CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Research scope and motivation

The continuous evolution of software systems has resulted in increasingly complex projects involving collaboration among multiple developers over extended periods. The version control systems play an important role in this collaboration by tracking the software changes made over time.

Logical dependencies, also known as logical coupling or co-changes, are extracted from patterns of co-evolution observed in version control systems. These dependencies capture relationships between software components that frequently change together [2]. Gall et al. [3, 4] introduced the concept of logical coupling and demonstrated that it could uncover hidden relationships between software entities. Unlike structural dependencies, which represent explicit relationships between code entities extracted from static code analysis, logical dependencies are derived from co-change patterns and can reveal hidden relationships that are not visible in the system's static structure [3, 4, 5].

Logical dependencies have been applied across various domains of software engineering. These dependencies have been used for tasks such as software change prediction [2, 6, 7], software change impact analysis [8], identifying ripple effects during software maintenance and evolution [9, 10, 11], and exploring their connections to defects [12, 2]. By capturing relationships that are not always visible in the static code structure, logical dependencies complement structural dependencies to provide a better understanding of software systems.

However, using logical dependencies presents some challenges: Co-changes in version control systems often include irrelevant relationships caused by refactorings, or non-functional updates. Gall et al. [3, 4] highlight the importance of filtering out co-changes resulting from non-functional commits, as these can reduce the effectiveness of logical dependency analysis. Ajienka and Capiluppi [13] confirmed that logical dependencies significantly outnumber structural dependencies, with an average ratio of approximately 12:1. This highlights the importance of filtering techniques to reduce the volume of logical dependencies and increase their validity.

Current research has investigated various filtering techniques, such as commit size thresholds [6], support and confidence metrics [2], and history length and age [7]. However, the methods and thresholds applied leave room for further investigation.

The scope of this thesis is to improve the filtering and integration of logical dependencies into dependency models, which will be applied in two software engineering tasks: key class detection and software clustering for architectural reconstruction.

Key Class Detection: Identifying the most important classes in a system, referred to as key classes, is an important task in software maintenance and evolution. Key classes encapsulate core functionality and manage system operations [14]. Previ-

ous research has mainly relied on structural dependencies, using relationships between classes, such as method invocations, inheritance, or field accesses [1, 14] to detect key classes.

By using logical dependencies, key class detection algorithms could achieve better results, as they can contain hidden relationships from the codebase or reinforce existing ones. This thesis investigates the usage of logical dependencies to complement or replace structural dependencies in key class detection algorithms.

Software Clustering: Software clustering is used to recover the architecture of a software system, mainly when the documentation is outdated or unavailable. It involves grouping software entities, such as files or classes, into cohesive modules based on their dependencies and interactions [15, 16, 17]. Related research has explored various types of dependencies to enhance software clustering and architectural reconstruction, including structural dependencies [16, 17], lexical dependencies (e.g., naming conventions and textual information) [18, 19] or logical dependencies [20].

This thesis also investigates whether different logical dependency association rules (filters) affect clustering results. To evaluate the impact of logical dependencies, two metrics will be used: Modularization Quality (MQ) [21] and MoJoFM (Move and Join Effectiveness Measure) [22], allowing the comparison of different approaches.

1.2. Objectives of the thesis

The objective of this doctoral thesis is to investigate and improve methods for filtering logical dependencies extracted from version control systems and to validate the effectiveness of these methods. The filtered logical dependencies are used to enhance key class detection in software systems and to improve software architecture reconstruction, two areas that mostly depend on structural dependencies.

This research proposes techniques for extracting, filtering, and integrating logical dependencies into dependency models. Various metrics will be used to evaluate the impact of incorporating logical dependencies and assess their impact on key class detection and software architecture reconstruction. The goal is to demonstrate that logical dependencies, when appropriately filtered, provide valuable information that can complement or, in some cases, replace structural dependencies. To achieve this, the following objectives have been established:

[O1]: *Study and analyze the current state of research on logical dependencies to improve their extraction and filtering methods.*

This objective involved the following tasks:

- **[T1.1]**: Review existing methods for logical dependency extraction from version control systems.
- **[T1.2]**: Identify and evaluate factors that influence the validity of logical dependencies, such as commit size thresholds, minimum co-change occurrences, and comment changes, to develop effective filtering methods.
- **[T1.3]**: Based on the observations from T1.2, propose filtering methods for logical dependencies.

- **[T1.4]**: Investigate the interplay between logical and structural dependencies to understand their overlap and differences better.

The outcomes of this objective [O1] have been published in [A1] and [A2].

[O2]: *Develop a tool for extracting and filtering logical dependencies.*

This objective involved the following tasks:

- **[T2.1]**: Designing a tool for extracting logical dependencies based on co-changing entities with different configurable filters.
- **[T2.2]**: Implementing and testing the filtering mechanisms by outputting the results in a commonly used format.

This objective [O2] has been published and described in more detail in [A1] and [A2], the developed tool being further used in [A3], [A4], and [A5].

[O3]: *Integrate logical dependencies in key class detection.* This objective involved the following tasks:

- **[T3.1]** Extract logical dependencies from version control systems and apply the proposed filter (connection strength).
- **[T3.2]** Modify the baseline key class detection tool to incorporate logical dependencies.
- **[T3.3]** Evaluate the impact of combining logical and structural dependencies on key class detection performance.
- **[T3.4]** Evaluate the impact of using only logical dependencies for key class detection.
- **[T3.5]** Investigate the effect of filtering strategies (connection strength with different thresholds) on the key class detection results.

The outcomes of this activity [O3] have been published in [A3].

[O4]: *Refine software clustering using logical dependencies.* This objective involved the following tasks:

- **[T4.1]** Review and select suitable clustering algorithms for the experiments.
- **[T4.2]** Use only logical dependencies as input for software clustering.
- **[T4.3]** Combine logical and structural dependencies as input for clustering algorithms.
- **[T4.4]** Investigate the impact of different connection strength filter thresholds on clustering results by using clustering evaluation metrics: MQ and MoJoFM.
- **[T4.5]** Analyze the results: logical dependencies alone vs. combined dependencies with varying filter thresholds.

The outcomes of this objective [O4] have been published in [A4] and [A5].

1.3. Structure of the thesis

To be done at the end

1.4. Main Contributions

The principal contributions of this thesis are:

- proposing methods for filtering logical dependencies extracted from version control systems to improve their reliability and usefulness; this includes introducing a new metric for filtering logical dependencies, called connection strength;
- developing a tool for extracting and filtering logical dependencies;
- integrating logical dependencies into key class detection methodologies and analyzing the impact of different filtering strategies when using logical dependencies both independently and in combination with structural dependencies;
- integrating logical dependencies into software clustering and analyzing the impact of different filtering strategies when using logical dependencies both independently and in combination with structural dependencies; this analysis involves using three distinct clustering algorithms and two evaluation metrics to study the results;
- providing evidence that logical dependencies when appropriately filtered, can improve key class detection and software architecture reconstruction.

2. SOFTWARE DEPENDENCIES: CONCEPTS, APPLICATIONS, AND CURRENT RESEARCH

2.1. Software dependencies overview

2.1.1. Structural dependencies

A dependency is created by two elements that are in a relationship and indicates that an element of the relationship, in some manner, depends on the other element of the relationship [23], [24].

Structural dependencies can be found by analyzing the source code [25], [26]. There are several types of relationships between these source code entities and all those create *structural dependencies*:

Data Item Dependencies. Data items can be variables, records or structures. A dependency is created between two data items when the value held in the first data item is used or affects the value from the second.

```
Example:

int a = 10;
int b = a + 5;  // b depends on a
a = 20;         // changing a doesn't impact b
```

Data Type Dependencies. Data items are declared to be of a specific data type. Besides the built-in data types that every programming language has, developers can also create new types that they can use. Each time the data type definition is changed it will affect all the data items that are declared to be of that type.

```
Example:

class Rectangle { int width, height; }
//Change posibility: class Rectangle { int width, height, color; }

Rectangle r = new Rectangle();
r.width = 5;
r.height = 10;
// Existing code will fail to handle color in case of change
```

Subprogram Dependencies. A subprogram is a sequence of instructions that performs a certain task. Depending on the programming language a subprogram

may also be called a routine, a method, a function or a procedure. When a subprogram is changed, the developer must check the effects of that change in all the places that are calling that subprogram. Subprograms may also have dependencies with the data items that they receive as input or the data items that they are computing.

Example:

```
// First implementation:
class Calculator {
    double calculateArea(double side) {
        return side * side;
    }
}

// Modified implementation:
class Calculator {
    double calculateArea(double radius) {
        return Math.PI * radius * radius;
    }
}

Calculator calc = new Calculator();
// Code expecting square area now gets circle area.
double area = calc.calculateArea(5);
```

2.1.2. Lexical dependencies

Lexical dependencies, similar to structural dependencies, are extracted from the source code. The difference lies in the fact that lexical dependencies focus on finding pairs of entities that are similar (and thus connected) from a linguistic point of view. This means they are based on the textual content and naming conventions used in the code rather than explicit structural relationships. The lexical information about an entity (such as a class or interface) can be extracted from class names, method names, parameter names, code comments, source code statements, and others [18], [27].

For example, a class named `StructuralDependenciesEvaluator` and a class named `LexicalDependenciesEvaluation` can be related even if they do not directly share dependencies or their connection is not visible from a structural point of view.

To extract lexical dependencies, the code is tokenized, breaking it down into lexical tokens (e.g., words or symbols). A document containing these tokens is created for each source code file. Each document is then split into different parts, such as a class names part, attribute names part, parameter names part, and others. For instance, the class names part will contain the names of all classes found in the source file.

After the tokenization process and splitting into various parts, the similarity between the two documents is estimated.

Various methods can be used for similarity calculation, such as:

- *Term Frequency-Inverse Document Frequency (TF-IDF):* Weighs the importance of words based on their frequency across documents [18], [19].
- *Cosine Similarity:* Measures the cosine of the angle between two vectors representing the documents [27].

2.1.3. Semantical dependencies

Podgurski and Clarke define semantic dependencies in source code as follows: if changes in a statement s_1 impact the behavior of another statement s_2, then s_1 and s_2 are *semantically dependent* [28].

Semantic dependencies can be identified from the source code by constructing control flow graphs (CFGs). A control flow graph is a directed graph $G = (V, E)$, where:

- V is the set of vertices representing program statements (e.g., assignments, method calls, conditions),
- E is the set of edges representing possible transfers of control between statements.

A vertex $u \in V$ is semantically dependent on a vertex $v \in V$ if changes in v determine changes in u.

In object-oriented (OO) systems, semantic dependencies are dependencies that are not visible in the static structure of the code. There is no direct reference between two entities (e.g., classes or interfaces), but changes in one entity impact the behavior of the other [29, 30, 11].

Example:

Let A a class that uses an object of type I, and let B a class that implements I. For example, class A calls a method `performAction()` via an interface I, while class B implements the `performAction()` method.

If B changes its implementation of `performAction()`, the behavior of A might also change when it uses B. This means that A and B are semantically dependent.

2.1.4. Logical dedependencies

Logical dependencies (also known as logical coupling) can be discovered through software history analysis. They reveal relationships between entities that are not always present in the source code's static structure (i.e., structural dependencies).

Software engineering practice has shown that sometimes modules without explicit structural dependencies still appear to be related. *Co-evolution* represents the phenomenon where one component changes in response to changes in another component [5, 24]. Such changes can be found in the software history maintained by version control systems. Gall et al. defined *logical coupling* as the occurrence of modules that *repeatedly* update together during the evolution of the software system [3, 4, 31].

The concepts of logical coupling and logical dependencies have been applied in different analysis tasks related to software changes: for software change impact analysis [8]; for identifying potential ripple effects caused by software changes during maintenance and evolution [9, 10, 11, 32]; and for exploring their link to defects [12, 2].

An in-depth analysis of how logical dependencies are identified within the versioning system and their applications is provided in Section 2.3.

2.1.5. Additional dependencies

Besides the dependencies mentioned above, there are many other types of dependencies, such as temporal, package, external, and several others.

Temporal dependencies represent a type of dependency where certain operations in the code must occur in a specific order. These operations may belong to the same software component or span across different parts of the system [24].

Example:

```
var file = new File();
file.open("example.txt");
file.readContents();
file.close();
```

In this example, there is a temporal dependency between the `open`, `readContents`, and `close` methods. The `readContents` method cannot function correctly unless the `open` method is called first to open the file. Similarly, the `close` method should be called after `readContents` to release the file resources.

Package dependencies refer to the relationships between different software packages, usually managed using package managers [33].

External dependencies are the relationships between a software system's code and components outside the system, such as third-party services, libraries, APIs, or the programming language itself [34].

2.2. Software change and version control systems

2.2.1. Software change

Software systems have distinctive stages during their life: initial development, evolution, servicing, phase out, and close down [35], [36].

In the *evolution stage*, iterative changes are made. By changes, we mean additions (new software features), modifications (changes of requirements or misun-

derstood requirements), or deletions. There are two main reasons for the change: the software team's learning process and new requests from the client.

Suppose new changes are no longer easy to be made or are very difficult and time-consuming. In that case, the software enters the *servicing stage*, also called aging software, decayed software, and legacy [35], [37].

The main difference between changes made in the evolution and servicing stages is the effort to make changes. In the evolution stage, software changes are made easily and do not require much effort, while in the servicing stage, only a limited number of changes are made and require a lot of effort, so they are time-consuming [38, 39, 40].

The change mini-cycle consists of the following phases [41]:

- Phase 1: The change request. This usually comes from the software users, and it can also be a bug report or a request for additional functionality.
- Phase 2: The planning phase includes program comprehension and change impact analysis. Program comprehension is a mandatory prerequisite of the change, while change impact analysis indicates how costly the change will be. [42]
- Phase 3: The change implementation, restructuring for change, and change propagation.
- Phase 4: Verification and validation.
- Phase 5: Re-documentation.

Understanding these phases is important for ensuring the software system remains maintainable and reliable throughout its lifecycle.

2.2.2. Version control systems

Software evolution implies change which can be triggered either by new feature requests or bug reports [43]. As presented also in section 2.2.1, one phase of the change mini-cycle consists of change implementation and propagation (changing source code files). Usually, developers use version control when it comes to software development. Version control is a system that records changes to a file or set of files over time so that developers can recall specific versions of those files later [44]. Distributed version control systems (such as Git, Mercurial, Bazaar or Darcs) allows many developers to collaboratively develop their projects [45].

Below, we will review some of the operations supported by versioning systems. The operation names in the following text are specific to Git, as Git is the version control system used for data collection in the current work. However, similar operations exist in other versioning systems; for example, Subversion (SVN) and Mercurial also provide operations such as branching, merging, and committing changes.

Repository

A repository serves as a centralized storage location for project files. The entire codebase of a project can be stored within a single repository or be split into a main repository and various modules stored in multiple repositories. Git repositories

can be hosted on numerous platforms, such as GitHub, GitLab, Bitbucket, and others.

It is important to differentiate between Git and Git hosting services. Git is a version control system that allows developers to download and modify code on their local devices while hosting services like GitHub provide platforms for teams to host projects that utilize Git [46], [47].

Commits

Committing is an operation that allows developers to record the latest changes to the codebase in the repository. A commit saves the current state of the code, including changes made since the last commit.

The changes tracked include:

- *Additions:* New files or lines of code created.
- *Modifications:* Updates to existing lines of code or files.
- *Deletions:* Removal of files or lines of code that are no longer needed.

The *push* operation uploads these changes to the remote repository on the hosting service.

A unique identifier, also known as a commit hash, is assigned to each commit. This hash allows developers to track specific changes or revert to earlier code versions. When a developer commits the changes, a commit message is required. This message serves as documentation for the changes, enhancing code traceability and maintainability [46].

Figure 2.1 illustrates how changes are made to the code over time through commits, starting from the initial commit to the latest version.

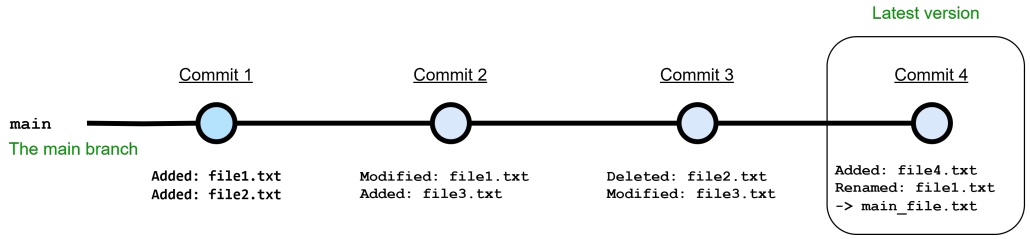

Figure 2.1: Tracking changes through commits.

Branches and Merging

By default, every repository comes with a main (or master) branch. This branch represents the central timeline for code changes and serves as the main integration point for stable code. Other branches can be created for parallel development from this branch.

Branches allow developers to encapsulate changes without affecting the main branch. In most software projects, it is standard practice for developers to use branches for development, with the main branch being locked for direct commits. Changes to the main branch can be integrated through merges from other branches.

Software projects often have multiple branches running in parallel, including branches from other branches. The main branch is not the only branch that supports branching; any branch can be a base for creating other branches.

The operation of integrating changes from one branch into another (usually into the main branch) is called *merging*.

There are three main types of merges in Git [46]:

- *Git Merge:* This is the most commonly used type of merging. It creates a new merge commit that combines all the changes from the branch being merged. Additionally, it retains the history of all the individual commits in the branch.

```
git checkout main
git merge feature-branch
```

- *Git Rebase and Merge:* This operation is typically used when the branch contains a single commit or a small number of commits. It moves all the commits from the source branch to the top of the target branch. The main disadvantage of this operation is that it rewrites the commit history.

```
git checkout feature-branch
git rebase main
git checkout main
git merge feature-branch
```

- *Git Squash and Merge:* This operation compresses all commits from the source branch into a single commit before merging it into the target branch. It results in a cleaner commit history but has the disadvantage of losing individual commits from the source branch.

```
git checkout main
git merge --squash feature-branch
git commit
```

Figure 2.2 illustrates the differences between these three types of merging.

Tagging

Another helpful operation in Git is tagging. Developers use the tagging operation to mark a specific code version at a particular commit. This is usually done for important milestones, such as a new release version or a stable build [46].

Tags provide a way to create a human-readable reference to a specific commit (e.g., `v1.0.0`), as the commit hash can be hard to remember (e.g., `a1b2c3d4e5f6g7h8i9`)

Git supports two types of tags:

- *Lightweight Tags:* Simple references to a commit that do not contain any additional metadata.
- *Annotated Tags:* These include additional metadata such as the author's name, date, and message, making them more suited for marking releases.

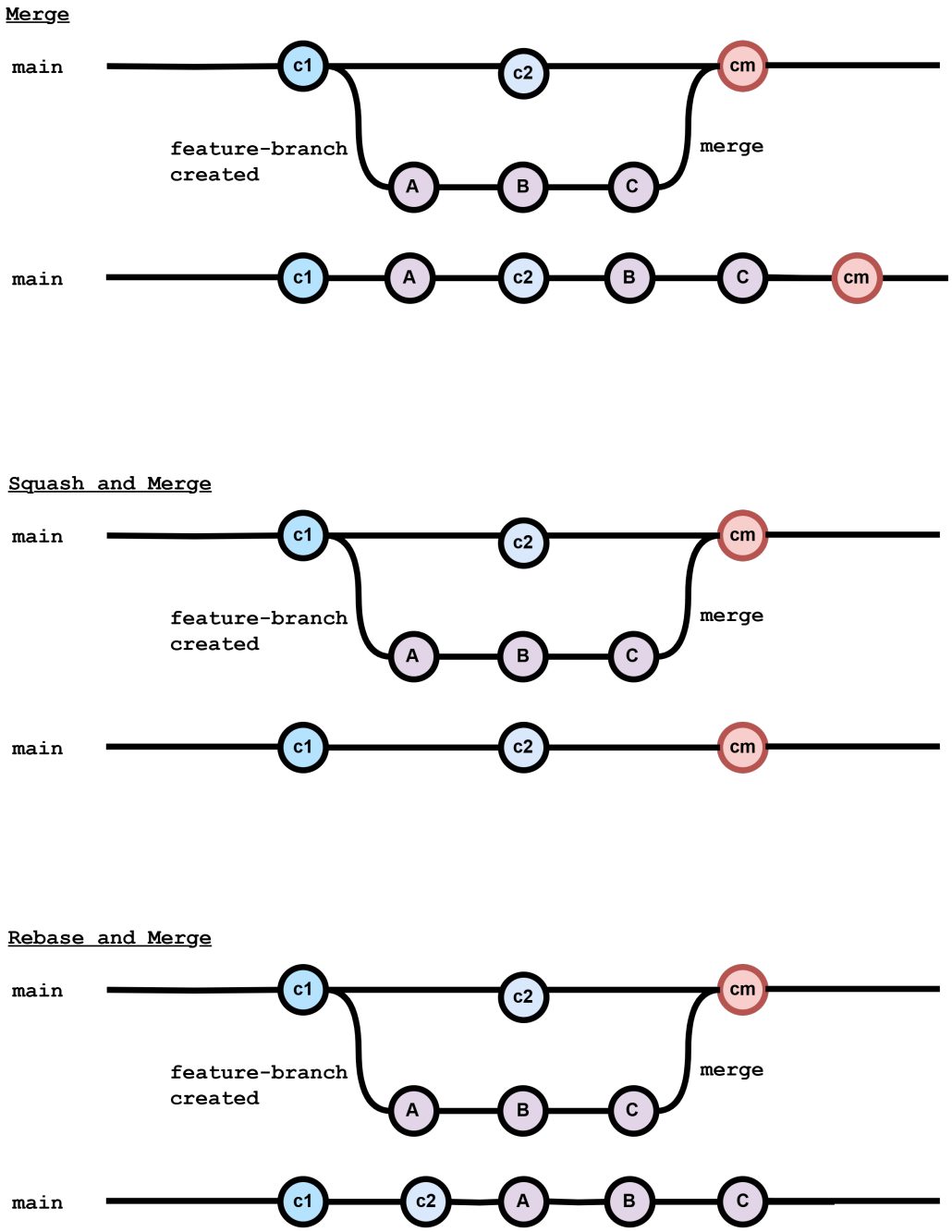

Figure 2.2: Comparison of Git merge types.

2.3. Current status of research on logical dependencies

2.3.1. Logical dependencies in software systems

The versioning system stores the long-term change history of every file in a

project. Each change made by an individual at a specific point in time is recorded as a commit [45]. This historical data provides insights into how software components evolve over time. Logical dependencies, also known as evolutionary coupling or co-changes, are derived from patterns of co-evolution, where one component changes in response to changes in another [5, 48].

Gall [3, 4, 31] identified *logical coupling* as the relationship between two modules that repeatedly change together during the historical evolution of a software system. Logical dependencies can be expressed at two levels of abstraction: module-level [49] and file/class-level [4, 50].

Several tools have been developed to detect and use logical dependencies. For instance, the **ROSE** tool, developed by Zimmermann et al. [2], mines version histories to suggest related code changes based on historical coupling between software entities. ROSE can predict 26% of further files to be changed and 15% of the functions or variables involved in future modifications.

Kagdi et al. [51, 52] developed **sqminer**, a tool based on the SPADE (Sequential Pattern Discovery Algorithm) to mine sequences of file changes for software change prediction. They also proposed an approach that combines conceptual couplings (derived from textual analysis of source code comments and identifiers) with logical dependencies (mined using the sqminer tool) to enhance software change impact analysis [51].

Moonen et al. [6, 7, 53] proposed an algorithm called **Tarmaq** that analyzes evolutionary coupling for predicting co-changes. Their results demonstrated that Tarmaq achieves better performance than the ROSE tool. Similarly, Mondal et al. [54] developed **HistoRank**, an algorithm that uses five different ranking strategies for prioritizing and filtering co-changes.

As shown above, logical dependencies have been widely used in co-change prediction. Additionally, they have been used for various other purposes, such as detecting code clones [55], identifying buggy code [56], evaluating the impact of software changes [8], and identifying potential ripple effects caused by software changes during maintenance and evolution [9, 10].

Despite their utility, logical dependency usage presents several challenges. One significant issue is the large volume of information generated during extraction, which must be processed and filtered to remove noise [57, 43, 58]. For instance, *Oliva and Gerosa* [10] found that the set of co-changed classes was much larger than the set of structurally coupled classes. At least 91% of logical dependencies involve files that are not structurally related. This implies that not all change dependencies are linked to structural dependencies and that other factors may cause software artifacts to be change-dependent.

Ajienka and Capiluppi also explored the interplay between the logical and structural coupling of software classes. In [13, 59], they conducted experiments on 79 open-source systems. For each system, they identified the sets of structural dependencies, logical dependencies, and the intersections of these sets. They studied the overlap between these dependencies, concluding that not all co-changed class pairs (logical dependencies) are also linked by structural dependencies. Another observation, which was not deeply investigated by the authors in [13, 59], is the ratio

between the total number of logical dependencies and structural dependencies in software systems. Based on their raw data, the average ratio across all analyzed projects is approximately 12.

Applications based on dependency analysis could benefit from incorporating non-structural dependencies alongside structural ones. For example, works that investigate various methods for architectural reconstruction [60, 61, 17], which rely mostly on structural dependency data, could enhance their dependency models by including logical dependencies.

Logical dependencies should integrate harmoniously with structural dependencies. Valid logical dependencies should not be excluded from the model, but structural dependencies should also not be overwhelmed by questionable logical dependencies. Therefore, to effectively incorporate logical dependencies into dependency models, co-changes must be filtered to remain only with a reduced and relevant set of valid logical dependencies.

2.3.2. Existing filtering techniques

Currently, there are no fixed rules for filtering extracted class co-changes to ensure they form a set of valid logical dependencies. Most studies using or investigating logical dependencies have applied various filtering techniques to the extracted co-changes. Below are some of the most commonly used filtering methods cited in the literature.

Commit Size

One of the most frequently used filters for co-change extraction is the commit size filter. Commit size refers to the number of code files modified in a specific commit. Large commits, which involve a significant number of files, are often the result of non-code-related tasks, such as branch merges or folder renaming. These operations can introduce irrelevant co-changing pairs of entities, adding noise. To solve this issue, commit size filtering is applied to the information extracted from version control systems.

Ajienka and Capiluppi [13] established a threshold of 10 files, discarding all commits with more than 10 modified files from their research.

Kagdi et al. also applied the same threshold, excluding commits with more than 10 source files from their analysis.

Ying et al. [62] set a higher threshold, excluding transactions involving more than 100 files.

Zimmermann et al. [2] configured the ROSE tool to exclude changes involving more than 30 files.

Moonen et al. [6] considered seven different transaction filtering sizes: 2, 4, 6, 8, 10, 20, and 30 (with 30 being the upper bound suggested by Zimmermann) and recommended excluding transactions larger than 8 files.

However, most of the works presented above do not discuss how the thresholds

were chosen, nor do they analyze the impact of different threshold values on the quantity and quality of the data extracted.

Support and confidence

Filtering based only on commit size is not enough. While this type of filtering can reduce the total number of extracted co-changes, it does so without guaranteeing that the remaining co-changes are more relevant.

Unrelated files can sometimes be updated in small commits due to human error. For instance, a file that was omitted from the current commit may be committed in the next commit together with some unrelated files. Such scenarios can introduce co-changing pairs that are not genuinely linked. To address this problem, a filter based on the occurrence rate of co-changing pairs should be applied. Co-changing pairs that occur multiple times are more likely to be dependent than those that appear only once [63].

Zimmermann et al. [2, 64] introduced the support and confidence metrics to measure the significance of co-changes based on the occurrence rate of co-changing pairs.

The *support metric* of a rule $(A \rightarrow B)$, where A is the antecedent and B is the consequent of the rule, is defined as the number of commits (transactions) in which both entities are changed together.

The *confidence metric* of $(A \rightarrow B)$, as defined in Equation (3.1), focuses on the antecedent of the rule and is the number of commits together of both entities divided by the total number of commits of (A).

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Nr. of commits containing } A \text{ and } B}{\text{Nr. of commits containing } A} \tag{2.1}$$

The support metric represents the frequency of changes for an association rule and can take values from 0 to infinity. On the other hand, the confidence metric is defined within the interval [0, 1]. As in Equation (3.1), the confidence metric measures the likelihood of a co-change. It can never exceed 1 because Commits$(A \cap B)$ is always a subset of Commits(A).

For example, consider the case where entity A is modified in 10 commits and entity B is modified together with A in 7 of those commits. The confidence is then:

$$\text{Confidence}(A \rightarrow B) = \frac{7}{10} = 0.7.$$

If A and B are always modified together in all 10 commits involving A, the confidence would be:

$$\text{Confidence}(A \rightarrow B) = \frac{10}{10} = 1.$$

Many studies have used the support and confidence metrics. *Zimmermann*

et al. [2], in their ROSE tool, used different combinations of minimum support and confidence thresholds. They applied three minimum support thresholds, 1, 3, and 5, together with confidence thresholds ranging from 0.1 to 0.9, to predict future changes in software systems.

Ying et al. [62] recommend potentially relevant source code to developers during modification tasks. The support thresholds in their study vary based on the analysis, ranging from 5 to 30 (5, 10, 15, 20, 25, 30). For confidence, Ying et al. chose not to use this metric, as they consider it misleading when some files are changed much more frequently than others.

Kagdi et al. [51], in their studies on software change prediction, used minimum support thresholds of 1, 2, 4, and 8. For confidence, they considered all possible values greater than zero.

Mandal et al. [55], in their study on detecting clones and analyzing their evolution, developed a tool called MARC (Mining Association Rules among Clones). The tool detects code clones and ranks their change-proneness based on support and confidence values. In their experiments, they applied a minimum support threshold of 1.

Oliva and Gerosa [10, 65] investigated the interplay between structural and logical dependencies, using a modified version of the support and confidence metrics introduced by Zimmermann, adjusted to a set of commits. The study found that the support interval [1, 7] contains 90% of all logical dependencies, while the interval [0, 10] contains 99.9% of logical dependencies. The interval [11, 31] represents only 0.1% of logical dependencies, with high variation observed across coupling levels.

The confidence metric was divided into three categories: low logical coupling (0.00–0.33), medium logical coupling (0.33–0.66), and high logical coupling (0.66–1.00). The study concluded that classes with high logical coupling were the least influenced by structural coupling.

Ajienka and Capiluppi [59], in their study on the overlappings between structural and logical dependencies used a support threshold of 0.1 and a confidence threshold of 0.01.

History length and age

Moonen et al. investigated the influence of history length (the number of analyzed transactions) and history age (the number of transactions that have occurred since the last co-change) when mining evolutionary coupling. Their study, which analyzed over 540,000 commits, found no evidence to support the idea that there is an upper limit to the amount of history that can be used for mining evolutionary coupling before outdated knowledge starts to negatively affect the quality of the extracted information [7].

2.4. Applications of software dependencies

This section reviews several applications of software dependencies (e.g., struc-

tural, lexical, semantical, logical dependencies). These dependencies play an important role in various software engineering tasks, architecture reconstruction, clone identification, and more.

Architecture reconstruction. Currently, software systems contain tens of thousands of lines of code and are updated daily by multiple developers. The software architecture is important for understanding and maintaining a system. Often, code updates are made without checking or updating the architecture.

These kinds of updates cause the architecture to drift from the reality of the code over time. Therefore, reconstructing the architecture and verifying if it still matches the actual implementation is important [66, 36, 67]. Architecture reconstruction has mainly been done using structural dependencies [68], [17], [69], but recent works also include semantical and logical dependencies [18], [19], [70].

Identifying Clones. Research suggests that a considerable portion (around 5-10%) of the source code in large-scale software is duplicate code ("clones"). Source code is often duplicated for a variety of reasons: programmers may simply reuse a piece of code by copying and pasting, or they may "reinvent the wheel" [71], [72]. Detection and removal of clones can significantly decrease software maintenance costs [73], [74].

Structural dependencies can be used to detect code clones. For example, Cordy and Roy created the NiCad tool, which receives the entire code base of a project as input to detect project clones [75]. The same authors also used versioning system information to link clones to bug-fix commits from the versioning system [76]. Other researchers have used semantic dependencies and machine learning techniques to identify clones [77], while others have used lexical dependencies extracted from code comments [78].

Code Smells. Fowler defined code smells citebookFowler as patterns generally associated with bad design and poor programming practices. Originally, code smells were used to identify areas in software that may require refactoring [79]. Studies have found that code smells can impact software comprehension and increase changes and faults in the system [80], [81], [82].

Examples of code smells include:

- **Large Class:** A class with many fields and methods, making it difficult to maintain or understand.
- **Feature Envy:** Methods that access more methods and fields of another class than of their class.
- **Data Class:** Classes that contain only fields and no meaningful functionality.
- **God Class:** A class that centralizes too much responsibility, often not respecting the Single Responsibility Principle.
- **Refused Bequest:** A subclass that inherits many fields or methods from its parent but leaves them unused.
- **Parallel Inheritance:** Every time a subclass is added to one class, a subclass must also be added to another class.
- **Shotgun Surgery:** A single change requires modifications in multiple classes.

Code smell detection approaches often use structural information extracted

from static code analysis. For example, Marinescu [83] proposed a method to identify smells like God Class and Data Class based on structural metrics. Recent works have used machine learning algorithms in combination with structural metrics to improve the detection of code smells [84, 85].

Certain smells, however, are more effectively detected using information from versioning systems, which capture logical dependencies. For instance, smells like *Parallel Inheritance* and *Shotgun Surgery* have been successfully identified by analyzing co-change patterns in version control systems [86].

Key Classes. The concept of key classes was first introduced by Zaidman et al. [14], referring to classes found in documents that provide an architectural overview of the system or an introduction to its structure. Tahvildari and Kontogiannis provided a more detailed definition of the key classes concept: *"Usually, the most important concepts of a system are implemented by very few key classes which can be characterized by specific properties. These classes, which we refer to as key classes, manage many other classes or use them in order to implement their functionality. The key classes are tightly coupled with other parts of the system."* [87].

Key class identification can be performed using different algorithms with various inputs. Most of the research is based on using structural dependencies [17], [88], [89], [1], [14], [90], class diagrams [91], or *dynamic dependencies* obtained through runtime analysis [14].

Comprehension. Software comprehension is the process of gaining knowledge about a software system. An increased understanding of the software system helps activities such as bug correction, enhancement, reuse, and documentation [92], [93].

Previous studies show that the proportion of resources and time allocated to maintenance activities can vary from 50% to 75% [94]. Within the maintenance process, the biggest effort is dedicated to understanding the system.

To support software comprehension, various tools have been developed to help this process. For example, the COSPEX tool developed by Gupta et al. uses source code analysis to help novice developers better comprehend their tasks [95]. Şora proposes a tool that uses structural dependencies and graph-based ranking to generate an executive summary highlighting the most important classes in the software system [88].

Fault Localization.

Debugging software is an expensive and mostly a manual process. Among all debugging activities, fault localization is the most time-consuming task [96].

Software developers typically locate faults in their programs through a manual process. This process begins when developers observe failures in the program. They select a dataset to inject into the system, which is a set of data likely to replicate previous failures or trigger new ones, and set breakpoints using a debugger. They then monitor the system's state until a failure occurs and backtrack from the failure state to identify the root cause of the fault [97, 98].

Several tools and methods have been developed to support in the fault local-

ization process. A commonly used technique is running coverage tests to evaluate the likelihood of faults in different parts of the source code. Code statements or methods having more failing tests than passing tests are labeled more likely to contain faults. An example of such a tool is GZoltar, developed by Campos et al. [99].

More recent studies have suggested enhancing fault localization techniques by incorporating information from versioning systems in addition to source code analysis. Wen et al. empirically demonstrated that versioning system data could improve fault localization. Their approach identifies code entities modified by more ”bug-inducing commits” than regular commits, labeling them as potential fault sources [100].

Defect Prediction.

Fault localization is the process of identifying elements responsible for software failures reported by users or discovered by developers. Defect prediction, on the other hand, predicts elements that are likely to be fault-prone before faults occur. Due to this difference, fault localization and defect prediction are studied as two separate research areas [101, 102].

Research has shown that modules or entities most likely to contain defects can be identified based on structural metrics (e.g., lines of code, cyclomatic complexity) and versioning system data (e.g., frequency of changes, modified code in a source file) [103], [104].

3. METHODOLOGY FOR EXTRACTING AND FILTERING LOGICAL DEPENDENCIES

3.1. Overview of the approach

This chapter investigates the process of extracting and refining logical dependencies. The process is organized into three steps: collecting data, extracting dependencies, and applying filters.

A set of 27 open-source projects were used to perform the investigations. These projects have different sizes, complexities, and programming languages (Java and C#). Section 3.2.1 provides an overview of the systems used.

The extraction process targets two types of dependencies:
- *Structural dependencies* are obtained by analyzing the static structure of the source code. Using the *srcML* tool, source code files are converted into an XML format to extract relationships between entities.
- *Logical dependencies* are derived from the version control history.

After extraction, filtering techniques are applied to refine the logical dependencies:
- *Commit size filtering* removes co-changing pairs from large commits, such as branch merges or formatting updates.
- *Support filtering* ensures co-changing pairs occur a minimum number of times.
- *Strength filtering* uses metrics like support, confidence, and a system factor to identify pairs with stronger relationships, scaling the results to the reality of the system.

Each filtering step helps reduce the number of co-changing pairs while keeping meaningful dependencies. Figure 3.2 shows the workflow of the tool, which handles data collection, dependency extraction, and filtering.

3.2. Data collection

3.2.1. Data set used

To investigate logical dependencies extraction and filtering techniques and their interplay with structural dependencies, 27 open-source projects from GitHub[1] were selected. Table 3.1 provides an overview of all the software systems analyzed. The columns in the table represent the following information:

- *ID*: A unique identifier assigned to each project.

[1] http://github.com/

- *Project name*: The name of the project as it appears on GitHub.
- *Number of entities*: The total number of entities (classes and interfaces) extracted from the source code.
- *Number of commits*: The total number of commits analyzed from the active branch (main/master) of each project.
- *Type*: The programming language used in the development of the project.

This selection includes projects implemented in two programming languages: Java and C#. The systems vary in size and complexity. From a structural perspective, the smallest project is *shipkit*, which contains only 639 entities, while the largest is *EntityFrameworkCore*, with 50,179 entities.

Regarding commit history, *shipkit* is again the smallest, with 1,563 commits analyzed, while *aeron* is the largest, with 5,977 commits. This selection of projects allows a better investigation of how filtering techniques affect systems of different sizes and commit histories.

Table 3.1: Summary of open source projects used for logical dependencies extraction and filtering.

ID	Project name	Number of entites	Number of commits	Type
1	bluecove	2685	894	Java
2	aima-java	5232	1006	Java
3	powermock	2801	949	Java
4	restfb	3350	1391	Java
5	rxjava	21097	4398	Java
6	metro-jax-ws	6482	2927	Java
7	mockito	5189	3330	Java
8	grizzly	10687	3113	Java
9	shipkit	639	1563	Java
10	OpenClinica	9655	3276	Java
11	robolectric	8922	5912	Java
12	aeron	4159	5977	Java
13	antlr4	4747	4431	Java
14	mcidasv	3272	4136	Java
15	ShareX	4289	5485	C#
16	aspnetboilerplate	9712	4323	C#
17	orleans	16963	3995	C#
18	cli	2063	4488	C#
19	cake	12260	2518	C#
20	Avalonia	16732	5264	C#
21	EntityFrameworkCore	50179	5210	C#
22	jellyfin	8764	5433	C#
23	PowerShell	2405	3250	C#
24	WeiXinMPSDK	7075	5729	C#
25	ArchiSteamFarm	702	2497	C#
26	VisualStudio	4869	5039	C#
27	CppSharp	17060	4522	C#

3.2.2. Extracting structural dependencies

A dependency is created between two elements that are in a relationship,

indicating that one element of the relationship, in some manner, depends on the other [23], [24].

Structural dependencies can be identified by analyzing the source code [105]. A structural dependency between two entities (e.g., a class and an interface) exists if one entity statically depends on the other, meaning it cannot be compiled without the dependent entity. In object-oriented systems, this dependency can be given by various types of relationships: one entity extends another (for classes), implements another (for interfaces), has attributes of the other entity's type, has methods with the other entity's type in their signature, uses local variables of the other entity's type, or calls methods of the other entity [25, 26, 93].

We use an external tool called *srcML* [106] to convert all source code files into XML files. We then extract all information about classes, interfaces, methods, or calls to other classes by parsing the XML files and building a dependency data structure.

Maletic and Collard [107, 108, 109] developed the *srcML* tool to offer an XML-based representation of the source code. The tool preserves all source code information, including comments and formatting. The tool supports languages such as C, C++, Java, and C#, and provides command-line utilities for converting an entire project to and from the *srcML* format.

Another advantage of the *srcML* tool is that it uses consistent markup for different programming languages, making it easier to extract structural dependencies from source code written in languages like Java, C++, Python, or C#.

3.2.3. Extracting Logical Dependencies

Logical dependencies (logical couplings) can be identified through software history analysis. Version control repositories (such as Git) store not only the source code of the system but also its change history. By examining both, we can extract structural and logical dependencies. The source code structure provides *structural dependencies*, while the system's change history reveals *logical dependencies* formed by pairs of files or components that co-evolve.

As illustrated in Figure 3.1, the `git clone` command retrieves the entire repository, including its code and commit history. The `git diff` command then highlights differences between two specific commits, generating a text file that contains the differences between the two commits: code differences, the number of files changed, and the names of changed files.

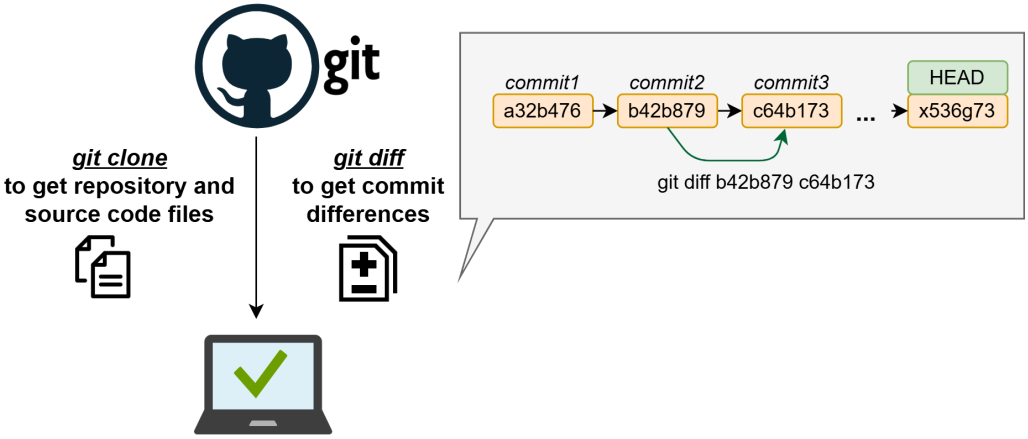

Figure 3.1: Using Git commands (`clone` and `diff`) to retrieve source code and commit changes.

Listing 3.1 provides an example of a `git diff` output for three files: `User.java`, `Calculator.java`, and `Utils.java`. From the diff output, we can identify the names of the changed files and information about added or deleted lines of code, but not the specific entities (e.g., classes or interfaces) from those files.

To determine the changed entities, we first extract the structural dependencies of the system and create a mapping between the entities defined in each file and the file. Later, we associate the changed files with their corresponding entities when processing the diff files using this mapping [110, 63, 111, 112].

For instance, in this example, the mapping can reveal that `User.java` contains the class `User`, `Calculator.java` contains the class `Calculator`, and `Utils.java` contains the class `Utils`. Even if Java typically enforces a standardized relationship between file and entity names, this is not always the case for other programming languages, so we need a uniform approach.

Using this mapping, we can say that the commit presented in Listing 3.1 generates three pairs of co-changed entities: `Utils-Calculator`, `Utils-User`, and `Calculator-User`.

```
commit 1a2b3c4d5e (HEAD -> main)
Author: Developer <developer@example.com>
Date:   Wed Dec 13 12:34:56 2024 +0000

    Refactored code and added new features.

diff --git a/src/User.java b/src/User.java
index abc1234..def5678 100644
--- a/src/User.java
+++ b/src/User.java
@@ -1,5 +1,6 @@
+    public User(String name) {
+        this.name = name;
+    }
```

```
15
16  -    public void greet() {
17  +    public void displayGreeting() {
18           System.out.println("Hello, " + name + "!");
19       }
20   }
21
22  diff --git a/src/Calculator.java b/src/Calculator.java
23  index 9876543..2345678 100644
24  --- a/src/Calculator.java
25  +++ b/src/Calculator.java
26  @@ -5,7 +5,7 @@
27  -    public int subtract(int a, int b) {
28  +    public int subtractNumbers(int a, int b) {
29           return a - b;
30       }
31   }
32
33  diff --git a/src/Utils.java b/src/Utils.java
34  index 56789ab..789abcd 100644
35  --- a/src/Utils.java
36  +++ b/src/Utils.java
37  @@ -3,7 +3,7 @@
38  -    public static String getCurrentTime() {
39  +    public static String getFormattedTime() {
40           return java.time.LocalTime.now().toString();
41       }
42   }
```

Listing 3.1: Example output of `git diff` between two commits.

3.3. Tool for measuring software dependencies

To extract structural and logical dependencies, we developed a tool that takes as input the source code repository URL of a given system and extracts the corresponding software dependencies [110, 112].

From a workflow perspective, the tool performs three main activities: downloading the necessary data from the repository, extracting structural dependencies from the source code, and identifying and filtering co-changing pairs from the repository's commit history. Figure 3.2 illustrates these activities, with each block representing a different step from the process.

To get the source code files and the change history, we first need to know the repository URL from GitHub (GitHub is a Git repository cloud-based hosting service). With the GitHub URL and a series of Git commands, the tool can download all the necessary data for dependencies extraction.

As presented in figure 3.1, the *"git clone"* command will download a repository, including the source code files. The *"git diff"* command will get the differences between two existing commits in the Git repository. The tool gets the Git repository and the source code files by executing the "clone" command. Afterward, it gets all the existing commits within the Git repository. The commits are ordered by date, beginning with

the oldest one and ending with the most recent one. The tool executes the ”diff” command between each commit and its parent (the previous commit). The ”diff” command generates a text file that contains the differences between the two commits: code differences, the number of files changed and changed file names.

The first step involves downloading the source code files and change history. This requires the GitHub repository URL, as GitHub is a cloud-based hosting service for Git repositories. Using this URL and Git commands, the tool downloads all the data necessary for dependency extraction.

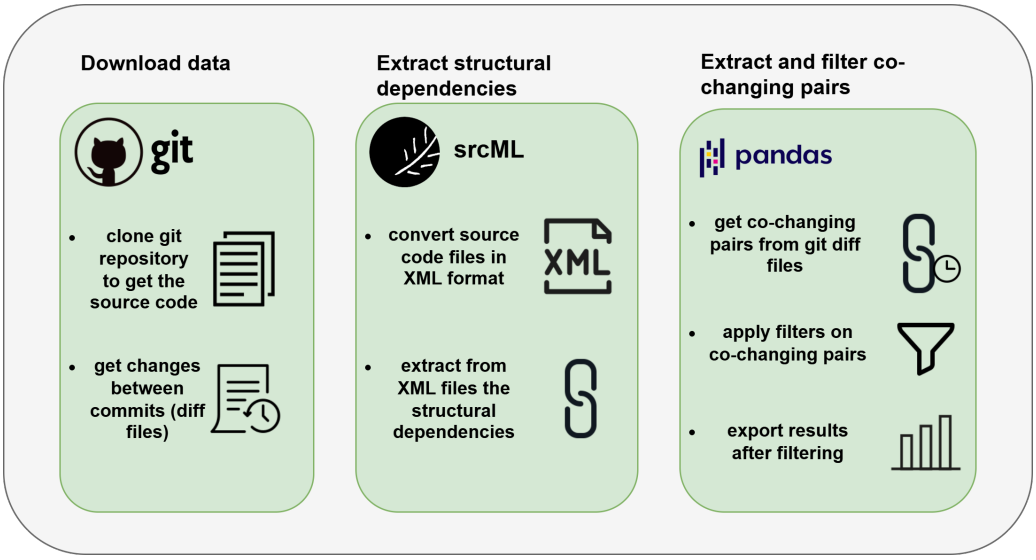

Figure 3.2: Tool workflow and major activities.

Extract structural dependencies.

To extract structural dependencies from the source code, the tool first converts each source file into the srcML format using the method introduced in subsection 3.2.2. The srcML format provides an XML representation of the source code, with markup tags identifying elements of the programming language syntax[106].

Once converted, the tool parses each file to identify all defined entities (such as classes, interfaces, and enums) from the file. It also detects all entities used by the defined entities. The connections between the defined and used entities form the structural dependencies.

Extract and filter co-changing pairs.

The process of extracting and filtering co-changing pairs is illustrated in Figure 3.3.

To analyze the changes between commits, the tool uses the `"git diff"` command. All existing commits in the repository are collected and chronologically ordered,

starting with the oldest and ending with the most recent. For each commit, the tool runs the "`git diff`" command to compare it with its parent (the preceding commit). This generates a text file containing the details of the changes between the two commits.

To extract co-changing pairs, the tool parses each generated diff file. From each file, it identifies the number of changed files and their names. Since the tool already knows the software entities contained in each file (from the structural dependencies extraction), it can determine co-changing pairs by linking entities from changed files. Once all the co-changing pairs for a diff file are extracted, the tool moves to the next diff file and repeats the process.

Explained in more detail in subsections 3.4.1, 3.4.2, and 3.4.3, not every extracted co-changing pair is considered a logical dependency. For a pair to be considered a logical dependency, it must pass some criteria. These criteria are implemented as filters in the tool. Each filter processes the extracted co-changing pairs and outputs the pairs that meet the filter requirements. The filters can also be combined to ensure that only meaningful logical dependencies remain after filtering.

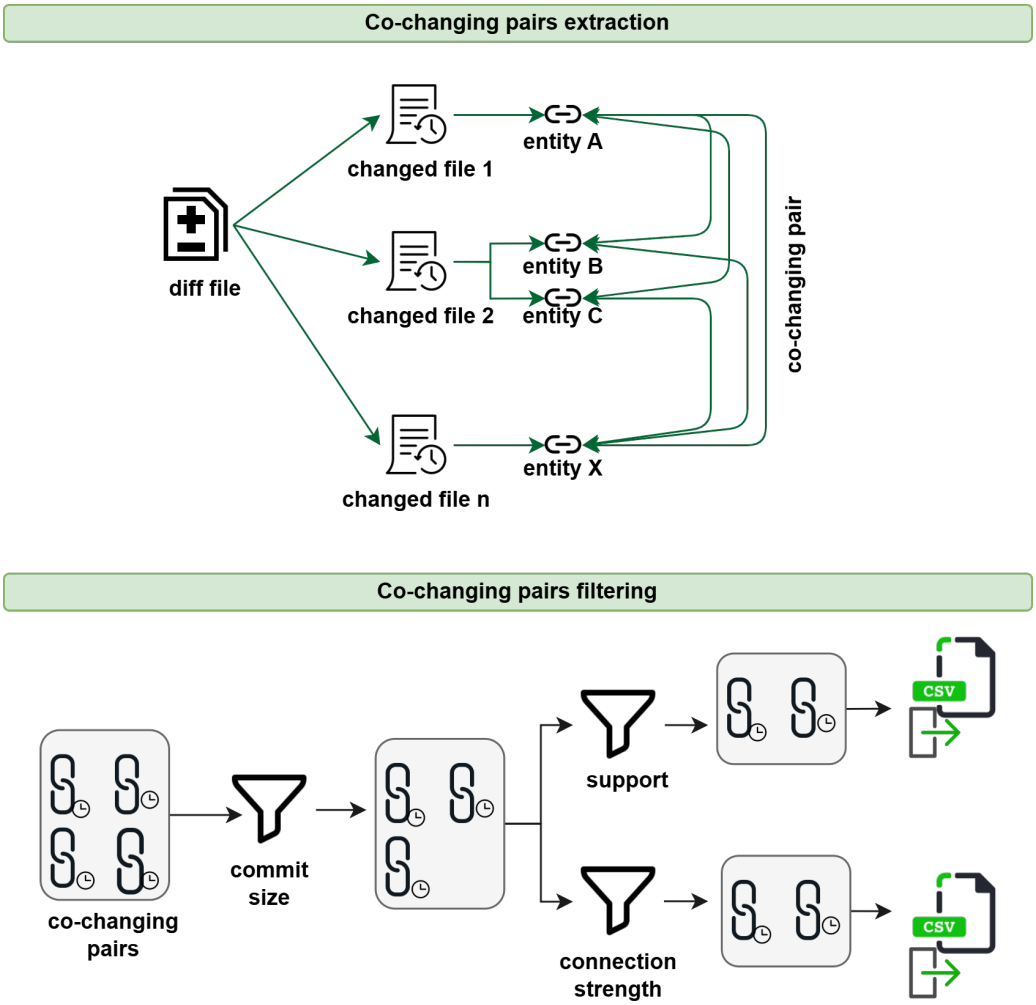

Figure 3.3: Co-changing pairs extraction and filtering.

3.4. Filtering logical dependencies

As discussed in Section 2.3, the number of co-changing pairs extracted from software repositories can be quite large, often reaching millions. This quantity, combined with noise in the data, makes it important to apply filtering techniques to identify meaningful co-changing pairs.

This section presents three filtering techniques used to refine the extracted co-changing pairs. The filter based on commit transaction size, discussed in Section 3.4.1, is always applied. The experiments are performed by combining this filter with one of the other two filters: the support filter presented in Section 3.4.2 or the connection strength filter discussed in Section 3.4.3. The commit size filter aims to reduce the

overall size of the co-changing pairs, while the support and connection strength filters aim to minimize noise in the data.

3.4.1. Filtering based on commit transaction size

As discussed in Section 2.3, the number of co-changing pairs extracted from software repositories can be quite large, often reaching millions. With this filtering approach, the goal is to reduce the total number of extracted co-changing pairs and move closer to identifying meaningful logical dependencies.

Large commits often involve many files due to non-functional changes like branch merges, folder restructuring, or formatting updates. These types of commits can introduce noise by creating irrelevant co-changing pairs. Dependencies are more meaningful when extracted from commits that involve feature development or bug fixes, where developers modify related code files. However, if multiple unrelated features or fixes are solved into a single commit, it can add false relationships between the entities.

One example of this issue is the initial commit when a system is migrated to a new versioning platform. These commits typically include many files but no functional changes. Another example is merge commits, created automatically during branch integrations. These commits combine changes from multiple smaller commits, which often address different issues or features. In such cases, analyzing the smaller, individual commits provides more accurate information than the overall merge commit [113].

Different studies have used various threshold values for filtering commit sizes. Cappiluppi and Ajienka [13, 59] only considered commits with fewer than 10 modified source code files. Similarly, Kagdi et al. used the same threshold, excluding commits with more than 10 source files. Ying et al. [62] took a different approach, excluding commits involving over 100 files. Zimmermann et al. [2] configured the ROSE tool to exclude commits larger than 30 files. Moonen et al. [6] explored seven different transaction filtering sizes (2, 4, 6, 8, 10, 20, and 30), recommending a threshold of 8 files.

We analyzed the overall transaction size trend for 27 open-source C# and Java systems with a total of 74,332 commits. The results are presented in Figure 3.4 and in Table 3.2. Based on the analysis, we observed that 90% of the total commit transactions involved fewer than 10 source code files. This percentage indicates that setting a threshold of 10 files for the maximum size of commit transactions will not affect too much the total number of commits available for extracting co-changing pairs, 90% of the transactions still remain available for analysis [110, 112].

Table 3.2 provides more details of the commit transaction size distribution for each system. The columns in the table represent the following information:

- $cs \leq 5$: The number of commits with a transaction size of 5 or fewer files.
- $cs \leq 10$: The number of commits with a transaction size of 10 or fewer files.
- $cs \leq 20$: The number of commits with a transaction size of 20 or fewer files.
- $cs < \infty$: The total number of commits for the system.
- Avg: The average transaction size for the system.

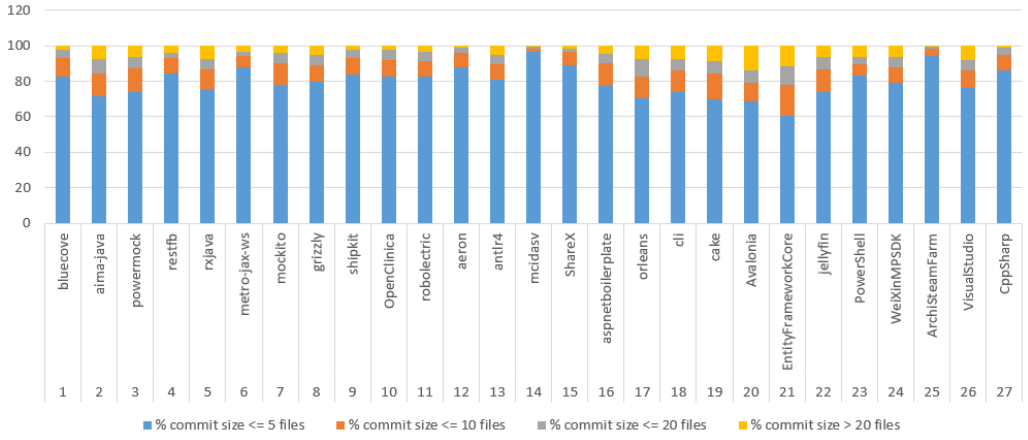

Figure 3.4: Commit transaction size(cs) trend in percentages.

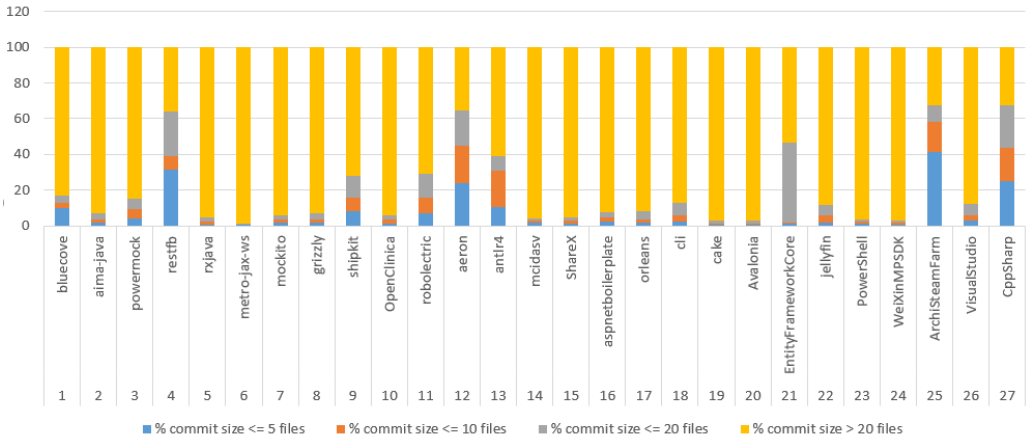

Figure 3.5: Percentages of co-changing pairs extracted from each commit transaction size(cs) group.

As we can see in Figure 3.5, even though only 5% of the commit transactions have more than 20 files changed ($20 < cs < \infty$), they generate, on average, 80% of the total amount of co-changing pairs extracted from the systems. The high number of co-changing pairs extracted from such a small number of commit transactions is caused by the number of files involved in those commit transactions.

One single commit transaction can lead to a large amount of co-changing pairs. For example, in RxJava, we have commit transactions with 1030 source code files. This means that those commits can generate

$$^{n}C_{k} = \frac{n!}{k!(n-k)!} = \frac{1030!}{2!(1028)!} = 529935$$

logical dependencies. By setting a threshold on the commit transaction size, we can avoid the introduction of those co-changing pairs into the system.

Filtering 10% of the total amount of commit transactions can significantly decrease the number of co-changing pairs. That is why we choose the value of 10 files as our fixed threshold for the maximum size of a commit transaction [110].

Table 3.2: Commit transaction size(cs) trend and average per system.

$Nr.$	$Project$	$cs \leq 5$	$cs \leq 10$	$cs \leq 20$	$cs < \infty$	Avg
1	bluecove	738	97	37	22	4.9
2	aima-java	733	134	74	65	7.24
3	powermock	685	128	66	70	9.61
4	restfb	1160	127	44	60	9.9
5	rxjava	3395	447	253	303	8.46
6	metro-jax-ws	2583	198	78	68	4.33
7	mockito	2522	433	222	153	6.33
8	grizzly	2487	302	180	144	5.28
9	shipkit	1311	151	64	37	4.26
10	OpenClinica	2837	250	119	70	3.31
11	robolectric	4827	503	264	318	7.43
12	aeron	4844	684	300	149	4.6
13	antlr4	3426	437	304	264	8.5
14	mcidasv	3996	81	35	24	2.47
15	ShareX	4731	529	145	80	4.69
16	aspnetboilerplate	3208	569	321	225	6.61
17	orleans	2780	518	369	328	8.95
18	cli	3377	551	308	252	6.43
19	cake	1785	359	174	200	9.89
20	Avalonia	3806	641	371	446	8.43
21	EntityFrameworkCore	2866	878	644	822	15.38
22	jellyfin	4007	662	419	345	6.25
23	PowerShell	2702	224	133	191	7.33
24	WeiXinMPSDK	4604	526	296	303	9.01
25	ArchiSteamFarm	2357	92	28	20	2.24
26	VisualStudio	3902	521	295	321	6.71
27	CppSharp	3870	390	203	59	3.28

3.4.2. Filtering based on support

In the previous section, we filtered the co-changing pairs based on commit size. Although this reduced the number of extracted co-changing pairs, this type of filtering does not guarantee that the remaining co-changing pairs are valid logical dependencies. A single occurrence of a co-changing pair could represent a valid logical dependency, but it could also be a coincidence.

To address this, the *support metric* is applied. The support metric of a rule $(A \rightarrow B)$, where A is the antecedent and B is the consequent, is defined as the number of commits (transactions) in which both entities are changed together. Many studies have used the support metric to filter logical dependencies. For example, *Zimmermann et al.* [2] applied minimum support thresholds of 1, 3, and 5 in their ROSE tool, while *Kagdi et al.* [51] used thresholds of 1, 2, 4, and 8.

Considering only co-changing pairs with a minimum support threshold can help improve accuracy. However, for projects with a smaller number of commits, it becomes less likely to find co-changing pairs with high support, which could end up filtering out all the extracted co-changes.

We performed a series of analyses on the test systems, incrementing the support threshold (*support*) from 1 to 4. Co-changing pairs were extracted only from commits with a transaction size of a maximum of 10 files. For each threshold mentioned above, the extracted co-changing pairs were then filtered based on the specified support threshold. Any co-changing pairs that did not meet the minimum support criteria were discarded.

Tables 3.4 and 3.3 provide detailed results for the support filtering applied to co-changing pairs. Table 3.3 presents the percentages of co-changing pairs that are also structural dependencies, and Table 3.4 presents the ratio of the number of co-changing pairs to the number of structural dependencies (SD). The columns in the tables represent the following information:

- $Project\ nr.$: The unique identifier assigned to each system.
- $support \geq 1$: Represents the results when the minimum support threshold is set to 1, the ratio of co-changing pairs to structural dependencies (Table 3.4) or the percentage of co-changing pairs that are also structural dependencies (Table 3.3).
- $support \geq 2$: Represents the results when the minimum support threshold is set to 2.
- $support \geq 3$: Represents the results when the minimum support threshold is set to 3.
- $support \geq 4$: Represents the results when the minimum support threshold is set to 4.

Based on Table 3.3, we observe that only a small percentage of the extracted co-changing pairs overlap with structural dependencies. This observation is consistent with findings from related works [13, 59]. The percentage of co-changing pairs that are structural dependencies increases as the minimum support threshold rises.

We calculate the overlap between co-changing pairs and structural dependencies not only to understand how many structural dependencies are reflected in the versioning system through co-changing pairs but also to filter out co-changing pairs that are structural dependencies, as they do not provide new information about the system.

We stopped the minimum support threshold at 4 because, as observed in Table 3.4, systems with IDs 2, 6, 10, and 16 show a ratio below 1, indicating that the number of structural dependencies exceeds the number of co-changing pairs. For systems with IDs 4, 11, 25, and 27, increasing the threshold to 4 does not significantly reduce the difference between the number of co-changing pairs and structural dependencies.

Raising the support threshold beyond 4 might cause filtering out all co-changing pairs for some systems. Therefore, while applying a threshold of 4 filters co-changing pairs to identify logical dependencies, the remaining number of co-changing pairs still significantly exceeds the number of structural dependencies for certain systems.

Table 3.3: Percentage of co-changing pairs that are also structural dependencies.

$Project\ nr.$	$support \geq 1$	$support \geq 2$	$support \geq 3$	$support \geq 4$
1	7,13	7,77	7,99	19,71
2	19,54	25,76	29,55	32,16
3	6,66	8,58	11,82	14,87
4	1,16	1,17	0,91	0,80
5	3,99	3,96	7,75	7,49
6	13,92	20,16	22,91	22,77
7	8,38	9,28	14,93	14,58
8	6,70	9,73	14,20	15,60
9	16,98	23,34	29,22	32,89
10	8,94	9,15	11,05	10,59
11	4,99	6,92	8,88	11,08
12	13,19	17,15	18,60	19,57
13	2,43	5,59	8,33	8,21
14	13,27	18,88	19,02	19,28
15	12,90	21,95	25,51	27,01
16	13,33	17,34	18,53	16,24
17	6,09	6,18	6,41	6,44
18	9,73	10,60	14,27	18,80
19	10,26	13,54	13,64	12,60
20	12,83	18,36	21,00	25,72
21	2,86	4,65	5,70	4,98
22	5,20	6,56	8,18	8,90
23	8,23	13,64	17,04	17,65
24	6,77	10,89	14,47	16,05
25	9,85	10,15	11,65	11,33
26	8,65	10,79	12,78	14,34
27	7,04	8,78	9,87	10,08
Avg	8,93	11,88	14,23	15,55

3.4.3. Filtering based on connection strength

In Section 3.4.1, we applied a filtering rule to exclude co-changing pairs extracted from commits involving more than 10 changed files. This decision was based on the results obtained from analyzing the commit size.

In Section 3.4.2, we introduced an additional filter based on the support metric of a co-changing pair. This filter was applied after the commit size filter. However, the results highlighted a challenge: using a fixed threshold for filtering may not be effective across systems of varying sizes. A threshold that works well for smaller systems may need to be increased for medium-sized systems and vice versa.

To address this issue, we introduced another filter complementary to the commit size filter described in Section 3.4.1. This new filter focuses on the connection strength of co-changing pairs. A co-changing pair that appears only once in a system's history may be less reliable than one that occurs more frequently [113].

Zimmermann et al. proposed the support and confidence metrics to measure the reliability of co-changes [2].

The *support metric* for a rule $(A \rightarrow B)$, where A is the antecedent, and B is the

Table 3.4: Ratio of number of co-changing pairs to number of structural dependencies.

$Project\ nr.$	$support \geq 1$	$support \geq 2$	$support \geq 3$	$support \geq 4$
1	4,13	1,94	1,23	0,26
2	0,81	0,33	0,16	0,10
3	5,12	1,93	0,78	0,38
4	53,36	42,00	38,31	36,30
5	4,27	2,90	0,88	0,72
6	1,07	0,46	0,30	0,23
7	4,09	2,38	0,99	0,73
8	4,06	1,57	0,76	0,49
9	3,64	2,03	1,14	0,77
10	1,41	1,01	0,47	0,34
11	7,91	4,47	2,93	2,03
12	3,92	2,15	1,47	1,07
13	10,15	3,18	1,22	1,03
14	3,07	1,53	1,16	0,97
15	2,34	0,84	0,48	0,33
16	1,21	0,47	0,26	0,19
17	2,99	1,83	1,11	0,84
18	2,26	1,37	0,67	0,40
19	2,32	1,38	0,76	0,67
20	1,24	0,58	0,35	0,18
21	5,33	2,12	1,27	1,05
22	3,38	1,88	0,99	0,74
23	3,62	1,22	0,76	0,37
24	2,57	1,22	0,67	0,46
25	7,47	5,36	4,16	3,73
26	4,03	2,16	1,50	1,15
27	7,46	4,26	2,99	2,43
Avg	5,67	3,43	2,51	2,15

consequent, is defined as the number of commits (transactions) in which both entities are modified together.

The *confidence metric* of the same rule $(A \rightarrow B)$, as expressed in Equation (3.1), measures the proportion of commits involving both entities relative to the total number of commits where A appears.

$$\text{Confidence}(A \rightarrow B) = \frac{\text{Nr. of commits containing } A \text{ and } B}{\text{Nr. of commits containing } A} \tag{3.1}$$

The confidence metric focuses on entities that are modified less frequently but consistently together, rather than entities that are modified more often but with multiple other entities.

For example, assuming that entity A was changed in 10 commits and, of these 10 commits, 9 also included changes to entity B, the confidence for the rule $(A \rightarrow B)$ would be 0.9. On the other hand, if entity C was changed in 100 commits and, of these 100 commits, 50 also included changes to entity D, the confidence for the rule $(C \rightarrow D)$ would be 0.5 [113]. This means we would have more confidence in the first pair $(A \rightarrow B)$ than in the second pair $(C \rightarrow D)$, even though the second pair has more than five times the number of updates together.

This observation was also made by *Ying et al.* [62], who analyzed co-change patterns to recommend potentially relevant source code to developers during development tasks. In their study, they did not use the confidence metric, as they considered it misleading when some files are changed much more frequently than others. Instead, they used only support thresholds from 5 to 30.

To favor entities that frequently appear in commits together, we calculated a *system factor*. This factor represents the average support metric value for all entity pairs within the system [113].

The system factor is then incorporated into the confidence metric calculation by multiplying the two values. To ensure the metric can also serve as a weight alongside structural dependency weights, we scale the resulting value by 100 to make it supraunitary and limit the range between 0 and 100.

This modified formula is referred to as the strength metric, which is defined in Equation (3.2).

$$\text{strength}(A \to B) = \text{confidence}(A \to B) \times 100 \times \text{system factor} \tag{3.2}$$

Table 3.5: Ratio of number of filtered co-changing pairs to number of SD, when strength A and strength B $\geq threshold$%

$Project\ nr.$	≥ 10%	≥ 20%	≥ 30%	≥ 40%	≥ 50%	≥ 60%	≥ 70%	≥ 80%	≥ 90%	≥ 100%
1	1.326	0.658	0.433	0.401	0.244	0.199	0.195	0.022	0.011	0.011
2	0.266	0.137	0.070	0.044	0.036	0.019	0.005	0.004	0.003	0.003
3	0.505	0.243	0.147	0.086	0.061	0.031	0.031	0.031	0.031	0.031
4	0.822	0.163	0.045	0.017	0.011	0.002	0.001	0.001	0.001	0.001
5	0.234	0.119	0.054	0.037	0.034	0.018	0.013	0.011	0.007	0.007
6	0.227	0.155	0.101	0.077	0.070	0.036	0.018	0.017	0.016	0.016
7	1.590	0.804	0.357	0.288	0.215	0.088	0.052	0.036	0.032	0.032
8	2.073	0.293	0.170	0.111	0.093	0.050	0.039	0.034	0.021	0.007
9	1.495	0.479	0.271	0.142	0.108	0.059	0.047	0.011	0.008	0.008
10	0.253	0.135	0.093	0.078	0.062	0.042	0.024	0.019	0.019	0.017
11	0.114	0.086	0.064	0.037	0.027	0.025	0.001	0.000	0.000	0.000
12	0.277	0.136	0.085	0.069	0.053	0.045	0.039	0.015	0.007	0.004
13	11.363	0.721	0.031	0.010	0.007	0.004	0.000	0.000	0.000	0.000
14	3.225	0.805	0.660	0.533	0.493	0.454	0.386	0.356	0.005	0.005
15	6.097	0.725	0.663	0.564	0.500	0.242	0.176	0.170	0.001	0.001
16	1.302	0.333	0.219	0.146	0.094	0.045	0.014	0.008	0.007	0.007
17	0.816	0.640	0.551	0.503	0.496	0.196	0.159	0.152	0.142	0.142
18	1.676	0.233	0.159	0.118	0.102	0.062	0.058	0.029	0.026	0.026
19	2.335	0.753	0.614	0.337	0.075	0.021	0.007	0.004	0.004	0.004
20	0.846	0.117	0.098	0.018	0.013	0.002	0.001	0.001	0.001	0.001
21	3.377	1.691	1.608	1.584	1.576	1.310	0.001	0.001	0.001	0.001
22	0.132	0.006	0.003	0.002	0.002	0.000	0.000	0.000	0.000	0.000
23	1.732	1.299	0.158	0.053	0.007	0.001	0.000	0.000	0.000	0.000
24	3.295	0.334	0.188	0.061	0.017	0.006	0.003	0.001	0.000	0.000
25	0.897	0.479	0.429	0.423	0.412	0.403	0.339	0.009	0.001	0.000
26	1.281	0.090	0.053	0.028	0.020	0.013	0.006	0.001	0.001	0.001
27	99.528	1.020	0.992	0.980	0.972	0.927	0.078	0.075	0.073	0.072

In Table 3.6, the columns represent the ratio between the number of structural dependencies and the number of co-changing pairs that remain after filtering out pairs where at least one factor falls below the specified threshold in the column header. In

Table 3.6: Ratio of number of filtered co-changing pairs to number of SD, when strength A or strength B $\geq threshold$%

Project nr.	≥ 10%	≥ 20%	≥ 30%	≥ 40%	≥ 50%	≥ 60%	≥ 70%	≥ 80%	≥ 90%	≥ 100%
1	1.312	1.181	0.700	0.599	0.419	0.235	0.219	0.046	0.045	0.045
2	0.430	0.280	0.176	0.118	0.103	0.056	0.022	0.020	0.020	0.020
3	0.508	0.328	0.234	0.179	0.150	0.092	0.091	0.091	0.091	0.091
4	0.662	0.336	0.122	0.067	0.059	0.016	0.015	0.015	0.015	0.015
5	0.279	0.206	0.145	0.100	0.099	0.047	0.044	0.039	0.034	0.034
6	0.271	0.261	0.204	0.172	0.160	0.106	0.082	0.081	0.080	0.080
7	2.481	1.521	0.904	0.623	0.411	0.199	0.128	0.107	0.101	0.101
8	1.332	0.838	0.515	0.320	0.288	0.142	0.117	0.106	0.090	0.076
9	1.376	1.083	0.725	0.515	0.424	0.191	0.149	0.105	0.094	0.094
10	0.830	0.434	0.314	0.256	0.217	0.130	0.093	0.082	0.080	0.072
11	0.366	0.122	0.088	0.046	0.031	0.027	0.003	0.002	0.002	0.002
12	0.781	0.449	0.265	0.190	0.160	0.096	0.062	0.031	0.021	0.018
13	11.363	0.798	0.055	0.022	0.011	0.007	0.002	0.002	0.002	0.002
14	1.932	1.203	0.858	0.682	0.579	0.473	0.396	0.365	0.013	0.013
15	2.681	1.292	0.916	0.730	0.593	0.287	0.210	0.201	0.017	0.017
16	1.055	0.759	0.493	0.364	0.273	0.130	0.067	0.050	0.046	0.046
17	1.120	0.962	0.849	0.750	0.744	0.559	0.482	0.476	0.466	0.466
18	1.676	0.762	0.560	0.434	0.375	0.269	0.237	0.149	0.142	0.142
19	1.883	1.197	1.001	0.541	0.185	0.103	0.019	0.013	0.013	0.013
20	0.510	0.224	0.138	0.037	0.028	0.011	0.006	0.003	0.003	0.003
21	2.636	1.888	1.695	1.623	1.608	1.317	0.006	0.006	0.006	0.006
22	0.132	0.030	0.016	0.011	0.008	0.003	0.002	0.002	0.002	0.002
23	3.454	1.648	0.232	0.081	0.021	0.004	0.003	0.003	0.003	0.003
24	1.342	0.603	0.327	0.144	0.080	0.047	0.015	0.008	0.007	0.007
25	5.472	1.416	0.830	0.677	0.575	0.450	0.353	0.023	0.016	0.014
26	1.281	0.236	0.142	0.092	0.060	0.040	0.031	0.020	0.019	0.019
27	55.038	1.343	1.106	1.044	1.030	0.983	0.449	0.443	0.441	0.439

Table 3.5, the columns represent the ratio between the number of structural dependencies and the number of co-changing pairs that remain after filtering out pairs where both factors fall below the specified threshold in the column header.

We calculate this ratio between co-changing pairs and structural dependencies to assess the size of the extracted co-changing pairs compared to the structural dependencies in the system. According to surveys [58], [68], the primary reason logical dependencies (i.e., filtered co-changes) are not used together with structural dependencies is their size. Therefore, it is important to evaluate the ratio between the sizes of co-changes and structural dependencies at each filtering step.

The results in Tables 3.5 and 3.6 show that the number of co-changing pairs is reduced. In most cases, the number of structural dependencies exceeds the number of co-changing pairs after filtering. However, the goal of filtering is not just to reduce the size of co-changing pairs but to ensure that the remaining co-changing pairs truly indicate logical dependencies.

We keep a manageable number of co-changing pairs by filtering out all co-changing pairs that do not co-occur at least 50% of the time (factor A and factor B ≥ 50%). Considering both the output size and the connection strength of these pairs, the remaining co-changing pairs can be regarded as logical dependencies at this stage.

4. COMBINING STRUCTURAL AND LOGICAL DEPENDENCIES

4.1. Overlaps between structural and logical dependencies

A logical dependency can be also a structural dependency and vice-versa, so studying the overlapping between logical and structural dependencies while filtering is important since the intention is to introduce those logical dependencies among with structural dependencies in architectural reconstruction systems. Current studies have shown a relatively small percentage of overlapping between them with and without any kind of filtering [13]. This means that a lot of non related entities update together in the versioning system, the goal here is to establish the factors that determine such a small percentage of overlapping [112].

Since we are first extracting co-changing pairs and only after various filters we call the remaining co-changing pairs logically dependent, we will be studying the overlapping between the remaining co-changing pairs after each filtering stage and the structural dependencies. For each system, we extracted the structural dependencies and the co-changing pairs and determined the overlap between the two dependencies sets, in various experimental conditions.

One variable experimental condition is whether changes located in comments contribute towards logical dependencies. This condition distinguishes between two different cases:

- with comments: a change in source code files is counted as a co-changing pair, even if the change is inside comments in all files
- without comments: commits that changed source code files only by editing comments are ignored

In all cases, we varied the following threshold values:

- commit size (cs): the maximum size of commit transactions which are accepted to generate co-changes. The values for this threshold were 5, 10, 20 and no threshold (infinity).
- number of occurrences (occ): the minimum number of repeated occurrences for a co-change to be counted as logical dependency. The values for this threshold were 1, 2, 3 and 4.

The six tables below present the synthesis of our experiments. We have computed the following values:

- the mean ratio of the number of co-changes to the number of structural dependencies (SD)
- the mean percentage of structural dependencies that are also co-changes (calculated from the number of overlaps divided to the number of structural depen-

dencies)
- the mean percentage of co-changes that are also structural dependencies (calculated from the number of overlaps divided to the number of co-changes)

In all the six tables, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 we have on columns the values used for the commit size cs, while on rows we have the values for the number of occurrences threshold occ. The tables contain median values obtained for experiments done under all combinations of the two threshold values, on all test systems. In all tables, the upper right corner corresponds to the most relaxed filtering conditions, while the lower left corner corresponds to the most restrictive filtering conditions.

Table 4.1: Ratio of number of co-changes to number of SD, case with comments

	$cs \leq 5$	$cs \leq 10$	$cs \leq 20$	$cs < \infty$
$occ \geq 1$	3,39	5,67	9,00	80,31
$occ \geq 2$	2,24	3,47	5,02	60,14
$occ \geq 3$	1,04	2,53	3,52	44,68
$occ \geq 4$	0,90	2,16	2,88	33,47

Table 4.2: Ratio of number of co-changes to number of SD, case without comments

	$cs \leq 5$	$cs \leq 10$	$cs \leq 20$	$cs < \infty$
$occ \geq 1$	3,24	5,33	7,90	67,16
$occ \geq 2$	1,35	3,27	4,72	47,39
$occ \geq 3$	1,00	1,67	2,49	32,39
$occ \geq 4$	0,43	1,26	1,93	22,15

Table 4.3: Percentage of SD that are also co-changes, case with comments

	$cs \leq 5$	$cs \leq 10$	$cs \leq 20$	$cs < \infty$
$occ \geq 1$	19,75	29,86	39,29	76,59
$occ \geq 2$	12,50	20,20	27,68	66,11
$occ \geq 3$	8,49	14,22	19,94	55,99
$occ \geq 4$	6,58	10,95	15,76	47,12

Table 4.4: Percentage of SD that are also co-changes, case without comments

	$cs \leq 5$	$cs \leq 10$	$cs \leq 20$	$cs < \infty$
$occ \geq 1$	18,88	28,47	37,44	71,12
$occ \geq 2$	11,87	19,03	25,93	59,58
$occ \geq 3$	8,00	13,09	18,15	48,65
$occ \geq 4$	5,85	9,94	14,27	39,07

In order to assess the influence of comments, we compare pairwise Tables 4.1 and 4.2, Tables 4.3 and 4.4 and Tables 4.5 and 4.6. We observe that, although there are some differences between pairs of measurements done in similar conditions with and without comments, the differences are not significant.

On the other hand, the overlap between structural and co-changes is given by the number of pairs of classes that have both structural and co-change dependencies.

Table 4.5: Percentage of co-changes that are also SD, case with comments

	$cs \leq 5$	$cs \leq 10$	$cs \leq 20$	$cs < \infty$
$occ \geq 1$	12,02	8,86	6,72	1,79
$occ \geq 2$	15,05	11,71	9,38	2,21
$occ \geq 3$	17,45	13,97	11,57	2,86
$occ \geq 4$	18,96	15,28	12,94	3,67

Table 4.6: Percentage of co-changes that are also SD, case without comments

	$cs \leq 5$	$cs \leq 10$	$cs \leq 20$	$cs < \infty$
$occ \geq 1$	12,05	9,02	6,98	1,93
$occ \geq 2$	15,08	12,03	9,66	2,42
$occ \geq 3$	17,78	14,37	12,24	3,28
$occ \geq 4$	19,22	15,59	13,30	4,21

Table 4.7: Percentage of SD that are also co-changing pairs after connection strength filtering.

Condition	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$
factor A and factor B	11.20	6.80	4.44	3.25	2.58	1.74	1.16	0.57	0.35	0.33
factor A or factor B	15.94	11.02	7.56	5.59	4.52	2.90	2.00	1.33	1.04	1.02

Table 4.8: Percentage of co-changing pairs that are SD after connection strength filtering.

Condition	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$
factor A and factor B	10.95	20.61	23.73	26.75	28.57	33.31	33.43	38.34	42.52	39.41
factor A or factor B	12.19	16.85	19.41	20.70	21.63	22.84	21.86	23.08	24.00	22.73

We evaluate this overlap as a percentage relative to the number of structural dependencies in Tables 4.3,4.4 and 4.7, respectively as a percentage relative to the number of co-changes in Tables 4.5,4.6, 4.8.

A first observation from Tables 4.3, 4.4, and 4.7 is that not all pairs of classes with structural dependencies co-change. The biggest value for the percentage of structural dependencies that are also co-changes is 76.5% obtained in the case when no filterings are done.

From Tables 4.5, 4.6, and 4.8 we notice that the percentage of co-changes which are also structural is always low to very low. This means that most co-changes are recorded between classes that have no structural dependencies to each other [112].

4.2. Weight Assignment

4.2.1. Structural dependencies weights

Structural dependencies are important for understanding the architecture of a software system because they reveal how different modules interact at the code level. In our research, we extract structural dependencies using a tool from our previous work [63]. This tool analyzes the source code to identify various relationships between software entities and exports them in CSV format.

Structural dependencies do not all have the same level of influence on a software system's architecture and behavior. For instance, the relationship between a variable and the class that uses it is not the same as the relationship between a class and the interface it implements. To reflect these differences, we assign different weights to each type of dependency.

The dependency types and weights were previously defined in related works on clustering [60], [1].

Table 4.9 shows the weights assigned to different categories of structural dependencies, as proposed in previous works.

Weight	Dependency types
4	Interface realization
3	Inheritance, parameter, return type, field, cast, type binding
2	Method call, field access, instantiation
1	Local variable

Table 4.9: Weights assigned to different structural dependency types. [1]

The weights are assigned based on the following considerations:

Weight 4 – Interface Realization: Assigned the highest weight because it signifies a strong architectural relationship. Implementing an interface means classes are expected to provide specific functionalities.

Weight 3 – Inheritance, Parameter, Return Type, Field, Cast, Type Binding: These dependencies represent significant connections between entities. They include inheritance relationships and shared data or types, which affect the behavior and properties of entities.

Weight 2 – Method Call, Field Access, Instantiation: These indicate interactions between classes but are less impactful than higher weights. They involve using methods or fields of other classes or creating instances. When a method call, field access, or instantiation occurs multiple times between the same pair of entities, the weight is multiplied by the number of occurrences. For example, if Class A calls a method in Class B three times, the assigned weight would be 6 (weight 2 multiplied by 3).

Weight 1 – Local Variable: Given the lowest weight, local variables are the

most basic level of interaction.

4.2.2. Logical dependencies weights

We refer to logical dependencies as the filtered co-changes between software entities. A co-change occurs when two or more software entities are modified together during the same commit in the version control system. Co-changes indicate that these entities are likely directly or indirectly related or dependent on each other.

Co-changes are associated with a degree of uncertainty. Compared to structural dependencies, where a dependency is certain, co-changes are less reliable. For example, if the system was migrated from one version control system to another, the first commit will include all the entities from the system at that point in time. Should we consider all these entities related to one another in this case? This would introduce false dependencies and reduce the likelihood of achieving accurate results when combining them with more reliable types of dependencies.

Even if we address the issue of the first commit, a developer can still resolve multiple unrelated issues in the same commit (even though development processes do not recommend this).

To solve this problem, in our previous works, we refined some filtering methods to ensure that the co-changes that remain after filtering are more reliable and suitable for use with other dependencies or individually [63], [110], [112]. Based on our previous results, the filters we decided to use further in our research are the commit size filter and the strength filter. Both filters are used together, and the result is the set of logical dependencies that we use to generate software clusters.

It is important to note that the strength metric is only used for filtering and is *not considered as a weight* of the dependencies. The *weight assigned to each dependency is the number of commits in which both entities were updated together*.

4.3. Integration techniques

When structural dependencies (SD) and logical dependencies (LD) are combined in software clustering, both types of relationships are represented within the same graph.

Each entity in the system is represented as a node in the graph, and the dependencies between them are represented as directed weighted edges.

SD and LD weights are combined when the same pair of entities appear in both dependencies. In this case, the weights from SD and LD are summed, giving more influence to those entity pairs. When a pair of entities appear only in SD or only in LD, the edge is added to the graph together with its corresponding weight.

Figure 4.1 illustrates combining structural and logical dependencies in the same dependency graph. The structural dependencies between `House`, `OrangeCat`, and `CatBehavior` entities are visible from the source code analysis.

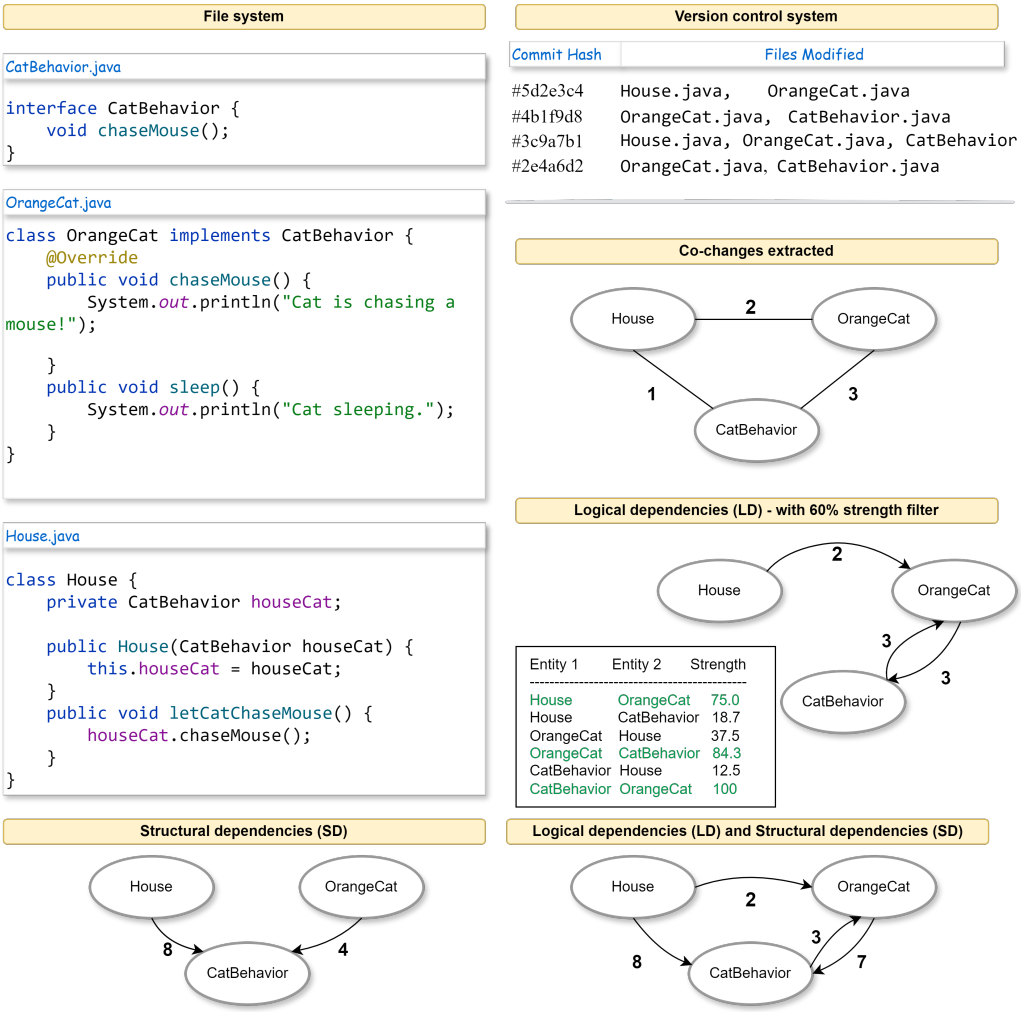

Figure 4.1: Dependency Graph: Combining structural and logical dependencies.

However, the combination of SD and LD reveals additional insights. One important observation is the logical dependency between `House` and `OrangeCat`, which is not observed from the structural analysis. This relation is extracted from version control and filtered using a 60% strength filter. The strength metric reveals that `House` and `OrangeCat` have a significant co-change value of 75.0, usually associated with a strong relationship.

When SD and LD overlap, such as between `OrangeCat` and `CatBehavior`, their weights are summed. This summation increases the weight of the dependency, making it more important in the dependency graph.

5. LOGICAL DEPENDENCIES IN KEY CLASS DETECTION

5.1. Introduction

Zaidman et al [14] were the first to introduce the concept of key classes and it refers to classes that can be found in documents written to provide an architectural overview of the system or an introduction to the system structure. Tahvildari and Kontogiannis have a more detailed definition regarding key classes concept: "Usually, the most important concepts of a system are implemented by very few key classes which can be characterized by the specific properties. These classes, which we refer to as key classes, manage many other classes or use them in order to implement their functionality. The key classes are tightly coupled with other parts of the system. Additionally, they tend to be rather complex, since they implement much of the legacy system's functionality" [87]. Also, other researchers use a similar concept as the one defined by Zaidman but under different terms like important classes [114] or central software classes [115].

The key class identification can be done by using different algorithms with different inputs. In the research of Osman et al., the key class identification is made by using a machine learning algorithm and class diagrams as input for the algorithm [91]. Thung et al. builds on top of Osman et al.'s approach and adds network metrics and optimistic classification in order to detect key classes [90].

Zaidman et al. use a webmining algorithm and dynamic analysis of the source code to identify the key classes [14].

Sora et al. use a page ranking algorithm for finding key classes and static analysis of the source code [17], [88], [89]. In [1] the authors use in addition to the previous research also other class attributes to identify important classes. The page ranking algorithm is a customization of PageRank, the algorithm used to rank web pages [116]. The PageRank algorithm works based on a recommendation system. If one node has a connection with another node, then it recommends the second node. In previous works, connections are established based on structural dependencies extracted from static code analysis. If A has a structural dependency with B, then A recommends B, and also B recommends A.

The ranking algorithm ranks all the classes from the source code of the system analyzed according to their importance. To identify the important classes from the rest of the classes a threshold for TOP classes from the top of the ranking is set. The TOP threshold value can go from 1 to the total number of classes found in the system.

Some researchers [14], [117], [118] consider that 15% of the total number of classes of the system is a suited value for the TOP threshold. Other researchers [1] consider that 15% of the total number of classes is a too high value for the TOP threshold and suggest that a value in the range of 20–30 is better.

5.2. Metrics for results evaluation

To evaluate the quality of the key classes ranking algorithm and solution produced, the key classes found by the algorithm are compared with a reference solution.

The reference solution is extracted from the developer documentation. Classes mentioned in the documentation are considered key classes and form the reference solution (ground truth) used for validation [119].

For the comparison between both solutions, is used a classification model. The quality of the solution produced is evaluated by using metrics that evaluate the performance of the classification model, such as Precision-Recall and Receiver Operating Characteristic Area Under Curve (ROC-AUC).

A classification model (or ”classifier”) is a mapping between expected results and predicted results [120], [121]. Both results can be labeled as positive or negative, which leads us to the confusion matrix from figure 5.1. The confusion matrix has the

Expected Result \ Predicted Result	Positive	Negative
Positive	***T**rue* ***P**ositive*	***F**alse* ***P**ositive*
Negative	***F**alse* ***N**egative*	***T**rue* ***N**egative*

Figure 5.1: Confusion matrix

following outcomes:

- *true positive*, if the expected result is positive and the predicted result is also positive.
- *false positive*, if the expected result is positive but the predicted result is negative.
- *false negative*, if the expected result is negative but the predicted result is positive.
- *true negative*, if the expected result is negative and the predicted result is also negative.

Precision-recall

Precision is the ratio of True Positives to all the positives of the result set.

$$precision = \frac{TP}{TP + FN} \tag{5.1}$$

The recall is the ratio of True Positives to all the positives of the reference set.

$$recall = \frac{TP}{TP + FP} \tag{5.2}$$

As mentioned in section 5.1, to distinguish the key classes from the rest of the classes a TOP threshold is used. Some researchers consider that 15% of the total classes is the best value for the TOP threshold and others consider that the value should be in the range of 20-30.

The precision-recall metric is suited if the threshold value is fixed. If the threshold value is variable, then metrics that capture the behavior over all possible values must be used. Such metric is the Receiver Operating Characteristic metric.

Receiver Operating Characteristic Area Under Curve

The ROC graph is a two-dimensional graph that has on the X-axis plotted the false positive rate and on the Y-axis the true positive rate. By plotting the true positive rate and the false positive rate at thresholds that vary between a minimum and a maximum possible value we obtain the ROC curve. The area under the ROC curve is called Area Under the Curve (AUC).

The true positive rate of a classifier is calculated as the division between the number of true positive results identified and all the positive results identified:

$$True\ positive\ rate(TPR) = \frac{TP}{TP + FN} \tag{5.3}$$

The false positive rate of a classifier is calculated as the division between the number of false positive results identified and all the negative results identified:

$$False\ positive\ rate(FPR) = \frac{FP}{FP + TN} \tag{5.4}$$

In multiple related works, the ROC-AUC metric has been used to evaluate the results for finding key classes of software systems. For a classifier to be considered good, its ROC-AUC metric value should be as close to 1 as possible, when the value is 1 then the classifier is considered to be perfect.

Osman et al. obtained in their research an average Area Under the Receiver Operating Characteristic Curve (ROC-AUC) score of 0.750 [91]. Thung et al. obtained an average ROC-AUC score of 0.825 [90] and Sora et al. obtained an average ROC-AUC score of 0.894 [1].

5.3. Data set used

In this section, we will look over all the systems studied in the baseline research presented in section 5.4.1, and we will try to identify the systems that could be used also in our current research involving logical dependencies.

The research of I. Sora et al [1] takes into consideration structural public dependencies that are extracted using static analysis techniques and was performed on the object-oriented systems presented in table 5.1.

The requirements for a system to qualify as suited for investigations using logical dependencies are: has to be on GitHub, has to have release tags to identify the version, and also has to have an increased number of commits. From the total of 14 object-oriented systems listed in the paper [1], 13 of them have repositories in Github 5.2. And from the found repositories we identified only 6 repositories that have the same release tag as the specified version from table 5.1. It is important to identify the correct release tag for each repository to limit the commits further analyzed by date. Only commits that were made until the specified release are considered and analyzed. The commits number found on the remaining 6 repositories varies from 19108 commits for Tomcat Catalina to 149 commits for JHotDraw. In order to have more accurate results, we need a significant number of commits, so we reached the conclusion that only 3 systems can be used for key classes detection using logical dependencies: Apache Ant, Hibernate, and Tomcat Catalina. From all the systems mentioned in table 5.1 Apache Ant is the most used and analyzed in other works [112], [122], [123], [124].

Table 5.1: Analyzed software systems in previous research paper.

ID	System	Description	Version
Sl	Apache Ant	Java library and command line tool that drive the build processes as targets and extension points depending upon each other	1.6.1
S2	Argo UML	UML modelling tool with support for all UML diagrams.	0.9.5
S3	GWT Portlets	Open source web framework for building GWT (Google Web Toolkit) Applications.	0.9.5 beta
S4	Hibernate	Persistence framework for Java.	5.2.12
S5	javaclient	Java distributed application for playing with robots	2.0.0
S6	jEdit	Java mature text editor for programmers.	5.1.0
S7	JGAP	Genetic Algorithms and Genetic Programming Java library.	3.6.3
S8	JHotDraw	JHotDraw is a two-dimensional graphics framework for structured drawing editors that is written in Java.	6.0b.1
S9	JMeter	JMeter is a Java application designed to load test functional behavior and measure performance	2.0.1
S10	Log4j	Logging Service	2.10.0
S11	Mars	The Mars Simulation Project is a Java project that models and simulates human settlements on Mars planet	3.06.0
S12	Maze	The Maze-solver project simulates an artificial intelligence algorithm on a maze	1.0.0
S13	Neuroph	Neuroph is a Java neural network framework.	2.2.0
S14	Tomcat Catalina	The Apache Tomcat project is an open-source implementation of JavaServlet and JavaServerPages technologies	9.0.4
S15	Wro4J	The Wro4J is a web resource (JS and CSS) optimizer for Java.	1.6.3

Table 5.2: Found systems and versions of the systems in GitHub.

ID	System	Version	Release Tag name	Commits number
Sl	Apache Ant	1.6.1	rel/1.6.1	6713
S2	Argo UML	0.9.5	not found	0
S3	GWT Portlets	0.9.5 beta	not found	0
S4	Hibernate	5.2.12	5.2.12	6733
S5	javaclient	2.0.0	not found	0
S6	jEdit	5.1.0	not found	0
S7	JGAP	3.6.3	not found	0
S8	JHotDraw	6.0b.1	not found	149
S9	JMeter	2.0.1	v2_1_1	2506
S10	Log4j	2.10.0	v1_2_10-recalled	634
S11	Mars	3.06.0	not found	0
S12	Maze	1.0.0	not found	0
S13	Neuroph	2.2.0	not found	0
S14	Tomcat Catalina	9.0.4	9.0.4	19108
S15	Wro4J	1.6.3	v1.6.3	2871

5.4. Measurements using logical dependencies

As we mentioned in the beginning the purpose is to check if the logical dependencies can improve key class detection.

As presented in section 5.4.1, and section 5.1 the key class detection was done by using structural dependencies of the system. In this section, we will use the same tool used in the baseline approach presented in section 5.4.1, and we will add a new input to it, the logical dependencies.

Below is a comparison between the new approach and baseline approach, how we collect the logical dependencies, the results obtained previously, and the new results obtained. The new results are separated into two categories, the results obtained by using structural and logical dependencies and the results obtained by using only logical dependencies.

5.4.1. Baseline approach

We use the research of I. Sora et al [1] as a baseline for our research involving the usage of logical dependencies to find key classes. The baseline approach uses a tool that takes as an input the source code of the system and applies ranking strategies to rank the classes according to their importance.

In order to rank the classes according to their importance, different class metrics are used [117], [14], [118]. Below are presented some of the class metrics used in the baseline approach in order to rank the classes according to their importance.

Class attributes that characterize key classes

The metrics used in the baseline research can be grouped into the following categories:

- class size metrics: number of fields (NoF), number of methods (NoM), global size (Size = NoF+NoM).
- class connection metrics, any structural dependency between two classes:
  - CONN-IN, the number of distinct classes that use a class;
  - CONN-OUT, the total number of distinct classes that are used by a class;
  - CONN-TOTAL, the total number of distinct classes that a class uses or are used by a class (CONN-IN + CONN-OUT).
  - CONN-IN-W, the total weight of distinct classes that use a class.
  - CONN-OUT-W, the total weight of distinct classes that are used by a class.
  - CONN-TOTAL-W, the total weight of all connections of the class (CONN-IN-W + CONN-OUT-W) [1].
- class pagerank values, previous research use pagerank values computed on both directed and undirected, weighted and unweighted graphs:
  - PR - value computed on the directed and unweighted graph;
  - PR-W - value computed on the directed and weighted graph;
  - PR-U - value computed on the undirected and unweighted graph;
  - PR-U-W - value computed on the undirected and weighted graph;
  - PR-U2-W - value computed on the weighted graph with back-recommendations [17], [88], [1], [89].

Based on the class attributes presented, all the classes of the system are ranked. To differentiate the important (key) classes from the rest of the classes, a TOP threshold for the top classes found is set. The threshold vary between 20 and 30 classes.

The baseline approach not only identifies the key classes but also evaluates the performance of the solution produced. The same approach as the one presented in section 5.2 is used for the evaluation of the results. The key classes found by the ranking algorithm are compared with a reference solution that is extracted from the developer documentation by using a classification model.

The true positives (TP) are the classes found in the reference solution and also in the top TOP ranked classes. False positives (FP) are the classes that are not in the reference solution but are in the TOP ranked classes. True Negatives (TN) are classes that are found neither in the reference solution nor in the TOP ranked classes. False Negatives (FN) are classes that are found in the reference solution but not found in the TOP ranked classes.

Due to the fact that the TOP threshold is varied, the Receiver Operating Characteristic Area Under Curve metric is used for the evaluation of the results.

The entire workflow of the baseline approach that was presented above is also presented in figure 5.2.

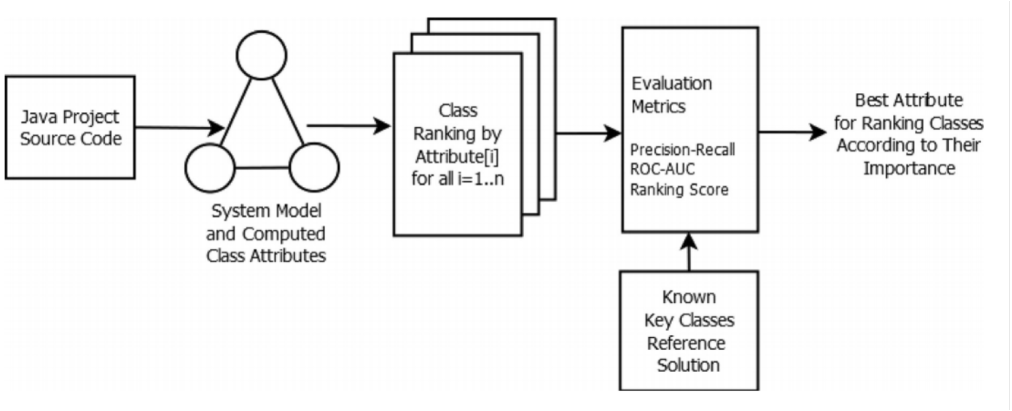

Figure 5.2: Overview of the baseline approach. Reprinted from "Finding key classes in object-oriented software systems by techniques based on static analysis." by Ioana Sora and Ciprian-Bogdan Chirila, 2019, Information and Software Technology, 116:106176. Reprinted with permission.

5.4.2. Comparison with the baseline approach

The baseline approach uses a tool that takes as input the source code of the system to identify the key classes and the reference solution to evaluate the quality of the solution. We modified the tool such that it can also take as input the logical dependencies.

In order to rank the classes according to their importance, the tool uses different class metrics. The list of the metrics used in the baseline approach is presented in section 5.4.1. The difference in the metrics used compared with the baseline approach is that we use a subset of those metrics. The reason why we are not using all the metrics is that the extracted logical dependencies are undirected. The metrics used by the current approach are CONN-TOTAL, CONN-TOTAL-W, PR-U, PR-U-W, and PR-U2-W.

We did not change the rest of the workflow of the tool. Meaning that the TOP threshold is varied between 20 and 30 and the resulting solution is evaluated by using the ROC-AUC metric. The goal being a ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) metric value as close to 1 as possible.

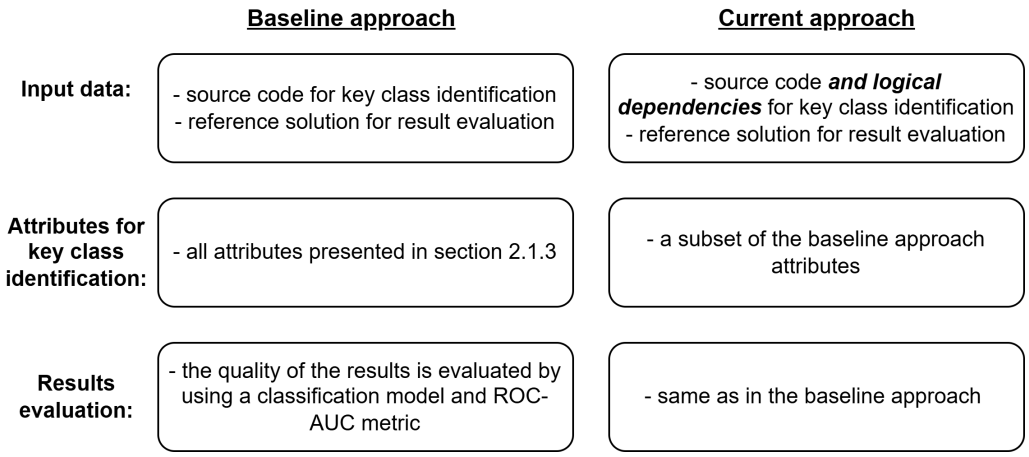

Figure 5.3: Comparison between the new approach and the baseline

5.4.3. Logical dependencies collection and current workflow used

The logical dependencies are those co-changing pairs extracted from the versioning system history that remain after filtering. The filtering part consists of applying two filters: the filter based on commit size and the filter based on connection strength.

To determine the connection strength of a pair, we first need to calculate the connection factors for both entities that form a co-changing pair. Assuming that we have a co-changing pair formed by entities A and B, the connection factor of entity A with entity B is the percentage from the total commits involving A that contains entity B. The connection factor of entity B with entity A is the percentage from the total commits involving B that contain also entity A.

$$connection\ factor\ for\ A = \frac{100 * commits\ involving\ A\ and\ B}{total\ nr\ of\ commits\ involving\ A} \tag{5.5}$$

$$connection\ factor\ for\ B = \frac{100 * commits\ involving\ A\ and\ B}{total\ nr\ of\ commits\ involving\ B} \tag{5.6}$$

We calculated the connection factor for each entity involved in a co-changing pair and filtered the co-changing pairs based on it. The rule set is that both entities had to have a connection factor with each other greater than the threshold value.

After the filtering part, the remaining co-changing pairs, now called logical dependencies, are exported in CSV files.

The entire process of extracting co-changing pairs from the versioning system, filter them, and export the remaining ones into CSV files is done with a tool written in Python.

The next step is to use the exported logical dependencies for key classes detection. In order to do that we used the same key class detection tool used in the previous research presented in section 5.4.1. We adapted the tool to be able to process also logical dependencies because previously the tool used only structural dependencies extracted from the source code of the software systems. The workflow is presented in figure 5.4

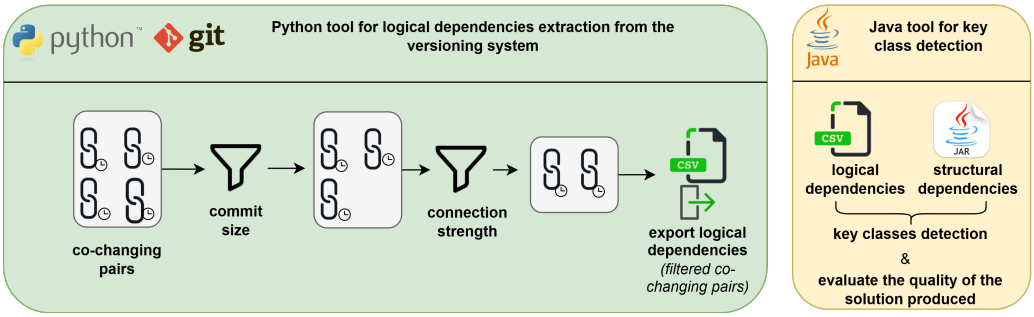

Figure 5.4: Workflow for key classes detection

5.4.4. Measurements using only the baseline approach

In table 5.3 are presented the ROC-AUC values for different attributes computed for the systems Ant, Tomcat Catalina, and Hibernate by using the baseline approach. We intend to compare these values with the new values obtained by using also logical dependencies in key class detection.

Table 5.3: ROC-AUC metric values extracted.

Metrics	Ant	Tomcat Catalina	Hibernate
PR_U2_W	0.95823	0.92341	0.95823
PR	0.94944	0.92670	0.94944
PR_U	0.95060	0.93220	0.95060
CONN_TOTAL_W	0.94437	0.92595	0.94437
CONN_TOTAL	0.94630	0.93903	0.94630

5.4.5. Measurements using combined structural and logical dependencies

The tool used in the baseline approach runs a graph-ranking algorithm. The graph used contains the structural dependencies extracted from static source code analysis. Each edge in the graph represents a dependency, the entities that form a structural dependency are represented as vertices in the graph. As mentioned in section 5.4.2, we modified the tool to read also logical dependencies and add them to the graph. In this section, we add in the graph the logical dependencies together with the structural dependencies.

In tables 5.4, 5.5, and 5.6, on each line, we have the metric that is calculated and on each column, we have the connection strength threshold that was applied to

the logical dependencies used in identifying the key classes. We started with logical dependencies that have a connection strength greater than 10%, which means that in at least 10% of the commits involving A or B, A and B update together. Then we increased the threshold value by 10 until we remained only with entities that update in all the commits together. The last column contains the results obtained previously by the tool by only using structural dependencies.

As for the new results obtained by combining structural and logical dependencies, highlighted with orange are the values that are close to the previously registered values but did not surpass them. Highlighted with green are values that are better than the previously registered values. At this step, we can also observe that for all three systems measured in tables 5.4, 5.5, and 5.6, the best values obtained are for connection strength between 40-70%.

Table 5.4: Measurements for Ant using structural and logical dependencies combined

Metrics	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$	Baseline
PR_U2_W	0.924	0.925	0.926	0.927	0.927	0.927	0.929	0.928	0.928	0.928	0.929
PR	0.914	0.854	0.851	0.866	0.876	0.882	0.887	0.854	0.852	0.852	0.855
PR_U	0.910	0.930	0.933	0.933	0.935	0.934	0.939	0.933	0.933	0.933	0.933
CON_T_W	0.924	0.928	0.931	0.932	0.933	0.934	0.936	0.934	0.934	0.934	0.934
CON_T	0.840	0.886	0.904	0.909	0.915	0.923	0.932	0.935	0.936	0.936	0.942

Table 5.5: Measurements for Tomcat using structural and logical dependencies combined

Metrics	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$	Baseline
PR_U2_W	0.910	0.917	0.923	0.924	0.924	0.924	0.924	0.924	0.924	0.924	0.923
PR	0.811	0.800	0.815	0.834	0.847	0.852	0.853	0.858	0.858	0.858	0.927
PR_U	0.910	0.921	0.931	0.933	0.933	0.932	0.933	0.932	0.932	0.932	0.932
CON_T_W	0.914	0.920	0.924	0.926	0.926	0.926	0.926	0.926	0.926	0.926	0.926
CON_T	0.868	0.906	0.930	0.936	0.937	0.938	0.938	0.938	0.938	0.938	0.939

Table 5.6: Measurements for Hibernate using structural and logical dependencies combined

Metrics	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$	Baseline
PR_U2_W	0.954	0.957	0.958	0.958	0.958	0.958	0.958	0.958	0.958	0.958	0.958
PR	0.929	0.929	0.933	0.939	0.939	0.946	0.947	0.947	0.947	0.947	0.949
PR_U	0.942	0.947	0.948	0.949	0.949	0.950	0.950	0.950	0.950	0.950	0.951
CON_T_W	0.939	0.942	0.943	0.944	0.944	0.945	0.945	0.945	0.945	0.945	0.944
CON_T	0.924	0.933	0.938	0.941	0.941	0.944	0.945	0.945	0.945	0.945	0.946

5.4.6. Measurements using only logical dependencies

In the previous section, we added in the graph based on which the ranking algorithm works the logical and structural dependencies. In the current section, we will add only the logical dependencies to the graph.

In tables 5.7, 5.8, and 5.9, are presented the results obtained by using only logical dependencies to detect key classes. The measurements obtained are not as good as using logical and structural dependencies combined or using only structural dependencies. But, all the values obtained are above 0.5, which means that a good part of the key classes is detected by only using logical dependencies. As mentioned in section 5.2, a classifier is good if it has the ROC-AUC value as close to 1 as possible.

One possible explanation for the less performing results is that the key classes may have a better design than the rest of the classes, which means that are less prone to change. If the key classes are less prone to change, this implies that the number of dependencies extracted from the versioning system can be less than for other classes.

Table 5.7: Measurements for Ant using only logical dependencies

Metrics	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$	Baseline
PR_U2_W	0.720	0.627	0.718	0.703	0.732	0.824	0.852	0.881	0.876	0.876	0.929
PR	0.720	0.627	0.718	0.703	0.732	0.824	0.852	0.881	0.876	0.876	0.855
PR_U	0.720	0.627	0.718	0.703	0.732	0.824	0.852	0.881	0.876	0.876	0.933
CON_T_W	0.722	0.581	0.644	0.676	0.727	0.819	0.842	0.874	0.876	0.876	0.934
CON_T	0.722	0.581	0.644	0.676	0.727	0.819	0.842	0.874	0.876	0.876	0.942

Table 5.8: Measurements for Tomcat using only logical dependencies

Metrics	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$	Previous
PR_U2_W	0.672	0.656	0.645	0.697	0.754	0.776	0.786	0.799	0.799	0.799	0.923
PR	0.685	0.643	0.642	0.697	0.754	0.776	0.786	0.799	0.799	0.799	0.927
PR_U	0.685	0.643	0.644	0.697	0.754	0.776	0.786	0.799	0.799	0.799	0.932
CON_T_W	0.694	0.636	0.636	0.697	0.754	0.776	0.786	0.799	0.799	0.799	0.926
CON_T	0.654	0.611	0.636	0.697	0.754	0.776	0.786	0.799	0.799	0.799	0.939

Table 5.9: Measurements for Hibernate using only logical dependencies

Metrics	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$	Baseline
PR_U2_W	0.657	0.564	0.601	0.619	0.622	0.650	0.653	0.654	0.654	0.654	0.958
PR	0.644	0.564	0.601	0.619	0.622	0.650	0.653	0.654	0.654	0.654	0.949
PR_U	0.644	0.564	0.601	0.619	0.622	0.650	0.653	0.654	0.654	0.654	0.951
CON_T_W	0.649	0.564	0.601	0.619	0.622	0.650	0.653	0.654	0.654	0.654	0.944
CON_T	0.644	0.564	0.601	0.619	0.622	0.650	0.653	0.654	0.654	0.654	0.946

5.5. Correlation between details of the systems and results

In this section, we discuss about the correlation between the details of the systems and the results obtained in section 5.4.

The reason why we are doing this correlation is to find if there are some links between the details of the systems and the results obtained.

The results obtained are presented in figures 5.5 - 5.10. We are using plots to display the results obtained to have a clearer view of how the results fluctuate over different thresholds values.

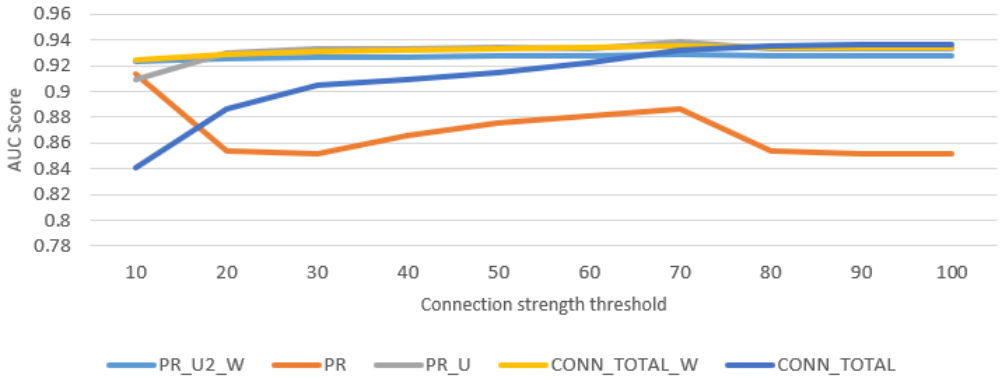

Figure 5.5: Variation of AUC score when varying connection strength threshold for Ant. Results for structural and logical dependencies combined.

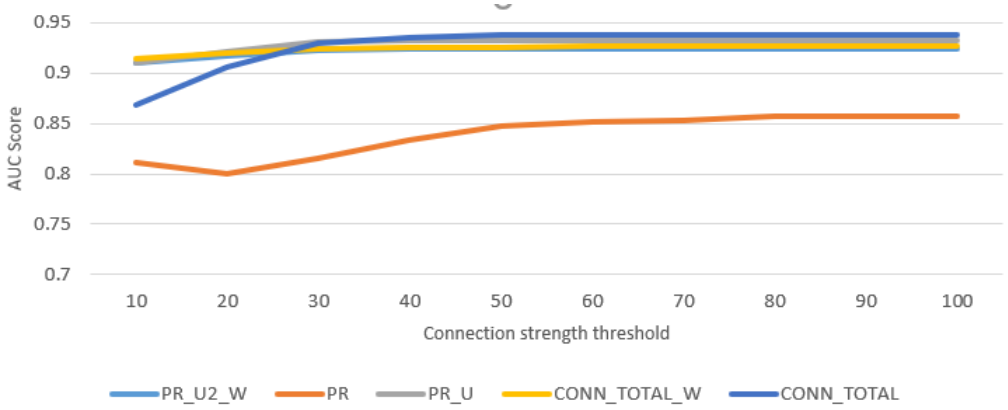

Figure 5.6: Variation of AUC score when varying connection strength threshold for Tomcat. Results for structural and logical dependencies combined.

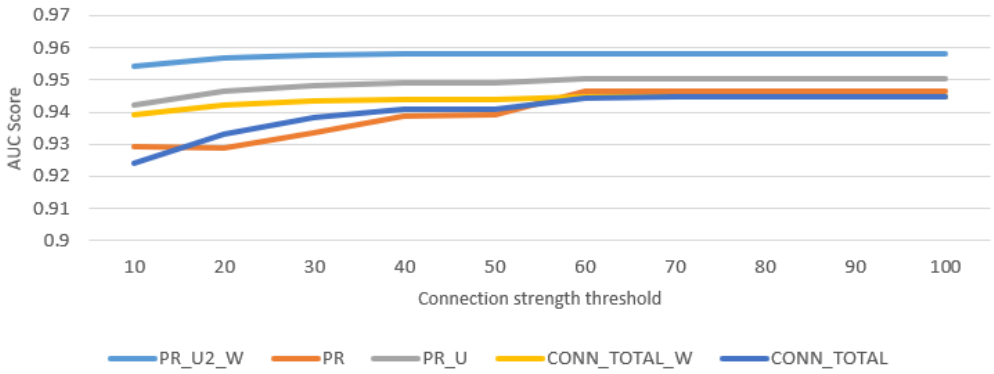

Figure 5.7: Variation of AUC score when varying connection strength threshold for Hibernate. Results for structural and logical dependencies combined.

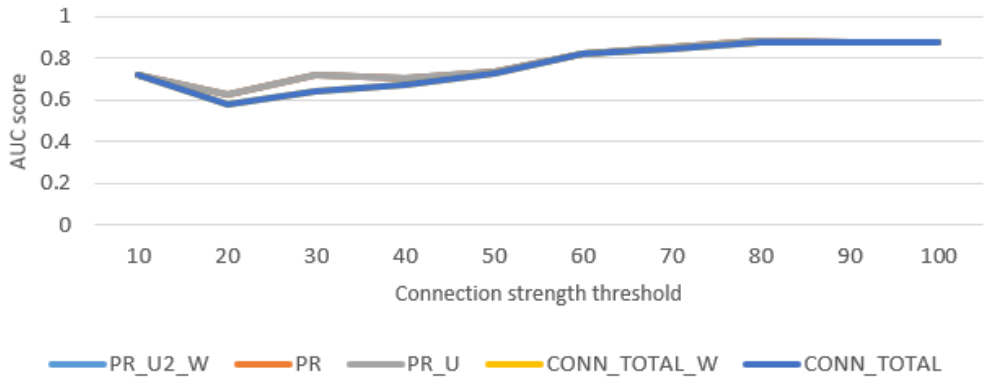

Figure 5.8: Variation of AUC score when varying connection strength threshold for Ant. Results for logical dependencies only.

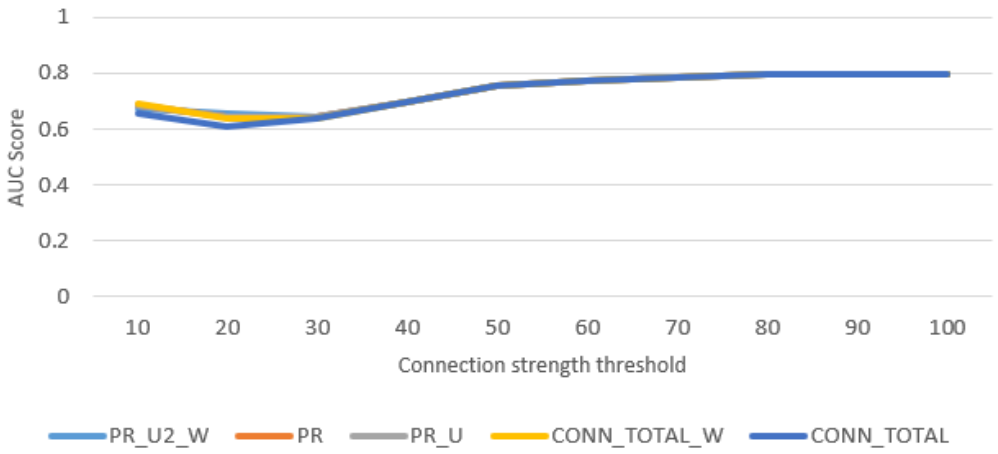

Figure 5.9: Variation of AUC score when varying connection strength threshold for Tomcat. Results for logical dependencies only.

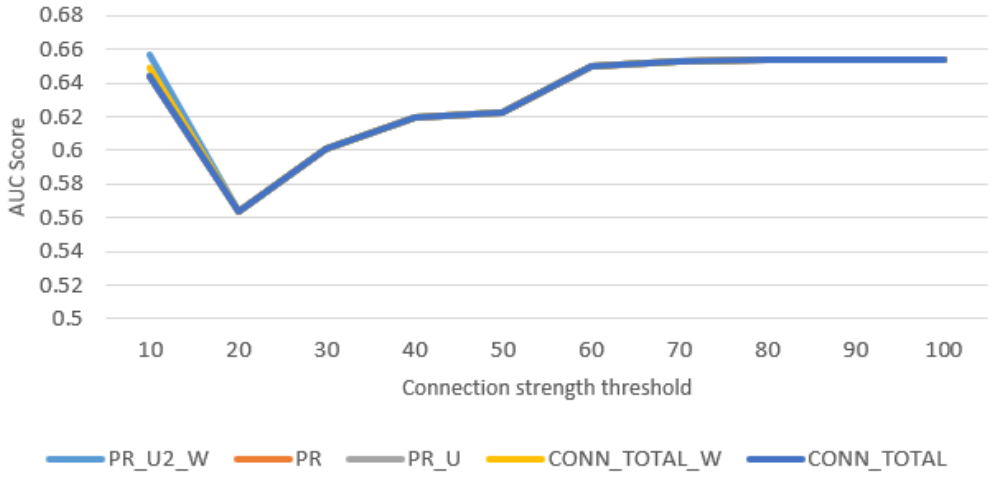

Figure 5.10: Variation of AUC score when varying connection strength threshold for Hibernate. Results for logical dependencies only.

The details of the systems are presented in two tables. In table 5.10 are the overlappings between structural and logical dependencies expressed in percentages. Each column represents the percentage of logical dependencies that are also structural, for each column the logical dependencies are obtained by applying a different connection strength filter. The connection strength filter begins at 10, meaning that in at least 10 % of the total commits involving two entities, the entities update together. We increase the connection strength filter by 10 up until we reach 100, meaning that in all the commits that involve one entity, the other entity is present also.

In table 5.11 are the ratio numbers between structural dependencies and logical dependencies. We added this table in order to highlight how different the total number of both dependencies is.

Table 5.10: Percentage of logical dependencies that are also structural dependencies

System	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$
Ant	25.202	34.419	36.385	34.656	33.528	33.333	28.659	33.333	35.294	35.294
Tomcat Catalina	4.059	22.089	25.000	25.758	25.926	37.525	47.368	55.285	75.000	76.923
Hibernate	6.546	26.607	29.565	32.374	32.543	45.170	44.980	42.473	42.473	42.473

Table 5.11: Ratio between structural and logical dependencies (SD/LD)

System	$\geq 10\%$	$\geq 20\%$	$\geq 30\%$	$\geq 40\%$	$\geq 50\%$	$\geq 60\%$	$\geq 70\%$	$\geq 80\%$	$\geq 90\%$	$\geq 100\%$
Ant	1.315	3.284	4.972	5.603	6.175	10.697	12.915	27.154	41.529	41.529
Tomcat Catalina	0.120	0.923	1.313	1.531	1.619	3.177	7.092	13.146	67.375	124.385
Hibernate	1.037	6.391	10.037	14.947	18.940	54.248	83.442	111.704	111.704	111.704

In figures 5.5, 5.6 and 5.7 are the measurements obtained by using structural and logical dependencies combined. In all three figures, the measurements at the beginning are smaller than the rest. Once with the increasing of the threshold value also the measurements begin to increase. Meaning that better results for key class detection are found. The best measurements are when the threshold value is between 40 and 60, after that, the measurements tend to decrease a little bit and stay at that fixed value.

A possible explanation of the results fluctuation and then capping is that if we are looking at table 5.11 we can see that at the beginning, the total number of logical dependencies used is close to the number of existing structural dependencies. The high volume of logical dependencies introduced might cause an erroneous detection of the key classes, in consequence, smaller measurements. When the threshold begins to be more restrictive and the total number of logical dependencies used begins to decrease, the key classes detection starts to improve. This improvement stops after the threshold value reaches 60%. If we look again at table 5.11 we can see that after 60% the number of structural dependencies outnumbers the number of logical dependencies up to 124 times in some cases. In addition, if we look at table 5.10 we can see that the remaining logical dependencies overlap a lot with the structural dependencies, so we are not introducing too much new information.

So, the number of logical dependencies used is so small that it doesn't influence the key class identification. Since the structural dependencies used don't change, we obtain the same results for different threshold values.

In figures 5.8, 5.9 and 5.10 are the measurements obtained by using only logical dependencies. Initially, we expected to see a Gaussian curve, but instead, we see a bell curve. We think that in the beginning, we use a high number of logical dependencies in key class detection, among those logical dependencies is an important number of key classes and also an important number of other classes. But the number of other classes does not influence the key classes detection. When we start to increase the value of the threshold and filter more the logical dependencies, we also filter some of the initial detected key classes and remain with a significant number of other classes. In this case, the other classes that remain influence the measurements, causing the worst-performing solutions. Some of the key classes are strongly

connected in the versioning system, and even for higher threshold values don't get filtered out. Meanwhile, the rest of the classes that are not key classes get filtered out for higher threshold values which leads to better performing measurements when the threshold value are above 60%.

5.6. Comparison of the extracted data with fan-in and fan-out metric

Fan-in and fan-out are coupling metrics. The fan-in of entity A is the total number of entities that call functions of A. The fan-out of A is the total number of entities called by A [125].

In tables 5.12, 5.13, and 5.14 we can find the metrics detalis for each documented key class of each system. The first column represents the name of each key class, the second column represents the fan_in values for each key class, the third column represents the fan_out values, the fourth column represents the number of entities that call functions of that key class plus the number of entities that are called by the key class (fan_in and fan_out combined), and the fifth column represents the number of logical dependencies in which an entity is involved.

For Ant, we can see in table 5.12 that all the key classes have logical dependencies with other classes. The LD_NUMBER means the number of logical dependencies of an entity. The key classes with the most LD number are Project and IntrospectionHelper, these two entities can be found also in table 5.15 in which we did a top 10 entities that have a logical dependency with other entities. This means that some key classes are involved in software change quite often and can be observed via system history.

Table 5.12: Measurements for Ant key classes

Nr.	Classname	FAN_IN	FAN_OUT	FAN_TOTAL	LD_NUMBER
1	Project	191	23	214	157
2	Target	28	6	34	78
3	UnknownElement	17	13	30	90
4	RuntimeConfigurable	17	13	30	118
5	IntrospectionHelper	18	24	42	143
6	Main	1	13	14	82
7	TaskContainer	11	1	12	21
8	ProjectHelper2$ElementHandler	1	12	13	30
9	Task	110	7	117	88
10	ProjectHelper	16	8	24	101

For Tomcat Catalina, same as for Ant, we can see in table 5.13 that all the key classes have logical dependencies. The key classes with the most LD number are StandardContext and Request, these two entities can also be found in table 5.16 in which we did a top 10 entities that have the most logical dependencies with other entities for Tomcat Catalina.

For Hibernate things are a little bit different, as we can see in table 5.14, key classes like Criterion, Projection, or Transaction have 0 logical dependencies, meaning that those key classes are not involved in any software change. One possible expla-

nation for this is that for Hibernate the architecture is designed in such way that the core is not often touched by change.

Table 5.13: Measurements for Tomcat Catalina key classes.

Nr.	Classname	FAN_IN	FAN_OUT	FAN_TOTAL	LD_NUMBER
1	Context	74	8	82	126
2	Request	48	28	76	215
3	Container	51	8	59	64
4	Response	38	12	50	90
5	StandardContext	11	38	49	216
6	FANector	23	9	32	89
7	Session	29	2	31	28
8	Valve	29	2	31	19
9	Wrapper	29	1	30	36
10	Manager	25	3	28	31
11	Host	26	1	27	44
12	Service	20	6	26	51
13	Engine	23	2	25	1
14	Realm	18	6	24	21
15	CoyoteAdapter	1	22	23	140
16	StandardHost	8	15	23	88
17	LifecycleListener	21	1	22	3
18	StandardEngine	2	19	21	57
19	Pipeline	19	2	21	20
20	Server	16	4	20	49
21	HostConfig	3	15	18	79
22	StandardWrapper	5	13	18	92
23	StandardService	3	12	15	81
24	Catalina	2	13	15	94
25	Loader	14	1	15	18
26	StandardServer	2	12	14	94
27	StandardPipeline	1	10	11	62
28	Bootstrap	3	3	6	41

Table 5.14: Measurements for Hibernate key classes.

Nr.	Classname	FAN_IN	FAN_OUT	FAN_TOTAL	LD_NUMBER
1	SessionFactoryImplementor	438	43	481	51
2	Type	444	5	449	0
3	Table	89	29	118	82
4	SessionImplementor	52	12	64	14
5	Criteria	45	12	57	15
6	Column	46	10	56	20
7	Session	31	21	52	52
8	Query	12	28	40	0
9	Configuration	1	38	39	115
10	SessionFactory	24	12	36	33
11	Criterion	30	3	33	0
12	Projection	11	3	14	0
13	FANectionProvider	12	2	14	0
14	Transaction	11	1	12	0

In tables 5.15, 5.16, and 5.17 we can find the top 10 entities with logical dependencies. The first column represents the name of each top 10 entity, the second column represents the fan_in values, the third column represents the fan_out values, the fourth column represents the fan_in and fan_out combined, and the fifth column represents the number of logical dependencies in which the entity is involved.

We did these top 10 tables to offer an overview of the highest registered numbers for LD for each system. As we mentioned before, some of the key classes are also present in these tables, but not all of them.

In table 5.17 we can find the top 10 measurements for Hibernate, most of the table is occupied by inner classes of AbstractEntityPersister. This is expected behavior since class AbstractEntityPersister is also present. This behavior is caused by the impossibility to separate the updates done for a class from its inner classes in the versioning system. So, each time AbstractEntityPersister records a change, also the inner classes are considered to have changed.

Table 5.15: Top 10 measurements for Ant.

Nr.	Classname	FAN_IN	FAN_OUT	FAN_TOTAL	LD_NUMBER
1	Project	191	23	214	157
2	Project$AntRefTable	1	2	3	157
3	Path	39	13	52	147
4	Path$PathElement	3	2	5	147
5	IntrospectionHelper	18	24	42	143
6	IntrospectionHelper$AttributeSetter	8	1	9	143
7	IntrospectionHelper$Creator	3	5	8	143
8	IntrospectionHelper$NestedCreator	7	1	8	143
9	Ant	2	15	17	136
10	Ant$Reference	3	1	4	136

Table 5.16: Top 10 measurements for Tomcat Catalina.

Nr.	Classname	FAN_IN	FAN_OUT	FAN_TOTAL	LD_NUMBER
1	StandardContext	11	38	49	216
2	StandardContext$ContextFilterMaps	0	0	0	216
3	StandardContext$NoPluggabilityServletContext	0	0	0	216
4	Request	48	28	76	215
5	Request$SpecialAttributeAdapter	0	0	0	215
6	ApplicationContext	3	22	25	158
7	ApplicationContext$DispatchData	0	0	0	158
8	ContextConfig	3	26	29	143
9	ContextConfig$DefaultWebXmlCacheEntry	0	0	0	143
10	ContextConfig$JavaClassCacheEntry	0	0	0	143

Table 5.17: Top 10 measurements for Hibernate.

Nr.	Classname	FAN_IN	FAN_OUT	FAN_TOTAL	LD_NR
1	AvailableSettings	1	0	1	205
2	AbstractEntityPersister	9	143	152	190
3	AbstractEntityPersister$CacheEntryHelper	0	0	0	190
4	AbstractEntityPersister$InclusionChecker	0	0	0	190
5	AbstractEntityPersister$NoopCacheEntryHelper	0	0	0	190
6	AbstractEntityPersister$ReferenceCacheEntryHelper	0	0	0	190
7	AbstractEntityPersister$StandardCacheEntryHelper	0	0	0	190
8	AbstractEntityPersister$StructuredCacheEntryHelper	0	0	0	190
9	Dialect	265	104	369	176
10	SessionFactoryImpl$SessionBuilderImpl	1	25	26	167

Overall, by looking at the comparisons between FAN_IN, FAN_OUT, FAN_TOTAL, and the logical dependencies in which a class is involved we could not determine a direct connection between them. Nither we can say that one influences the other.

We consider that even though the metrics are not related directly, they could be all used together to get a better view of the system connections.

6. LOGICAL DEPENDENCIES IN ARCHITECTURAL RECONSTRUCTION

We explore using code co-changes as input for software clustering for architectural reconstruction. Since structural dependencies are the most commonly used dependencies in software clustering, we investigate whether integrating them with code co-changes provides better results than using either dependency type alone.

Our experiments are applied to four open-source Java projects from GitHub. For each project, we apply three distinct clustering algorithms (Louvain, Leiden, and DBSCAN) and evaluate their performance using two clustering evaluation metrics. These metrics allow a comparison between clustering based solely on code co-changes and clustering that integrates both co-changes and structural dependencies, offering a better understanding of how these co-changes influence software architecture reconstruction.

6.1. Introduction

Software systems often need more documentation. Even if there was original documentation at the beginning of development, it may become outdated over the years. Additionally, the original developers may leave the company, taking with them knowledge about how the software was designed. This situation challenges the teams when it comes to maintenance or modernization. In this context, recovering the system's architecture is essential. Understanding the system's architecture helps developers evaluate better and understand the nature and impact of changes they must make. One technique to help in reconstructing the system architecture is software clustering. Software clustering involves creating cohesive groups (modules) of software entities based on their dependencies and interactions.

Among the dependencies that can be used for software clustering are structural dependencies (relationships between entities based on code analysis), lexical dependencies (relationships based on naming conventions), and code co-changes/logical dependencies (relationships between entities extracted from the version control system), and others.

This paper assesses the impact of logical dependencies in software clustering alone and combination with structural dependencies. The structural dependencies are used as they are extracted from static code analysis, while the logical dependencies are filtered co-changes obtained from the version control system [126]. The co-changes are filtered to enhance their reliability and remove noise caused by large commits with many files unrelated to development activities (e.g., formatting changes) or rare co-changes that may not indicate a true dependency [**?**].

The following research questions guide our investigation:

- **RQ1:** Does using structural dependencies (SD) combined with logical dependencies (LD) improve software clustering results compared to traditional approaches using only structural dependencies (SD)?
- **RQ2:** Can using only logical dependencies (LD) produce good software clustering results?
- **RQ3:** How do different filtering settings for logical dependencies (LD) impact clustering results, and which filtering settings provide the best performance?

To answer these research questions, we apply three different clustering algorithms (Louvain, Leiden, and DBSCAN) to different open-source projects. We then evaluate the results using two metrics: MQ (Modularization Quality) [21] and MoJoFM (Move and Join e**F**fectiveness **M**easure) [22]. The MoJoFM metric is used for external evaluation, evaluating against the perspective of the system's architect or developers. The MQ metric is used for internal evaluation based on the software structure itself. These two metrics allow us to compare the effectiveness of using structural and logical dependencies alone and combined. This comparison helps clarify how different dependencies and filtering choices affect clustering results.

6.2. Related work

Several studies have explored the use of different types of dependencies in software clustering, applying different algorithms to improve clustering results and using various metrics to evaluate the results obtained.

Tzerpos and Holt developed ACDC (Algorithm for Comprehension-Driven Clustering). This pattern-driven clustering algorithm uses subsystem structures such as source file patterns, directory patterns, system graph patterns, and support library patterns to detect similarities and create clusters [15]. For result evaluation, the authors introduced the MoJo metric, which counts the minimum number of move and join operations required to transform one clustering result into another, assessing how close one clustering solution is to another [127], [128]. Later, Wen and Tzerpos introduced the MoJoFM metric, an enhanced version of the original MoJo distance metric for more effective measurements, as presented in more detail in subsection 6.3.2 [22].

Corazza et al. [18], [19] used lexical dependencies derived from code comments, class names, attribute names, and parameter names, applying Hierarchical Agglomerative Clustering (HAC) to group-related entities. For evaluating the results, the authors used a metric based on the MoJo distance metric and NED (Non-Extremity Cluster Distribution), which measures that the formed clusters are not too large or too small.

Andritsos and Tzerpos [128] used structural dependencies and nonstructural attributes, such as file names and developer names, and proposed the LIMBO algorithm, a hierarchical clustering algorithm for clustering software systems. They used the MoJo distance metric to evaluate the algorithm's output.

Anquetil et al. [129] also used lexical information, including file names, routine names, included files, and comments. They applied an n-gram-based clustering approach to detect semantic similarities between entities and evaluated the results using precision and recall metrics.

Maletic and Marcus [70] propose an approach to software clustering that uses semantic dependencies extracted using Latent Semantic Indexing (LSI), a technique for identifying similarities between software components. They apply the minimal spanning tree (MST) algorithm for clustering and evaluate the results using metrics based on both semantic and structural information.

Wu et al. [130] conducted a comparative study of six clustering algorithms using structural dependencies on five software systems. Four of the algorithms are based on agglomerative clustering, one on program comprehension patterns, and one algorithm is a customized version of Bunch [21]. The performance of these algorithms was evaluated using the MoJo metric and NED (Non-Extreme Distribution).

Mancoridis and Mitchell [21], [131], [16] developed the Bunch tool for software clustering and used structural dependencies as input. The tool applies clustering algorithms to the structural dependency graph and outputs the system's organization. For evaluation, the authors introduced the Modularization Quality (MQ) metric, described in more detail in Section 6.3.2, and is also used in our current experiments as an evaluation metric.

Prajapati et al. [20] propose a many-objective SBSR (search-based software remodularization) approach with an improved definition of objective functions based on lexical, structural, and change-history dependencies. The authors evaluate their approach on several open-source software systems using the MoJoFM metric for external evaluation and the MQ metric for internal evaluation.

Şora et al. [61], [60] developed the ARTs (Architecture Reconstruction Tool Suite) for their experiments on improving software architecture reconstruction through clustering. The tool suite implements various clustering algorithms, such as minimum spanning tree-based, metric-based, search-based, and hierarchical clustering, primarily using structural dependencies as input. The research focuses on identifying the right factors for direct coupling between classes, indirect coupling, and layered architecture. The results of applying these different factors are evaluated using the MoJo distance metric.

Silva et al. [132] investigated using solely co-change dependencies as input for the Chameleon algorithm, an agglomerative hierarchical clustering method, to identify clusters. For evaluation, the authors used distribution maps to compare the clusters generated from co-change dependencies with the system's package structure.

6.3. Methodology and implementation

In this section, we present the methodology used to evaluate the impact of logical dependencies on the quality of software clustering solutions.

First, we describe the clustering algorithms used in our experiments: Louvain, Leiden, and DBSCAN. Next, we introduce the evaluation metrics used to assess the quality of the clustering results. Finally, we present the workflow and implementation of the tool developed for this research, which is built to process structural and logical dependencies, apply the selected clustering algorithms, and compute the evaluation metrics.

6.3.1. Clustering algorithms

Louvain

The Louvain algorithm was originally developed by Blondel et al. and is used to find community partitions (clusters) in large networks. The algorithm begins with a weighted network of N nodes, initially assigning each node to its own cluster, resulting in N clusters. For each node, the algorithm evaluates the modularity gained from moving the node to the cluster of each of its neighbors. Based on the results, the node is moved to the cluster with the maximum positive modularity gain. This process is repeated for all nodes until no further improvement in modularity is possible [133], [134].

Leiden

The Leiden algorithm, developed by Traag et al., is an improvement over the Louvain algorithm for community detection in large networks. Like Louvain, the Leiden algorithm begins with each node assigned to its own cluster and iteratively moves nodes between clusters to optimize modularity. However, the Leiden algorithm addresses some problems of the Louvain method, particularly regarding poorly connected communities and runtime performance issues [135] [136].

The Leiden algorithm introduces a refinement phase that ensures communities are locally optimally clustered and well-connected. This refinement step distinguishes the Leiden algorithm from Louvain.

DBSCAN

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, introduced by Ester et al., is a density-based clustering algorithm for identifying clusters of arbitrary shape and detecting noise in data [137], [136].

DBSCAN operates based on two main parameters:

- **Eps**: It defines the radius within which to search for neighboring points.
- **MinPts**: The minimum number of points required for a dense region. It determines the minimum number of neighbors a point should have to be considered a core point.

The algorithm classifies points into three categories:

1. **Core Points**: Points that have at least *MinPts* neighbors within a radius of *Eps*. These points are located in the interior of a cluster.
2. **Border Points**: Points that have fewer than *MinPts* neighbors within a radius of *Eps* but are in the Eps-neighborhood of a core point. They are located on the edge of a cluster.

3. **Noise**: Points that are neither core points nor border points.

The DBSCAN algorithm starts by visiting an arbitrary point in the dataset. If the point is a core point, the algorithm starts a new cluster and retrieves all reachable points from this core point. All points are then marked as part of the cluster. If the point is a border point, it moves to the next point in the dataset. This process is repeated until all points have been visited.

DBSCAN can be applied for software clustering by considering software entities as data points. A distance measure based on dependency weights can be used to compute the neighborhood between entities.

6.3.2. Clustering result evaluation

We evaluate the clustering results using two metrics: the Modularity Quality (MQ) metric and the Move and Join Effectiveness Measure (MoJoFM) metric. Each provides a different perspective on the quality of the clustering solutions.

Modularity Quality metric

Mancoridis et al. introduced the Modularity Quality (MQ) metric to evaluate the modularization quality of a clustering solution based on the interaction between modules (clusters) [**?**],[21]. It evaluates the difference between connections within clusters and connections between different clusters.

The MQ of a graph partitioned into k clusters, where A_i is the Intra-Connectivity of the i-th cluster and E_{ij} is the Inter-Connectivity between the i-th and j-th clusters, is calculated using Equation (6.1) [131].

$$MQ = \left(\frac{1}{k}\sum_{i=1}^{k} A_i\right) - \left(\frac{1}{k(k-1)}\sum_{i,j=1}^{k} E_{ij}\right) \tag{6.1}$$

The MQ metric's value ranges between -1 and 1. A value of -1 means that the clusters have more connections between the clusters than within the clusters, while a value of 1 means that there are more connections within clusters than between clusters. A good clustering solution should have an MQ value close to 1, since this indicates that the clusters are more cohesive internally and have fewer connections to other clusters.

The MQ metric is useful because it does not require additional input besides the clustering result. It relies on the structure of the clustered entities and their interactions.

MoJoFM metric

Wen and Tzerpos introduced the MoJoFM metric to evaluate the similarity between two different software clustering results [22]. The metric is based on the MoJo metric, which measures the absolute minimum number of *Move* and *Join* operations required to transform one clustering solution into another [127], [22]. However, MoJoFM provides a similarity measure ranging between 0% and 100%, where 100% indicates identical clustering solutions.

The MoJoFM metric is calculated using Equation (6.2):

$$\text{MoJoFM}(A, B) = \left(1 - \frac{\text{mno}(A, B)}{\text{max}(\text{mno}(\forall A, B))}\right) \times 100\% \tag{6.2}$$

Where:

- $\text{mno}(A, B)$ is the minimum number of *Move* and *Join* operations required to transform clustering solution A into clustering solution B.
- $\text{max}(\text{mno}(\forall A, B))$ is the maximum possible number of such operations required to transform any clustering A into clustering B.

To use the metric, we first need to generate a reference clustering solution for comparison. We manually created this reference based on our analysis of the codebase.

Using the MoJoFM metric, we can evaluate the similarity between the generated and reference clustering solutions. This metric is useful when combining multiple dependencies because it measures the similarity between the obtained clustering solutions and the same reference.

6.3.3. Tool workflow for software clustering and evaluation

To evaluate how logical dependencies impact the quality of clustering solutions, we developed a Python tool capable of using any type of dependency, either alone or combined with other types of dependencies, as long as it is provided in CSV format. The tool clusters and evaluates software clustering solutions using the MQ or MoJoFM metrics.

Input

The tool takes one or multiple dependency CSV files as input and the reference solution required for the MoJoFM metric. We designed the tool to accept multiple dependency files so that we can generate clustering solutions based on either a single type of dependency (structural or logical) or a combination of both.

Since the MoJoFM metric requires a reference solution to evaluate the obtained clustering solutions, we manually inspected the code and created reference clustering solutions, which we then provided as input for the tool.

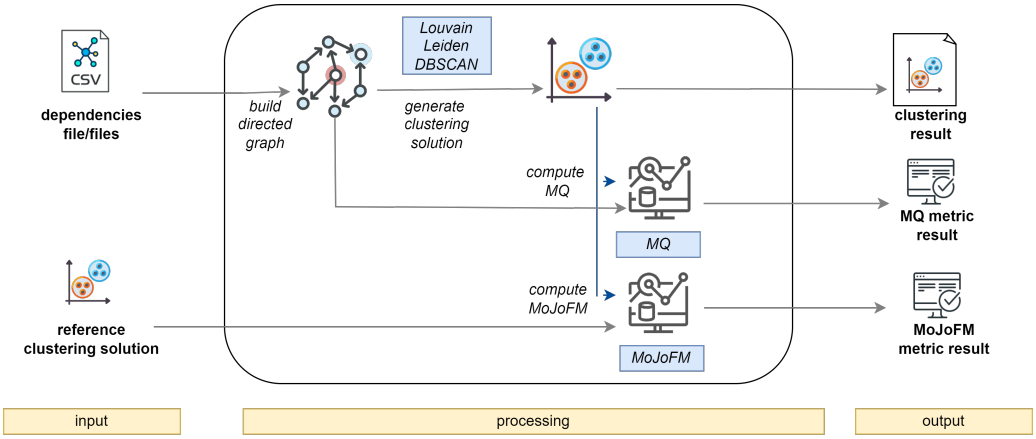

Figure 6.1: Tool workflow overview: input, processing and output.

Processing

The dependencies are saved in the CSV file in the following format: `antecedent of a dependency, consequent of a dependency, weight`. The tool reads each line, adds the antecedent and consequent as nodes in a directed graph, and creates an edge between them, with the weight from the CSV file becoming the edge weight. The edge weights are summed if multiple dependency files are processed and the same dependency is found in multiple files.

The workflow of applying the clustering algorithms and performing the evaluations is shown in Figure 6.1. After all dependencies are read, the directed graph is passed to the clustering algorithms: Louvain, Leiden, and DBSCAN. Each algorithm generates its own clustering result. The results from each algorithm are then evaluated using the MQ metric and the MoJoFM metric. The MQ metric requires the directed graph and the clustering result, while the MoJoFM metric requires the reference clustering solution provided as input and the clustering result.

Output

After applying each clustering algorithm and completing both evaluations, we export the clustering result, the number of clusters from the clustering solution, and the MQ and MoJoFM metrics values.

6.4. Data set used in experimental analysis

In Table 6.1, we have synthesized all the information about the four projects used in our experiments. The 'Project Name' column contains the names of the software projects sourced from GitHub. The 'Release Tag' column contains the specific release

Table 6.1: Overview of projects used in experimental analysis

Project Name	Release Tag	Commits	GitHub Repository Link	Repository Description
Apache Ant	1.10.13	14,917	https://github.com/apache/ant	Java build tool for automating software tasks.
Apache Tomcat	8.5.93	22,698	https://github.com/apache/tomcat	Java web server and servlet container.
Hibernate ORM	6.2.14	16,609	https://github.com/hibernate/hibernate-orm	Java ORM framework for database management.
Gson	gson-parent-2.10.1	1,772	https://github.com/google/gson	Java library for JSON serialization and deserialization.

tag of the project that was analyzed. We processed all the commits for logical dependency extraction, from the first commit to the commit associated with the specified tag. We extracted the dependencies from the code of that specific tag for structural dependencies. The 'Number of Commits' column provides the total number of commits used for logical dependencies extraction. The 'GitHub Repository Link' column includes the URL link to the project's repository on GitHub. Finally, the 'Repository Description' column briefly describes the project's purpose and functionality.

We mainly chose projects with more than 10,000 commits in their commit history so that the logical dependencies extraction can be done on a more extensive information base. However, we selected Gson, which has a relatively small commit history (1,772 commits), to determine if our experiments work with a smaller information base.

Table 6.2 presents the commit statistics for the studied projects. The columns represent the percentage of commits with under 5 files modified, between 5 and 10 files, between 10 and 20 files, and above 20 files modified. We can observe that most commits have under 5 files changed, with Apache Tomcat having more than 90% of the commits with less than 5 files changed. On the opposite side, only a few commits involve more than 20 files changed, Hibernate ORM having the highest percentage at 8.39%.

Table 6.2: Commit statistics for studied projects

Project Name	Number of files changed			
	Under 5	5-10	10-20	Above 20
Apache Ant	83.83%	7.50%	4.17%	4.50%
Apache Tomcat	90.95%	5.44%	2.04%	1.58%
Hibernate ORM	71.74%	12.37%	7.50%	8.39%
Gson	83.63%	9.85%	3.70%	2.81%

6.5. Experimental plan and results

6.5.1. Experimental plan

Tool runs

To assess the impact of logical dependencies and to answer the research questions from section 6.1, we run the tool presented in Section 6.3.3 in three different scenarios for all the projects from table 6.1. All three scenarios are illustrated in Fig. 6.2.

In the first scenario, we run the tool once, providing only the system's structural dependencies as input for the clustering algorithm.

In the second scenario, we run the tool ten times, using only logical dependencies as input. We perform ten runs because we generate logical dependencies with different threshold values for the strength filter. We start with a threshold of 10 and increase it in steps of 10 up to 100, where 100 is the maximum value for the threshold.

In the third scenario, we combine logical with structural dependencies. Similar to the second scenario, we ran the tool ten times using structural and logical dependencies generated with different strength thresholds.

6.5.2. Results

The experimental results are presented in this subsection in four tables, each corresponding to a different project. Table 6.3 presents the results for Apache Ant, Table 6.4 presents the results for Apache Tomcat, Table 6.5 presents the results for Hibernate ORM, and Table 6.6 presents the results for Gson.

Each table includes the following columns:

- **Dependency Type:** The types of dependencies used are as follows: SD for Structural Dependencies, LD for Logical Dependencies, and SD+LD for their combination. The strength threshold used is specified in parentheses right after LD.
- **Entities Count:** The total number of software entities (such as classes, interfaces, enums) involved in clustering.
- **System Coverage:** Considering that the total number of entities extracted from the codebase — which represents the entities forming structural dependencies (SD) — constitutes the entire set of entities in the system (the first line of each table), we calculated the percentage of entities present in the filtered logical dependencies (LD) relative to the total number of known codebase entities.
- **Louvain/ Leiden/ DBSCAN:** The clustering algorithms used in the experiments.
- **Nr. of Clusters:** The number of clusters from the clustering solution.
- **MQ (Modularization Quality):** The result obtained when applying the MQ evaluation metric to the clustering solution.
- **MoJoFM:** The result obtained when applying the MoJoFM evaluation metric to

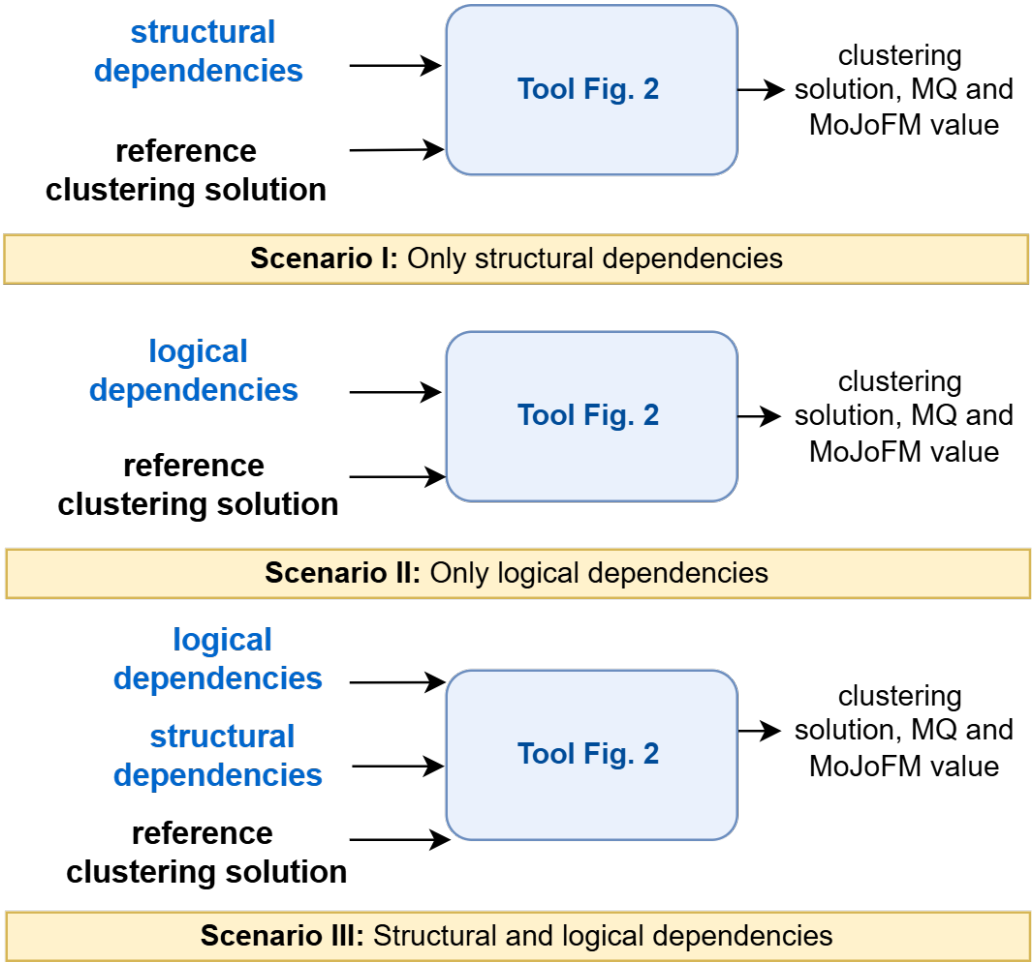

Figure 6.2: Experimental scenarios for analyzing the impact of logical dependencies on clustering quality

the clustering solution.

The rows in each table represent different dependency types and strength filter thresholds used in the clustering experiments.

To better understand the impact of different dependency types on software clustering, we also analyzed the average weights assigned to structural dependencies (SD) and logical dependencies (LD) across the studied projects. Table 6.7 presents these average dependency weights. The first row shows the average weights for SD, which remain constant across all strength thresholds, while the other rows show the average weights for logical dependencies at different strength thresholds.

Table 6.3: Clustering results based on different dependency types and strength filter thresholds for repository: https://github.com/apache/ant

Dependency Type (strength threshold)	Entities Count	System Coverage (%)	Louvain			Leiden			DBSCAN		
			Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM
SD	517	100.00	14	0.114	46.02	14	0.101	52.99	34	0.144	25.1
LD (10)	320	61.89	55	0.506	65.57	55	0.506	65.57	30	0.435	39.02
LD (20)	215	41.58	53	0.547	68	53	0.547	68	23	0.505	53.5
LD (30)	174	33.65	44	0.558	71.7	44	0.558	71.7	19	0.585	50
LD (40)	152	29.40	40	0.580	71.53	40	0.580	71.53	19	0.602	53.06
LD (50)	138	26.69	35	0.604	73.98	35	0.604	73.98	17	0.633	56.1
LD (60)	120	23.21	34	0.587	70.48	34	0.587	70.48	14	0.650	51.43
LD (70)	106	20.50	32	0.577	71.43	32	0.577	71.43	11	0.661	51.65
LD (80)	92	17.79	29	0.576	70.13	29	0.576	70.13	9	0.709	50.65
LD (90)	79	15.28	24	0.606	71.88	24	0.606	71.88	8	0.705	56.6
LD (100)	64	12.37	19	0.611	75.51	19	0.611	75.51	6	0.691	56.93
SD+LD (10)	517	100.00	18	0.355	55.18	15	0.254	54.98	37	0.147	25.9
SD+LD (20)	517	100.00	17	0.318	52.39	19	0.365	53.78	32	0.149	26.49
SD+LD (30)	517	100.00	16	0.282	53.19	16	0.265	54.78	30	0.159	24.5
SD+LD (40)	517	100.00	17	0.340	51.99	17	0.317	53.19	31	0.146	24.7
SD+LD (50)	517	100.00	15	0.248	52.59	19	0.298	56.77	31	0.146	24.7
SD+LD (60)	517	100.00	16	0.244	50.8	16	0.271	54.38	32	0.155	25.1
SD+LD (70)	517	100.00	15	0.238	51.00	18	0.281	52.99	32	0.155	25.1
SD+LD (80)	517	100.00	13	0.246	45.22	15	0.255	45.82	32	0.155	25.1
SD+LD (90)	517	100.00	14	0.258	46.02	16	0.268	47.01	32	0.155	25.1
SD+LD (100)	517	100.00	15	0.214	50.8	15	0.227	50.4	32	0.155	25.1

Table 6.4: Clustering results based on different dependency types and strength filter thresholds for repository: https://github.com/apache/tomcat

Dependency Type (strength threshold)	Entities Count	System Coverage (%)	Louvain			Leiden			DBSCAN		
			Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM
SD	662	100.00	26	0.186	77.76	24	0.184	76.99	43	0.142	73.31
LD (10)	406	61.33	42	0.505	72.47	42	0.505	72.47	60	0.393	67.93
LD (20)	303	45.77	45	0.538	68.26	45	0.538	67.24	41	0.510	72.7
LD (30)	249	37.61	46	0.532	69.87	46	0.532	69.87	32	0.561	80.33
LD (40)	208	31.42	42	0.590	69.70	42	0.591	70.71	28	0.572	83.84
LD (50)	198	29.91	44	0.604	70.21	44	0.604	70.21	22	0.631	85.11
LD (60)	177	26.74	45	0.601	70.66	45	0.601	70.66	18	0.662	85.63
LD (70)	164	24.77	45	0.598	75.32	45	0.598	75.32	17	0.676	88.96
LD (80)	127	19.18	36	0.618	79.49	36	0.618	79.49	15	0.713	89.74
LD (90)	116	17.52	32	0.623	81.13	32	0.623	81.13	14	0.718	89.62
LD (100)	110	16.62	30	0.640	85.00	30	0.640	85.00	13	0.735	89.00
SD+LD(10)	662	100.00	28	0.324	78.99	28	0.324	78.99	40	0.161	74.23
SD+LD(20)	662	100.00	31	0.287	78.22	30	0.320	80.06	50	0.189	73.31
SD+LD(30)	662	100.00	32	0.296	79.92	32	0.277	75.77	45	0.209	73.47
SD+LD(40)	662	100.00	34	0.292	79.91	32	0.326	78.22	43	0.198	73.47
SD+LD(50)	662	100.00	33	0.294	76.53	35	0.301	76.23	43	0.196	73.31
SD+LD(60)	662	100.00	35	0.304	77.15	33	0.286	76.84	41	0.177	73.62
SD+LD(70)	662	100.00	34	0.282	76.69	34	0.292	77.45	41	0.166	73.62
SD+LD(80)	662	100.00	34	0.283	76.23	33	0.282	76.38	42	0.153	73.47
SD+LD(90)	662	100.00	31	0.311	78.99	31	0.311	78.99	43	0.153	73.31
SD+LD(100)	662	100.00	31	0.311	78.83	31	0.305	78.37	43	0.153	73.31

Table 6.5: Clustering results based on different dependency types and strength filter thresholds for repository: https://github.com/hibernate/hibernate-orm

https://github.com/hibernate/hibernate-orm

Dependency Type (strength threshold)	Entities Count	System Coverage (%)	Louvain			Leiden			DBSCAN		
			Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM
SD	4414	100.00	30	0.09	52.23	23	0.071	52.44	373	0.128	46.32
LD (10)	1450	32.85	44	0.389	57.22	45	0.39	58.22	99	0.395	57.08
LD (20)	1325	30.02	66	0.397	62.66	66	0.397	62.66	151	0.378	63.36
LD (30)	1222	27.68	66	0.38	62.45	67	0.38	63.04	148	0.378	65.42
LD (40)	915	20.73	84	0.417	63.68	85	0.412	63.56	110	0.382	66.9
LD (50)	900	20.39	84	0.409	64.56	84	0.409	64.56	105	0.386	67.02
LD (60)	848	19.21	82	0.406	63.26	81	0.41	63.39	104	0.379	65.13
LD (70)	459	10.40	89	0.516	69.08	89	0.516	69.08	41	0.467	58.21
LD (80)	450	10.19	91	0.506	68.64	91	0.506	68.64	39	0.479	60.49
LD (90)	432	9.79	92	0.492	66.93	92	0.492	66.93	40	0.473	58.66
LD (100)	356	8.07	81	0.524	65.92	81	0.524	65.92	29	0.537	58.2
SD+LD (10)	4414	100.00	19	0.096	53.93	19	0.099	52.28	282	0.121	46.01
SD+LD (20)	4414	100.00	21	0.126	52.85	23	0.122	56.21	309	0.135	47.4
SD+LD (30)	4414	100.00	26	0.121	55.76	26	0.15	54.54	317	0.135	49.45
SD+LD (40)	4414	100.00	27	0.182	54.57	28	0.163	55.89	350	0.134	49.35
SD+LD (50)	4414	100.00	26	0.16	52.37	24	0.147	53.31	350	0.134	49.37
SD+LD (60)	4414	100.00	26	0.161	52.35	27	0.153	53.19	352	0.135	49.31
SD+LD (70)	4414	100.00	28	0.139	52.78	29	0.154	54.34	366	0.13	47.13
SD+LD (80)	4414	100.00	28	0.142	52.83	28	0.147	53.35	366	0.13	47.72
SD+LD (90)	4414	100.00	28	0.136	52.62	30	0.153	53.83	365	0.13	47.72
SD+LD (100)	4414	100.00	30	0.128	52.78	28	0.114	55.23	365	0.128	47.75

Table 6.6: Clustering results based on different dependency types and strength filter thresholds for repository: https://github.com/google/gson

https://github.com/google/gson

Dependency Type (strength threshold)	Entities Count	System Cover (%)	Louvain			Leiden			DBSCAN		
			Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM
gson SD	210	100.00	10	0.139	53.47	9	0.129	55.94	23	0.127	51.88
gson LD (10)	66	31.43	10	0.565	62.07	9	0.572	60.34	19	0.399	68.97
gson LD (20)	50	23.81	11	0.547	64.29	11	0.547	64.29	9	0.523	59.52
gson LD (30)	41	19.52	12	0.544	63.64	12	0.544	63.64	6	0.606	66.67
gson LD (40)	31	14.76	8	0.635	69.57	8	0.635	69.57	6	0.612	69.57
gson LD (50)	31	14.76	8	0.600	69.57	8	0.600	69.57	6	0.565	60.87
gson LD (60)	28	13.33	8	0.552	65.00	8	0.552	65.00	5	0.584	60.00
gson LD (70)	26	12.38	7	0.579	66.67	7	0.579	66.67	5	0.586	55.56
gson LD (80)	18	8.57	5	0.590	60.00	5	0.590	60.00	4	0.544	40.00
gson LD (90)	18	8.57	5	0.590	60.00	5	0.590	60.00	4	0.544	40.00
gson LD (100)	18	8.57	5	0.590	60.00	5	0.590	60.00	4	0.544	40.00
gson SD+LD(10)	210	100.00	11	0.317	64.36	11	0.317	64.36	20	0.172	63.86
gson SD+LD(20)	210	100.00	11	0.259	61.39	11	0.259	61.39	17	0.136	53.96
gson SD+LD(30)	210	100.00	11	0.277	61.39	11	0.277	61.39	20	0.136	55.94
gson SD+LD(40)	210	100.00	10	0.277	61.39	10	0.277	61.39	20	0.135	55.94
gson SD+LD(50)	210	100.00	10	0.270	60.40	11	0.270	60.89	20	0.135	55.94
gson SD+LD(60)	210	100.00	9	0.296	61.39	10	0.290	61.88	20	0.135	55.94
gson SD+LD(70)	210	100.00	8	0.295	59.41	8	0.295	59.41	20	0.135	55.94
gson SD+LD(80)	210	100.00	7	0.267	58.91	8	0.263	59.41	21	0.134	55.45
gson SD+LD(90)	210	100.00	7	0.267	58.91	7	0.267	58.91	21	0.134	55.45
gson SD+LD(100)	210	100.00	7	0.267	58.91	8	0.263	59.41	21	0.134	55.45

Dependency type	Ant	Tomcat	Hibernate	Gson
SD	5.91	6.91	5.41	5.24
LD(10)	11.17	12.82	2.45	14.15
LD(20)	16.01	19.65	3.00	19.10
LD(30)	18.08	23.56	3.27	27.58
LD(40)	19.08	25.57	4.63	29.85
LD(50)	19.94	26.31	4.80	29.97
LD(60)	24.26	28.91	5.14	33.93
LD(70)	26.70	30.35	9.53	34.37
LD(80)	30.83	35.33	10.18	43.00
LD(90)	32.11	36.90	10.47	43.00
LD(100)	33.93	37.04	12.00	43.00

Table 6.7: Average weights of Structural Dependencies (SD) and Logical Dependencies (LD).

6.6. Evaluation

The overall analysis of all the results from subsection 6.5.2 indicates that combining structural and logical dependencies (SD+LD) provides better clustering solutions than using structural dependencies (SD) alone, covering 100% of the system, meaning that no entity is missed during cluster generation. On the other hand, logical dependencies (LD) alone result in better clustering quality metrics compared to both SD and SD+LD, but they do not cover the entire system.

The best results for SD+LD are observed with a strength threshold between 10-40%. For LD only, the best results are obtained at a 100% strength threshold. The overall trend shows that for LD only, the MQ metric increases in value with a higher strength threshold, indicating more cohesive clusters, while the MoJo metric decreases, indicating that fewer transformations are needed to reach the expected clustering. For SD+LD, the best MQ and MoJo values are obtained at lower strength thresholds, and then both metrics indicate a less effective clustering solution obtained with higher strength thresholds.

We analyze each project in detail in the sections below and address the research questions.

6.6.1. Detailed evaluation

Apache Ant

The clustering results for Apache Ant (Table 6.3) show that the combined structural and logical dependencies (SD+LD) achieved the best values with a strength threshold between 10% and 30%. The highest value for the MQ metric is reached with Leiden at a strength threshold of 20%, and the highest value for MoJoFM is also reached with Leiden at a strength threshold of 10%.

Compared with the SD-only results, all SD+LD clustering solutions for all algorithms show better MQ metric values, with the highest MQ value for SD+LD being more than three times greater than the corresponding SD-only value. Similarly, the MoJoFM metric shows better results than SD-only. However, it does not always outperform the MoJoFM metric applied to SD-only data.

Logical dependencies (LD) alone produced the highest MQ and MoJoFM values at the 100% strength threshold for both Leiden and Louvain, with the obtained metric values being higher than those of SD-only and SD+LD. However, the percentage of entities covered is significantly lower (LD(100) covers only 12.37% of the system). If we look at LD(10), where there is a 61.89% coverage of the system, which is more compared to LD(100), both metrics still perform better than SD-only and SD+LD(10). However, there is still a gap until 100% coverage.

From the clustering algorithm performance point of view, Leiden obtains the best evaluation metrics for all scenarios, followed by Louvain and DBSCAN.

One interesting observation is that, based on the LD-only results, where the metrics results improved with higher strength thresholds, the SD+LD results did not

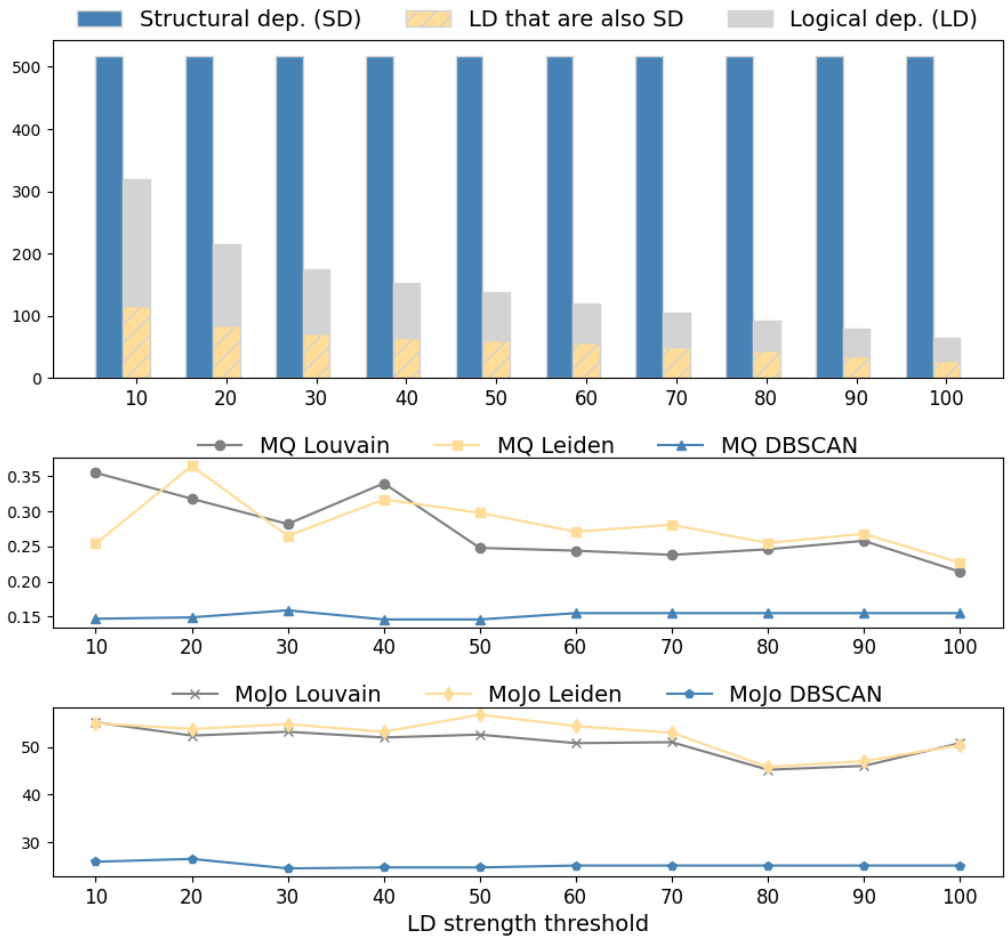

Figure 6.3: Apache Ant: Overlap between structural and logical dependencies and its correlation with clustering metrics.

follow the same pattern. On the opposite, the SD+LD metric results decline with a higher strength threshold. An explanation for this behavior may lie in the overlap between structural and logical dependencies. As presented in Figure 6.3, the number of LD decreases with a stricter strength threshold compared to the number of SD, and the overlap between the two types of dependencies increases.

In our previous works, we studied how these two types of dependencies overlap [63], [110]. The reason behind those studies was to check how much new information we can get from using logical dependencies and how much is already present via structural dependencies.

Our overall findings were that with stricter filtering of logical dependencies, we obtain a higher percentage of overlap between the two dependencies, reaching at

most 50% of logical dependencies that are also structural dependencies.

So, we consider that the reason why SD+LD clustering solutions decline in performance with a higher strength threshold is that less and less new additional information is added to the system (logical dependencies that are not structural dependencies), causing the clustering solution to start resembling the performance of the SD-only solution. In Figure 6.3, we can see that LD(10) represents 61% of the quantity of SD, while LD(100) is only at 12%, with half of them being duplicated with SD.

Apache Tomcat

For Apache Tomcat (Table 6.4), the best results for SD+LD were obtained with strength thresholds between 10% and 40% across all algorithms. The Leiden algorithm achieved the best result for the MQ metric at a strength threshold of 40%, while the best MoJoFM result was obtained at a threshold of 20%, also with the Leiden algorithm. Compared with the SD-only results, the peak MQ values almost double the SD-only values. Like Apache Ant, the MQ values for all strength thresholds are higher than those for SD-only. While MoJoFM is not better for all thresholds, it still improves compared with the SD-only results.

The LD-only results show the highest MQ and MoJoFM values at LD(100) for the Louvain and Leiden algorithms. However, as with the Apache Ant results, coverage remains an issue. LD(100) covers only 16.62% of the system, lower than the coverage from SD-only or SD+LD combinations. On the other hand, LD(10), which covers 61.33% of the system, still has better clustering solutions compared to SD-only, based on both MQ and MoJoFM results.

We observe the same decline in results with a stricter strength threshold for SD+LD. As with the previous system, these results can again be connected to the percentage of LD that also overlaps with SD and the decreasing number of LD compared to SD once the strength threshold becomes stricter. As shown in Figure 6.4, LD filtered with a 10% strength threshold overlaps with SD by approximately 22%, while at a 100% strength threshold, the overlap increases to approximately 39%.

To ensure that the decline in performance for SD+LD with a stricter strength threshold is indeed caused by the fact that LD are significantly fewer than SD, and SD duplicates that part of them at higher thresholds, we added an additional experiment to our study. In this experiment, whose results can be found in Table 6.8, we increased the weights associated with LD(100) for Apache Tomcat to confirm that we are dealing with an LD quantity problem rather than a weight problem.

Therefore, in this experiment, we increased the weight assigned to each logical dependency filtered with a 100% strength threshold from the Apache Tomcat system by values ranging from 1 to 5 and re-ran scenario III from Fig. 6.2.

To maintain consistency, we used the same columns as in the other result tables (6.3, 6.4, 6.5, 6.6), with the addition of two new columns:

- **Multiplication Factor:** The value by which each logical dependency weight is

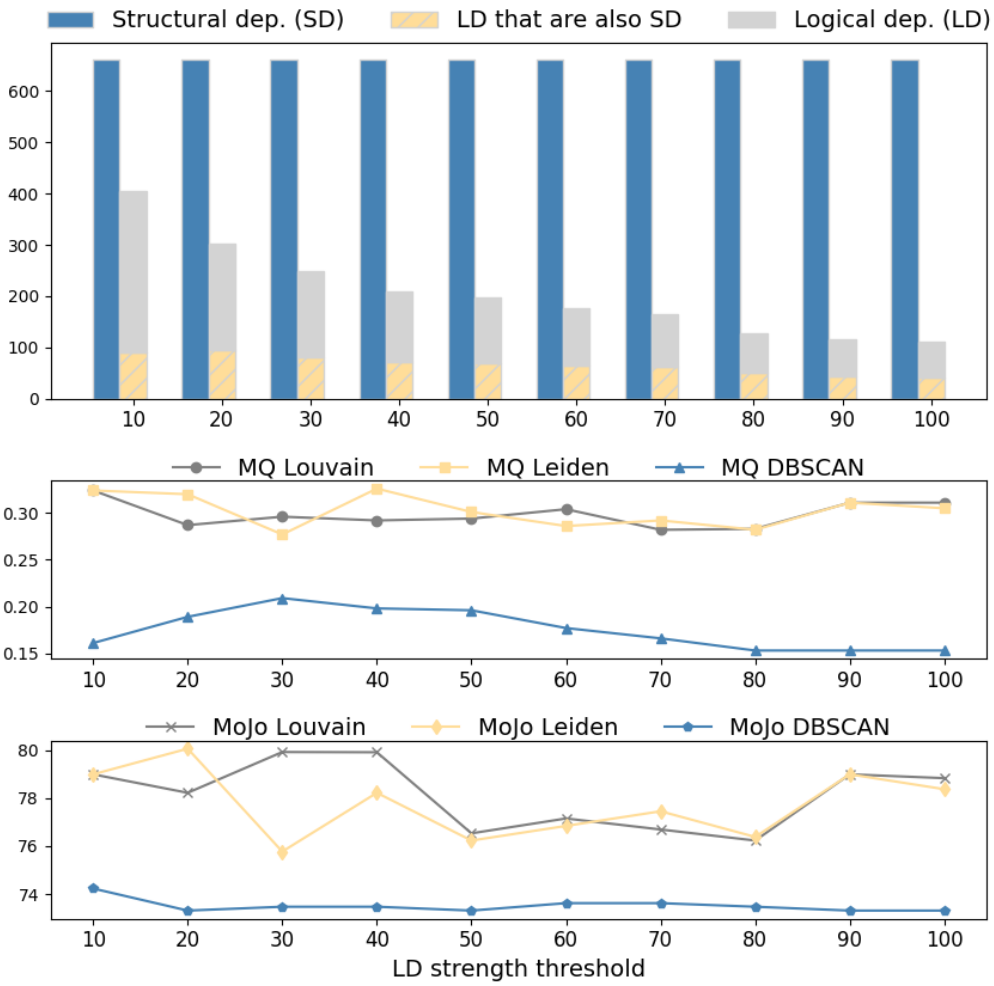

Figure 6.4: Apache Tomcat: Overlap between structural and logical dependencies and its correlation with clustering metrics.

multiplied.

- **Avg Weight:** The average weight assigned to each type of dependency used.

In Table 6.7, which presents the average weights associated with the dependencies across all systems, we can see that for Tomcat, the average weight for LD(10) is already almost double the SD average weight. For LD(100), the average weight is approximately five times higher than that of SD.

Based on the metric values obtained for multiplication factors of 2 to 5, we can see that after increasing the weights assigned to LD, the metric values improve only slightly, with changes recorded at the second decimal: a 0.02 improvement for Louvain and 0.07 for Leiden. The results for DBSCAN remain unchanged due to the

Table 6.8: Impact of multiplication factors on clustering results for LD(100) in Apache Tomcat

Multiplication Factor	Avg. weight		Louvain			Leiden			DBSCAN		
	SD	LD	Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM	Nr. of clusters	MQ	MojoFM
1	6.91	37.04	31	0.311	78.83	31	0.305	78.37	43	0.153	73.31
2	6.91	74.08	33	0.295	73.57	30	0.301	72.33	43	0.153	73.31
3	6.91	111.12	34	0.313	74.19	33	0.309	72.80	43	0.153	73.31
4	6.91	148.16	34	0.312	73.88	33	0.312	72.49	43	0.153	73.31
5	6.91	185.20	34	0.306	73.88	33	0.308	72.18	43	0.153	73.31

fixed values of MinPts and Eps and the already high LD weights for LD(100).

We can conclude from the experiment with weights that the issue is the quantity of dependencies. SD outnumbers LD, making LD information less impactful on the clustering solution.

Hibernate ORM

Hibernate ORM is the second largest system after Apache Tomcat regarding the number of commits analyzed, with 16,609 commits considered for LD extraction. Additionally, it is the largest in terms of system size, with 4,414 entities (Table 6.1).

Based on the results from Table 6.5, the SD+LD combination with a strength threshold of 40 performs best for this system. Louvain achieves the best MQ metric at this threshold, while Leiden achieves the best MoJoFM metric.

LD-only produced the best MQ values at 100% strength and the best MoJoFM values at 70% strength for both Louvain and Leiden. Compared to the previous systems, where both best values were recorded at the same strength threshold, Hibernate shows an earlier peak for MoJoFM. The system coverage is likely a factor contributing to this. Hibernate LD[100] covers only 8.07% of the system, the lowest percentage among all systems studied. This low percentage can be linked to the number of commits compared to the number of entities. For Apache Tomcat, there were 22,698 commits and 662 entities, while for Hibernate, there were 16,609 commits and 4,414 entities. This indicates that not all entities had a chance to be updated in the version control system.

This observation is also reflected in Table 6.7, where Hibernate has the lowest average weights for LD compared to SD across all systems. In other systems, LD(10) starts with almost double the average weight compared to SD, while Hibernate's LD(10) average weight is less than half of the SD average weight.

Hibernate has the lowest overlap percentage between LD and SD, as shown in Figure 6.5. Similar to the other systems, the performance for MQ and MoJoFM decreases for SD+LD as the strength threshold becomes stricter.

The results for Hibernate highlight the challenge of achieving better clustering in larger systems with fewer commits relative to their size.

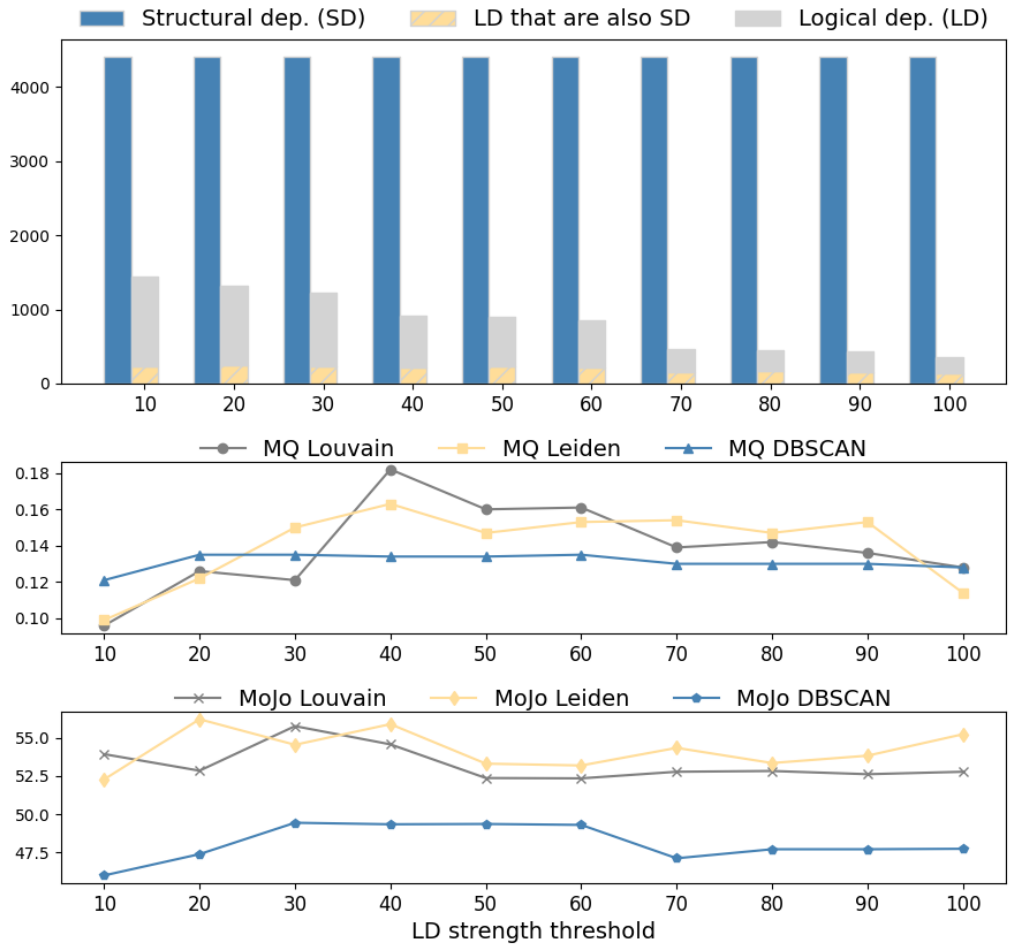

Figure 6.5: Hibernate ORM: Overlap between structural and logical dependencies and its correlation with clustering metrics.

Google Gson

Gson has the smallest number of commits analyzed, with 1,772 commits considered for LD extraction, and it is also the smallest in terms of system size, with 210 entities involved in clustering (Table 6.1).

Based on the results from Table 6.6, the SD+LD combination with a strength threshold of 10 achieved the best results for both MQ and MoJoFM. Like Apache Ant and Tomcat, all SD+LD combinations achieve better MQ values than SD-only.

LD-only produced the best results for MQ at 40% strength for both Louvain and Leiden, and the best MoJoFM value was also observed at the same threshold for

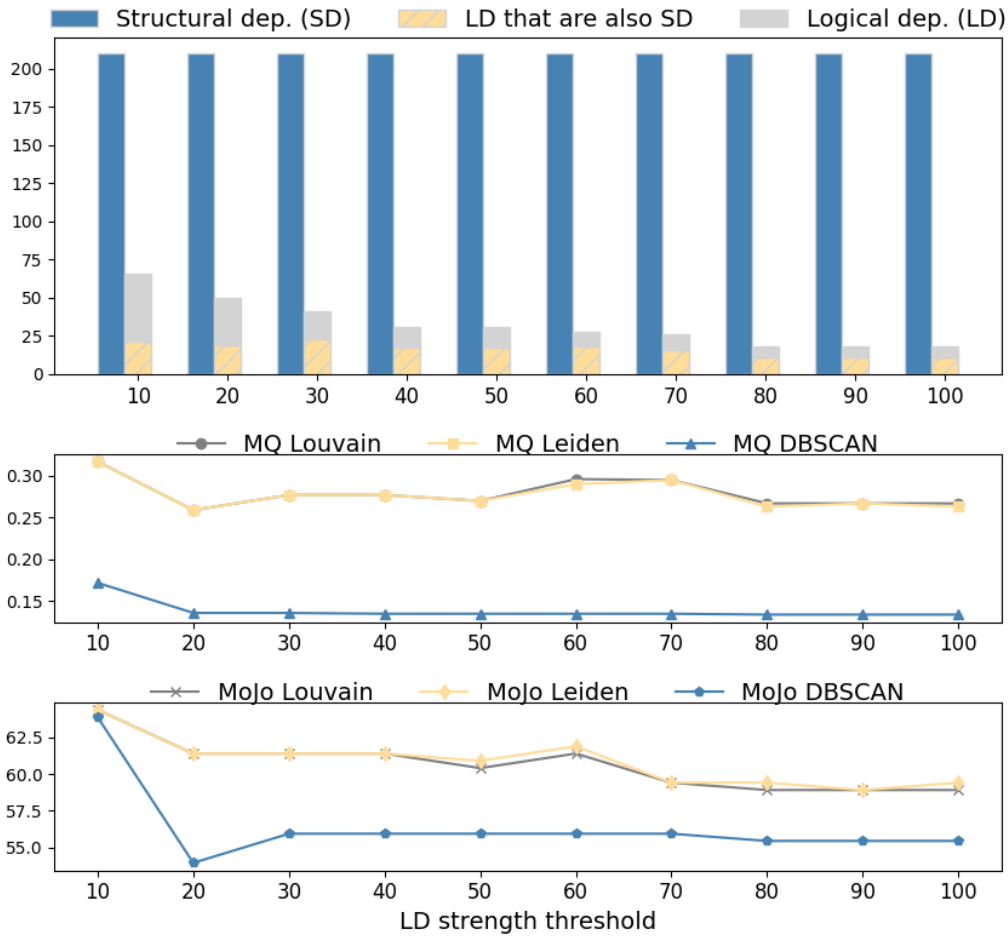

Figure 6.6: Google Gson: Overlap between structural and logical dependencies and its correlation with clustering metrics.

both algorithms. It is the only system where the best MQ result for LD-only occurs at a lower strength threshold than 100%. This is due to the very low number of entities remaining in the system at 100% (only 18 out of 210).

In this particular system, it is more visible that the Leiden clustering algorithm does not improve the Louvain algorithm in some scenarios. This observation is based on the fact that the values obtained for both MQ and MoJoFM metrics are the same in most cases for the Gson system for both algorithms.

It can also be observed that Gson has identical metric values for MQ and MoJoFM across multiple strength thresholds. Again, the small number of commits and the size of the system contribute to the stability of these metrics.

Gson also has relatively high overlap rates between LD and SD compared to the other systems, as shown in Figure 6.6. Despite the constant values, the trend of decreasing performance for SD+LD with stricter strength filtering for LD is also present in Gson.

The results for this system highlight the difficulty of achieving better clustering solutions using logical dependencies in smaller systems with fewer commits. However, even for a small system like Gson, an improvement is still visible when using logical dependencies.

6.6.2. Discussion on Ant clustering

Based on the results from Table **??**, we can observe that the combined approach of structural dependencies and logical dependencies gives the highest Modularity Quality (MQ) metric of 0.227 with a strength threshold of 30%, which is an improvement over the 0.08 MQ metric obtained when considering only structural dependencies.

Beyond the positive result indicated by the MQ metric, we searched for further validation by human software engineering expert opinion. After thoroughly studying and understanding the analyzed system source code and documentation, we evaluated the remodularization proposals resulting from the two clustering solutions.

The two clustering solutions compared are the clustering solution obtained only from structural dependencies, in comparison to the clustering solution obtained from using both structural and logical dependencies, filtered with a threshold of 30% for strength.

The clustering solution relying solely on structural dependencies consists of 12 clusters, while the solution using both structural and logical dependencies consists of 15 clusters. Both solutions involve the same number of entities (517). The entities listed below are placed in different clusters:

- taskdefs.Available$FileDir
- taskdefs.Concat and its inner classes taskdefs.Concat$1, taskdefs.Concat$MultiReader, taskdefs.Concat$ TextElement
- taskdefs.Javadoc$AccessType
- util.WeakishReference and its inner class util.WeakishReference$HardReference
- taskdefs.Replace and its inner classes taskdefs.Replace$ NestedString, taskdefs.Replace$Replacefilter

The migration of entities between clusters is illustrated in Figure 6.7. Given that the inner classes are shifted from one cluster to another in the same way as the outer class, we omitted the inner classes from the diagram.

As the cluster number itself is not significant and may vary across different script runs (labels might vary), we will refer to each individual cluster resulting from structural dependencies as *Cluster A* and to the ones resulting from both logical and structural dependencies as *Cluster B*.

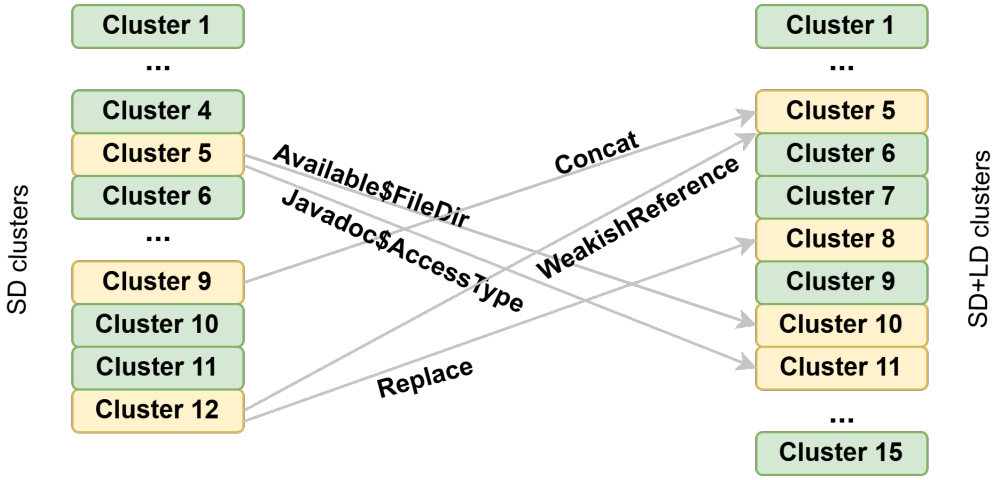

Figure 6.7: Migration of entities between clusters

taskdefs.Concat and its inner classes

To have a better overview of how and why entities are transferred between clusters, we depicted Concat's logical and structural connections in Figure 6.8. Additionally, Figure 6.9 illustrates connections within Cluster A, while Figure 6.10 does the same for Cluster B.

In Cluster A, the Concat class and its inner classes (Concat$1, Concat$MultiReader, Concat$TextElement) are placed together with conditions like Available, And, Or, IsTrue, Equals, IsReference, Contains.

On the other hand, in Cluster B, they are placed with classes associated with file manipulation and archive operations such as Ear, Jar, War, and Zip, as well as utility classes for file handling like FileUtils and JavaEnvUtils, and entities for zip file processing (ZipEntry, ZipFile). This placement is due to the logical dependencies that Concat class has with FileUtils and FileSet in the versioning system.

To assess whether the placement of Concat in Cluster B is better than in Cluster A, we referred to the official Ant documentation. According to the documentation: "This class contains the 'concat' task, used to concatenate a series of files into a single stream" [**?**]. Therefore, judging by its usage and purpose according to the documentation, positioning the Concat class along with its inner classes in Cluster B is more suitable than in Cluster A.

taskdefs.Available$FileDir

In Cluster A the entity 'taskdefs.Available$FileDir' is in the same cluster with entities that are related to the build process (ProjectHelper, TaskAdapter, ComponentHelper), but not with entities that have any relation to condition checks or file existence eval-

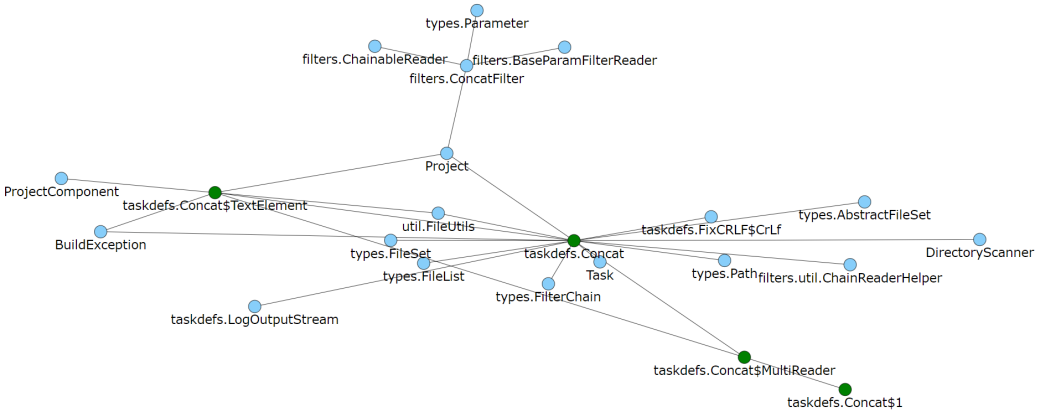

Figure 6.8: Dependencies (LD and SD) of Concat class

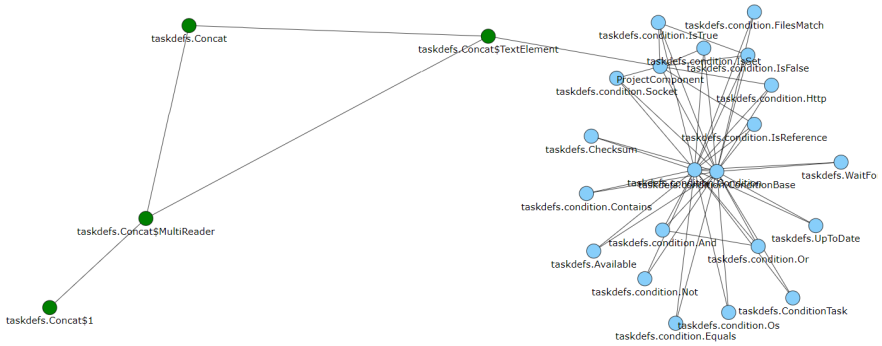

Figure 6.9: Placement of Concat in ClusterA (SD); cluster size: 25

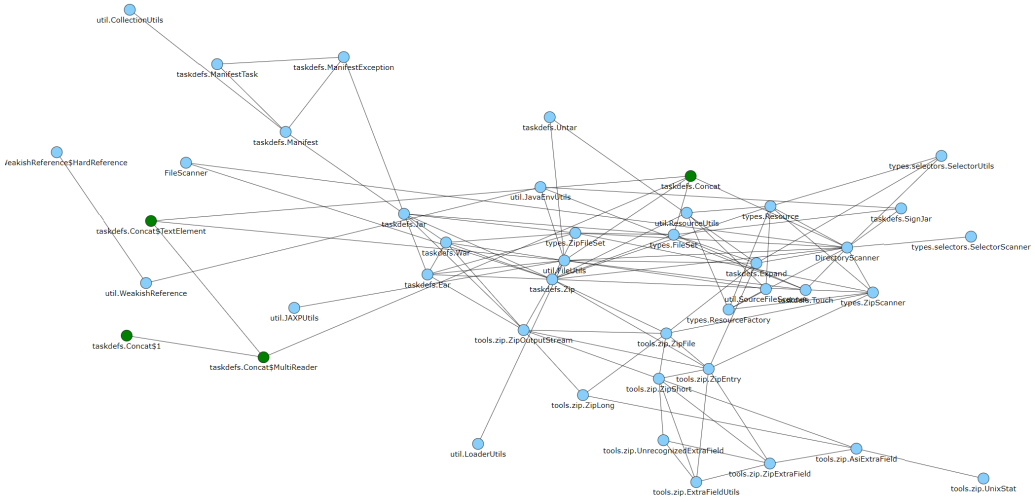

Figure 6.10: Placement of Concat in ClusterB (SD and LD); cluster size: 52

uations or with its outer class.

Cluster B contains entities that are related to condition checking (And, Contains, Equals, Or, UpToDate), and also with the outer class of 'taskdefs.Available$FileDir', 'taskdefs.Available'. The movement of entities from Cluster A to Cluster B was influenced by the logical dependencies between 'taskdefs.Available' and 'taskdefs.Available$FileDir'.

The documentation description for 'taskdefs.Available' states: "Will set the given property if the requested resource is available at runtime. This task may also be used as a condition by the condition task."[**?**]. This reinforces its placement in Cluster B, which is centered around task definitions and conditions related to build processes.

taskdefs.Replace and its inner classes

We depicted the logical and structural connections of Replace in Figure 6.11. Figure 6.12 illustrates connections within Cluster A, while Figure 6.13 does the same for Cluster B.

Replace and its inner classes are placed in Cluster B with entities with which they share common functionality and purpose, such as Copydir, Delete, DependSet, or MatchingTask. These entities are involved in similar tasks and operations. On the other hand, the placement in Cluster A, with WeakishReference and its inner class doesn't seem good, as these entities are completly unrelated.

The movement of entities from Cluster A to Cluster B was influenced by the strong logical dependency between 'taskdefs.Replace' and 'taskdefs.MatchingTask'.

The cluster from figure 6.12 is the 12th cluster from the clustering solution

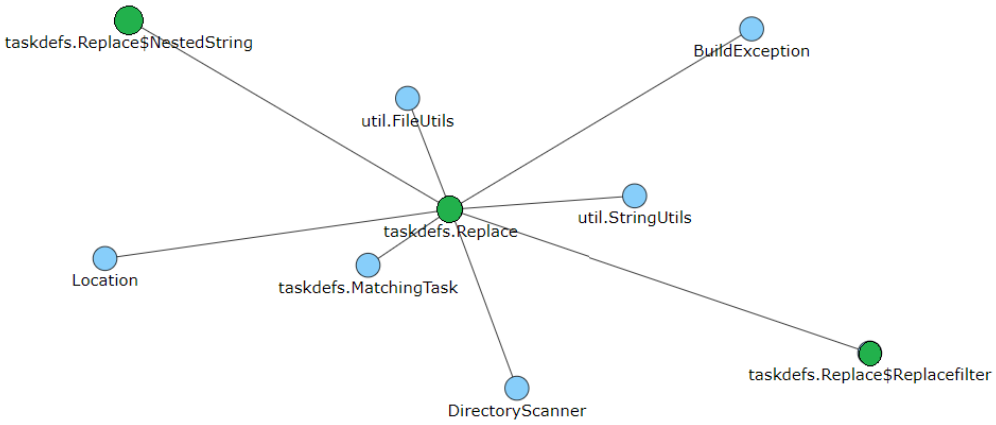

Figure 6.11: Ant dependencies (LD and SD) of Replace and its inner classes

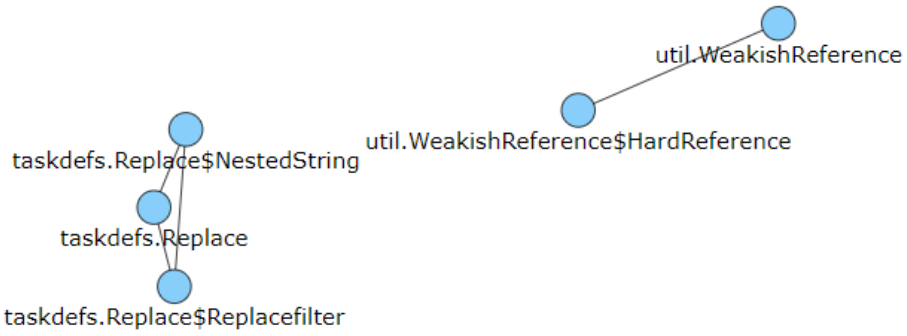

Figure 6.12: Placement of Replace in ClusterA (SD); cluster size: 5

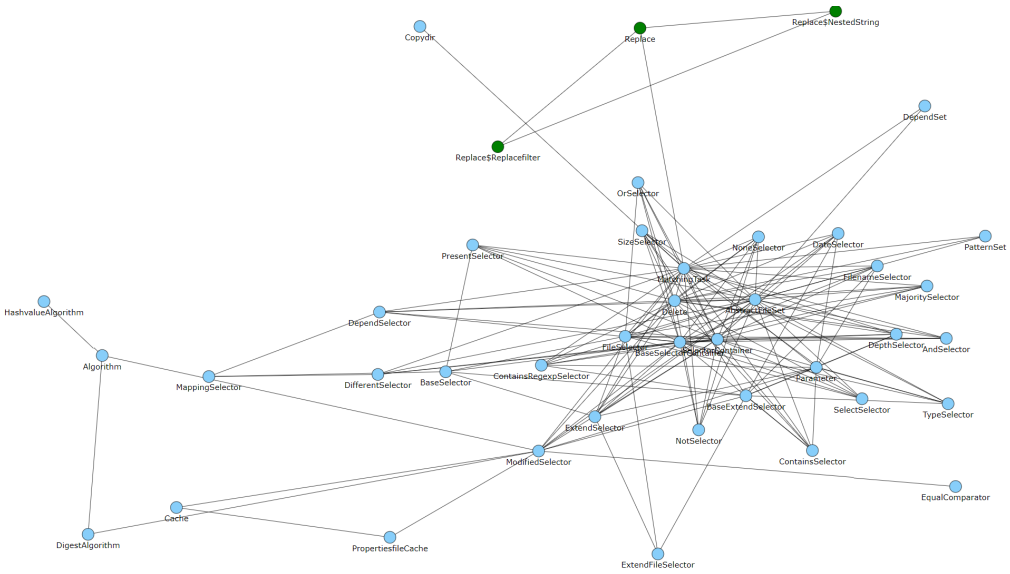

Figure 6.13: Placement of Replace in ClusterB (SD and LD); cluster size: 42

based on structural dependencies. This particular cluster is essentially comprised of two unrelated entities: Replace and WeakishReference. Due to their logical dependencies, both entities are grouped within larger clusters that share similar functionalities in the SD and LD solution [111].

6.6.3. Research questions and findings

In this section, we will answer our research questions based on the results of our experiments.

RQ1: *Does using structural dependencies (SD) combined with logical dependencies (LD) improve software clustering results compared to traditional approaches using only structural dependencies (SD)?*

Based on the results from all four systems analyzed, the combination of SD and LD performed better than SD-only according to the MQ and MoJoFM metrics, confirming that using SD+LD improves software clustering results. Across different strength thresholds, SD+LD achieved higher MQ and MoJoFM values for all clustering algorithms, indicating better modularization. The filtering thresholds applied to logical dependencies influence the effectiveness of combining SD and LD. Lower strength thresholds (10% to 40%) resulted in better MQ and MoJoFM values across all clustering algorithms compared with higher thresholds.

At higher strength thresholds, the MQ and MoJoFM results decline. This decline occurs because stricter strength thresholds significantly reduce the amount of logical dependencies. This reduction leads to fewer new relationships introduced into the clustering process. Our previous research on the overlap between SD and LD showed

that higher strength thresholds correlate with increased overlap between the two dependencies. This overlap indicates that at higher strength thresholds, not much new information is added to the system besides what is already introduced by structural dependencies, reducing the impact of logical dependencies on clustering results.

However, when considering a lower strength threshold, the relationship between the system size and the number of commits in the version history becomes an important factor.

In the case of Hibernate, a stricter strength threshold was needed to achieve the best metric values compared to other systems. Although the metrics obtained at a 10% strength threshold for LD are better than SD-only, the system reached peak metric values at 40% threshold across all algorithms.

With 16,609 commits and 4,414 entities, Hibernate has a significantly lower average number of commits per entity than Ant (14,917 commits and 517 entities) or Tomcat (22,698 commits and 662 entities). This lower ratio means that each entity in Hibernate is, on average, involved in fewer commits. As a result, the co-change data extracted for logical dependencies are sparser and contain more noise at lower strength thresholds.

A stricter strength threshold (e.g., 40%) filters out these weaker logical dependencies.

In contrast, systems like Ant and Tomcat, with higher ratios of commits to entities, obtain better results with logical dependencies at a 10% strength threshold because entities participate in more commits on average.

In conclusion, with an appropriate threshold, combining SD with LD leads to better clustering results than using SD alone.

RQ2: *Can using only logical dependencies (LD) produce good software clustering results?*

Using LD-only produced good clustering results, especially at higher strength thresholds. LD(100) produced the highest MQ and MoJoFM values for most systems compared with SD-only and SD+LD. However, the coverage of LD-only is significantly lower at these higher thresholds. For all systems, after filtering with 100% strength threshold, the system coverage of the remaining logical dependencies is less than 17% of the total known dependencies in the system.

Thus, while LD(100) provides the highest metric results compared to SD+LD and SD-only, it represents only a small subset of the system's dependencies.

On the other hand, LD(10) has, in most of the cases, better metric results than those for SD-only and SD+LD with better system coverage. Apache Ant and Tomcat LD(10) cover more than 60% of the system, while for Hibernate and Gson, the coverage is slightly above 30%.

Therefore, LD(10) can be an alternative to SD-only or SD+LD, especially if structural dependencies are not available.

It is also important to consider that LD-only performance at higher strength

thresholds depends on the system's characteristics, such as the number of commits and entities. For Gson project, the performance at a 100% strength threshold is not as good as for the other systems, reaching its peak at a 40% threshold. This is due to the low number of entities remaining at higher thresholds, with only 18 entities at the 80% to 100% strength thresholds, and the relatively small number of commits considered.

In summary, while LD-only can produce good clustering results, especially at higher strength thresholds, its limited coverage reduces its usability, as the clustering is intended for the entire system, not just a small subset. LD-only offers a good alternative to SD-only at lower strength thresholds, providing acceptable coverage.

RQ3: *How do different filtering settings for logical dependencies (LD) impact clustering results, and which filtering settings provide the best performance?*

The impact of different filtering settings on clustering performance was observed across all systems. For LD-only clustering, lower strength thresholds like 10% provided good system coverage but had lower MQ and MoJoFM values compared to higher thresholds like 100%, where the best metric results were often achieved. However, at these higher thresholds, the system coverage was significantly reduced.

The best performance was generally observed with strength thresholds between 10% and 40% for the combination of SD and LD (SD+LD). At these thresholds, the clustering solutions achieved higher MQ and MoJoFM values than SD alone.

It is important to select the optimal filtering threshold. Logical dependencies filtered with lower strength thresholds include more relationships, introducing more knowledge that could improve clustering results. However, a too-low threshold may sometimes introduce noise, especially in systems with a low commits-to-entities ratio. On the other hand, higher thresholds significantly reduce noise but can exclude valuable dependencies and decrease coverage.

The optimal filtering threshold may vary depending on system characteristics (number of commits, size of the system), so it is important to consider these factors when filtering logical dependencies.

7. CONCLUSION AND FUTURE WORK

7.1. Summary of research findings

7.2. Contributions

7.3. Future work

BIBLIOGRAPHY

[1] Ioana Şora and Ciprian-Bogdan Chirila. Finding key classes in object-oriented software systems by techniques based on static analysis. *Information and Software Technology*, 116:106176, 2019.

[2] Thomas Zimmermann, Peter Weisgerber, Stephan Diehl, and Andreas Zeller. Mining version histories to guide software changes. In *Proceedings of the 26th International Conference on Software Engineering*, ICSE '04, pages 563–572, Washington, DC, USA, 2004. IEEE Computer Society.

[3] Harald Gall, Karin Hajek, and Mehdi Jazayeri. Detection of logical coupling based on product release history. In *Proceedings of the International Conference on Software Maintenance*, ICSM '98, pages 190–, Washington, DC, USA, 1998. IEEE Computer Society.

[4] Harald Gall, Mehdi Jazayeri, and Jacek Krajewski. Cvs release history data for detecting logical couplings. In *Proceedings of the 6th International Workshop on Principles of Software Evolution*, IWPSE '03, pages 13–, Washington, DC, USA, 2003. IEEE Computer Society.

[5] Liguo Yu. Understanding component co-evolution with a study on linux. *Empirical Softw. Engg.*, 12(2):123–141, April 2007.

[6] Leon Moonen, Stefano Di Alesio, David Binkley, and Thomas Rolfsnes. Practical guidelines for change recommendation using association rule mining. In *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 732–743, 2016.

[7] Leon Moonen, Thomas Rolfsnes, Dave Binkley, and Stefano Di Alesio. What are the effects of history length and age on mining software change impact? *Empirical Software Engineering*, 23, 08 2018.

[8] Xiaoxia Ren, B. G. Ryder, M. Stoerzer, and F. Tip. Chianti: a change impact analysis tool for java programs. In *Proceedings. 27th International Conference on Software Engineering, 2005. ICSE 2005.*, pages 664–665, May 2005.

[9] Gustavo Ansaldi Oliva and Marco Aurélio Gerosa. Experience report: How do structural dependencies influence change propagation? an empirical study. In *26th IEEE International Symposium on Software Reliability Engineering, ISSRE 2015, Gaithersbury, MD, USA, November 2-5, 2015*, pages 250–260, 2015.

[10] Gustavo Ansaldi Oliva and Marco Aurelio Gerosa. On the interplay between structural and logical dependencies in open-source software. In *Proceedings of the 2011 25th Brazilian Symposium on Software Engineering*, SBES '11, pages 144–153, Washington, DC, USA, 2011. IEEE Computer Society.

[11] Denys Poshyvanyk, Andrian Marcus, Rudolf Ferenc, and Tibor Gyimóthy. Using information retrieval based coupling measures for impact analysis. *Empirical Software Engineering*, 14(1):5–32, Feb 2009.

[12] Igor Scaliante Wiese, Rodrigo Takashi Kuroda, Reginaldo Re, Gustavo Ansaldi Oliva, and Marco Aurélio Gerosa. An empirical study of the relation between strong change coupling and defects using history and social metrics in the apache aries project. In Ernesto Damiani, Fulvio Frati, Dirk Riehle, and Anthony I. Wasserman, editors, *Open Source Systems: Adoption and Impact*, pages 3–12, Cham, 2015. Springer International Publishing.

[13] Nemitari Ajienka and Andrea Capiluppi. Understanding the interplay between the logical and structural coupling of software classes. *Journal of Systems and Software*, 134:120–137, 2017.

[14] Andy Zaidman and Serge Demeyer. Automatic identification of key classes in a software system using webmining techniques. *Journal of Software Maintenance and Evolution: Research and Practice*, 20(6):387–417, 2008.

[15] V. Tzerpos and R. C. Holt. Accd: an algorithm for comprehension-driven clustering. In *Proceedings of the Seventh Working Conference on Reverse Engineering*, pages 258–267, 2000.

[16] B. S. Mitchell and S. Mancoridis. On the automatic modularization of software systems using the bunch tool. *IEEE Transactions on Software Engineering*, 32(3):193–208, March 2006.

[17] Ioana Şora. Helping program comprehension of large software systems by identifying their most important classes. In *Evaluation of Novel Approaches to Software Engineering - 10th International Conference, ENASE 2015, Barcelona, Spain, April 29-30, 2015, Revised Selected Papers*, pages 122–140. Springer International Publishing, 2015.

[18] Anna Corazza, Sergio Di Martino, Valerio Maggio, and Giuseppe Scanniello. Investigating the use of lexical information for software system clustering. In *2011 15th European Conference on Software Maintenance and Reengineering*, pages 35–44, 2011.

[19] Anna Corazza, Sergio Di Martino, and Giuseppe Scanniello. A probabilistic based approach towards software system clustering. In *15th European Conference on Software Maintenance and Reengineering (CSMR)*, pages 88–96, 2010.

[20] Amarjeet Prajapati, Anshu Parashar, and Amit Rathee. Multi-dimensional information-driven many-objective software remodularization approach. *Frontiers of Computer Science in China*, 17(3):173209, 2023.

[21] S. Mancoridis, B. S. Mitchell, Y. Chen, and E. R. Gansner. Bunch: a clustering tool for the recovery and maintenance of software system structures. In *Proceedings of the IEEE International Conference on Software Maintenance (ICSM'99)*, pages 50–59, 1999.

[22] Zhihua Wen and V. Tzerpos. An effectiveness measure for software clustering algorithms. In *Proceedings of the 12th IEEE International Workshop on Program Comprehension*, pages 194–203, 2004.

[23] Grady Booch. *Object-Oriented Analysis and Design with Applications (3rd Edition)*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 2004.

[24] Marcelo Cataldo, Audris Mockus, Jeffrey A. Roberts, and James D. Herbsleb. Software dependencies, work dependencies, and their impact on failures. *IEEE Transactions on Software Engineering*, 35:864–878, 2009.

[25] Neeraj Sangal, Ev Jordan, Vineet Sinha, and Daniel Jackson. Using dependency models to manage complex software architecture. In *Proceedings of the 20th Annual ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications*, OOPSLA '05, pages 167–176, New York, NY, USA, 2005. ACM.

[26] Trosky B. Callo Arias, Pieter van der Spek, and Paris Avgeriou. A practice-driven systematic review of dependency analysis solutions. *Empirical Software Engineering*, 16(5):544–586, Oct 2011.

[27] Amarjeet Prajapati and Jitender Chhabra. Improving modular structure of software system using structural and lexical dependency. *Information and Software Technology*, 82, 10 2016.

[28] A. Podgurski and L.A. Clarke. A formal model of program dependences and its implications for software testing, debugging, and maintenance. *IEEE Transactions on Software Engineering*, 16(9):965–979, 1990.

[29] Alberto Costa Neto, Marcio de Medeiros Ribeiro, Marcos Dosea, Rodrigo Bonifacio, and Paulo Borba. Semantic dependencies and modularity of aspect-oriented software. In *Proceedings of the 29th International Conference on Software Engineering Workshops*, ICSEW '07, page 171, USA, 2007. IEEE Computer Society.

[30] Cezar Sas and Andrea Capiluppi. Using structural and semantic information to identify software components. 02 2021.

[31] G. Bavota, B. Dit, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A. De Lucia. An empirical study on the developers' perception of software coupling. In *2013 35th International Conference on Software Engineering (ICSE)*, pages 692–701, May 2013.

[32] H. Kagdi, M. Gethers, D. Poshyvanyk, and M. L. Collard. Blending conceptual and evolutionary couplings to support change impact analysis in source code. In *2010 17th Working Conference on Reverse Engineering*, pages 119–128, Oct 2010.

[33] Guang-yi Tang and Hong-wei Xuan. Research on measurement of software package dependency based on component. *Journal of Software*, 7, 09 2012.

[34] Dharmalingam Ganesan. Adam: External dependency-driven architecture discovery and analysis of quality attributes. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 23, 03 2014.

[35] Franz Lehner. Software life cycle management based on a phase distinction method. *Microprocessing and Microprogramming*, 32:603–608, 08 1991.

[36] K H. Bennett, Dh Le, and Vaclav Rajlich. The staged model of the software lifecycle: A new perspective on software evolution. 05 2000.

[37] K. Bennett. Legacy systems: coping with success. *IEEE Software*, 12(1):19–23, Jan 1995.

[38] Keith H. Bennett and Vaclav Rajlich. Software maintenance and evolution: a roadmap. pages 73–87, 05 2000.

[39] Vaclav Rajlich. Modeling software evolution by evolving interoperation graphs. *Ann. Software Eng.*, 9:235–248, 05 2000.

[40] Gerardo Canfora and Massimiliano Di Penta. New frontiers of reverse engineering. pages 326 – 341, 06 2007.

[41] S. S. Yau, J. S. Collofello, and T. MacGregor. Ripple effect analysis of software maintenance. In *The IEEE Computer Society's Second International Computer Software and Applications Conference, 1978. COMPSAC '78.*, pages 60–65, Nov 1978.

[42] S A. Bohner and R S. Arnold. Software change impact analysis. *IEEE Computer Society*, 1, 01 1996.

[43] Hongji Yang and Martin Ward. Successful evolution of software systems. 01 2003.

[44] Ben Collins-Sussman, Brian W. Fitzpatrick, and C. Michael Pilato. *Version Control With Subversion for Subversion 1.6: The Official Guide And Reference Manual*. CreateSpace, Paramount, CA, 2010.

[45] S. Li, H. Tsukiji, and K. Takano. Analysis of software developer activity on a distributed version control system. In *2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 701–707, March 2016.

[46] Git Contributors. *Git Documentation*, 2024.

[47] GitHub, Inc. Github, 2024.

[48] Fabian Beck and Stephan Diehl. On the congruence of modularity and code coupling. In *Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering*, ESEC/FSE '11, pages 354–364, New York, NY, USA, 2011. ACM.

[49] Noriko Hanakawa. Visualization for software evolution based on logical coupling and module coupling. In *14th Asia-Pacific Software Engineering Conference (APSEC'07)*, pages 214–221, 2007.

[50] Jacek Ratzinger, Michael Fischer, and Harald Gall. Improving evolvability through refactoring. volume 30, 07 2005.

[51] Huzefa Kagdi, Malcom Gethers, and Denys Poshyvanyk. Integrating conceptual and logical couplings for change impact analysis in software. *Empirical Software Engineering*, 18, 10 2012.

[52] Huzefa Kagdi, Shehnaaz Yusuf, and Jonathan I. Maletic. Mining sequences of changed-files from version histories. In *Proceedings of the 2006 International Workshop on Mining Software Repositories*, MSR '06, page 47–53, New York, NY, USA, 2006. Association for Computing Machinery.

[53] Thomas Rolfsnes, Stefano Di Alesio, Razieh Behjati, Leon Moonen, and Dave W. Binkley. Generalizing the analysis of evolutionary coupling for software change impact analysis. In *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, volume 1, pages 201–212, 2016.

[54] Manishankar Mondal, Banani Roy, Chanchal K. Roy, and Kevin A. Schneider. Historank: History-based ranking of co-change candidates. In *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 240–250, 2020.

[55] Manishankar Mandal, Chanchal K. Roy, and Kevin A. Schneider. Automatic ranking of clones for refactoring through mining association rules. In *2014 Software Evolution Week - IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE)*, pages 114–123, 2014.

[56] Chakkrit Tantithamthavorn, Akinori Ihara, and Ken-Ichi Matsumoto. Using co-change histories to improve bug localization performance. In *2013 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 543–548, 2013.

[57] Marco D'Ambros and Michele Lanza. Reverse engineering with logical coupling. In *2006 13th Working Conference on Reverse Engineering*, pages 189–198, 2006.

[58] Mark Shtern and Vassilios Tzerpos. Clustering methodologies for software engineering. *Adv. Soft. Eng.*, 2012:1:1–1:1, January 2012.

[59] Nemitari Ajienka, Andrea Capiluppi, and Steve Counsell. An empirical study on the interplay between semantic coupling and co-change of software classes. *Empirical Software Engineering*, 23(3):1791–1825, 2018.

[60] Ioana Şora, Gabriel Glodean, and Mihai Gligor. Software architecture reconstruction: An approach based on combining graph clustering and partitioning. In *Computational Cybernetics and Technical Informatics (ICCC-CONTI), 2010 International Joint Conference on*, pages 259–264, May 2010.

[61] Ioana Şora. Software architecture reconstruction through clustering: Finding the right similarity factors. In *Proceedings of the 1st International Workshop in Software Evolution and Modernization - Volume 1: SEM, (ENASE 2013)*, pages 45–54. INSTICC, SciTePress, 2013.

[62] A.T.T. Ying, G.C. Murphy, R. Ng, and M.C. Chu-Carroll. Predicting source code changes by mining change history. *IEEE Transactions on Software Engineering*, 30(9):574–586, 2004.

[63] Adelina-Diana Stana and Ioana Şora. Logical dependencies: Extraction from the versioning system and usage in key classes detection. *Computer Science and Information Systems*, 20:25–25, 2023.

[64] T. Zimmermann, S. Diehl, and A. Zeller. How history justifies system architecture (or not). In *Sixth International Workshop on Principles of Software Evolution, 2003. Proceedings.*, pages 73–83, 2003.

[65] Gustavo Ansaldi Oliva and Marco Aurélio Gerosa. Experience report: How do structural dependencies influence change propagation? an empirical study. In *2015 IEEE 26th International Symposium on Software Reliability Engineering (ISSRE)*, pages 250–260, 2015.

[66] Kamran Sartipi. Software architecture recovery based on pattern matching. 09 2003.

[67] Eirini Kalliamvakou, Georgios Gousios, Kelly Blincoe, Leif Singer, Daniel M. German, and Daniela Damian. An in-depth study of the promises and perils of mining github. *Empirical Software Engineering*, 21(5):2035–2071, Oct 2016.

[68] S. Ducasse and D. Pollet. Software architecture reconstruction: A process-oriented taxonomy. *IEEE Transactions on Software Engineering*, 35(4):573–591, July 2009.

[69] L Bass, P Clements, and Rick Kazman. Software architecture in practice 2nd edition. 01 2003.

[70] J. I. Maletic and A. Marcus. Supporting program comprehension using semantic (lexical) and structural information. In *Proceedings of the 23rd International Conference on Software Engineering (ICSE 2001)*, pages 103–112, 2001.

[71] Jean Mayrand, Claude Leblanc, and Ettore M. Merlo. Experiment on the automatic detection of function clones in a software system using metrics. pages 244–, 01 1996.

[72] Andrian Marcus and J.I. Maletic. Identification of high-level concept clones in source code. pages 107– 114, 12 2001.

[73] Ira Baxter, Andrew Yahin, Leonardo de Moura, Marcelo Sant'Anna, and Lorraine Bier. Clone detection using abstract syntax trees. volume 368-377, pages 368–377, 01 1998.

[74] Toshihiro Kamiya, Shinji Kusumoto, and Katsuro Inoue. Ccfinder: A multilinguistic token-based code clone detection system for large scale source code. *Software Engineering, IEEE Transactions on*, 28:654– 670, 08 2002.

[75] James R. Cordy and Chanchal K. Roy. The nicad clone detector. In *2011 IEEE 19th International Conference on Program Comprehension*, pages 219–220, 2011.

[76] Md Saidur Rahman and Chanchal K. Roy. On the relationships between stability and bug-proneness of code clones: An empirical study. In *2017 IEEE 17th International Working Conference on Source Code Analysis and Manipulation (SCAM)*, pages 131–140, 2017.

[77] Abdullah Sheneamer, Hanan Hazazi, Swarup Roy, and Jugal Kalita. Schemes for labeling semantic code clones using machine learning. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 981–985, 2017.

[78] Arianna Blasi and Alessandra Gorla. Replicomment: Identifying clones in code comments. In *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)*, pages 320–3203, 2018.

[79] Eva Van Emden and Leon Moonen. Java quality assurance by detecting code smells. 11 2002.

[80] M. Abbes, F. Khomh, Y. Gueheneuc, and G. Antoniol. An empirical study of the impact of two antipatterns, blob and spaghetti code, on program comprehension. In *2011 15th European Conference on Software Maintenance and Reengineering*, pages 181–190, March 2011.

[81] F. Khomh, M. Di Penta, and Y. Gueheneuc. An exploratory study of the impact of code smells on software change-proneness. In *2009 16th Working Conference on Reverse Engineering*, pages 75–84, Oct 2009.

[82] Foutse Khomh, Massimiliano Di Penta, Yann-Gaël Guéhéneuc, and Giuliano Antoniol. An exploratory study of the impact of antipatterns on class change- and fault-proneness. *Empirical Software Engineering*, 17:243–275, 06 2012.

[83] R. Marinescu. Detection strategies: metrics-based rules for detecting design flaws. In *20th IEEE International Conference on Software Maintenance, 2004. Proceedings.*, pages 350–359, 2004.

[84] Francesca Arcelli Fontana, Marco Zanoni, Alessandro Marino, and Mika V. Mäntylä. Code smell detection: Towards a machine learning-based approach. In *2013 IEEE International Conference on Software Maintenance*, pages 396–399, 2013.

[85] Fabio Palomba, Gabriele Bavota, Massimiliano Di Penta, Fausto Fasano, Rocco Oliveto, and Andrea De Lucia. A large-scale empirical study on the lifecycle of code smell co-occurrences. *Information and Software Technology*, 99:1–10, 2018.

[86] F. Palomba, G. Bavota, M. D. Penta, R. Oliveto, D. Poshyvanyk, and A. De Lucia. Mining version histories for detecting code smells. *IEEE Transactions on Software Engineering*, 41(5):462–489, May 2015.

[87] L. Tahvildari and K. Kontogiannis. Improving design quality using meta-pattern transformations: a metric-based approach. *J. Softw. Maintenance Res. Pract.*, 16:331–361, 2004.

[88] Ioana Şora. Finding the right needles in hay - helping program comprehension of large software systems. In *Proceedings of the 10th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: ENASE,*, pages 129–140. INSTICC, SciTePress, 2015.

[89] Ioana Şora. A PageRank based recommender system for identifying key classes in software systems. In *2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 495–500, May 2015.

[90] Ferdian Thung, David Lo, Mohd Hafeez Osman, and Michel R. V. Chaudron. Condensing class diagrams by analyzing design and network metrics using optimistic classification. In *Proceedings of the 22nd International Conference on Program Comprehension*, ICPC 2014, page 110–121, New York, NY, USA, 2014. Association for Computing Machinery.

[91] M. H. Osman, M. R. V. Chaudron, and P. v. d. Putten. An analysis of machine learning algorithms for condensing reverse engineered class diagrams. In *2013 IEEE International Conference on Software Maintenance*, pages 140–149, 2013.

[92] Spencer Rugaber. Program comprehension. 08 1997.

[93] M. L. Collard, H. H. Kagdi, and J. I. Maletic. An XML-based lightweight C++ fact extractor. In *11th IEEE International Workshop on Program Comprehension, 2003.*, pages 134–143, May 2003.

[94] Bennet Lientz, E Burton Swanson, and Gerry E. Tompkins. Characteristics of application software maintenance. *Communications of the ACM*, 21:466–471, 06 1978.

[95] Nakshatra Gupta, Ashutosh Rajput, and Sridhar Chimalakonda. Cospex: A program comprehension tool for novice programmers. In *2022 IEEE/ACM 44th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 41–45, 2022.

[96] Iris Vessey. Expertise in debugging computer programs. 12 1984.

[97] James A. Jones and Mary Harrold. Empirical evaluation of the tarantula automatic fault-localization technique. pages 273–282, 01 2005.

[98] Holger Cleve and Andreas Zeller. Locating causes of program failures. pages 342– 351, 06 2005.

[99] José Campos, André Riboira, Alexandre Perez, and Rui Abreu. Gzoltar: an eclipse plug-in for testing and debugging. In *2012 Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, pages 378–381, 2012.

[100] Ming Wen, Junjie Chen, Yongqiang Tian, Rongxin Wu, Dan Hao, Shi Han, and Shing-Chi Cheung. Historical spectrum based fault localization. *IEEE Transactions on Software Engineering*, 47(11):2348–2368, 2021.

[101] Jeongju Sohn. Bridging fault localisation and defect prediction. In *2020 IEEE/ACM 42nd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 214–217, 2020.

[102] W. Eric Wong, Ruizhi Gao, Yihao Li, Rui Abreu, and Franz Wotawa. A survey on software fault localization. *IEEE Transactions on Software Engineering*, 42(8):707–740, 2016.

[103] R. W. Selby and V. R. Basili. Analyzing error-prone system structure. *IEEE Transactions on Software Engineering*, 17(2):141–152, Feb 1991.

[104] V. Y. Shen, Tze-jie Yu, S. M. Thebaut, and L. R. Paulsen. Identifying error-prone software—an empirical study. *IEEE Transactions on Software Engineering*, SE-11(4):317–324, April 1985.

[105] David Binkley. Source code analysis: A road map. pages 104–119, 06 2007.

[106] srcml; www.srcml.org.

[107] Michael Collard, Michael Decker, and Jonathan Maletic. srcml: An infrastructure for the exploration, analysis, and manipulation of source code: A tool demonstration. pages 516–519, 09 2013.

[108] Michael L. Collard, Michael J. Decker, and Jonathan I. Maletic. Lightweight transformation and fact extraction with the srcML toolkit. In *Proceedings of the 2011 IEEE 11th International Working Conference on Source Code Analysis and Manipulation*, SCAM '11, pages 173–184, Washington, DC, USA, 2011. IEEE Computer Society.

[109] Michael Collard. Addressing source code using srcml. 01 2005.

[110] Stana Adelina and Şora Ioana. Analyzing information from versioning systems to detect logical dependencies in software systems. In *International Symposium on Applied Computational Intelligence and Informatics (SACI)*, May 2019.

[111] Adelina-Diana Stana and Ioana Şora. Integrating logical dependencies in software clustering: A case study on apache ant. pages 113–118, 10 2024.

[112] Adelina Diana Stana. and Ioana Şora. Identifying logical dependencies from co-changing classes. In *Proceedings of the 14th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: ENASE,*, pages 486–493. INSTICC, SciTePress, 2019.

[113] Adelina-Diana Stana and Ioana Şora. Refining software clustering: The impact of code co-changes on architectural reconstruction. *Access*, 20:25–25, 2025.

[114] P. Meyer, H. Siy, and S. Bhowmick. Identifying important classes of large software systems through k-core decomposition. *Adv. Complex Syst.*, 17, 2014.

[115] D. Steidl, B. Hummel, and E. Juergens. Using network analysis for recommendation of central software classes. In *2012 19th Working Conference on Reverse Engineering*, pages 93–102, 2012.

[116] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[117] Yi Ding, B. Li, and Peng He. An improved approach to identifying key classes in weighted software network. *Mathematical Problems in Engineering*, 2016:1–9, 2016.

[118] Weifeng Pan, Beibei Song, Kangshun Li, and Kejun Zhang. Identifying key classes in object-oriented software using generalized k-core decomposition. *Future Generation Computer Systems*, 81:188–202, 2018.

[119] X. Yang, D. Lo, X. Xia, and J. Sun. Condensing class diagrams with minimal manual labeling cost. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 22–31, 2016.

[120] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[121] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[122] L. do Nascimento Vale and M. de A. Maia. Keecle: Mining key architecturally relevant classes using dynamic analysis. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 566–570, 2015.

[123] A. Zaidman, T. Calders, S. Demeyer, and J. Paredaens. Applying webmining techniques to execution traces to support the program comprehension process. In *Ninth European Conference on Software Maintenance and Reengineering*, pages 134–142, 2005.

[124] M. Kamran, M. Ali, and B. Akbar. Identification of core architecture classes for object-oriented software systems. *Journal of Applied Computer Science & Mathematics*, 10:21–25, 2016.

[125] A. Mubarak, S. Counsell, and R. M. Hierons. An evolutionary study of fan-in and fan-out metrics in oss. In *2010 Fourth International Conference on Research Challenges in Information Science (RCIS)*, pages 473–482, 2010.

[126] Gustavo Oliva and Marco Aurelio Gerosa. A method for the identification of logical dependencies. In *Proceedings of the 7th International Conference on Global Software Engineering (ICGSEW)*, pages 70–72, 2012.

[127] V. Tzerpos and R. C. Holt. Mojo: a distance metric for software clusterings. In *Proceedings of the Sixth Working Conference on Reverse Engineering*, pages 187–193, 1999.

[128] P. Andritsos and V. Tzerpos. Information-theoretic software clustering. *IEEE Transactions on Software Engineering*, 31(2):150–165, February 2005.

[129] Nicolas Anquetil and Timothy Lethbridge. File clustering using naming conventions for legacy systems. *Proceedings of the Second Working Conference on Reverse Engineering*, 1998.

[130] A. E. Hassan Wu and R. C. Holt. Comparison of clustering algorithms in the context of software evolution. In *Proceedings of the 21st IEEE International Conference on Software Maintenance (ICSM'05)*, pages 525–535, 2005.

[131] S. Mancoridis, B. Mitchell, C. Rorres, Y. Chen, and E. Gansner. Using automatic clustering to produce high-level system organizations of source code. In *Proceedings of the 6th International Workshop on Program Comprehension (IWPC'98)*, pages 45–52, 1998.

[132] Luciana Silva, Marco Valente, and Marcelo Maia. Co-change clusters: Extraction and application on assessing software modularity. *Transactions on Aspect-Oriented Software Development*, 2015.

[133] Vincent Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

[134] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, 80(5), 2009.

[135] V. Traag, L. Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

[136] T. Bonald, N. de Lara, Q. Lutz, and B. Charpentier. Scikit-network: Graph analysis in python. *Journal of Machine Learning Research*, 21(185):1–6, 2020.

[137] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, pages 226–231, 1996.

List of published papers

[A1]. **Stana, Adelina-Diana**, and Şora, Ioana. (2019). *Analyzing Information from Versioning Systems to Detect Logical Dependencies in Software Systems*. In Proceedings of the 2019 International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 15–20. DOI: 10.1109/SACI46893.2019.9111582.

[A2]. **Stana, Adelina-Diana**, and Şora, Ioana. (2019). *Identifying Logical Dependencies from Co-Changing Classes*. In Proceedings of the 2019 International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE), pp. 486–493. DOI: 10.5220/0007758104860493.

[A3]. **Stana, Adelina-Diana**, and Şora, Ioana. (2023). *Logical Dependencies: Extraction from the Versioning System and Usage in Key Classes Detection*. International Conference on System Theory, Control and Computing (ComSIS), pp. 25–25. DOI: 10.2298/CSIS220518025S.

[A4]. **Stana, Adelina-Diana**, and Şora, Ioana. (2024). *Integrating Logical Dependencies in Software Clustering: A Case Study on Apache Ant*. In Proceedings of the 2024 International Conference on Control Systems and Computer Science (ICSTCC), pp. 113–118. DOI: 10.1109/ICSTCC62912.2024.10744671.

[A5]. **Stana, Adelina-Diana**, and Şora, Ioana. (2024). *Refining Software Clustering: The Impact of Code Co-Changes on Architectural Reconstruction*. IEEE Access, pp. xx–xx. DOI: 10.xxxx/ACCESS.xxxxxxx.