Playing Roles in Design Patterns: An Empirical Descriptive and Analytic Study

Foutse Khomh[1], Yann-Gaël Guéhéneuc[1], and Giuliano Antoniol[2]
[1] Ptidej Team—DGIGL, École Polytechnique de Montréal, Québec, Canada
[2] SOCCER Lab.—DGIGL, École Polytechnique de Montréal, Québec, Canada

E-mail: {foutsekh,guehene}@iro.umontreal.ca,antoniol@ieee.org

Abstract

This work presents a descriptive and analytic study of classes playing zero, one, or two roles in six different design patterns (and combinations thereof). First, we answer three research questions showing that (1) classes playing one or two roles do *exist in programs and are not negligible and that there are significant differences among the (2) internal (class metrics) and (3) external (change-proneness) characteristics of classes playing zero, one, or two roles. Second, we revisit a previous work on design patterns and changeability and show that its results were, in a great part, due to classes playing two roles. Third, we exemplify the use of the study results to provide a ranking of the occurrences of the design patterns identified in a program. The ranking allows developers to balance precision and recall.*

1. Introduction

Design patterns are proven solutions to recurrent design problems in object-oriented software design. Their design motifs [10] describe *ideal* solutions that will be either used to generate an architecture [3] or superimposed [12] on designed (or already existing) classes of a program. Consequently, classes in a program may play n roles in m motifs, with $n > 0$, $m > 0$.

Yet, to the best of our knowledge, previous work considered that classes either play *no* role or *some* role(s) in *some* motif(s), without distinguishing classes playing one or more roles in one or more motifs. It neglected that classes may play many different roles and considered role playing as a *all-or-nothing* characteristic of classes. This coarse-grained perspective prevents studying *finely* the impact of motifs on classes.

A reason for the current coarse-grained perspective is the lack of a method and manually-validated data to identify and evaluate the characteristics of classes playing one, two, or more roles with respect to classes playing no role. Therefore, we present a descriptive and analytic study of the impacts of playing one or two role(s) in motifs on classes *wrt.* playing zero role. We also show that previous work on design patterns benefit from this novel fine-grain perspective.

We exemplify these benefits on previous work that studied (1) design motif changeability and (2) motif identification. First, Bieman [4], Di Penta [8], and others (see Section 2) showed that classes playing some role(s) in one design motif are more complex and–or change-prone than classes playing *no* roles. They did not distinguish classes playing different numbers of roles. Consequently, they could only conclude generally on role change-proneness. A fine-grain perspective allows us to revisit these previous works and show that classes playing two roles represent, in average, 56% of all the changes that occurred before the studied release date while classes playing one role only 33%.

Second, Tsantalis *et al.* [23], Guéhéneuc and Antoniol [10], and others (see Section 2) proposed approaches to identify occurrences of design motifs in programs, which return unordered sets of occurrences. Yet, a ranking would help developers focus on the most relevant occurrences first. A fine-grain perspective allows us to sketch an approach to assign ranks to occurrences in function of the numbers of roles played by their classes. Applying this approach on JHotDraw v5.1, the Decorator design motif, and the occurrences obtained from DeMIMA [10] leads to 100% precision and recall on the first occurrence, to be compared to the previously reported 7.7% precision and 100% recall.

Thus, the contributions of this paper are: (1) a descriptive study showing that a non-negligible proportions of classes play one or two roles; (2) an analytic study showing that internal (class metrics) and external (change-proneness) characteristics of classes are im-

978-1-4244-4828-9/09/$25.00 2009 IEEE Proc. ICSM 2009, Edmonton, Canada

pacted differently by playing one and two roles; (3) a revisit of previous works confirming the soundness of our study and showing that they should be reexamined with a finer-grain perspective; and, (4) a revisit of a design pattern identification approach illustrating the ranking of occurrences and the possible improvements in precision and recall.

Section 3 presents the study definition and design while Section 4 its implementation: the method, programs, and motifs to build the samples; the metrics and their computations. Section 5 provides the study results. Section 6 present possible threats. Section 7 revisit previous works by describing two uses of the study results. Section 8 concludes with future work.

2. Related Work

Many pieces of work are related to design patterns, from their definition [16] to their identification [10]. We present here works related to the impact of design motifs on object-oriented quality and to the identification of occurrences of motifs in programs.

Motif Impacts. Bieman and McNatt [19] performed a qualitative study of the coupling between motifs and claimed that, when motifs are loosely composed and abstracted, maintainability, modularity, and reusability are well supported by the motifs. They concluded on a need for further studies to examine different motif compositions and their impact on quality.

Di Penta *et al.* [8] studied the change-proneness of roles and the kinds of changes affecting roles. Their results confirmed the expected, theoretical impact of motifs, *e.g.*, in Abstract Factory, classes playing concrete roles change more often than these playing abstract roles. They also highlighted deviations from the intuition, *e.g.*, in Composite, classes playing the role of Composite can be complex and undergo many changes.

Hannemann and Kiczales [12] studied the use of aspect-oriented programming and show that 17 of the 23 design patterns in [9] benefits from their "aspectisation" to overcome: the influence of motifs on programs and of programs on motifs; the loss of motif modularity and of traceability; the invasiveness of motifs; the difficulty to reason about classes involved in several motifs.

Khomh and Guéhéneuc [21] performed an empirical study of the impact of the 23 design patterns from [9] on ten different quality characteristics and concluded that patterns do not necessarily promote reusability, expandability, and understandability, as advocated by Gamma *et al.* They also studied patterns with respect to object-oriented principles and concluded that patterns do not necessarily lead to programs with good quality. Overall, their study advocate a considered use of patterns during development and maintenance.

Vokac *et al.* [24] analysed the corrective maintenance of a large commercial program over three years and studied the defect rates of classes playing roles in design motifs. Classes in motifs were less defect prone than others. He also noticed that the Observer and Singleton motifs are correlated with larger classes; classes playing roles in Factory Method were more compact, less coupled, and less defect prone than others classes; and, no clear tendency exists for Template Method.

Motif Identification and Ranking of Occurrences. We refer the kind reader to our recent survey for a complete overview of design motif identification approaches [10]. This survey shows that most approaches do not rank the identified occurrences. For example, Tsantalis *et al.* [23] proposed an approach based on similarity scoring to identify classes potentially playing a role in the design motif. This approach is fast and has reasonable precision and recall. It is exemplified on three programs and 10 design motifs. The occurrences are not ranked by their similarity score.

Our approach, DeMIMA [10], uses explanation-based constraint programming to provide approximations and explanations on the occurrences. It assigns a weight to each occurrence but this weight is subjective: it essentially depends on the weight assigned to each constraint and on the user's choice of the relaxed constraints; it does not consider the probability of a class to play zero, one, or more roles.

To the best of our knowledge, only Jahnke *et al.* [15] provide ranked occurrences. They used fuzzy-reasoning nets to identify design motifs. Their approach computes, for example, the probability of a class to be a Singleton. The main advantage of their approach is that fuzzy-reasoning nets deal with inconsistent and incomplete knowledge and that each occurrence is assigned a probability. However, their approach requires the description of all possible approximations of a design motif and users' assumptions.

Our study builds on this previous work, in particular Bieman and McNatt's work, to understand the impact on classes of playing one role in a motif or two roles in two different motifs. We use the study results to revisit previous works on design motif change-proneness and to rank identified occurrences of motifs. Spinellis' study [20] of four OS kernels also inspired us.

3. Study Definition and Design

Following GQM [2], the *goal* of our study is to study classes playing zero, one, or two roles in some design motifs. Our *purpose* is to bring generalisable, quantitative evidence on the impact of playing roles on classes.

The *quality focus* is that playing zero, one, or two roles impact differently classes. The *perspective* is that both researchers and practitioners should be aware of the impact of playing roles on classes to make inform design and implementation choices and to understand and forecast the characteristics of classes. The *context* of our study is both development and maintenance.

3.1. Research Questions and Hypotheses

Descriptive Question. The first research question is descriptive and assesses the extent of classes playing zero, one, or two roles in a general population of classes.

- **RQ1**: What is the proportion of classes playing zero, one, or two roles in some motif(s)?

Analytic Questions. The two following questions are analytic and divide in two sets of null hypotheses.

- **RQ2**: What are the internal characteristics of a class that are the most impacted by playing one or two roles *wrt.* zero role?
- **RQ3**: What are the external characteristics of a class that are the most impacted by playing one or two roles *wrt.* zero role?

For any metric m measuring some internal or external characteristics of a class, we test the set of null hypotheses: $H_{0mi/j}$: the distribution of the values of metric m for the classes playing $i \in [1, 2]$ role(s) is similar to that of classes playing $j \in [0, 1] \wedge j \neq i$ role.

We relate the following independent and dependent variables to assess the proportions of classes playing different roles and to test the previous null hypotheses.

3.2. Independent Variables

In an ideal situation, we would know the general population of *all possible classes* and know the number of roles played by any class. Then, we would use the sub-populations of classes playing zero, one, or more roles to answer the research questions. However, this situation is impossible because the population of all possible classes is so large and, in general, a class does not know if it plays any roles.

Therefore, the independent variables are three samples of classes playing zero, one, and two roles in design motifs. We limit our study to two roles and will consider more roles in future work. We name these samples the 0-, 1-, and 2-role samples. The samples must be large enough to be statistically representative but small enough to make it possible for manual inspection. The method to build these samples along with its implementation are presented in Section 4.

3.3. Dependent Variables

The dependent variables are the metrics measuring classes internal and external characteristics. We choose to study a large number of metrics, as previous work [20], to assess all the possible impacts of role playing.

Internal Characteristics are related to class themselves and are measured using 56 different metrics from the literature, including Briand *et al.*'s class-method import and export coupling [5]; Chidamber and Kemerer's Coupling Between Objects (CBO), Lack of Cohesion in Methods (LCOM5), and Weighted Method Count (WMC) [6]; Hitz and Montazeri 'C' connectivity of a class [13]; Lorenz and Kidd numbers of new, inherited, and overridden methods and total number of methods [17]; McCabe's Cyclomatic Complexity Metric (CC) [18]; Tegarden *et al.*'s numbers of hierarchical levels below a class and class-to-leaf depth [22]. The definitions of all the metrics is available on-line[1].

External Characteristics are limited in this study to the change-proneness of classes. A class is change-prone if, at a given time, it has been changed more than other classes. Change-proneness is assessed by computing the numbers and frequencies of past and future changes per class. Future work will study issue-proneness as well as other external characteristics.

The computation of the internal and external characteristics is described in Section 4. In Section 5, we report the metrics that proved to be significantly impacted by the number of roles played by classes and also discuss the not-impacted metrics.

3.4. Descriptive and Analytic Analyses

We use the following analyses to answer the research question with independent and dependent variables.

RQ1. Given a population of classes from 6 programs, we computed the classes playing zero, one, and two roles with our identification approach DeMIMA. Then, we compute the accuracy of our approach for one and two roles by manually validating classes playing roles in the identified occurrences. With this precision, we extrapolate the proportions of classes playing zero, one, and two roles in the general population.

RQ2 and RQ3. We use the Wilcoxon rank-sum test to compute for each metric and each pair of samples (0-role, 1-role), (0-role, 2-roles), and (1-role, 2-role), the p-values for the corresponding null hypotheses. The Wilcoxon rank-sum test is a non-parametric statistical hypothesis test that assesses whether two samples

[1] `http://wiki.ptidej.dyndns.org/research/pom`.

come from a same distribution or not. It allows us to attempt rejecting the null hypotheses while making no assumptions on the normality of the samples.

4. Study Implementation

The following subsections detail the building of the samples and the computation of the metrics.

4.1. Size of the Samples

We compute the sample size in two steps: (1) we assume the normality of the population and we compute the sample size needed for a two-sample t-test; and, (2) we adjust this size based on the Asymptotic Relative Efficiency (ARE) [14] of the two-sample Wilcoxon test.

We choose a *typical power of* 0.8, *i.e.*, we seek 80% chance of finding statistical significance if the specified effect exists. We also choose a *typical significance level of* 0.05 because we seek to reduce the possibility that the probability is due to chance alone.

With this power and significance level, we study the relation between effect size and sample size to choose the adequate sample size for a two-sample t-test, assuming the normality of the distribution. Following Cohen's work [7], we chose a *medium effect size of 0.58* that corresponds to a sample size of 50 classes.

The ARE represents the asymptotic limit of the ratio of the sample sizes needed to achieve equal power for two statistical tests: given a sample size for a statistical test A achieving a power p, the sample size needed for a test B to achieve the same power p is obtained from the ARE of A *wrt.* B. We compute the sample size for the two-sample Wilcoxon test that ensures the same power as the t-test, with no assumption of the distribution. The ARE for the two-sample Wilcoxon test is never less than 0.864 [14], we choose to be conservative and therefore divide the sample size for a t-test by 0.864. We obtain a *sample size of 58 classes.*

4.2. Selection of the General Population

We choose six programs to form the general population of classes from which to build the n-role samples: ArgoUML v0.18.1, Azureus v2.1.0.0, Eclipse JDT Core plug-in v2.1.2 (JDT Core v2.1.2), JHotDraw v5.4b2, Xalan v2.7.0, and Xerces v1.4.4. These programs are written in Java and open source. They are of different domains, sizes, complexity, and maturity. Table 1 summarises facts on these programs.

ArgoUML is a full-fledged UML modelling tool with code generation and reverse-engineering capabilities. It provides the user with a set of views and tools to model

Programs	NOC	KLOC	Release Dates	Past Changes	Future Changes
ArgoUML v0.18.1	1,267	203	30/04/05	20,290	12,617
Azureus v2.1.0.0	591	84	1/06/04	18,304	483
JDT Core v2.1.2	669	185	3/11/03	23,243	26,923
JHotDraw v5.4b2	413	45	1/02/04	5,793	51
Xalan v2.7.0	734	259	8/08/05	12,298	1,714
Xerces v1.4.4	306	87	13/10/03	5,213	1,209
Total	3,980	862	6 releases	85,141	42,997

Table 1. Statistics for the six programs. (Future refers to the time between the release dates and 31/01/09.)

Programs	Expected	Zero Role	One Role	Two Roles
ArgoUML v0.18.1	17	17	10	10
Azureus v2.1.0.0	9	9	9	9
JDT Core v2.1.2	10	10	17	17
JHotDraw v5.4b2	6	6	6	6
Xalan v2.7.0	11	11	11	11
Xerces v1.4.4	5	5	5	5
Total	58	58	58	58

Table 2. Distribution of the sample size among the programs of our strata.

programs using UML diagrams, to generate the corresponding code skeletons and to reverse-engineer diagrams from existing code. Azureus (now called Vuze) is a bit-torrent client. Bit torrent is a protocol to exchange data among peers across a network. Azureus provides advanced user-interface and implementation of the protocol. JDT Core is an Eclipse plug-in that implements the infrastructure for the Java IDE of the Eclipse platform. It provides a Java model and capabilities to parse, manipulate, and rewrite Java programs. JHotDraw is a graphic framework for drawing 2D graphics. It was created in October 2000 by Beck and Gamma with the purpose of illustrating the use of design patterns. Xalan is an XSLT processor for transforming XML documents into other document types (HTML, text, and so on). It implements the XSLT and XPath standards. Xerces is a Java XML parser which supports XML, DOM, and SAX.

4.3. Selection of the Motifs and their Roles

We select six design motifs used in previous work [8, 23]: Command, Composite, Decorator, Observer, Singleton, and State. We follow [8] in their choice of the motifs *main* roles. We only study main roles be-

Patterns	Descriptions	Main Roles
Command	Encapsulates a request as an object, thereby letting you parameterize clients with different requests, queue or log requests, and support undoable operations	Command, Invoker
Composite	Composes objects into tree structures to represent part-whole hierarchies. Composite lets clients treat individual objects and compositions of objects uniformly	Component, Composite
Decorator	Attaches additional responsibilities to an object dynamically. Decorators provide a flexible alternative to subclassing for extending functionality	Component, Decorator
Observer	Defines a one-to-many dependency between objects so that when one object changes state, all its dependents are notified and updated automatically	Observer, Subject
Singleton	Defines a mechanism that ensure that the same instance of a class is used throughout a program execution	Singleton
State	Allows an object to alter its behavior when its internal state changes	Context, State

Table 3. Chosen design patterns and the main roles of their motifs.

cause (1) they are most likely to impact classes, as confirmed by the following results, and (2) they allow us to concentrate on a fewer number of roles during the manual validation. In the following, roles are named using the notation <Pattern Name> . <Role Name>.

In addition to choosing the roles of interest, we must also select pairs of roles for classes playing two roles. We exclude pairs with the same role because identical roles in different motifs must have similar characteristics, *e.g.*, among the six motifs, Component is the only role that appears twice with similar structure albeit slightly different semantics. We exclude pairs involving roles from the same motif because a class playing both the roles of Composite.Component and Composite.Composite must be a degenerated case. Consequently, we retain 45 possible pairs.

4.4. Building of the Samples

Building the n-role sample, with $n \in [0, 2]$, consists of searching in the general population for three sets of 58 classes playing n roles. We reduce the search space using our DeMIMA approach because it ensures 100% recall by automatically relaxing appropriate constraints and has up to 80% of precision, with an average of 40% for the six design motifs in Table 3 in a set of programs different from these used in this study.

We applied DeMIMA on the classes in the general population and obtain candidate classes playing (at least) one role in the selected motifs. We automatically divided this set in two 1- and 2-role subsets.

Then, for each subset, we studied each class (its code source, comments, hierarchy, relationships) to decide whether it plays one role (respectively two roles) using a voting process: the authors and a post-doc. student marked independently each class as *true* when a class played one role (respectively, two roles) or $false$ else. Each class was marked by only three persons to avoid ties. Then, a class was assigned to the 1-role sample (respectively, 2-role sample) if the majority marked it as *true*, else it was excluded. We stopped the voting process as soon as the samples were completed.

In total, 238 classes were manually validated: 81 classes where false positives, *i.e.*, classes playing no role but belonging to occurrences identified by DeMIMA; 88 classes played 1 role; and, 69 classes played 2 roles. Finally, from the classes *not* included in any of the occurrences identified by DeMIMA, we selected randomly and validated manually 58 classes playing 0 role.

The distribution in the samples of the classes from the general population must be representative of the population. We distributed the 58 classes per sample along the strata formed by the six programs. We computed stratified sample sizes so that each stratum reflected the proportional size of one program with respect to the others. For example, JHotDraw v5.4b2 makes up 10.38% of the general population. So, it had to provide 10.38% of the 58 classes in each sample. Thus, we ensured that the results equally reflect the six programs. The second column in Table 2 shows the expected size of each stratum, *i.e.*, the expected numbers of classes of each program in each sample.

We could not find enough 1- and 2-role classes in ArgoUML. Therefore, we made up for the reduced number of classes in ArgoUML by using more classes from JDT Core. The fourth and fifth columns in Table 2 show the actual repartitions of classes in the 1- and 2-role samples. We replicated our study on the general population without JDT Core and on JDT Core exclusively and noticed the same trends.

4.5. Computing Dependent Variables

We compute the dependent variables using two different frameworks.

Internal Characteristics are computed using the PADL meta-model and parsers and the POM framework [11]. PADL models of programs are obtained using the Java parser and the metric values are computed by applying each metric on each class of the models.

External Characteristics are computed using the Ibdoos framework. Ibdoos extracts commit information from any CVS, GIT, or SVN repository and stores this in a database. We implemented queries to count

Programs	Candidates	One Role	Two Roles
ArgoUML v0.18.1	21	3	10
		14.28%	47.61%
Azureus v2.1.0.0	64	22	19
		34.37%	29.68%
JDT Core v2.1.2	67	30	22
		44.77%	32.83%
JHotDraw v5.4b2	30	11	13
		36.66%	43.33%
Xalan v2.7.0	55	23	11
		41.81%	20.00%
Xerces v1.4.4	29	10	11
		34.48%	37.93%
Total	266	99	86
		37.21%	32.33%

Table 4. Validated precisions of DeMIMA.

Programs	Total	One Role	Two Roles
ArgoUML v0.18.1	1,267	51	316
	100%	4.02%	24.94%
Azureus v2.1.0.0	591	67	75
	100%	11.33%	12.69%
JDT Core v2.1.2	669	46	178
	100%	6.88%	26.60%
JHotDraw v5.4b2	413	24	101
	100%	5.81%	24.45%
Xalan v2.7.0	734	36	104
	100%	4.90%	14.16%
Xerces v1.4.4	306	94	56
	100%	30.72%	18.30%
Total	3,980	318	830
	100%	7.99%	20.85%

Table 5. Extrapolated numbers and percentages of classes playing no, one, or two roles.

the numbers and frequencies of changes for each class before and after the release dates of the six programs.

5. Study Results

We analyse the metrics values computed on the classes in the samples to answer the research questions.

RQ1. To answer our first research question, "What is the proportion of classes playing zero, one, or two roles in some motifs in a program?", we extrapolate, for each program and each motif, the number of classes playing zero, one role, and two roles in the motifs.

First, from the class subsets, we compute the accuracy of DeMIMA as the number of classes in a subset *indeed* playing zero, one, or two roles with respect to the total numbers of classes in the subsets. Table 4 summarises this accuracy using all manually validated classes (reported in the Candidates column) and shows that it varies across motifs and programs. It highlights the need for more detailed studies of the accuracy of identification approaches. Indeed, current approaches report their precision and recall in function of the motifs but not of the programs. Reporting variations in terms of programs could help the community to focus on programs in which the identification is difficult. Such a focus would lead to a better understanding of the impact of program design and implementation on analysis tools and to a collection of difficult programs.

Second, we extrapolate in Table 5 the numbers of classes playing one and two roles from the previous accuracy and the numbers of candidate classes in each program. Table 5 shows that the percentage of classes playing one or two roles in any of the six selected design motif varies from 4.02% to 30.72%.

The answer to RQ1 is that classes playing one or two roles do *exist in programs and are not negligible, which confirms the need to understand the characteristics of classes playing different numbers of roles.*

RQ2. To answer RQ2, "What are the internal characteristics of a class that are the most impacted by playing one or two roles?", we test the null hypotheses $H_{0mi/j}, i \in [1,2], j \in [0,1] \wedge j \neq i$ for the 56 metrics. We first discuss unchanged metrics and then metrics whose distributions vary between each pair of samples.

There are 8 metrics whose distributions did not change significantly between the three samples: ANA, connectivity, CP, DSC, MFA, NOH, PP, and RPII. These metrics are therefore unlikely to be of interest when assessing the impact of role playing and could be excluded from future studies on design motifs. This finding was predictable for CP, PP, RPII because these metrics measure the structure of the packages of a system rather than the structure of its classes. The same explanation applies to DSC and NOH, which count respectively the total number of classes and the number of class hierarchies in a system. The finding for ANA, connectivity, and MFA is surprising because we expected that classes playing roles in design motifs would inherit more from and would be more "connected" to other classes. We explain this finding by the specific definitions of these three metrics because the values of other metrics related to inheritance and coupling significantly change between the samples.

There is a statistically significant difference between classes playing zero and one role for 29 metrics. These metrics characterise coupling, cohesion, inheritance, size and polymorphism, and complexity. The trends are a decrease in metric values for only four metrics: LCOM1, LCOM2, WMC1, and RRTP. This finding is explained again by the implementations of the metrics:

	1 role vs. 0 role	2 role vs. 0 role	2 role vs. 1 role
Unsignificant change	ANA, connectivity, CP, DSC, MFA, NOH, PP, and RPII		
Significant ↗	CIS, CLD, DCAEC, DCMEC, EIC, EIP, ICHClass, LCOM5, MOA, NCM, NCP, NMA, NMD, NMDExtended, NMO, NOC, NOD, NOM, NOParam, NOPM, PIIR, REIP, RFP, SIX, WMC	ACAIC, ACMIC, AID, CAM, CBO, CBOin, CBOout, CIS, CLD, cohesionAttributes, DAM, DCAEC, DCC, DCMEC, DIT, EIC, EIP, ICHClass, LCOM1, LCOM2, LCOM5, McCabe, MOA, NAD, NADExtended, NCM, NCP, NMA, NMD, NMDExtended, NMI, NMO, NOA, NOC, NOD, NOM, NOP, NOParam, NOPM, PIIR, REIP, RFP, RTP, SIX, WMC, WMC1	ACMIC, CAM, CBO, CBOin, CBOout, cohesionAttributes, DAM, DCC, ICHClass, LCOM1, LCOM2, LCOM5, McCabe, MOA, NAD, NADExtended, NMD, NMO, NOA, NOM, NOP, SIX, WMC, WMC1
Significant ↘	LCOM1, LCOM2, RRTP, WMC1	RRFP, RRTP	CLD

Table 6. Metrics Trends. (↗ or ↘ represent a significant increase (respectively, decrease) of, for example, the metrics values of 1-role classes compared to these of 0-role classes. Metrics that do not appear in a cell are those which values changed with no statistical significance.)

LCOM1 and 2 have been superseded by LCOM5, which changes significantly, while WMC1 weighs each method by 1 and RRTP is related to packages. The others metrics see a statistically significant increase in their values. Among these, we can quote: CBO, DCAEC, LCOM5, McCabe, SIX, WMC. We explain this finding by the fact that playing roles implies responsibilities, thus classes playing one role have more responsibilities than classes playing zero role, which results in classes being more complex (McCabe, SIX, WMC), more coupled (CBO, DCAEC), and less cohesive (LCOM5), as examples. We conclude that playing one role impact classes *wrt.* playing zero role.

There is a statistically significant difference between classes playing zero and two roles for 48 metrics, with, for each metric, an increase of its values for classes playing two roles, except for RRFP and RRTP. This finding was expected because RRFP and RRTP concern packages. For the 46 other metrics, the argument of added responsibilities with each role can also help explain the impact of 2-role classes on metric values in comparison to the impact of classes playing zero role. Having more responsibilities, classes become more complex (McCabe, WMC, WMC1, SIX), more coupled (CBO, DCAEC, DCC, DCMEC), inherit more from their superclasses (CLD, DIT, NOC, NOD), and use more polymorphism (MOA, NMA, NMD). Therefore, we conclude that playing two roles has a major impact on classes, in particular in comparison to the impact of playing zero role. Playing two roles should be carefully considered during design and implementation.

The change in the distributions of the metrics values between classes in the 2- and 1-role samples is significant for 26 metrics, among which: CAM, CLD, DCC, LCOM5, McCabe, SIX, WMC. We observe that the more they play roles, the more classes are complex (McCabe, SIX, WMC, WMC1), are coupled (CBO, DCC), inherit (NOP), and use polymorphism (MOA, NAD, NMO). The values of CLD decrease significantly, possibly hinting at more shallow inheritance tree thank to the elegant solutions provided by the motifs. We conclude that, indeed, playing two roles has a significant impact on classes that cannot be accounted for by the fact that they play two different one roles.

Consequently, the answer to RQ2 is that, wrt. the studied metrics, playing two roles has a major impact on classes when compared to playing zero or one role.

RQ3. We answer the last research question, "What are the external characteristics of a class that are the most impacted by playing one or two roles?", by carrying null hypothesis tests on the numbers and frequencies of past and future changes in the three samples, extracted from the version repositories of the programs.

We can reject the null hypotheses related to the external metrics for 1-role and 2-role classes *wrt.* 0-role classes with statistical significance. We cannot reject the null hypotheses for 2-role classes when compared to 1-role classes.

These results confirm previous works on the change-proneness of classes playing roles in some design motifs, for example [4, 8]. We perform in Section 7 a deeper analysis that shows that 2-role classes are the cause of the greater parts of the changes (56%) with 1-role classes causing only 33% of changes.

The answer to RQ3 is that playing roles do impact the number of changes as well as the frequencies of the changes. It confirms that playing roles has a major impact on change-proneness.

6. Threats to Validity

The results of any empirical studies are subject to the following threats to their validity.

Construct Validity. There is actually no agreed-upon definition of motif composition. In this study, we define a motif composition as the implementation

of two different roles in two different motifs by a same class. We only considered pairs of roles and ignored the effect of the particular roles on a class. We also explicitly excluded auto-composition, *i.e.*, a class playing two different roles in a same motif. Future work should distinguish compositions based on their roles and further study auto-compositions. Also, we purposefully studied only main roles of design patterns. Future work includes extending our study to all roles.

Internal Validity. Our approach relies on the precision of the automatic detection technique DeMIMA. The results include false positive. We try to limit the number of false positive through a manual validation. However, the manual validation is a tedious task that leads to resilience and the experimenter bias: some false positive class may pass the validation because it "looks like" a motif. An approach that would provide a better precision is to use a manually validated repository of motifs such as P-MARt [11]. However, P-MARt does not contain enough data as of now to perform such a study. We used as a baseline for our study of classes playing 1-role and 2-role, the 0-role population of classes playing none of the 11 roles considered in our study. However, among these classes, some may be playing one or two roles in *other* design motifs. Future work should extend this study to cover the 23 patterns from Gamma *et al.* [9]

External Validity. We studied six programs of different sizes, domains, maturity, and complexity. However, these programs are all open-source programs written in Java. We choose six design patterns among the many available. The results could be different with industrial programs, other object-oriented programming languages, and different design patterns.

Reliability validity. This threat concern the possibility of replicating this study. We attempted to provide all the necessary details to replicate our study. Moreover, both Eclipse source code repository and issue-tracking system are available to obtain the same data. Finally, the data from which our statistics have been computed is available on-line[2].

Statistical Validity. In Section 4, we presented the process to build the sample size of our study. We could not find enough classes playing one role and two roles in ArgoUML and, therefore, used more classes from JDT Core. We assess the impact of this selection on the conclusions of our study by replicating the study on the population without JDT Core and on JDT Core exclusively. We obtained for these two additional studies the same trends on the results.

Conclusion Validity. There is no threat to the validity of the conclusion of this study as there is a direct relation between the chosen metrics and the overall internal quality of a class.

[2] `http://www.ptidej.net/downloads/experiments/icsm09/`

7. Discussions

We now exemplify the use of our study results by revisiting previous works and sketching an approach to rank occurrences of identified design motifs.

7.1. Previous Work Comparison and Revisit

Bieman and McNatt's Work. We observe that playing one or more roles in a design motif decreases the cohesion of classes (increases of the LCOM$\star$ metrics) while increasing their coupling (increase of the coupling metrics). This result confirm Bieman and McNatt's claim [19] that design motifs impact the cohesion and coupling of programs.

Hannemann and Kicazles' Work. We explain the decrease in cohesion and increase in coupling by suggesting that design motif-related methods may be orthogonal to the responsibilities of the classes and thus reduce their cohesion. Therefore, our study confirms that design motifs are often "cross-cutting concern" that could benefit from being "separated" from the program using, for example, aspect-oriented programming. We thus bring quantitative support to previous work on rewriting design motifs as aspects [12].

Di Penta *et al.*'s Work. We revisit Di Penta *et al.*'s study of the numbers and frequencies of changes of classes playing roles. We compare the set of classes playing some roles, as identified by DeMIMA, which is the union of the samples of 1- and 2-role classes with the sample of false positive classes, noted 0^{FP}, with the set of classes playing *really* zero role: 0-role sample vs. (0^{FP}-role $\cup$ 1-role $\cup$ 2-role) sample. This comparison yields a p-value of **1.973e-14** < 0.05, thus confirming the previous work as well as the statistical validity of our three samples.

It appears from our study that, in average, the numbers of changes prior to the releases of the studied program for classes playing two roles accounts for 56% of the total number of past changes. Also, classes playing two roles change 1.52 times more than classes playing one role. Classes playing zero and one role account respectively for 33% and 11% of past changes. Classes playing one role change more than two role classes after the studied release of the programs. They change 1.46 times more than the 2-role classes and they account for 61.53% of the total number of future changes. We explain this result by the fewer numbers of future changes, shown in Table 1: in total, there are twice as much past changes than future changes. Therefore, we

bring evidence that the results found by Di Penta *et al.* was largely due to classes playing two roles.

We conclude that developers should be careful with classes playing roles, in particular 2-role classes, because they have internal and external metric values that are significantly higher than these of other classes: they are more change-prone, less cohesive, more coupled, more complex, and more issue-prone.

7.2. Ranking Design Motif Occurrences

We get inspiration from previous works by Antoniol *et al.* [1], Guéhéneuc *et al.* [11], and Jahnke *et al.* [15] to use the study results to rank occurrences.

First, we assign to each class in a program its probability to play one or more roles in a design motif using its metrics values. We select a set of discriminating metrics for the 0-, 1- and 2-role classes from Table 6. Then, we plot the distributions of these metrics for the 0-, 1- and 2-role samples. Finally, we find the thresholds characterising these samples for each selected metrics by superposing the curves of each selected metrics.

Second, the probability of a class is computed by interpolation as the distance between the values of its metrics and the thresholds characterising each samples. We aggregate these probabilities with the min and max fuzzy logic operators. Finally, from the probability of classes, we assign a probability to an occurrence as:

$$p_O = \frac{\sum_{i=1}^{n} \alpha_i \times p_{C_i}}{\sum_{i=1}^{n} \alpha_i}$$

where p_O is the probability of the occurrence to be a true positive; p_{C_i} is the probability of the class playing the i^{th} role in the occurrence to play one or more roles; and, α_i is a weight to discriminate roles.

We apply this naive approach using the metrics CAM, CBO, LCOM5, McCabe, MOA, NAD, NMO, SIX, and WMC. These metrics have proven in this study to be the most discriminating of classes playing 0, 1, and 2 roles. We choose $\forall i \in [1, n], \alpha_i = 1$. We apply this approach on the occurrences identified by DeMIMA in JHotDraw v5.1. We choose JHotDraw v5.1 to be able to compare with our previous work [10] ans also to show that our naive approach can be applied successfully on a different set of programs.

Figure 1(a) shows that, in the case of Decorator, our naive approach assigns the higher rank to the true positive occurrence. The precision and recall are therefore 100% when considering the first occurrence. These are to be contrasted to the 7.7% precision and 100% recall obtained by DeMIMA with no ranking [10].

Figure 1(b) shows that, in the case of State (or Strategy), our approach rank occurrences with less efficiency. Still, the precision of 33.3% with 100% recall obtained on the 18^{th} occurrence must be compared to the DeMIMA precision of 28.6%. Also, if a recall of 100% is not mandatory, precision reaches 38.5% on the 13^{th} occurrence.

We obtain results for the other four motifs in-between those presented for Decorator and State. Therefore, this naive approach allows reducing the developers' efforts by presenting true positive occurrences first and modulating precision and recall.

We conclude that our study results allow ranking the occurrences obtained from a design pattern identification approach using the number of roles likely to be played by classes. This ranking reduces the developers' efforts and allows developers to balance precision and recall as they see fit.

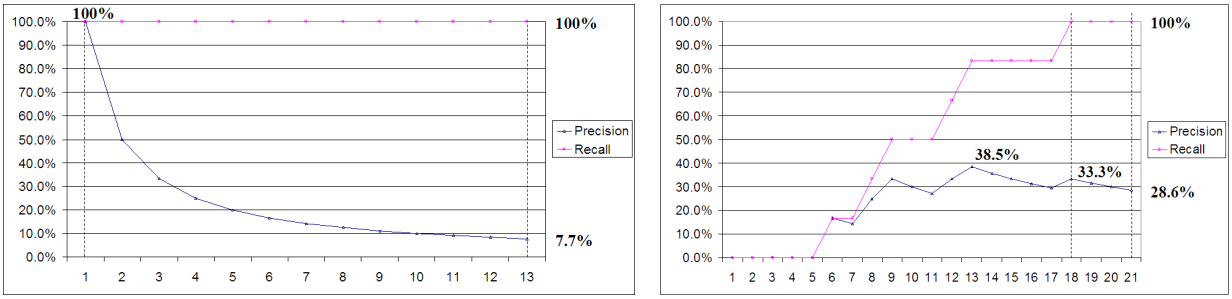
8. Conclusion

In this paper, we presented a study of the impact of playing one or two roles in a(some) motif(s) for a class. We answered the following research questions: **RQ1.** In average, 8.24% (respectively 17.81%) of the classes of the six studied programs played one role (respectively two roles) in some motifs. These percentages are not negligible and therefore justify *a posteriori* the interest in design motif identification and *a priori* future studies on the impact of motifs on programs. **RQ2.** There is a significant increase in many metric values, in particular for classes playing two roles. These increases confirm *a posteriori* the warning addressed to the community by Bieman, Beck, and others on the use of design patterns. **RQ3.** There is a significant increase in the frequencies and numbers of changes of classes playing two roles. We thus confirmed on new samples the previous results by Di Penta *et al.*

We justify the usefulness of this study by revisiting previous work and proposing a naive approach to rank occurrences. We show that developers should be wary of classes playing two roles because they have significantly higher metric values and represent 56% of changes while 1-role classes only 33%.

We also sketched a naive approach to illustrate the possibility of ranking occurrences using the metrics characterising 1- and 2-role classes. This approach leads to a precision and recall of 100% for the first occurrence of the Decorator. Extending on this naive approach, a new family of design pattern identification approaches could be designed to include the knowledge of the numbers of roles played by classes.

As future work, we plan to further study the impact of unique design motif on metrics values with the intuition that some motifs actually *do* fulfill (part of) the



(a) Precision and recall of the identification of Decorator.

(b) Precision and recall of the identification of State.

Figure 1. Precision and Recall.

intrinsic responsibilities of classes. We will also replicate this study on other motifs and programs as well as study classes playing three roles and more to confirm its generalisability. We also plan to further study the ranking of occurrences using other a more sophisticated approach, other identification approaches, and other programs. Also, the use of Bayesian beliefs networks to assign probabilities presents a great potential of obtaining better ranking and thus improving further the precision of identification approaches.

Acknowledgements. We thank Simon Denier for fruitful discussions and his help extracting the data. This work has been partly funded by Égide Lavoisier (France) and the NSERC and CFI (Canada).

References

[1] G. Antoniol, R. Fiutem, and L. Cristoforetti. Design pattern recovery in object-oriented software. In *Proceedings of the* 6^{th} *International Workshop on Program Comprehension*, pages 153–160. IEEE Computer Society Press, June 1998.

[2] R. Basili and D. M. Weiss. A methodology for collecting valid software engineering data. In *IEEE Transactions on Software Engineering*, 10(6):728–738, November 1984.

[3] K. Beck and R. E. Johnson. Patterns generate architectures. In *Proceedings of* 8^{th} *European Conference for Object-Oriented Programming*, pages 139–149. Springer-Verlag, July 1994.

[4] J. M. Bieman, D. Jain, and H. J. Yang. OO design patterns, design structure, and program changes: An industrial case study. In *Proceedings of the International Conference on Software Maintenance*, pages 580–589, http://www.dsi.unifi.it/icsm2001/, November 2001. IEEE Computer Society.

[5] L. Briand, P. Devanbu, and W. Melo. An investigation into coupling measures for C++. In *Proceedings of the* 19^{th} *International Conference on Software Engineering*, pages 412–421. ACM Press, May 1997.

[6] S. R. Chidamber and C. F. Kemerer. A metrics suite for object-oriented design. Technical Report E53-315, MIT Sloan School of Management, December 1993.

[7] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, 2nd edition, 1988.

[8] M. Di Penta, Luigi Cerulo, Y.-G. Guéhéneuc, and G. Antoniol. An empirical study of the relationships between design pattern roles and class change proneness. In *Proceedings of the* 24^{th} *International Conference on Software Maintenance (ICSM)*. IEEE Computer Society Press, September–October 2008. 10 pages.

[9] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns – Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1^{st} edition, 1994.

[10] Y.-G. Guéhéneuc and G. Antoniol. DeMIMA: A multi-layered framework for design pattern identification. In *Transactions on Software Engineering (TSE)*, 34(5):667–684, September 2008. 18 pages.

[11] Y.-G. Guéhéneuc, H. Sahraoui, and Farouk Zaidi. Fingerprinting design patterns. In *Proceedings of the* 11^{th} *Working Conference on Reverse Engineering (WCRE)*, pages 172–181. IEEE Computer Society Press, November 2004. 10 pages.

[12] J. Hannemann and G. Kiczales. Design pattern implementation in Java and AspectJ. In *Proceedings of the* 17^{th} *Conference on Object-Oriented Programming, Systems, Languages, and Applications*, pages 161–173. ACM Press, November 2002.

[13] M. Hitz and B. Montazeri. Measuring coupling and cohesion in object-oriented systems. In *Proceedings of the* 3^{rd} *International Symposium on Applied Corporate Computing*, pages 25–27. Texas A & M University, October 1995.

[14] M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. John Wiley and Sons, inc., 2nd edition, 1999.

[15] J. H. Jahnke and A. Zündorf. Rewriting poor design patterns by good design patterns. In *Proceedings the* 1^{st} *ESEC/FSE workshop on Object-Oriented Reengineering*. Distributed Systems Group, Technical University of Vienna, September 1997. TUV-1841-97-10.

[16] H. Kampffmeyer and S. Zschaler. Finding the pattern you need: The design pattern intent ontology. In *Proceedings of the* 10^{th} *International Conference on Model Driven Engineering Languages and Systems*, pages 211–225. Springer, September–October 2007.

[17] M. Lorenz and J. Kidd. *Object-Oriented Software Metrics: A Practical Approach*. Prentice-Hall, 1^{st} edition, July 1994.

[18] T. J. McCabe and C. W. Butler. Design complexity measurement and testing. In *Communications of the ACM*, 32(12):1415–1425, December 1989.

[19] W. B. McNatt and J. M. Bieman. Coupling of design patterns: Common practices and their benefits. In *Proceedings of the* 25^{th} *Computer Software and Applications Conference*, pages 574–579. IEEE Computer Society Press, October 2001.

[20] D. Spinellis. A tale of four kernels. In *Proceedings of the* 30^{th} *International Conference on Software Engineering*, pages 381–390. ACM Press, May 2008.

[21] Foutse Khomh and Y.-G. Guéhéneuc. Do design patterns impact software quality positively? In *Proceedings of the* 12^{th} *Conference on Software Maintenance and Reengineering (CSMR)*. IEEE Computer Society Press, April 2008. Short Paper. 5 pages.

[22] D. P. Tegarden, S. D. Sheetz, and D. E. Monarchi. A software complexity model of object-oriented systems. In *Decision Support Systems*, 13(3–4):241–262, March 1995.

[23] N. Tsantalis, A. Chatzigeorgiou, G. Stephanides, and S. Halkidis. Design pattern detection using similarity scoring. In *Transactions on Software Engineering*, 32(11), November 2006.

[24] M. Vokác. Defect frequency and design patterns: An empirical study of industrial code. In *Transactions on Software Engineering*, 30(12):904–917, 2004.