

Measuring and Mitigating Gaps in Structural Testing

Soneya Binta Hossain, Matthew Dwyer, Sebastian Elbaum, Anh Nguyen-Tuong





Code Coverage





FasterXML/jackson-dataformat-xml codecov 61% E README.md Overview Contributors 30 This projects contains Jackson extension component for reading and writing XML encoded data. 🛠 🗱 🔛 🗃 🍘 🚯 🚳 Further, the goal is to emulate how JAXB data-binding works with "Code-first" approach (no support is added for "Schema-first" approach). Support for JAXB annotations is provided by JAXB annotation nodule; this module provides low-level abstractions (JsonParser , JsonGenerator , JsonFactory) as well as small number of higher + 19 contributors level overrides needed to make data-binding work. Status Type Status Build and Deploy Snapshot passing Build (CI) Artifact mayen central 2.15.0-rc2 15 lifted! OSS Sponsorship Javadocs javadoc 2.15.0-rc2 codecov 615 Code coverage (2.15) Fuzzing oss-fuzz fuzzing

Based on <u>statement coverage</u> CSV seems to be better tested than jackson-dataformat.

Would we say the same thing if we considered the quality of the <u>test oracles</u> in addition to coverage?

Code Coverage, Test Oracle and Fault-detection

Code coverage is *essential* but *insufficient*

		Diffe	Different Testing Profiles				
Code Empirica	Coverage and Test Suite al Study with Real Bugs i	Effectiveness: n Large Systems	Xia Cal Orge, of Computer Scence and I The Onlinese University of Hur scali@cse.cu/R.adu/	Xia Cali Gept. of Computer Science and Engineering The Onlinear University of Hung King activities curve, suffic.edu.nk		Michael R. Lyu Dept. of Computer Science and Engineering The Denese University of Heng Kong SputRices.cu/W.edu.Nk	
Parsard Engli Xockidas, Yandian Thang, and David Lu. Engagene Hangarene Cuintrais, Engagene (Includers, E.L. Jonda J.H., andref d'emandres)			ABSTRACT Infrares toxing in a key procedure to reason and relability of adfract programs. The two more tarting is the selections and real-action of space. Toxic commany that been proposed to b	Kigh spailty Inne is soft Offerent last i to college	 INTRODUCTION As the main fash mesonic technique, adhes me of the main distributions articular to development [2]. The kine has a distribution to development [2]. The kine has a distribution and end of the metadation. As adhesive tend as a 	en inering is ing anti-ner givens ner	
council - Parting Selly to assess a or assess With the parts address to do concerning is off- address of test and dote is offer assess dotte is the second dotte is the second	Using Simulation for Assessing the R	ni Impact of Test Coverage of	 Introduction of comparing the second s	ten. In the old profile in condition of condition operations to operate to op	surface To Support the time sample of Control cases. To Support the time sample of Control and Completeness for the period of the case, and completeness for the period of the case, which are the 13-12 Control complete in teach fracture of program online that are excited thating the rank. Variant colds compare when the time of the case.	offer converge effectives and served as the strategies over a base tono a base tono	
the second p matches of the s in the paper of the second have been by our have been	Lionel Briand, Diemme Phill Franklight Juniter, for Experimental Software Exploreting (103) Sinerviewe 1: D-6704 Entervisioners, Garmany (Menal, Jahl) Warra Jie and		Coverage is Not Strongly Correlated with Test Suite Effectiveness			11111	
nit, te bin per	Abstract	methods. One may in bolier could improve the response difference in the second	Laura incuendas School of C University	vs and Reid imputer Science of Wighterton ON, Camata	Holmes #	And Row Accession Chail Ter-	
Access Color	in counted the software next process has because an increasingly connexe process. This is performed by the assungences that higher next converge leads a schere higher defail converge and directive improves software guides, in converge and directive improves software guides.	colored to as the severage) of the descend during testing testimal to: large method of test correspond progressid and statist. They map 1 constitute the progress theolis correspondence.	(mincremythol ABSTRACT	L INTR	rice.ca IODECTION	the self-	
To many 6 us many more and complexity of the information of the order of the second	before a lenser, it is notice which if is provided intervent point on the shifts of a point of the shifts of the shifts of the shifts of the shift of the shif	More of these two assumption is assumption of a simulation of the source of the sourc	<text><text><text><text></text></text></text></text>	nest service a s	are as an horder or discrossing blocks then discretized as a horder of a discrossing blocks that and the strength of the strength of the strength of the or physiciness of its absorber of the strength of th	Rag (24) endpaperation of the second	
	Toning is one of the same effect-connecter activities during software development (2). A great-line of research is directed severals development of same, supported test	president will done here it can yield. The operator of the paper is as b precise definition of the prioritients and constigne measures for control	Mananana Kaywardh Comage tan anin afferiment, tan anin aning	Roman' correlated o public multi correcto Per	 Quantizer 3. de the affectiveness of a tool ands with me utuatement eccurage, decrares converge deal conditions according place bits reacher of tool webs in general? 		
			Names a los depresas por esta de la seria	Remote and to the matrix of the The paper + 3 cm	a (program 8: A data officienciase of a tost main data data data data data data data data		

last of Code Counses on Early Detection

Code Coverage, Test Oracle and Fault-detection

 Code coverage is *essential* but insufficient

 Test oracles and faultdetection are *strongly correlated*

Assertions Are Strongly Correlated with Test Suite Effectiveness

Yucheng Zhang Electrical and Computer Engineering University of British Columbia Vancouver, BC, Canada vuchengz@Pece.ubc.ca

ABSTRACT

2.

Code coverage is a popular test adopasty office/on in practice. Code roverage, however, remains controvendal as there is a lack of solutions empirical evidence for its relation with built ruits effectiveness. More recently, test suits size has been shown to be highly correlated with effectiveness. However, previous studies trust test methods as the smallest unit of stevest, and ignore potostial factors influencing this relationality. We propose to go beyond two auto size, by trenatigating test american mode test methods. We empirically valuate the relationship between a test suite's effectiveness and the (1) number of assertions, (2) assertion coverage. and (3) different types of assertions. We compose 6,700 test mittee in total, using \$4,000 assertions of five read-world lines. projects. We find that the number of assertions is a test sale strongly correlates with its effectiveness, and this factor directly influences the relationship hotoson test onto size and effectiveness. Our results also indicate that assertion coverage is strongly correlated with effectiveness and different types of assertions can influence the effectiveness of their containing test makes.

Categories and Subject Descriptors

D.2.5 [Software Engineering]: Testing and Debugging: D.2.8 [Software Engineering]: Metrics

General Terms

Experimentation, Measurement

Keywords

Test subs effectiveness, assertions; coverage;

1. INTRODUCTION

Software testing has become an integral part of software development. A software product cannot be confidently released unless it is adequately tested. Code coverage is the

Protections to make aligned to least copies of dis to part of this work in generating the characteristic are parted without the protocol fills requires an end blanks in the information of the state and the particular state and the copies are followed by the state of the state and the state of the state is followed by each of the state of the sta

ATECOVER 23, August 30 - September 4, 2013, Bergann, Baly ACM, 976 1 - ANN, MISA, AUXAR, ADA 20 September 40, org/10.1145/2784405.2786408 Ali Mesbah Electrical and Computer Engineering University of Britain Columbia Yahnooving IBC, Canada amesbah@ecce.ubc.ca

must popular test adoptacy attention in practice. Reserver, reverage alone is not the goal of adbuue testing, since overage without checking its correctness is semaningless. A more meaningful adoptacy metric is the built detection adulty of a test mith, also haves an est mith effectiveness.

At these embes treat test neededs as the inadiat said of atternst. However, we before ratio course gassife shades are not reflected to deve the nain bettern influencing a test said of embedded and insetting or which are not and insetting with effectiveness. To that end, we have not needed and studied deviationase. To that end, we have not needed shades are methods. Their american are statements in test methods through suith denset operations are obtained agained actual program beforement. As web, assertions are at the one of test methods.

We hypothesise that assertions² have a strong influence on tost wile reflectiveness, and this influence, is turn, is the underlying reason belond file strong correlation between text subs size, node coverage, and not make effectiveness The the base of conditional was seen the first to conduct a large-scale empirical study on the diver relationship between survivane and to native directiveness.

In this paper, we conduct a series of experiments to questitatively study the relationship heresen next suits effectiveness and the (1) similar of americans, (2) american coverage, and (2) different types of americans.

This paper makes the following main contributions:

 The first large-scale study analysing the relation. Intreast team sources and test white effectiveness. Our study composes 0,200 test waites in total, from 5.802 test same and 24.701 saverthese of five real-world lives projects in different sizes and dismain.

"We use the terms "assertion" and "test assertion" interchangeably in this paper.

Coverage Based on Test Oracles

- Considers program execution and test oracles
 - Support statement criterion
 - Only *assess* test suite

- We build on Checked Coverage
 by Schuler and Zeller
 - We support stronger criterion
 - We introduce and study the concept of <u>Coverage Gap</u>

State Coverage: Software Validation Metrics beyond Code Coverage Dries Vanoverberghe^{1,*}, Jonathan de Halleux², Nikolai Tillmann², 2011 Foorth HEE International Conference on Sufficient Toping, Verification and Validation and Frank Piessens¹ Katholaka Universiteit Leaven, Leaven, Belgium Assessing Oracle Quality with Checked Coverage (dries.vanoverberghe.frank.plessens)#cs.kslevven.be David Schuler Andrew Zolley Rarting Daternity - Conner Lines. Sarlad Enternity - Company Science **Jasebrücket**, Gentury Sautombra Groups defining the statement of teller Box and teached at State Coverage: A Structural Test Adequacy Criterion for **Behavior Checking** erital behalts in the late int. Internet and Lot. Checked coverage: an indicator for oracle quality Key Kinder David Kaci ils - gatlettt: a + Antimiterant garantingthis kenk@lagtar.com dive Bagtar of Aptar Software Laboratories David Schuler".7 and Andreas Zeller 450 National Avenue -----Maurian Vice California India's Department of Computer Science, Approach (Science), Spartmented, Compare ABSTRACT tale that becalizates metaally wite makin, a frequently proposed approach lot. Adding the assortion may He proper a new insperse entry-print, reported not alsing Materian mating smalls artificial servers its take presentation projet for famous basis surger othering salled state towersay. Hule constrain tenthe code and assertants whether the test main STATISTICS. stinia value | and then the lat and whether and level into check the points and side ow scient of detected metaerts implies how others of a perigram. A basive problem of multional coverage metrics is that they do not assess over 5r gaulity-that is, whether the materia was not encouncil or low practic status at it Hait course filler is avoid reports from existing tor 10.12.4.67 the comparation result is actually checked against expectations. In this paper, we introduce the concept of afters and not checked). Unknowneds adequals entering such as only coverage and motation adeperiods a + -1; checked concerns the density day of concerned exemption that which induces an aracle the a is costly and with more advancements most. Unlike other communicational tohering mate constraint 1.11 reportments on arrest spon-source presents show that checked provings is a sure indicater for seadle guality mad work michael to wood out appivation more lie even of their of program behavior. And us PRIME IN and even more consistor than management maring. Copyright © 2013 John Wiley & Sons, Laf. Its couring had-based criteria cart as inclusion advances. date screenge has been designed to be readily assumed an introduce on observative, und officiant and its present upon with early antitudeed lost leadingsate Received 11 Separation 2011, Revised 23 November 2012, Accepted 13 Maril 2012 stack goality Using dynamic sloring, we shills total sandhalist Arrived company, stanmarts that want not An experiment down i strong positive completions in such and it is shart the faires and prostor which KEY WORDS: use sale quality coverage matrice, dynamic chering maturies evering he die actually contribute to the muchthe number of behavior obsets and both mate corresponded who he Planet it the checked coverage in similar administ one of the results and three land a run time Estapiries and Subject Descriptors: 0.25 Sufreen sense to be 87% traditional screenged Engineering) Testing and Dalongging turing Audo 5121.8 their company threads such 1. INTRODUCTION g a single position that he much is not more the checked comment to \$750 within General Terms Measurement, Relative e als obots. Its the stat one How can I couser that my programme is well-tensed? The most widespread matrix to assess test on on the properties of the small further Keymorike Corresp. had based autoitat batter inny of the report chestury. By rewiked powerspit quality is coverage. Test orwerage measures the percentage of code formers such as natements or terms of stationarity, dominant renings, strational batting, tail allegancy criteria, and testad universate as a meetin rather than regular beigned to be as may he and branches that are executed during a test. The nationale is that the higher the coverage, the higher significant advantages as dollarshit, farmerik communthe chances of catching a code feature that causes a failure --- a rationale that is easy to explain, but will a si fares characteristic reaction a low which reflex on an important commution. This construction is that we actually are able to deter the L INTRODUCTION enge, giving a most realisti auensensti 2. EXISTING ADDQUA Her throughly based is a program? This is the main failure. It does not suffice to cover the atom, we also need a means to detect it question that the advance official strange to server. Each official scorers this spectrum difference according to have litter of the partner and who As incoments-detecting advictory errors, it does not make a difference whether an error is detected. the an excession, has also a obscenet an meanly manually any hope by the text (e.g. a minimuch between actual and expected small), by the programme itself (e.g. a sA. homework be considered advocated As a searches program "die" and how much of that due to on, not would improve the smalles to actually a failing avertion) or by the run-time system (e.g. a development and pointer). To validate automotive of the a loss. Was between in broad-& constraint increasesses ion even the next basis of a arest, Toranella noder normage: computation results, however, we need checks in the inst code and in the programme code-checks a program by the monitor of branches it contains and informaa through an exercise or much only as nere with of these insuches were not taken during test to man estimate it. 10 10. Al henceforth summarized as oracles. A high coverage does not tell arplicing about reache quality it is velopers would know on therking as many major lists promiting also contains. Hash contract, the test advanced counting ar perfectly possible to achieve a 100% coverage and still not have any result checked by even a single suble, onlying new orors in for process. property in the paper, evaluate task providing to believe standing and using advances of oracle. The fact that the run time system did are discover are errors indicates robustness, but does checked overlage, out only seads to ran that dollars output and alls offset variables and judges roots The of all instruments but and not tell arothing about functional properties. according to effective day them these saturation. The campain 2010' nor being, tastabuilt/shoutdawd As an example of such a miseratch between coverage and oracle quality, consider the test for the criteria of these nategories and series_ Part Certifurner class from the JAXEN XP/CH library, shown in Figure 1. This test case pusses don not that program laborate anti-spatically. House to a set of paths through the parset. Intotestingly enough, it never checks any of the parset results. The terrestion is required to contain the respect of the last to ates by its position of commit statistic for the late being an with the convertions of the behavior of the program. The (1), setur hand others are parari mex return complete nomenoe, and this test would never notice it. inclination of the could be nelected by not compliant equivalent antidistants of a lat Why are the parson results unchecked? Actually, the purpose of this test is not to next the purson, but hand others, (of which metals Capacity to bell to do ashed the individual paths in the plat. It a array; if any of these costain a somethical error, the parset would them as churchen. If the stor means, all make are face. Exemine this toot however, also achiever, a statement coverage of KPS in the parset, and one may thus assume the parset is well-insted.

Focus of Our Paper

- Measuring the gap between code that is executed and code that is checked by test oracles – we call this the *coverage gap*
- Evaluating the impact of the coverage gap on fault-detection
- Mitigating coverage gaps by enhancing test suites to achieve better fault detection



public class Triangle {

int s1, s2, s3, p, color; Triangle(int a1, int a2, int a3, int c) {
 s1 = a1;
 s2 = a2;
 s3 = a2;
 color = c;
 setPerimeter();
}

private void setPerimeter() {
 p = s1 + s2 + s3;

public int getPerimeter() {
 return p;

public int getColor() {
 return color;------



Covered: 100% Checked: 25% In Gap: 75%



Evaluation: Artifacts

TABLE I DESCRIPTION OF ARTIFACTS

13 Java Applications

- 16K tests
- ✤ 51.6K test assertions

Artifact (version)	Description	Program Size(SLOC) ¹	Test Size(SLOC) ¹	Tests(#)2	Assertions(#)2
Commons-Cli (1.4) [14]	Command line option parsing	2,699	3,932	372	573
Commons-Codec (1.2) [15]	Common encodings	8,352	12,182	887	1,793
Commons-Csv (1.5) [16]	CSV utilities	1,615	4,467	296	934
Commons-Lang (3.6) [17]	Java helper utilities	27,265	48,172	2,908	15,424
Commons-Validator (1.6) [18]	Data validation	7,409	8,352	536	2,486
Gson (2.8.0) [22]	JSON support	7,815	13,762	1,014	1,780
Jackson-Dataformat-Xml (2.9.10) [27]	XML processing	4,945	5,728	185	556
Jaxen (1.2.0) [30]	XPath engine	10,760	8,042	716	587
JFreeChart (1.5.0) [3]	2D Charts	97,350	39,348	2,174	5,506
Joda-Time (2.10.11) [32]	Date and time library	28,811	55,849	4,238	17,973
Jsoup (1.10.1) [33]	HTML parsing	10,785	5,499	510	1,645
Plexus-Utils (3.1.0) [10]	Utility classes	18,496	6,337	304	799
XStream (1.14.15) [1]	XML serialization	21,741	25,518	1,830	1,554
	Total:	248K	237K	16K	51.6K

¹ Source lines of code (SLOC) are non-comment, non-blank lines reported by the IntelliJ statistic plugin.

² Tests are JUnit test cases annotated with @Test, and assertions are JUnit assertions.

Research Questions

RQ1: Gaps in studied artifacts

Finding: Gaps range from 19-51 percentage points (pp), with an average of 35pp

- RQ2: Impact of gaps on fault detection
- RQ3: Recommender performance
- RQ4: Recommended assertions and fault detection effectiveness

Finding: Fault detection improved as much as 57pp and on avg. 13pp

RQ2: Impact of Gaps on Fault Detection

Study Design:

- ✤ Generate 180 test suites by manipulating the gap size
- Generated 96K mutants to evaluate fault detection effectiveness
- Measure the correlation between gaps and kill scores

RQ2: Impact of Gaps on Fault Detection

Granularity: Application, Package Criteria: Statement, Object branch



Statement Coverage Gap (pp)

RQ2: Impact of Gaps on Fault Detection

Findings: Faults can hide in the coverage gap and there is a *strong negative* and *statistically-significant* correlation between gap size and fault-detection effectiveness.

RQ3: Recommender Performance

Study design:

- Remove developer written assertions from test suites
- Compute the resulting gap
- Analyze the SUT and the gap to recommend focus methods
- Compare recommended focus methods to focus methods in removed assertions

RQ3: Recommender Performance

TABLE III

PERCENTAGE OF ASSERTION FOCUS METHODS RECOMMENDED WITHIN THE TOP-K RECOMMENDATIONS ACROSS ALL 13 ARTIFACTS

Artifacts	Assert(#)	Top 1(%)	Top 5(%)	Top 10(%)
Commons-Cli	332	16	51	70
Commons-Codec	532	84	96	97
Commons-Csv	602	69	84	90
Commons-Lang	9843	80	96	98
Commons-Validator	1441	50	77	89
Jackson-Dataformat-Xml	83	33	43	63
Jaxen	134	11	30	37
JFreeChart	3240	82	93	97
Joda-Time	15775	17	43	53
Jsoup	1098	21	31	38
Gson	871	53	82	87
Plexus-Utils	365	55	75	78
XStream	578	38	56	59
Summary	Total	Average	Average	Average
	34894	46	67	73

RQ3: Recommender Performance

Finding: On average, 67% of the focus methods in the original test suites are suggested within the top-5 recommendations. Restricting to the top-1 recommendation, nearly half of the developer-written focus methods are present.

In summary:

- Traditional coverage can mislead.
- Gaps better reflect the under-tested codes.

Moving forward:

- Scale forms of assertion-based coverage.
- Leverage gaps for test suite improvement.

Artifact: https://github.com/soneyahossain/hcc-gap-recommender



Acknowledgement: DARPA ARCOS FA8750-20-C-0507, Air Force Office of Scientific Research FA9550-21-0164, and Lockheed Martin Advanced Technology Laboratories