. The 2nd and 5th column show the numbers and percentages of *InSep* and *InCol* at pair level. On average, 164 *InSep* pairs were identified in each project. For *InCol*, the average number of smells could be as high as 306 (over 4%). In certain projects, *e.g.*, P_2 , only a few *PairSmells* were identified, indicating that the MR design in these projects tends to be structurally sound. Overall, the presence of *PairSmell* is noteworthy across the 20 projects.

From the 3rd and 6th column, both *InSep* and *InCol* are widespread among software entities in the projects. About 15% of entities in each project are affected by *InSep*, while a higher average is observed for *InCol*. That is, a substantial proportion of entities are impacted by *PairSmell* in these projects.

Columns 4 and 7 present the smell *density* among affected entities. Results show that each ‘smelly’ entity is involved, on average, in 1.44 *InSep* pairs, and 3.40 *InCol* pairs.

To explore the differences between *InSep* and *InCol*, Fig. 5 shows the number of separated and collocated pairs aggregated across all projects, and how they overlap with the apt MRs. The two overlapping parts constitute the sets of *InSep* and *InCol* ‘smelly’ pairs respectively. For example, among the 194 pairs where all modularization tools design them to be collocated (*i.e.*, apt collocated), 164 (84.5%) are actually implemented as separated (*i.e.*, *InSep*). Such structuring into different modules increases inter-module coupling. In contrast, only 2.0% apt separated pairs are actually implemented as collocated (306 out of 15,375), suggesting developers’ caution for structuring responsibilities into modules.

RQ1 Summary: Both *InSep* and *InCol* entities are prevalent in the dataset. Developers seem more inclined to organize highly relevant entities into separate modules (*i.e.*, incur *InSep*), thus introducing inter-module coupling, than grouping responsibilities within the same module (*i.e.*, incur *InCol*).

C. RQ2: Impact on Software Maintenance

Code revision history, is frequently used as a benchmark to investigate the impact of generic smells, *e.g.*, how smells impact fault- and change-proneness [16], [50], smells’ impact on maintainability [51]–[53] or file co-change [54]. Given that *PairSmell* describes a problematic relationship between entities, our evaluation focuses on its impacts on file co-change relation. The underlying principle is, within a healthy modular structure, files in the same module should change together, while files from different modules should change independently. This RQ compares the co-change of smelly versus non-smelly pairs, within and between modules, to explore if *PairSmell* disrupts the expected healthy structure.

1) *Setup*: File co-change in prior studies is typically captured by the absolute frequency, *i.e.*, the number of commits that a file pair change together [34], [54], which is inefficient for comparing co-change extents among different pairs [55]. Additionally, various types of evidence, not just commits, have been used to measure software maintenance and changes [16], [56]. To robustly assess *PairSmell*’s impact, we propose a suite of measures based on relative measurement theory [15], [57], [58], utilizing commonly used evidence, as illustrated in Fig. 8.

For any two entities e_i and e_j , their commit sets [34] during a specific time period can be represented as Cmt_i and Cmt_j . If two entities changed in completely different commits, as

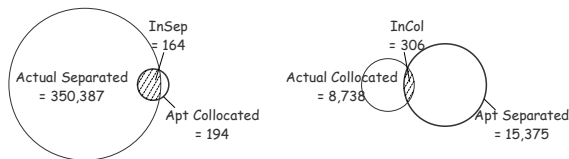


Fig. 5. The sets of *InSep* and *InCol*, averaged over 20 projects.

Fig. 8 (a), they are likely independent and can change without affecting each other. On the contrary, if e_i and e_j shared exactly the same commit sets, as Fig. 8 (b), these entities co-changed consistently. Based on these observations, we define three measures to quantify the extent of co-change between a pair:

1. **Commit Overlap Rate (COR)**: measures the extent changes made to two entities overlap. $COR = \frac{2 * |Cmt_{i,j}|}{|Cmt_i| + |Cmt_j|}$, where Cmt_i is the commit set that changed entity e_i , $Cmt_{i,j}$ is the commit set where entities e_i and e_j changed together. A larger *COR* means more overlap between two entities’ commits, indicating that these entities are more relevant.

2. **Code Change Overlap (CCO)**: measures the likelihood that code changes [59], [60] to two entities occurred simultaneously. $CCO = \frac{|Ch_{i,j}|}{|Ch_i| + |Ch_j|}$, where Ch_i is the lines of code changed in entity e_i , $Ch_{i,j}$ is the lines of code changed in either e_i or e_j that occurred together in the same commit. $Ch_{i,j}$ is counted once because Ch_i and Ch_j do not intersect. The larger the *CCO*, the more often two entities undergo simultaneous code changes, indicating a more relevant pair.

3. **Developer Overlap Rate (DOR)**: measures to what extent the sets of developers [34], [36] changing two entities overlap. $DOR = \frac{2 * |Dev_{i,j}|}{|Dev_i| + |Dev_j|}$, where Dev_i indicates the developer set changing entity e_i , $Dev_{i,j}$ is the developer set changing both e_i and e_j . The higher the value, the more likely the two entities were changed by the same developers, suggesting a possibly greater relevance between them.

We use K_{COR} , K_{CCO} , and K_{DOR} to comprehensively assess the relative co-change of a smelly pair as compared with that of a non-smelly pair, similar to the work of Mo et al. [60]. Our hypothesis is that an *InSep* pair is more likely to be related than other separated pairs, and thus more co-changed; in contrast, an *InCol* pair is less likely to be related than other collocated pairs and therefore less co-changed. The detailed measures are as follows:

$$K_{mtr} = \frac{mtr \text{ of Smelly pairs (avg.)}}{mtr \text{ of Non-Smelly pairs (avg.)}} \quad (8)$$

where *mtr* can be *COR*, *CCO*, and *DOR*. For *InSep*, *Smelly* pairs denote pairs in the *InSep* set, and *Non-Smelly* pairs are those in the set of *Separated-InSep*. For *InCol*, these are the sets of *InCol* and *Collocated-InCol*. For a project, a K_{COR} value (or K_{CCO} , K_{DOR}) exceeding 1 means that *InSep* pairs co-changed more frequently than other separated pairs. Conversely, a value less than 1 suggests that *InCol* pairs co-changed less frequently than other collocated pairs.

2) *Results*: Fig. 9 shows the values of K_{COR} , K_{CCO} , K_{DOR} regarding *InSep* and *InCol*. These values were calculated by mining 100, 200, and 300 commits before the current version (*Delta*) of each project, to ensure an evaluation with a sufficient evolution history [16]. We did not mine a project’s revision history from its beginning since an identified smell might not be smelly in the initial stages. Each point in the figure denotes the result for a single project. Some projects have no points in specific analyses, because the corresponding

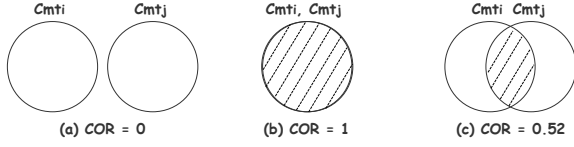


Fig. 8. The commit sets of two entities overlap differently.

smell sets are empty (as shown for 6 projects in Table II) or no smelly pairs were changed during the analyzed commits.

Considering the K_{COR} score for *InSep* as Fig. 9 (a), most of their values are greater than 1, except for 3 values below $K_{COR} = 1$ line. Similar results can be observed from other scores. This indicates that, although belonging to different modules, *InSep* pairs are more likely to be changed together than other separated pairs. The *Independent Samples T-test* is conducted to compare the mean numbers of the co-change between smelly and non-smelly pairs. Table 3 presents the *Effect size* (Cohen's d [61]) calculated for each T-test to show the strength of the differences, indicating medium to large differences for most deltas for *InSep*. As for the K_{COR} values for *InCol*, 14 out of 20 (70%) values are less than 1 (analyzed using 300 commits). Similar results are observed for other metrics, indicating that despite collocation, *InCol* pairs are less co-changed in their evolution than other collocated pairs, possibly suggesting a responsibility overload in the modules. Interestingly, the differences are significant only in the analysis using 100 commits but not in that with longer history length. This could be attributed to the variability of *InCol* smells across time (cf. Section III-D), implying that some *InCol* instances might not be smelly in a previous version.

On average, the differences of K values for *InSep* are larger than that for *InCol*. In the analysis using 100 commits, the average K values for *InSep* are $K_{COR} = 2.86$, $K_{CCO} = 3.00$, and $K_{DOR} = 2.83$ across all projects. That is, a *InSep* pair is on average $\frac{(2.86-1)+(3.00-1)+(2.83-1)}{3} = 190\%$ more likely to co-change than a separated pair without smell. For *InCol*, the averaged values are $K_{COR} = 0.68$, $K_{CCO} = 0.55$, and $K_{DOR} = 0.71$. The likelihood of *InCol* pairs co-changing is $\frac{(1-0.68)+(1-0.55)+(1-0.71)}{3} = 35\%$ lower than that of other collocated pairs. We assume that the difference

TABLE III
CO-CHANGE DIFFERENCES BETWEEN SMELLY AND NON-SMELLY PAIRS.
GRAY RESULTS ARE significant differences with $p < .05$.

Delta	InSep			InCol		
	K_{COR}	K_{CCO}	K_{DOR}	K_{COR}	K_{CCO}	K_{DOR}
300	.91	.61	.88	.09	.06	.06
200	.97	.67	.95	.08	-.01	.06
100	.64	.49	.64	-.57	-.91	-.45

between *InSep* and *InCol* stems from the fact that separated pairs are generally rarely (if ever) modified simultaneously; as a result, the frequent co-changes among *InSep* pairs appear more evident and detrimental by comparison. In fact, in over 50% projects, the average COR values of other separated pairs (i.e., *Separated* – *InSep* in Fig. 10) are close to 0.

RQ2 Summary: The pairs identified as *InSep* are 190% more likely to co-change compared to separated pairs without smells, whereas *InCol* pairs exhibit 35% less co-change than proper collocated pairs. Both of these observations indicate that the modular structure is significantly undermined, and software maintenance is adversely affected.

D. RQ3: Evolution of PairSmell

In this question, we analyze how the amount of smells changes across time to explore whether *PairSmell* will proliferate in a system if left unaddressed.

1) *Setup*: To answer RQ3, we gather all smell instances for each project across its evolution history. Considering that each commit may alter the architecture and affect the smell instances, it would be strenuous to analyze each commit in the history. Instead, we opt to analyze snapshots by selecting

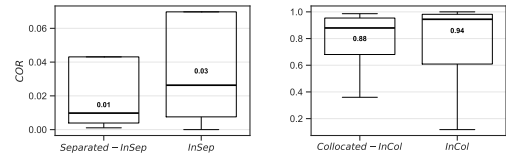


Fig. 10. COR distribution of different pairs (100 commits).

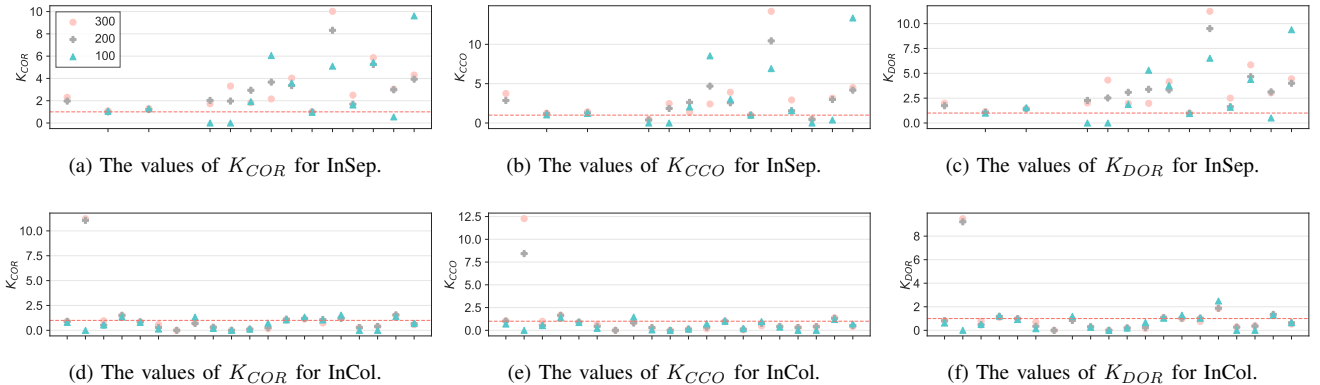


Fig. 9. The distributions of K_{COR} , K_{CCO} , K_{DOR} for *InSep* and *InCol* in all projects. Red lines “- -” mean $K_{mtr} = 1$.

one commit every two weeks before the current version in Table I. Our goal is to capture the evolution activities over approximately a year, which results in 25 snapshots for each project (including the current version). However, some projects may not experience changes during certain periods; therefore our analysis ultimately covers a total of 473 distinct snapshots.

To conduct a global analysis of *InSep* and *InCol*, we aggregate the percentages of smells at both the pair and entity levels across all projects and then compute the average values. We choose not to analyze the absolute number of smells, as the increase of this value could be attributed to the growing system size according to prior studies [62], [63]. We represent the average percentages at each level as a time series: s_1, \dots, s_{25} , where s_i is the averaged percentage for that level across all projects at the i -th snapshot. We collect time series for *InSep* and *InCol* respectively.

For each smell form, we determine the overall evolution trend for the percentage of smells: increase, decrease, or stable. We notice a non-monotonic trend in the percentage of smells, *i.e.*, the value increases and decreases at different time intervals. To account for such a non-monotonic trend, we fit a simple linear regression model, denoted as *lm*, and determine the trend by examining the sign of the *slope* of the regression line, similar to the work of Soto-Valero et al. [63].

2) *Results*: Fig. 11 shows the evolution trend of *InSep* and *InCol* at pair level across all analyzed snapshots. Each data point represents an average percentage measured for each snapshot. The lines are linear regression functions, fitted to show the trend of *InSep* and *InCol* at a 95% confidence interval. From Fig. 11, the average percentages of *InSep* remain stable across time. For example, the percentage of *InSep* in snapshot s_1 is 0.03%, and by snapshot s_{25} this value is still near to 0.03%. For *InCol*, although we observe a slight decreasing tendency as systems grow, we find that such a tendency is not statistically significant (with *slop* near to 0 and $p = 0.26$). Moreover, *Chow Test* shows that the slopes of the two linear regressions in Fig. 11 are not significantly different ($p = 0.14$). Thus, we conclude that overall, the percentages of smelly pairs for both forms remain stable over time, indicating that developers did not effectively intervene in *PairSmell* issues within the analyzed time span.

Fig. 12 shows the evolution trend of entities involved in *InSep* and *InCol*. Interestingly, we observe a clear and significant increasing tendency from the percentages of entities affected by *InSep*, despite the stable trend at pair level. Specifically, the proportion of entities affected by *InSep* is

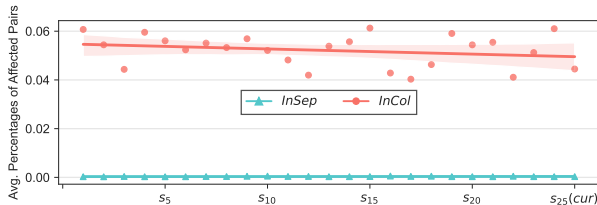


Fig. 11. Stable evolutionary trends of *InSep* and *InCol* at pair level, averaged across all projects.

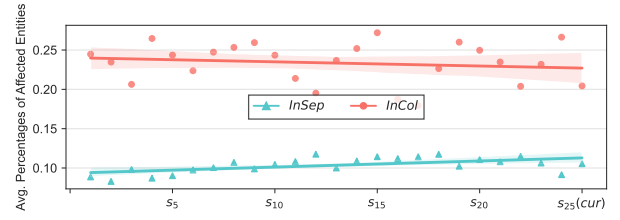


Fig. 12. Evolutionary trends at entity level: increasing for *InSep* and stable for *InCol*, averaged across all projects

8.87% in snapshot s_1 and 10.54% in snapshot s_{25} (increase = 1.19x). This indicates that the number of entities newly affected by *InSep* is generally higher than that of entities previously affected but no longer smelly, as systems grow. On the other hand, a slight decreasing tendency can be observed from the percentages of entities with *InCol*. Despite this, statistics show that such a tendency is not significant ($p = 0.47$). We notice that the percentage of entities affected by *InCol* is more variable ($SD = 0.03$) and represents a larger share in comparison with that affected by *InSep* ($SD = 0.01$). *Chow Test* indicates that the slopes of the two linear regressions in Fig. 12 are significantly different ($p < 0.01$).

RQ3 Summary: The percentages of both *InSep* and *InCol* pairs do not diminish across the analyzed time, suggesting that *PairSmell* increases with system growth. Moreover, the percentages of entities affected by *InSep* grow more noticeably, indicating a widespread and concerning phenomenon.

IV. DISCUSSION AND IMPLICATIONS

Based on our empirical findings, this section discusses the discovery, management and further study of *PairSmell*.

A. Discovery of *PairSmell*

The discovery of new software smells, since Fowler's seminal work [12] on code smells, generally follows inductive and deductive approaches, as summarized in Table IV. In inductive approach, recurring observations lead to the generalization of new smells [32], [59], [64], [66], while in deductive approach, new smells are derived from theoretical premises [68], [71]. These approaches differ in their characteristics and processes of smell discovery, particularly in definition, detection, and assessment. By reflecting on our research process and integrating methodologies reported in previous smell discovery studies (*e.g.*, [66], [67], [72]), we derived Table IV. The characteristics and processes outlined in this table serve provide a reference and guide for future researchers and practitioners in proposing new smells.

PairSmell, proposed in this study, is based on the premise that decisions violating appropriate or ideal ones could be problematic. By focusing on the MR perspective, it offers a granular and fundamental aspect for inspecting modularity principles. Unlike the inductive approach, this deductive method (1) broadens the scope of smell knowledge by identifying potential issues previously unrecognized within the

TABLE IV
OVERVIEW, CHARACTERISTICS, AND PROCESSES OF INDUCTIVE AND DEDUCTIVE APPROACHES TO DISCOVERING NOVEL SMELLS

	Inductive Approach	Deductive Approach
Overview	Smells are generalized from recurring observations in practice.	Smells are inferred from established premises.
Perspective	Now and past (problems observed in existing artifacts)	Now and future (possible problems based on theoretical premises)
Initiators	Practitioners, or researchers collaborating with practitioners	Researchers
Definition	<ul style="list-style-type: none"> Observe and gather instances where the problem manifests. <i>e.g.</i>, Configuration smells in [64] are discovered based on vulnerable packages. Identify the recurring characteristics across different instances. <i>e.g.</i>, Authors [65] observed recurring coding patterns as security smells. Formulate a rule encapsulating the characteristics and justify its impacts. <i>e.g.</i>, The impact of <i>flaky test</i> is elucidated using real-world cases [66]. 	<ul style="list-style-type: none"> Formulate a theoretical premise that logically suggests specific problems. <i>e.g.</i>, Our premise is that decisions violating the apt ones could be problematic. Describe what the problem looks like (<i>e.g.</i>, analysis units, problematic structure). <i>e.g.</i>, PairSmell focuses on MRs and their corresponding deviations (Section II-A). Justify the problem as a smell by highlighting its negative impacts on quality. <i>e.g.</i>, Section II-B illustrates how PairSmell could impair a healthy modular structure.
Process		
Detection	<ul style="list-style-type: none"> Define detection criteria (targets, indicators) based on inductive insights. <i>e.g.</i>, <i>Manual execution</i> is a configuration smell, except in deploy stages [67]. Develop logical mechanisms (<i>e.g.</i>, algorithms) based on inductive data. <i>e.g.</i>, Authors [67] set the <i>Retry Failure</i> threshold based on known causes. Implement and evaluate the detection tool with validation datasets. <i>e.g.</i>, Known or labeled smells [67], [68] can serve as validation dataset. 	<ul style="list-style-type: none"> Translate the smell definition into quantifiable metrics based on the premise. <i>e.g.</i>, PairSmell considers the MR between a pair as a key metric (Section II-B). Develop logical mechanisms with theoretical consistency. <i>e.g.</i>, Apt MRs, actual MRs, and their discrepancies help identify PairSmells. Implement and evaluate the tool using validation datasets or manual review. <i>e.g.</i>, Premise ensures the detection results align with expectations (Section II-C).
Method		
Design		
Usefulness	<p>Prevalence: Investigating its occurrences in practice to show its prevalence and importance.</p> <ul style="list-style-type: none"> Detect the smell (using the developed tool) in real software projects and observe its frequencies and percentages. Observations can guide project selection; <i>e.g.</i>, Jafari et al. [64] excluded projects without “package.json” as it hinders pinpointing dependency smells. <p>Premises help framing interpretation; based on our premise, Section III-B presents the number of apt MRs, actual MRs, and detected deviations (smells).</p> <p>Consequences: Provide empirical evidence demonstrating its impact on software quality.</p> <ul style="list-style-type: none"> Collect quantitative data to show how the smell affects key quality metrics (<i>e.g.</i>, change-proneness), by comparing code artifacts with and without smells [69]. <p>Impacts can be hypothesized based on observations or premises. Our hypothesis (Section III-C) stemmed from the unhealthy structure of <i>PairSmells</i> vs. other pairs.</p> <ul style="list-style-type: none"> Gather qualitative feedback from developers on the smell’s impact on their workflow and codebase, <i>e.g.</i>, via issue reporting [65], [67] and surveys [64], [70]. <p>Both inductive insights [70] and established premises can inform the questions posed to developers during the evaluation.</p>	
Assessment		
Benefits	1) Enhanced practitioner acceptance; 2) Easy verification.	1) Broadened scope of smell knowledge; 2) Hastened smell discovery.
Challenges	Delayed problem discovery could lead to higher maintenance costs.	Practitioners need to invest time in understanding the smell beforehand.

community, and (2) accelerates the discovery of new smells by proactively uncovering problems.

B. Management of PairSmell

Our findings indicate that *PairSmell* is significant for inspecting software modular structure (RQ1 and RQ2) and remains inadequately addressed (RQ3). This section discusses its management from three critical aspects (Fig. 13): identification, resolution, and training (prevention).

1) Early and continuous identification. A paramount benefit of *PairSmell* is its ability to be detected automatically and pinpoint specific modular issues at the pair level (Section II). This capability, requiring minimal developer effort [73], can be effectively integrated into IDE plugins for continuous assessment (coding, operating, and monitoring stages), similar to code smell tools such as SAT [74] and DARTS [72]. We envision that early and continuous identification of *PairSmell* will improve the modular structure by enabling developers to uncover and address suboptimal modular decisions promptly.

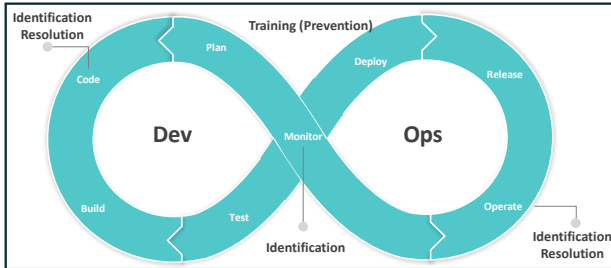


Fig. 13. Management of PairSmell in DevOps development.

2) Granular, intermittent, and selective refactoring.

Compared to prior smells, *PairSmell* provides a granular but fundamental perspective for inspecting software modules. To better manage it, we suggest the following refactoring strategies: Regarding *how to refactor*, for an *InSep* pair, developers could identify a single module to house both entities, *e.g.*, simply merging the entities into the module with the strongest connections to them. For an *InCol* pair, developers should examine the interactions between the two entities with other parts of the module, potentially separating the current module to establish clear boundaries. Considering the main issue involves two entities, the corresponding refactoring operations, such as a single move class operation, are more actionable than those for coarse-grained smells. Regarding *the timing of refactor*, engaging in *intermittent floss refactoring* [75]—consistently integrating refactoring activities throughout the development process, with particular focus on the coding and operating stages—is recommended due to the relatively low resolution costs. In addition, it is advisable to address *PairSmell* particularly during the *vibrant* and *growing* phases of projects to prevent detrimental co-changes in later development stages (Section III-C). Regarding *which smell to refactor first*, we suggest developers balance the severity of each smell instance—considering dependencies among the pair—and the remediation effort, such as the involved lines of code, following Darius Sas’ theoretical model [76].

3) Whole-process modular training. The widespread occurrence of *PairSmells* in numerous projects underscores the need for improved training in software design. We acknowledge that teaching design concepts is a challenge [19]; after all, many design principles such as SOLID [23], DRY [77], and

SoC [78] can often seem too abstract. However, educators can demystify these concepts with practical examples of modular smells, such as *PairSmells*, as demonstrated in Section II. Developers can also enhance their understanding of modular design by actively identifying *PairSmells* in their projects and conducting detailed analyses to mitigate these issues.

C. Further Study of *PairSmell*

Empirical evidence suggests that *PairSmell* could undermine the ideal modular structure of a project during software maintenance. Future studies should continue to gather feedback from practitioners on *PairSmell*.

As we find in RQ2, *InSep* correlates with increased co-changes across modules; while *InCol* pairs exhibit fewer, suggesting reduced module coherence. Both observations violate the modular design principles [23], [39] and undermine software maintenance and change. Our findings provide preliminary insights into the usefulness of *PairSmell* for inspecting software modular quality. Nevertheless, future studies should further evaluate *PairSmell* by examining its relevance to developers, to better understand its impact from the developers' perspective. One possible method is collecting developers feedback by opening *PairSmell* issues in issue trackers, similar to Vassallo et al.'s approach [67]. The validity of *PairSmell* could be confirmed if developers not only agree with the issues but also take actions to address them.

V. THREATS TO VALIDITY

Internal validity could be threaten by factors that influence smell identification. A relevant threat is that the inferred apt MRs might be biased by individual tools. Research on consensus clustering find that low-quality base clusterings can degrade the quality of final ensemble solutions [79]. The modularization tools we selected might yield poorly structured solutions and potentially unreasonable MR design. To minimize this threat, we follow a set of rigorous criteria to select tools that have demonstrated promising results. We avoid non-deterministic tools (e.g., Bunch [11]) which could introduce their own chance factors. Additionally, we choose 4 distinct tools to further reduce the likelihood of biases by individual tools while maintaining an acceptable overhead. On the other hand, we use a system's lowest level folders as its existing modules for extracting actual MRs. While folders can reflect the development architectural view, not all folders are meaningful from architecture's perspective [80]. For example, many C/C++ projects organize header files and cpp files into separate folders, and smells suggesting to move them into the same folder can be examples of *false positives*.

External validity could be threatened, impairing the generalization of our findings. We are aware that our results may not be generalizable to other projects since all 20 studied projects are open source. To minimize this threat, a set of criteria are used to select projects varying widely across different domains and project characteristics. Future studies are encouraged to replicate our research on other projects in different settings.

Construct validity could be threatened by possible imprecision in our measurements. This can be related to possible mistakes in our tool's implementation, beyond what we could discover by testing. We performed extensive manual examination to mitigate this threat. In addition, the dependent variables used to measure co-changes, i.e., COR, CCO, and DOR, are defined based on the evidence commonly used for studying co-change relations [34], [54] and software maintenance [36], and thus can be considered constructively valid.

VI. RELATED WORK

Software Modularization Techniques. Over the past two decades, numerous techniques have been developed to restructure a large software system into smaller, and more manageable subsystems [9]. These techniques typically conceptualize modularization as an optimization problem, seeking an optimal solution to refactor the original modules. The most commonly used optimization objectives are intraconnectivity (*high cohesion*) and interconnectivity (*low coupling*), e.g., in [6], [7], [10], [11]. Additionally, some researchers incorporate refactoring effort, such as the number of changes [5], as an objective to minimize the effort required for modularization. However, an industrial case study [5] reveals that completely modularizing an entire system remains prohibitively expensive and thus impractical, given the extensive size of the code base. *Instead of seeking to restructure an entire system, this paper aims to integrate the intelligence of multiple modularization techniques to deduce promisingly appropriate MR designs and identify opportunities that necessitate refactoring.*

A few modularization techniques also focus on MR decisions. Erdemir and Buzluca [81] calculated the probability of two entities being within the same module and utilized this information to promote subsequent modularization. Chong and Lee [82] obtained two constraints—an entity pair *must be* and *must not be* within a module—as the foundation for constraint-based clustering to enhance modularization solutions. *Our study diverges from these research not only in objectives but also in the methodologies used to determine apt MRs.*

Metrics and Smells for Identifying Modularity Issues. Identifying and alarming modularity-related 'issues' is an essential objective for many architecture analysis activities, such as architectural quality measurement [83] and architectural smell detection [47], [60]. Architecture metrics, including modularity and maintainability measures [36], aim to assess the extent to which a software system is maintainable. In addition, numerous metrics of coupling [24] and cohesion [84] can be employed to identify quality issues at the module level, for example, MCI [15] for microservice coupling.

Architectural smells represent structural problems that negatively influence software evolution [16], [60] to indicate refactoring opportunities in subsequent development. Since Joshua Garcia's definition [32], numerous types of architectural smells have been introduced within the community. To the best of our knowledge, smells relevant to *PairSmell* include: *Modularity Violation* [71], referring to two components that consistently change together but belong to separate modules; *Implicit*

Cross-module Dependency [59], indicating two structurally independent modules that frequently change together in the revision history; *Co-change Coupling* [85], where changes to one component require changes in another component. Compared to these smells, the novelty of *PairSmell* manifests in two aspects: (1) *PairSmell* is defined at the fine-grained pair level, thus providing more actionable insights to enhance existing software modules than those targeting the module or component level; (2) while the above smells focus on the deviation between modular structure and historical revisions, *PairSmell* concerns deviation in the modular structure from the apt or ideal design decisions, offering a broader perspective than the existing smells.

VII. CONCLUSION

Focusing on the granular pair-level, we introduce a novel architectural smell that reveals modular issues due to deviations from consensus modular decisions. With the empirical study on 20 open source projects, we explore the prevalence and consequences of such smells. Our study presents solid evidence that the impact of such smells is nontrivial, but unordinarily high in practice by comparing the pairs with and without smells.

Our study benefits software research and practice by: (1) introducing a novel type of smell for inspecting software modular structure, (2) providing empirical evidence of its prevalence and consequences, and (3) suggesting how software modular activities can be enhanced—augmented with *PairSmell*'s identification, resolution, and training. This smell envisions contributing to software engineering by enabling more targeted and effective module enhancements.

ACKNOWLEDGMENTS

He Zhang is the corresponding author of this paper. This work is supported by the Key Research and Development Program of Jiangsu Province (No.BE2021002-2), the National Natural Science Foundation of China (No.62302210, No.62202219), the Natural Science Foundation of Jiangsu Province (No.BK20241195), the Innovation Projects and Overseas Open Projects of State Key Laboratory for Novel Software Technology (Nanjing University) (ZZKT2024A18, ZZKT2024B07, KFKT2023A09, KFKT2023A10, KFKT2024A02, KFKT2024A13, KFKT2024A14), and China Scholarship Council.

REFERENCES

- [1] C. Y. Baldwin and K. B. Clark, *Design rules: The power of modularity*, vol. 1. MIT press, 2000.
- [2] J. Krüger and T. Berger, "An empirical analysis of the costs of clone-and platform-oriented software reuse," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 432–444, 2020.
- [3] Y. Abgaz, A. McCarren, P. Elger, D. Solan, N. Lapuz, M. Bivol, G. Jackson, M. Yilmaz, J. Buckley, and P. Clarke, "Decomposition of monolith applications into microservices architectures: A systematic review," *IEEE Transactions on Software Engineering*, 2023.
- [4] X. Wang, R. Hu, C. Gao, X.-C. Wen, Y. Chen, and Q. Liao, "Reposvul: A repository-level high-quality vulnerability dataset," in *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings*, pp. 472–483, 2024.
- [5] C. Schröder, A. van der Feltz, A. Panichella, and M. Aniche, "Search-based software re-modularization: A case study at adyen," in *Proceedings of the 43rd International Conference on Software Engineering: Software Engineering in Practice*, pp. 81–90, IEEE, 2021.
- [6] N. Teymourian, H. Izadkhah, and A. Isazadeh, "A fast clustering algorithm for modularization of large-scale software systems," *IEEE Transactions on Software Engineering*, vol. 48, no. 4, pp. 1451–1462, 2022.
- [7] B. Pourasghar, H. Izadkhah, A. Isazadeh, and S. Lotfi, "A graph-based clustering algorithm for software systems modularization," *Information and Software Technology*, vol. 133, p. 106469, 2021.
- [8] I. Candela, G. Bavota, B. Russo, and R. Oliveto, "Using cohesion and coupling for software remodularization: Is it enough?," *ACM Transactions on Software Engineering and Methodology*, vol. 25, no. 3, pp. 1–28, 2016.
- [9] Q. I. Sarhan, B. S. Ahmed, M. Bures, and K. Z. Zamli, "Software module clustering: An in-depth literature analysis," *IEEE Transactions on Software Engineering*, vol. 48, no. 6, pp. 1905–1928, 2022.
- [10] K. Yang, J. Wang, Z. Fang, P. Wu, and Z. Song, "Enhancing software modularization via semantic outliers filtration and label propagation," *Information and Software Technology*, vol. 145, p. 106818, 2022.
- [11] B. S. Mitchell and S. Mancoridis, "On the automatic modularization of software systems using the bunch tool," *IEEE Transactions on Software Engineering*, vol. 32, no. 3, pp. 193–208, 2006.
- [12] M. Fowler, *Refactoring*. Addison-Wesley Professional, 2018.
- [13] J. Garcia, E. Kouroshfar, N. Ghorbani, and S. Malek, "Forecasting architectural decay from evolutionary history," *IEEE Transactions on Software Engineering*, vol. 48, no. 7, pp. 2439–2454, 2022.
- [14] G. Bavota, A. De Lucia, A. Marcus, and R. Oliveto, "Using structural and semantic measures to improve software modularization," *Empirical Software Engineering*, vol. 18, pp. 901–932, 2013.
- [15] C. Zhong, H. Zhang, C. Li, H. Huang, and D. Feitosa, "On measuring coupling between microservices," *Journal of Systems and Software*, p. 111670, 2023.
- [16] L. Xiao, Y. Cai, R. Kazman, R. Mo, and Q. Feng, "Detecting the locations and predicting the costs of compound architectural debts," *IEEE Transactions on Software Engineering*, vol. 48, no. 9, pp. 3686–3715, 2022.
- [17] H. Mumtaz, P. Singh, and K. Blincoe, "A systematic mapping study on architectural smells detection," *Journal of Systems and Software*, vol. 173, p. 110885, 2021.
- [18] I. Griffith and C. Izurieta, "Design pattern decay: The case for class grime," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pp. 1–4, ACM, 2014.
- [19] Y. Cai and R. Kazman, "Software design analysis and technical debt management based on design rule theory," *Information and Software Technology*, vol. 164, p. 107322, 2023.
- [20] F. A. Fontana, I. Pigazzini, R. Roveda, D. Tamburri, M. Zanoni, and E. Di Nitto, "Arcan: A tool for architectural smells detection," in *Proceedings of the 2017 IEEE International Conference on Software Architecture Workshops*, pp. 282–285, IEEE, 2017.
- [21] T. D. Oyetoyan, D. S. Cruzes, and C. Thurmman-Nielsen, "A decision support system to refactor class cycles," in *Proceedings of the 31th IEEE International Conference on Software Maintenance and Evolution*, pp. 231–240, IEEE, 2015.
- [22] R. C. Martin, *Agile software development: Principles, patterns, and practices*. Prentice-Hall, 2003.
- [23] R. C. Martin, *Clean architecture: A craftsman's guide to software structure and design*. Pearson Education, 2018.
- [24] S. Almugrin, W. Albattah, and A. Melton, "Using indirect coupling metrics to predict package maintainability and testability," *Journal of Systems and Software*, vol. 121, pp. 298–310, 2016.
- [25] I. G. Czibula, G. Czibula, D.-L. Miholca, and Z. Onet-Marian, "An aggregated coupling measure for the analysis of object-oriented software systems," *Journal of Systems and Software*, vol. 148, pp. 1–20, 2019.
- [26] R. Benkoczi, D. Gaur, S. Hossain, and M. A. Khan, "A design structure matrix approach for measuring co-change-modularity of software products," in *Proceedings of the 15th International Conference on Mining Software Repositories*, pp. 331–335, 2018.
- [27] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, 2005.

- [28] M. Zhang, "Weighted clustering ensemble: A review," *Pattern Recognition*, vol. 124, p. 108428, 2022.
- [29] D. Tsoukalas, N. Mittas, A. Chatzigeorgiou, D. Kehagias, A. Ampatzoglou, T. Amanatidis, and L. Angelis, "Machine learning for technical debt identification," *IEEE Transactions on Software Engineering*, vol. 48, no. 12, pp. 4892–4906, 2021.
- [30] Y. Yang, L. Lyu, Q. Yang, Y. Liu, and W. An, "Trust-based consensus reaching process for product design decision-making with heterogeneous information," *Advanced Engineering Informatics*, vol. 56, p. 101934, 2023.
- [31] H. Muccini *et al.*, "Group decision-making in software architecture: A study on industrial practices," *Information and Software Technology*, vol. 101, pp. 51–63, 2018.
- [32] J. Garcia, D. Popescu, G. Edwards, and N. Medvidovic, "Identifying architectural bad smells," in *Proceedings of the 13th European Conference on Software Maintenance and Reengineering*, pp. 255–258, IEEE, 2009.
- [33] Anonymous, "Replication package." <https://figshare.com/s/0a27c85b83bbfc69b5fc>, 2024.
- [34] W. Jin, Y. Cai, R. Kazman, G. Zhang, Q. Zheng, and T. Liu, "Exploring the architectural impact of possible dependencies in python software," in *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, pp. 758–770, 2020.
- [35] O. Maqbool and H. A. Babri, "The weighted combined algorithm: A linkage algorithm for software clustering," in *Proceedings of the 8th European Conference on Software Maintenance and Reengineering*, pp. 15–24, IEEE, 2004.
- [36] R. Mo, Y. Cai, R. Kazman, L. Xiao, and Q. Feng, "Decoupling level: A new metric for architectural maintenance complexity," in *Proceedings of the 38th International Conference on Software Engineering*, pp. 499–510, IEEE, 2016.
- [37] Y. Jia, S. Tao, R. Wang, and Y. Wang, "Ensemble clustering via co-association matrix self-enhancement," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [38] P. B. Kruchten, "The 4+ 1 view model of architecture," *IEEE software*, vol. 12, no. 6, pp. 42–50, 1995.
- [39] E.-M. Arvanitou, A. Ampatzoglou, A. Chatzigeorgiou, and P. Avgeriou, "Introducing a ripple effect measure: A theoretical and empirical validation," in *Proceedings of the 9th International Symposium on Empirical Software Engineering and Measurement*, pp. 1–10, IEEE, 2015.
- [40] J. Garcia, I. Ivkovic, and N. Medvidovic, "A comparative analysis of software architecture recovery techniques," in *Proceedings of the 28th IEEE/ACM International Conference on Automated Software Engineering*, pp. 486–496, IEEE, 2013.
- [41] P. Andritsos, P. Tsaparas, R. J. Miller, and K. C. Sevcik, "Limbo: Scalable clustering of categorical data," in *Proceedings of the 2004 International Conference on Extending Database Technology*, pp. 123–146, Springer, 2004.
- [42] V. Tzerpos and R. C. Holt, "Accd: An algorithm for comprehension-driven clustering," in *Proceedings of the Seventh Working Conference on Reverse Engineering*, pp. 258–267, IEEE, 2000.
- [43] "Depends." <https://github.com/multilang-depends/depends>, 2022.
- [44] D. J. Kim, B. Yang, J. Yang, and T.-H. Chen, "How disabled tests manifest in test maintainability challenges?," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1045–1055, 2021.
- [45] S. Boslaugh, *Statistics in a nutshell: A desktop quick reference*. O'Reilly Media, 2012.
- [46] B. A. Muse, M. M. Rahman, C. Nagy, A. Cleve, F. Khomh, and G. Antoniol, "On the prevalence, impact, and evolution of sql code smells in data-intensive systems," in *Proceedings of the 17th International Conference on Mining Software Repositories*, pp. 327–338, 2020.
- [47] A. Liu, J. Lefever, Y. Han, and Y. Cai, "Prevalence and severity of design anti-patterns in open source programs—a large-scale study," *Information and Software Technology*, vol. 170, p. 107429, 2024.
- [48] W. Mendes, O. Pinheiro, E. Santos, L. Rocha, and W. Viana, "Dazed and confused: Studying the prevalence of atoms of confusion in long-lived java libraries," in *Proceedings of the 38th IEEE International Conference on Software Maintenance and Evolution*, pp. 106–116, IEEE, 2022.
- [49] J. Y. Khan and G. Uddin, "Automatic detection and analysis of technical debts in peer-review documentation of r packages," in *Proceedings of the 29th IEEE International Conference on Software Analysis, Evolution and Reengineering*, pp. 765–776, IEEE, 2022.
- [50] F. Khomh, M. Di Penta, and Y.-G. Gueheneuc, "An exploratory study of the impact of code smells on software change-proneness," in *Proceedings of the 16th Working Conference on Reverse Engineering*, pp. 75–84, IEEE, 2009.
- [51] D. I. Sjøberg, A. Yamashita, B. C. Anda, A. Mockus, and T. Dybå, "Quantifying the effect of code smells on maintenance effort," *IEEE Transactions on Software Engineering*, vol. 39, no. 8, pp. 1144–1156, 2012.
- [52] A. Yamashita and L. Moonen, "Do code smells reflect important maintainability aspects?," in *Proceedings of the 28th IEEE International Conference on Software Maintenance*, pp. 306–315, IEEE, 2012.
- [53] W. Jin, Y. Dai, J. Zheng, Y. Qu, M. Fan, Z. Huang, D. Huang, and T. Liu, "Dependency facade: The coupling and conflicts between android framework and its customization," in *Proceedings of the IEEE/ACM 45th International Conference on Software Engineering*, pp. 1674–1686, IEEE, 2023.
- [54] R. Mo, Y. Zhang, Y. Wang, S. Zhang, P. Xiong, Z. Li, and Y. Zhao, "Exploring the impact of code clones on deep learning software," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 6, pp. 1–34, 2023.
- [55] C. P. Chambers and A. D. Miller, "Inefficiency measurement," *American Economic Journal: Microeconomics*, vol. 6, no. 2, pp. 79–92, 2014.
- [56] S. Chowdhury, R. Holmes, A. Zaidman, and R. Kazman, "Revisiting the debate: Are code metrics useful for measuring maintenance effort?," *Empirical Software Engineering*, vol. 27, no. 6, p. 158, 2022.
- [57] M. J. Allen and W. M. Yen, *Introduction to measurement theory*. Waveland Press, 2001.
- [58] T. Zimmermann, A. Zeller, P. Weissgerber, and S. Diehl, "Mining version histories to guide software changes," *IEEE Transactions on Software Engineering*, vol. 31, no. 6, pp. 429–445, 2005.
- [59] R. Mo, Y. Cai, R. Kazman, and L. Xiao, "Hotspot patterns: The formal definition and automatic detection of architecture smells," in *Proceedings of the 12th Working IEEE/IFIP Conference on Software Architecture*, pp. 51–60, IEEE, 2015.
- [60] R. Mo, Y. Cai, R. Kazman, L. Xiao, and Q. Feng, "Architecture anti-patterns: Automatically detectable violations of design principles," *IEEE Transactions on Software Engineering*, vol. 47, no. 5, pp. 1008–1028, 2019.
- [61] R. Rosenthal, H. Cooper, L. Hedges, *et al.*, "Parametric measures of effect size," *The Handbook of Research Synthesis*, vol. 621, no. 2, pp. 231–244, 1994.
- [62] Y. Gil and G. Lalouche, "On the correlation between size and metric validity," *Empirical Software Engineering*, vol. 22, no. 5, pp. 2585–2611, 2017.
- [63] C. Soto-Valero, T. Durieux, and B. Baudry, "A longitudinal analysis of bloated java dependencies," in *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1021–1031, 2021.
- [64] A. J. Jafari, D. E. Costa, R. Abdalkareem, E. Shihab, and N. Tsalis, "Dependency smells in javascript projects," *IEEE Transactions on Software Engineering*, vol. 48, no. 10, pp. 3790–3807, 2021.
- [65] A. Rahman, M. R. Rahman, C. Parnin, and L. Williams, "Security smells in ansible and chef scripts: A replication study," *ACM Transactions on Software Engineering and Methodology*, vol. 30, no. 1, pp. 1–31, 2021.
- [66] Y. Yang, X. Hu, X. Xia, and X. Yang, "The lost world: Characterizing and detecting undiscovered test smells," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 3, pp. 1–32, 2024.
- [67] C. Vassallo, S. Proksch, A. Jancso, H. C. Gall, and M. Di Penta, "Configuration smells in continuous delivery pipelines: A linter and a six-month study on gitlab," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 327–337, 2020.
- [68] Q. Chen, R. Cămară, J. Campos, A. Souto, and I. Ahmed, "The smelly eight: An empirical study on the prevalence of code smells in quantum computing," in *Proceedings of the 2023 IEEE/ACM 45th International Conference on Software Engineering*, pp. 358–370, IEEE, 2023.
- [69] M. Abidi, M. S. Rahman, M. Openja, and F. Khomh, "Are multi-language design smells fault-prone? an empirical study," *ACM Transactions on Software Engineering and Methodology*, vol. 30, no. 3, pp. 1–56, 2021.
- [70] V. Nardone, B. Muse, M. Abidi, F. Khomh, and M. Di Penta, "Video game bad smells: What they are and how developers perceive them," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 4, pp. 1–35, 2023.

- [71] S. Wong, Y. Cai, M. Kim, and M. Dalton, "Detecting software modularity violations," in *Proceedings of the 33rd International Conference on Software Engineering*, pp. 411–420, 2011.
- [72] S. Lambiase, A. Cupito, F. Pecorelli, A. De Lucia, and F. Palomba, "Just-in-time test smell detection and refactoring: The darts project," in *Proceedings of the 28th International Conference on Program Comprehension*, pp. 441–445, 2020.
- [73] S. Kalhor, M. R. Keyvanpour, and A. Salajegheh, "A systematic review of refactoring opportunities by software antipattern detection," *Automated Software Engineering*, vol. 31, no. 2, pp. 1–65, 2024.
- [74] S. Romano, F. Zampetti, M. T. Baldassarre, M. Di Penta, and G. Scan-niello, "Do static analysis tools affect software quality when using test-driven development?," in *Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pp. 80–91, 2022.
- [75] S. Noei, H. Li, S. Georgiou, and Y. Zou, "An empirical study of refactoring rhythms and tactics in the software development process," *IEEE Transactions on Software Engineering*, vol. 49, no. 12, pp. 5103–5119, 2023.
- [76] D. Sas and P. Avgeriou, "An architectural technical debt index based on machine learning and architectural smells," *IEEE Transactions on Software Engineering*, vol. 49, no. 8, pp. 4169–4195, 2023.
- [77] D. Thomas and A. Hunt, *The Pragmatic Programmer: Your journey to mastery*. Addison-Wesley Professional, 2019.
- [78] P. A. Laplante and M. Kassab, *What every engineer should know about software engineering*. CRC Press, 2022.
- [79] K. Golalipour, E. Akbari, S. S. Hamidi, M. Lee, and R. Enayatifar, "From clustering to clustering ensemble selection: A review," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104388, 2021.
- [80] Y. Zhang, Z. Xu, C. Liu, H. Chen, J. Sun, D. Qiu, and Y. Liu, "Software architecture recovery with information fusion," in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 1535–1547, 2023.
- [81] U. Erdemir and F. Buzluca, "A learning-based module extraction method for object-oriented systems," *Journal of Systems and Software*, vol. 97, pp. 156–177, 2014.
- [82] C. Y. Chong and S. P. Lee, "Automatic clustering constraints derivation from object-oriented software using weighted complex network with graph theory analysis," *Journal of Systems and Software*, vol. 133, pp. 28–53, 2017.
- [83] J. Al Dallal and L. C. Briand, "A precise method-method interaction-based cohesion metric for object-oriented classes," *ACM Transactions on Software Engineering and Methodology*, vol. 21, no. 2, pp. 1–34, 2012.
- [84] D. Athanasopoulos, A. V. Zarras, G. Miskos, V. Issarny, and P. Vassiliadis, "Cohesion-driven decomposition of service interfaces without access to source code," *IEEE Transactions on Services Computing*, vol. 8, no. 4, pp. 550–562, 2014.
- [85] D. M. Le, D. Link, A. Shahbazian, and N. Medvidovic, "An empirical study of architectural decay in open-source software," in *Proceedings of the 15th International Conference on Software Architecture*, pp. 176–17609, IEEE, 2018.