Contents

1 Presentation of the research topic **3**

2 Software evolution and maintainability **4**

2.1 Software evolution 4

2.1.1 Stages 4

2.1.2 Software change 7

2.1.3 Large software systems 7

2.2 Software maintenance 7

2.3 Software quality 9

3 Dependencies in software systems **10**

3.1 Structural dependencies 11

3.2 Logical dependencies 12

3.2.1 Co-change filtering 12

4 Usage of dependencies **13**

4.1 Fault detection 13

4.2 Architectural reconstruction 13

4.3 Software metrics 14

5 Problems in Dependency Tracing **15**

6 Current status of research **16**

7 Co-changes filtering **18**

8 Research content and stages of research **21**

8.1 Proposed research stages 21

8.2 Gantt chart 23

8.3 Proposed contents of the thesis 24

Chapter 1

Presentation of the research topic

The domain of the proposed thesis is Automated Software Engineering. The thesis will develop methods for the analysis of legacy software systems, focusing on using historical information describing the evolution of the systems extracted from the versioning systems. The methods for analysis will integrate techniques based on computational algorithms as well as data-mining. As proof-of-concept, tool prototypes will implement the proposed methods and validate them by extensive experimentation on several cases of real-life systems.

Chapter 2

Software evolution and maintainability

2.1 Software evolution

2.1.1 Stages

Software has distinctive stages during its life:

Initial development In this stage, the first functional version of the system is developed. Also, many of the main features are developed and documented and, the architecture is defined: components that compose the system, their interactions, and properties.

In this stage, the architecture is defined and must be made in such a manner that the software will easily support new additions. Changes made in the evolution stages must change very little the architecture defined in the initial stage [12]. The architecture must:

- allow unanticipated change in the software without compromising system integrity.
- evolve in controlled ways. [13]

Evolution The evolution stage begins only after the initial stage is passed successfully. In this stage, iterative changes are made. By changes, we mean additions (new software features), modifications (changes of requirements or misunderstood requirements) or deletions. There are two main reasons for the change: the learning process of the software team and new requests from the client.

Servicing - Like we mentioned in the description of the initial stage, the architecture must be made in such a manner that the software may easily embrace changes. But, if not, then new changes are no longer easy to be made, are very difficult and time-consuming. At this point, the software enters the servicing stage (also called aging software, decayed software, and legacy) [15], [2].

Phase out - This stage is also called decline [15]. If in the servicing stage limited changes are implemented, in the phase-out stage the entire development is frozen. In this way the software becomes outdated. The users must make workarounds in order to cover the deficiencies of the software.

Close down - In this stage, the software is completely shut down, and the users redirected to a replacement system if this exists.

The stages presented above are represented in Figure 2.1.1.

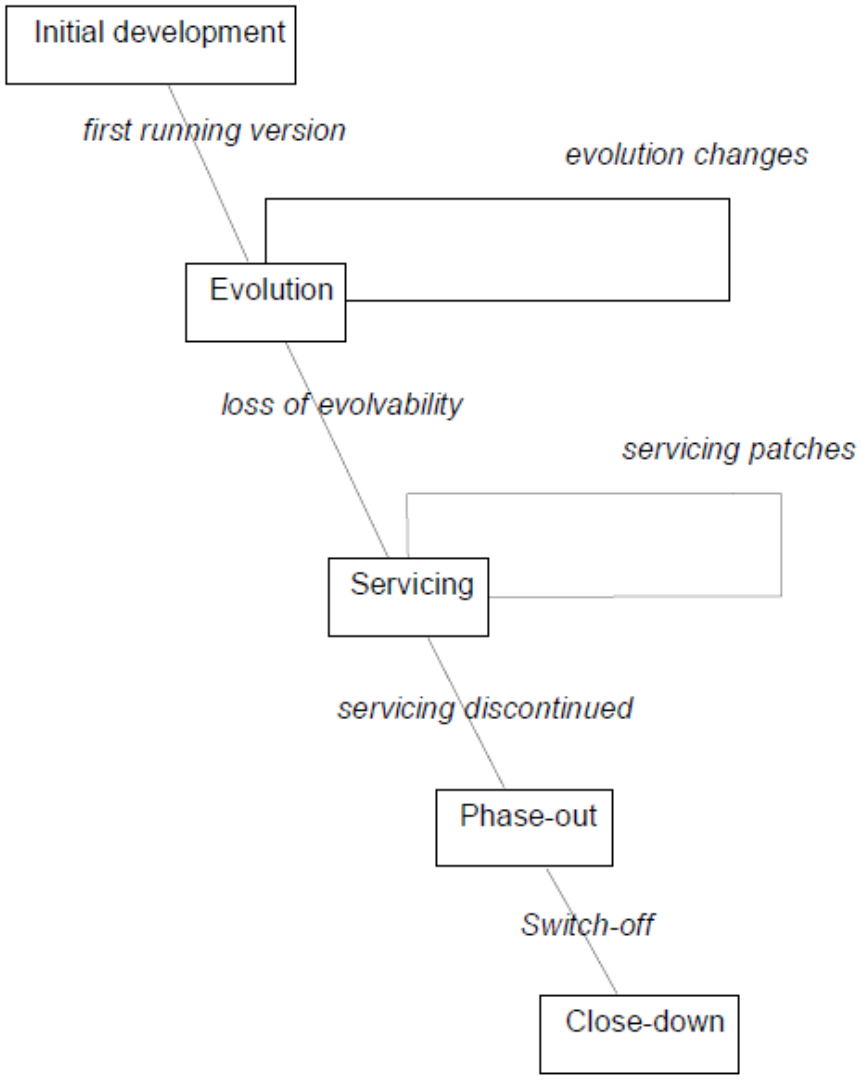


Figure 2-1: The versioned staged model [12]

2.1.2 Software change

As mentioned in section 2.1.1, software change in made in evolution and servicing stages. The difference between changes made in the evolution stage and changes made in the servicing stage is the effort of making 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 are really time-consuming.

The change is a process that either introduces new requirements into an existing system, or modifies the system if the requirements were not correctly implemented, or moves the system into a new operating environment. It is called the change mini-cycle [ ] and consists of the following phases [13]: Request for change Planning phase Program comprehension Change impact analysis Change implementation Restructuring for change Change propagation Verification and validation Redocumentation

2.1.3 Large software systems

P. Brooks [4] notes the following properties of large software systems: [10]

2.2 Software maintenance

In early days software maintenance was a small part of the software life cycle. As time was passing and more software was created, people realized that software does not die. Even if the actual development of the software may be frozen, requests regarding bug fixing and compatibility with new operating systems may appear over time. Sometimes software maintenance requires more effort than to build the system

from scratch [27].

Lientz and Swanson categorised maintenance activities into four classes [16]:

- Adaptive maintenance: changes in the software environment.
- Perfective maintenance: new user requirements and documentation improvement.
- Corrective maintenance: debugging anf bug fixing.
- Preventive maintenance: prevent future problems.

They also made a survey on the problems of application software maintenance in 487 organizations. The survey showed that most of the maintenance effort was on the first two types (adaptive and perfective maintenance). Many other studies suggest similar results.[13] All these studies underline one thing: *the incorporation of new user requirements is the main problem for software evolution and maintenance* [16].

One may ask why maintenance requires so much effort? There are some main problem factors when it comes to maintenance:

- User knowledge: lack of user training and/or lack of understanding of the system.
- Programmer effectiveness: lack of skills or experience for maintenance, development.
- Product quality: quality of the original system, quality of documentation and specs.

- Machine requirements: increasing storage or decreasing processing time requirements.
- System reliability: data integrity, hardware, and software reliability.

2.3 Software quality

Chapter 3

Dependencies in software systems

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 [3], [5].

A developer gains information about a software system by looking at it in two different ways:

- top-down: starting from the highest level of abstraction of the system and going down to the code.
- bottom-up: starting from the code and going up to the highest level of abstraction.

Both ways imply gradually gaining new information about the system and in both, the developer traces the dependencies and connections of the software system [26], [17].

3.1 Structural dependencies

Structural dependencies can be found by analysing the source code. [22]. 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.

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.

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.

3.2 Logical dependencies

3.2.1 Co-change filtering

Chapter 4

Usage of dependencies

4.1 Fault detection

past changes are good predictors of future faults

4.2 Architectural reconstruction

Currently, the software systems contain tens of thousands of lines of code and are updated multiple times a day by multiple developers. In order to successfully accomplish a task as a developer in such a system, you need to have a clear overview of the entire system and its connections. If for small systems a short look through the code can give the developer a good overview of the data-flow and functionalities of the system, in systems with tens of thousands of lines of code that's impossible. That is one of the reasons why the system's architecture is needed.

The software architecture also changes its characteristics from one stage to another. During the initial development, the architecture of software is established.

The architecture represents a significant commitment, and it will determine the future ease of evolution. In the projects we studied, the architecture changed only very little and hence the architectural commitments made during initial development are decisive factors for the future of the software. As the changes accumulate, the architecture can lose its original lucidity and integrity. There may be episodes of restructuringwhere the architecture is being partially rebuilt in order to facilitate future evolution. In the stage of servicing, the architecture becomes out of step with the needs of evolution and becomes an obstacle, limiting the scope of possible changes. Wrapping is a common process in this stage and it further damages the architecture, making it cryptic and very hard to understand. Where code changes are undertaken, they need to be very tactical and have minimal impact on other components. When the deterioration reaches a certain point, the architecture is no longer serviceable and the phase-out stage is the only option. The software decay can be explained as a positive feedback between loss of staff expertise, and loss of coherence in the program architecture. As the system architecture degrades, there is a stronger demand for greater and more sophisticated expertise to recognize and exploit the core design, to satisfy the main architectural constraints, and possibly to try to improve the architectural design where it has degraded too much. The real expert can understand when a change is tactical, and where it has profound effects on the architecture. The expert is able to spot danger signals. It is almost impossible to document or codify such expertise; it is almost always tacit. citemodel-bennett

4.3 Software metrics

Chapter 5

Problems in Dependency Tracing

Chapter 6

Current status of research

The current trend recommends that general dependency management methods and tools should also include logical dependencies besides the structural dependencies [18], [1].

Software engineering practice has shown that sometimes modules which do not present structural dependencies still appear to be related. Co-evolution represents the phenomenon when one component changes in response to a change in another component [28]. Those changes can be found in the software history maintained by the versioning system. Gall [11] identified as logical coupling between two modules the fact that these modules *repeatedly* change together during the historical evolution of the software system. *Logical dependencies* (a.k.a logical coupling) can be found by software history analysis and can reveal relationships that are not always present in the source code (structural dependencies).

The concepts of logical coupling and logical dependencies were first used in different analysis tasks, all related to changes: for software change impact analysis [21], for identifying the potential ripple effects caused by software changes during software

maintenance and evolution [19], [18], [20], [14] or for their link to deffects [25], [29].

Different applications based on dependency analysis could be improved if, beyond structural dependencies, they also take into account the hidden non-structural dependencies. For example, works which investigate different methods for architectural reconstruction [8], [6], [7], all of them based on the information provided by structural dependencies, could enrich their dependency models by taking into account also logical dependencies. However, a thorough survey [9] shows that historical information has been rarely used in architectural reconstruction.

Another survey [23] mentions one possible explanation why historical information have been rarely used in architectural reconstruction: the size of the extracted information. One problem is the size of the extraction process, which has to analyze many versions from the historical evolution of the system. Another problem is the big number of pairs of classes which record co-changes and how they relate to the number of pairs of classes with structural dependencies.

Chapter 7

Co-changes filtering

The software architecture is important in order to understand and maintain a system. Often code updates are made without checking or updating the architecture. This kind of updates cause the architecture to drift from the reality of the code over time.[9] So reconstructing the architecture and verifying if still matches the reality is important.

Surveys show that architectural reconstruction is mainly made based on structural dependencies [23] [9], the main reason why historical information is rarely used in architectural reconstruction is the size of the extracted information.

Logical dependencies should integrate harmoniously with structural dependencies in an unitary dependency model: valid logical dependencies should not be omitted from the dependency model, but structural dependencies should not be engulfed by questionable logical dependencies generated by casual co-changes. Thus, in order to add logical dependencies besides structural dependencies in dependency models, class co-changes must be filtered until they remain only a reduced but relevant set of valid logical dependencies.

Currently there is no set of rules or best practices that can be applied to the extracted class co-changes and can guarantee their filtering into a set of valid logical dependencies. This is mainly because not all the updates made in the versioning system are code related. For example a commit that has as participants a big number of files can indicate that a merge with another branch or a folder renaming has been made. In this case, a series of irrelevant co-changing pairs of entities can be introduced. So, in order to exclude this kind of situations the information extracted from the versioning system has to be filtered first and then used.

Other works have tried to filter co-changes [18], [1]. One of the used co-changes filter is the commit size.The commit size is the number of code files changed in that particular commit. Ajienka and Capiluppi established a threshold of 10 for the maximum accepted size for a commit [1]. This means that all the commits that had more than 10 code files changed where discarded from the research. But setting a harcoded threshold for the commit size is debatable because in order to say that a commit is big or small you have to look first at the size of the system and at the trends from the versioning system. Even thought the best practices encourage small and often commits, the developers culture is the one that influences the most the trending size of commits from one system.

Filtering only after commit size is not enought, this type of filtering can indeed have an impact on the total number of extracted co-changes, but will only shrink the number of co-changes extracted without actually guaranteeing that the remaining ones have more relevancy and are more logical linked.

Although, some unrelated files can be updated by human error in small commits, for example: one file was forgot to be commited in the current commit and will be commited in the next one among some unrelated files. This kind of situation can introduce a set of co-changing pairs that are definetly not logical liked. In order

to avoid this kind of situation a filter for the occurrence rate of co-changing pairs must be introduced. Co-changing pairs that occur multiple times are more prone to be logically dependent than the ones that occur only once. Currently there are no concrete examples of how the threshold for this type of filter can be calculated. In order to do that, incrementing the threshold by a certain step will be the start and then studying the impact on the remaining co-changing pairs for different systems.

Taking into account also structural dependencies from all the revisions of the system was not made in previous works, this step is important in order to filter out the old, out-of-date logical dependencies. Some logical dependencies may have been also structural in previous revisions of the system but not in the current one. If we take into consideration also structural dependencies from previous revisions then the overlapping rate between logical and structural dependencies could probably increase. Another way to investigate this problem could be to study the trend of concurrences of co-changes: if co-changes between a pair of classes used to happen more often in the remote past than in the more recent past, it may be a sign that the problem causing the logical coupling has been removed in the mean time.

Also, 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 [1]. 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.

Chapter 8

Research content and stages of research

8.1 Proposed research stages

The research will be made by following the next stages of implementation:

A. DEVELOPMENT OF CONTENT AND TOOLS

Stage 1: Build tool to extract structural dependencies from code and co-changes from git for a given set of projects.

Stage 2: Find filters for the co-changes extracted, the filters can be the ones already mentioned in previous works or new ones. Establish different thresholds for those filters.

Stage 3: Study the impact of those filters and the coresponding thresholds on the remaining quantity of co-changes for each system. Study the overlappings between the remaining pairs of co-changing entites and the structural dependencies extracted. [24]

Stage 4: Establish a dynamic way to determine the thresholds for filters in order to fit the best each studied sistem. Main focus on the threshold for number of occurences of co-changing pairs. Use plots or other visual instruments in order to see the highests and the lowest rates for the numbers of occurrences among co-changing pairs. Also filter those rates into normal and abnormal ones and study what was the cause of the highest rates (code or human related).

Stage 5: Take into account also structural dependencies from all the revisions of the system to filter out the old, out-of-date logical dependencies. Study how this affects the remaining number of logical dependencies.Here an extra check is needed, it can be a case in which old structural dependencies that were also logically linked to continue to be logically linked even after the structural dependency was removed.

B. USAGE

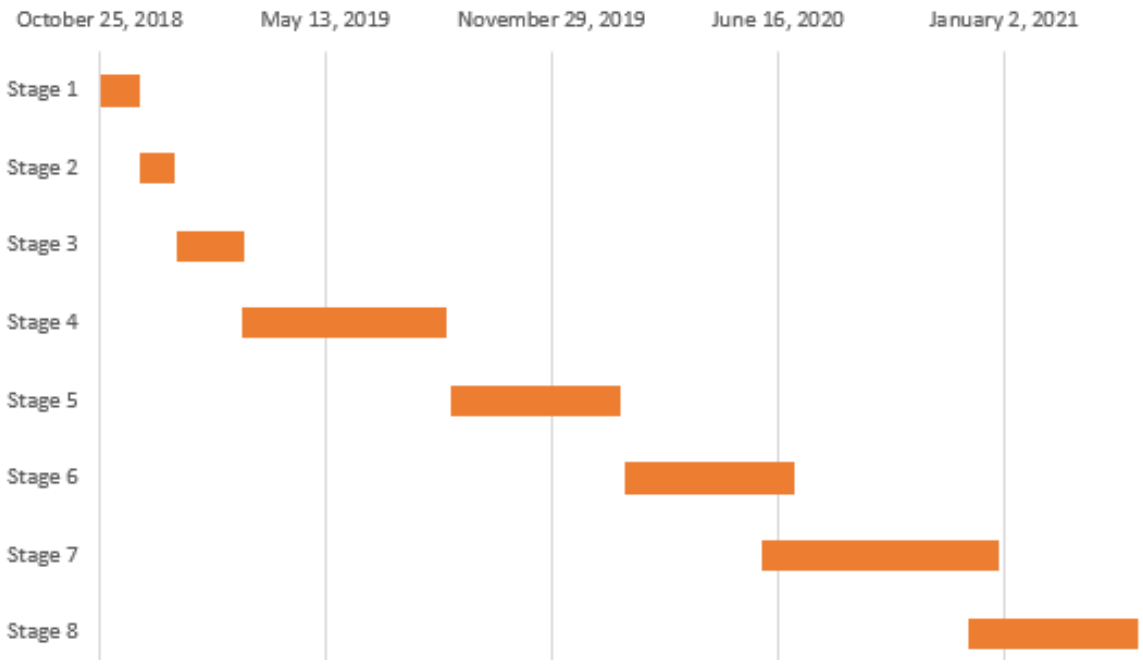
Stage 6: Export the remaining co-changes whom at this step we can call logical dependencies and use them among structural dependencies in tools for architectural reconstruction to evaluate the improvement.

Stage 7: Compare the number of logical dependencies with metrics like Fan Out, Fan In, Efferent Coupling (Ce), Afferent Coupling (Ca) and study their connections.

Stage 8: Identify other tools that use historical information and evaluate the impact of co-changes filtering into logical dependences for them.

8.2 Gantt chart

Gantt Chart				
START DATE	END DATE	DESCRIPTION	PAPER OUTPUT	DURATION (days)
10/25/18	11/30/18	Stage 1	No	35
11/30/18	12/30/18	Stage 2	No	30
1/1/19	3/1/19	Stage 3	Yes: ENASE, SACI	60
3/1/19	9/1/19	Stage 4	No	180
9/1/19	2/1/20	Stage 5	Yes: ICSME	150
2/1/20	7/1/20	Stage 6	Yes: ENASE	150
6/1/20	1/1/21	Stage 7	Yes: ICSE	210
12/1/20	5/1/21	Stage 8	No	150



8.3 Proposed contents of the thesis

CHAPTER I - theoretical notions

CHAPTER II - dependencies in software systems

CHAPTER III - co-changes

CHAPTER IV - co-changes filtering

CHAPTER V - usage of dependencies

Bibliography

[1] 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.

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

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

[4] Frederick P. Brooks, Jr. No silver bullet essence and accidents of software engineering. *Computer*, 20(4):10–19, April 1987.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] Steven Fraser, Frederick Brooks, Jr, Martin Fowler, Ricardo Lopez, Aki Namioka, Linda M. Northrop, David Parnas, and Dave Thomas. "no silver bullet" reloaded: retrospective on "essence and accidents of software engineering". pages 1026–1030, 01 2007.

[11] 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.

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

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

[14] 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.

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

[16] Bennet P. Lientz and E. Burton Swanson. Problems in application software maintenance. *Commun. ACM*, 24(11):763–769, November 1981.

[17] P. K. Linos and V. Courtois. A tool for understanding object-oriented program dependencies. In *Proceedings 1994 IEEE 3rd Workshop on Program Comprehension- WPC '94*, pages 20–27, Nov 1994.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] Adelina Diana Stana. and Ioana ora. 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.

[25] 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.

[26] Norman Wilde. Understanding program dependencies, 1990.

[27] Hongji Yang and Martin Ward. *Successful Evolution of Software Systems*. Artech House, Inc., Norwood, MA, USA, 2003.

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

[29] 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.