2011-04-30

Modeling Complex System Testing: Characterizing Test Coverage to Improve Information Return

Karl Pfeiffer

NPS-AM-11-C8P10R01-042

EXCERPT FROM THE

PROCEEDINGS

OF THE

EIGHTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM WEDNESDAY SESSIONS VOLUME I

Modeling Complex System Testing: Characterizing Test Coverage to Improve Information Return

Karl Pfeiffer, Valery Kanevsky, and Thomas Housel, NPS

Published: 30 April 2011

Prepared for the Naval Postgraduate School, Monterey, California 93943

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

The research presented at the symposium was supported by the Acquisition Chair of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request Defense Acquisition Research or to become a research sponsor, please contact:

NPS Acquisition Research Program
Attn: James B. Greene, RADM, USN, (Ret.)
Acquisition Chair
Graduate School of Business and Public Policy
Naval Postgraduate School
555 Dyer Road, Room 332
Monterey, CA 93943-5103
Tel: (831) 656-2092
Fax: (831) 656-2253
E-mail: jbgreene@nps.edu

Copies of the Acquisition Sponsored Research Reports may be printed from our website www.acquisitionresearch.net

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Preface & Acknowledgements

During his internship with the Graduate School of Business & Public Policy in June 2010, U.S. Air Force Academy Cadet Chase Lane surveyed the activities of the Naval Postgraduate School's Acquisition Research Program in its first seven years. The sheer volume of research products—almost 600 published papers (e.g., technical reports, journal articles, theses)—indicates the extent to which the depth and breadth of acquisition research has increased during these years. Over 300 authors contributed to these works, which means that the pool of those who have had significant intellectual engagement with acquisition issues has increased substantially. The broad range of research topics includes acquisition reform, defense industry, fielding, contracting, interoperability, organizational behavior, risk management, cost estimating, and many others. Approaches range from conceptual and exploratory studies to develop propositions about various aspects of acquisition, to applied and statistical analyses to test specific hypotheses. Methodologies include case studies, modeling, surveys, and experiments. On the whole, such findings make us both grateful for the ARP's progress to date, and hopeful that this progress in research will lead to substantive improvements in the DoD's acquisition outcomes.

As pragmatists, we of course recognize that such change can only occur to the extent that the potential knowledge wrapped up in these products is put to use and tested to determine its value. We take seriously the pernicious effects of the so-called "theory–practice" gap, which would separate the acquisition scholar from the acquisition practitioner, and relegate the scholar's work to mere academic "shelfware." Some design features of our program that we believe help avoid these effects include the following: connecting researchers with practitioners on specific projects; requiring researchers to brief sponsors on project findings as a condition of funding award; "pushing" potentially high-impact research reports (e.g., via overnight shipping) to selected practitioners and policy-makers; and most notably, sponsoring this symposium, which we craft intentionally as an opportunity for fruitful, lasting connections between scholars and practitioners.

A former Defense Acquisition Executive, responding to a comment that academic research was not generally useful in acquisition practice, opined, "That's not their [the academics'] problem—it's ours [the practitioners']. They can only perform research; it's up to us to use it." While we certainly agree with this sentiment, we also recognize that any research, however theoretical, must point to some termination in action; academics have a responsibility to make their work intelligible to practitioners. Thus we continue to seek projects that both comport with solid standards of scholarship, and address relevant acquisition issues. These years of experience have shown us the difficulty in attempting to balance these two objectives, but we are convinced that the attempt is absolutely essential if any real improvement is to be realized.

We gratefully acknowledge the ongoing support and leadership of our sponsors, whose foresight and vision have assured the continuing success of the Acquisition Research Program:

- Office of the Under Secretary of Defense (Acquisition, Technology & Logistics)
- Program Executive Officer SHIPS
- Commander, Naval Sea Systems Command
- Army Contracting Command, U.S. Army Materiel Command
- Program Manager, Airborne, Maritime and Fixed Station Joint Tactical Radio System

- Program Executive Officer Integrated Warfare Systems
- Office of the Assistant Secretary of the Air Force (Acquisition)
- Office of the Assistant Secretary of the Army (Acquisition, Logistics, & Technology)
- Deputy Assistant Secretary of the Navy (Acquisition & Logistics Management)
- Director, Strategic Systems Programs Office
- Deputy Director, Acquisition Career Management, US Army
- Defense Business Systems Acquisition Executive, Business Transformation Agency
- Office of Procurement and Assistance Management Headquarters, Department of Energy

We also thank the Naval Postgraduate School Foundation and acknowledge its generous contributions in support of this Symposium.

James B. Greene, Jr.
Rear Admiral, U.S. Navy (Ret.)

Keith F. Snider, PhD
Associate Professor

Panel 10 – New Testing Protocols for the Open Architecture Era

Wednesday, May 11, 2011	
3:30 p.m. – 5:00 p.m.	**Chair: Captain Brian Gannon, USN**, Program Manager, Naval Open Architecture, PEO IWS ***Modeling Complex System Testing: Characterizing Test Coverage to Improve Information Return*** Karl Pfeiffer, Valery Kanevsky, and Thomas Housel, NPS ***Test Reduction in Open Architecture via Dependency Analysis*** Valdis Berzins, Peter Lim, and Mohsen Ben Kahia, NPS ***Utilizing Statistical Inference to Guide Expectations and Test Structuring During Operational Testing and Evaluation*** Joy Brathwaite, Georgia Institute of Technology, Alton Wallace and Robert Holcomb, Institute for Defense Analyses

Captain Brian Gannon—CAPT Gannon was born in Chicago, Illinois and received a commission in 1985 through the Naval Reserve Officer Training Corps program at the Illinois Institute of Technology. His formal education includes a Bachelor of Science in Mechanical Engineering from the Illinois Institute of Technology, a Master of Science in Astronautical Engineering from the Naval Postgraduate School, and a Master of Business Administration from the University of Phoenix.

His service tours include Electronics Readiness Officer, ASW Officer and CIC Officer onboard *USS Gary* (FFG-51) from 1986 to 1989; Combat Systems Instructor at the Surface Warfare Officer's School in Coronado, CA, from 1989 to 1992; Student in the Space Systems Engineering curriculum at the Naval Postgraduate School from 1992 to 1994; Aegis Project Officer at the Port Hueneme Division, Naval Surface Warfare Center from 1994 to 1998; AEGIS LEAP Intercept (ALI) Project Officer in the Navy Theater Wide Program Office (PMS 452) from 1998 to 2002; TBMD Section Head in the Aegis Combat System Engineering Program Office (PMS 400B) from 2002 to 2003; Combat Systems Officer on the Fleet Maintenance staff for Commander, United States Pacific Fleet from 2003 to 2005; Technical Representative for Surface Naval Weapons (PEO IWS 3.0) and Aegis Ballistic Missile Defense (PD 452) portfolio of programs at Raytheon Missile Systems in Tucson, AZ.

CAPT Gannon assumed his present duties as Major Program Manager Future Combat Systems and Open Architecture (PEO IWS 7.0) in October 2008.

Captain Gannon's personal awards include the Meritorious Service Medal (four awards), Navy Commendation Medal and the Navy Achievement Medal in addition to various service awards. He is married to the former Jean Raup of Alexandria, VA. He has three children: Brittany (18), Timothy (15), and Christopher (13).

Modeling Complex System Testing: Characterizing Test Coverage to Improve Information Return

Karl Pfeiffer—Visiting Assistant Professor, Information Sciences, Naval Postgraduate School. His current research interests include decision-making under uncertainty, particularly with regard to command and control (C2) systems; stochastic modeling of environmental impacts to weapons and communication systems; and probability modeling and numerical simulation in support of search, identification, and pattern recognition applications (e.g., complex system testing, allocation of effort for reconnaissance). [kdpfeiff@nps.edu]

Valery Kanevsky—Research Professor, Information Sciences, Naval Postgraduate School. His research interests include probabilistic pattern recognition; inference from randomly distributed inaccurate measurements, with application to mobile communication; patterns and image recognition in biometrics; computational biology algorithms for microarray data analysis; and Kolmogorov complexity, with application to value allocation for processes without saleable output. Another area of interest is in the so-called needle-in-a-haystack problem: searching for multiple dependencies in activities within public communication networks as predictors of external events of significance (e.g., terrorist activities).

Thomas Housel—Professor, Information Sciences, Naval Postgraduate School. Professor Housel specializes in valuing intellectual capital, knowledge management, telecommunications, information technology, value-based business process re-engineering, and knowledge value measurement in profit and non-profit organizations. His current research focuses on the use of knowledge-value added (KVA) and real options models in identifying, valuing, maintaining, and exercising options in military decision-making. His work on measuring the value of intellectual capital has been featured in a *Fortune* cover story (October 3, 1994) and *Investor's Business Daily*, numerous books, professional periodicals, and academic journals (most recently in the *Journal of Intellectual Capital*, 2005). [tjhousel@nps.edu]

Abstract

Effective, cost-efficient testing is critical to the long-term success of open architecture within the Navy's Integrated Warfare System. In previous research we have developed a simple, effective framework to examine the testing of complex systems. This model and its prototype decision aid provide a rigorous yet tractable approach to improve system testing, and to better understand and document the system and component interdependencies across the enterprise. An integral part of this model is characterizing test coverages on modules. Using idealized simulations of complex systems, we investigate the sensitivity of test selection strategy to the precision with which these coverages are specified. Monte Carlo analysis indicates that best-test selection strategies are somewhat sensitive to the precision of test coverage specification, suggesting significant impact on testing under fixed-cost constraint. These results provide significant insight as we extend this work with further study of real-world systems by applying, and refining, the mathematical analysis and computer simulation within this framework. The current decision-aid software will be further developed using these operational test and evaluation data, improving the fidelity of the current modeling while making available to program managers and system designers a usable and relevant tool for test-retest decisions.

Overview

In previous research we have developed a framework for describing the performance of a test suite for assessment of a complex system under repair or under routine maintenance (Pfeiffer, Kanevsky, & Housel, 2009a, 2009b, 2010). This model was then implemented in a decision support tool to investigate strategies for test selection under fixed cost or fixed reliability constraints. Construction of the model for simulation required the characterization of test coverages on modules; that is, we needed an a priori estimate of how much of the module was exercised by a particular test.

For hardware modules, this test coverage is reasonably simple to estimate (see, for example, Barford, Kanevsky, & Kamas, 2004). For software systems, however, the notion of test coverage is more problematic and may require more knowledge of the internal structure of the modules and their interdependencies (Zhu, Hall, & May, 1997). Although the notion of software testing is well studied, the characterization of test coverage can vary widely among investigators (compare, for example, Leung & White, 1991; White & Leung, 1992; Weyuker, 1998; Tsai, 2001; Rothermel, Untch, & Harrold, 2001; and Mao & Lu, 2005). Often, internal knowledge of hardware and software modules will not be available to developers of integrated test suites, particularly with commercial-off-the-shelf (COTS) technologies. The increasing use of COTS in current weapons systems (Caruso, 1995; Dalcher, 2000), coupled with the complexity of end-to-end systems (Athans, 1987; Brazet, 1993), suggests that characterizing test coverages in an open architecture system will remain a significant challenge.

How important are these test coverages to developing an effective test strategy? That is, how precisely and how accurately must we specify these coverages to evaluate effective test strategies? Extending our previous work, we investigate the sensitivity of test selection strategy to the characterization of test coverages within the system under test. Using an analytic approach to inform further modeling and simulation work, we seek to better understand how well we must specify these a priori coverage estimates in order to derive useful testing strategies.

The rest of this paper is organized as follows: Background will briefly review the framework we have developed for investigating testing strategies; Analytic Modeling and Computer Simulation Approach will discuss the analytic background and simulation approach in examining the sensitivity of information returned to the coverage specification; Simulation Results will describe simulation results and significant findings; and Conclusions and Future Work will discuss future avenues for research.

Background

In the present discussion, we define testing as the mechanism by which we trade some fixed cost (e.g., time, money) for information about the state of subcomponents and overall reliability of our system (Figure 1). In general, we seek the maximum information available for the minimum cost.

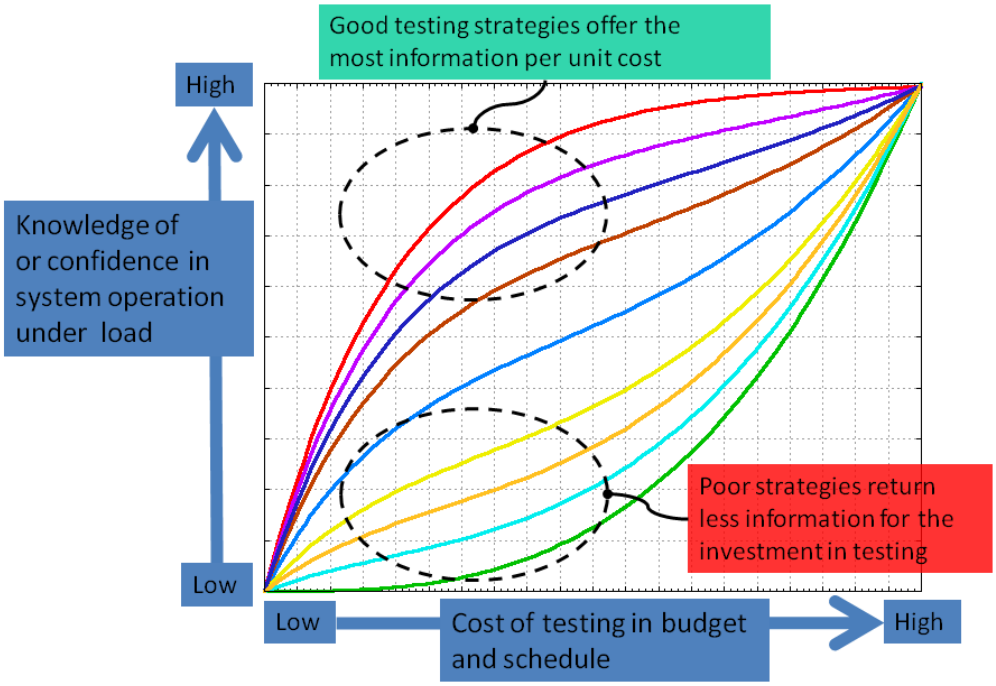

Figure 1. Information Returned for Cost of Testing Executed

Note. An idealized representation of testing strategies in terms of information returned for testing accomplished. Each solid line represents a particular testing strategy, with better strategies distinguished by steeper ascent or greater information return per unit cost.

Mathematical models proposed by von Neumann (1952) and Moore and Shannon (1956a, 1956b) shaped much of the early work on component and system reliability. An early focus on fault diagnosis, particularly in electro-mechanical systems, characterized work by Sobel and Groll (1966), Butterworth (1972), Garey (1972), Fishman (1990), and others, in what is often known as the test-sequencing problem. That is, which test sequence most cost-efficiently arrives at a correct diagnosis in a failed system?

In software engineering, we are most often faced not with a failed system, but with a large system undergoing maintenance. Testing in this situation is used to establish that no defect has been added to the system by these engineering upgrades. This regression testing, or test-retest dilemma, can be more difficult than diagnostic testing of a failed system, because by its nature, testing cannot absolutely demonstrate that no defect exists (Dijkstra, 1972). A good test suite and good testing strategy, however, can often demonstrate that a defect is *highly unlikely* in the system under test (Zhu et al., 1997).

In previous work (Pfeiffer et al., 2009a, 2009b, 2010) we have developed a unified modeling framework with risk and cost as the common tension regulating the degree of testing required. The cost of testing can be evaluated in terms of dollars, or time, or both, with an assumption that more testing is generally more costly; it is not true in general, however, that more testing always increases our knowledge of the state of our system. This knowledge is tied to risk. In this context, risk refers to the degree of certainty we can achieve (or ambiguity we can eliminate) within a fixed cost constraint or within the power or sensitivity of a given test suite.

We characterize our system under test **S** as a collection of modules $\{M_i\}$, and a suite of tests $\{T_x\}$ used to interrogate these modules (Figure 2). These tests are our means to identify defective modules, or, in the case of test-retest, to establish with high probability that no defects exist. We assume that tests return ambiguous information about the state of modules within the system; that is, no single test is likely to return perfect knowledge about a particular module.

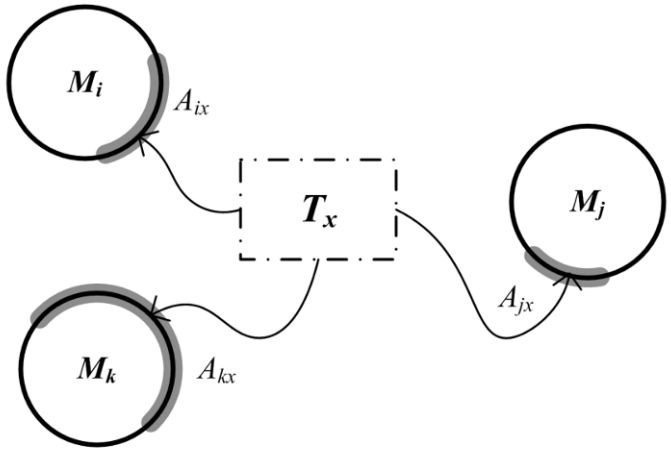

Figure 2. Simple Test Coverage on Modules

Note. Notional depiction of the coverage of T_x on S, with multiple modules exercised by this test. A FAIL result from T_x indicates that at least one of the subset $\{M_i, M_j, M_k\}$ has failed.

Each test is assumed to exercise several modules (Figure 2), and several tests may exercise the same module. In the case of several tests covering a particular module, the framework easily accounts for overlapping and disjoint coverages (Figure 3).

The model framework described in Pfeiffer et al. (2010, 2009a, 2009b) presents a useful and realistic ambiguity in two aspects. The first is that a test is rarely assumed to cover or exercise all functionality of a module. This means that when a particular test T_x passes, we know only that the region exercised by the test does not contain a defect; a defect may still exist in those regions not inspected by the test (Figure 2). A second ambiguous aspect is that when test T_x fails, several modules may be at fault (Figure 2), though such a result should significantly reduce the number of suspect modules in **S**.

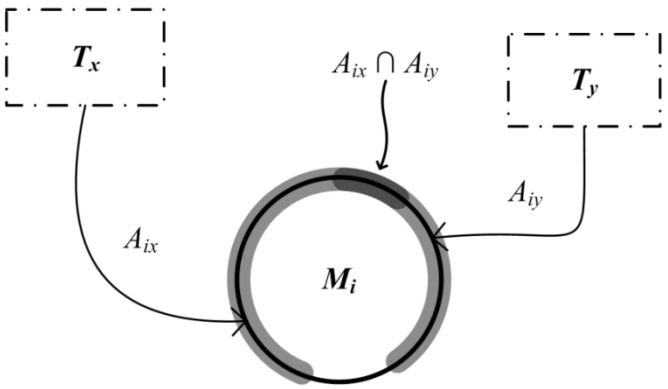

Figure 3. Overlapping Coverage Between Tests

Note. The overlapping coverage between tests T_x and T_y are characterized with the arcs A_{ix} and A_{iy}. The joint coverage is computable as the intersection of these arcs.

The vector arcs specifying test coverage are intended to lend precision to the model specification and implementation. With these vector artifacts, the overlap among tests on a single module can be precisely specified, and the disjoint regions can be similarly specified (Figure 3). In the original work (Pfeiffer et al., 2009a), we proposed that subject matter experts could estimate these coverage data as a starting point for further modeling and simulation work. In the present study, we further examine how precise these estimates should be to deliver meaningful decision support for cost-effective test strategies.

Analytic Modeling and Computer Simulation Approach

We characterize our knowledge of the system under test **S** as a vector of probabilities $\{b_i\}$ that any given module M_i is bad. Our knowledge of the system is perfect when every b_i is either 0 (absolutely good) or 1 (absolutely bad). In practice, we are unlikely to see perfect results (e.g., $b_i = 0$ or 1), though we can, with a well-designed test suite and an effective test strategy, minimize the residual information entropy of the vector (Pfeiffer et al., 2009a). This entropy is defined following Shannon (1948):

$$h_i = -b_i \log_2 b_i - (1-b_i)\log_2(1-b_i) \quad (1)$$

The initial values for $\{b_i\}$ are assumed to be available from subject matter experts or a priori failure rate estimates. Our simulation work has demonstrated that test strategy outcomes are relatively insensitive to these initial $\{b_i\}$ because of the iterative nature of this approach. That is, after a few tests have been executed, the initial vector $\{b_i\}$ moves significantly towards lower entropy (Pfeiffer et al., 2010). The test coverages connecting the tests $\{T_x\}$ to the modules $\{M_i\}$ appear to be the more relevant initial criteria in these simulations of system testing. This is another motivation for the present study.

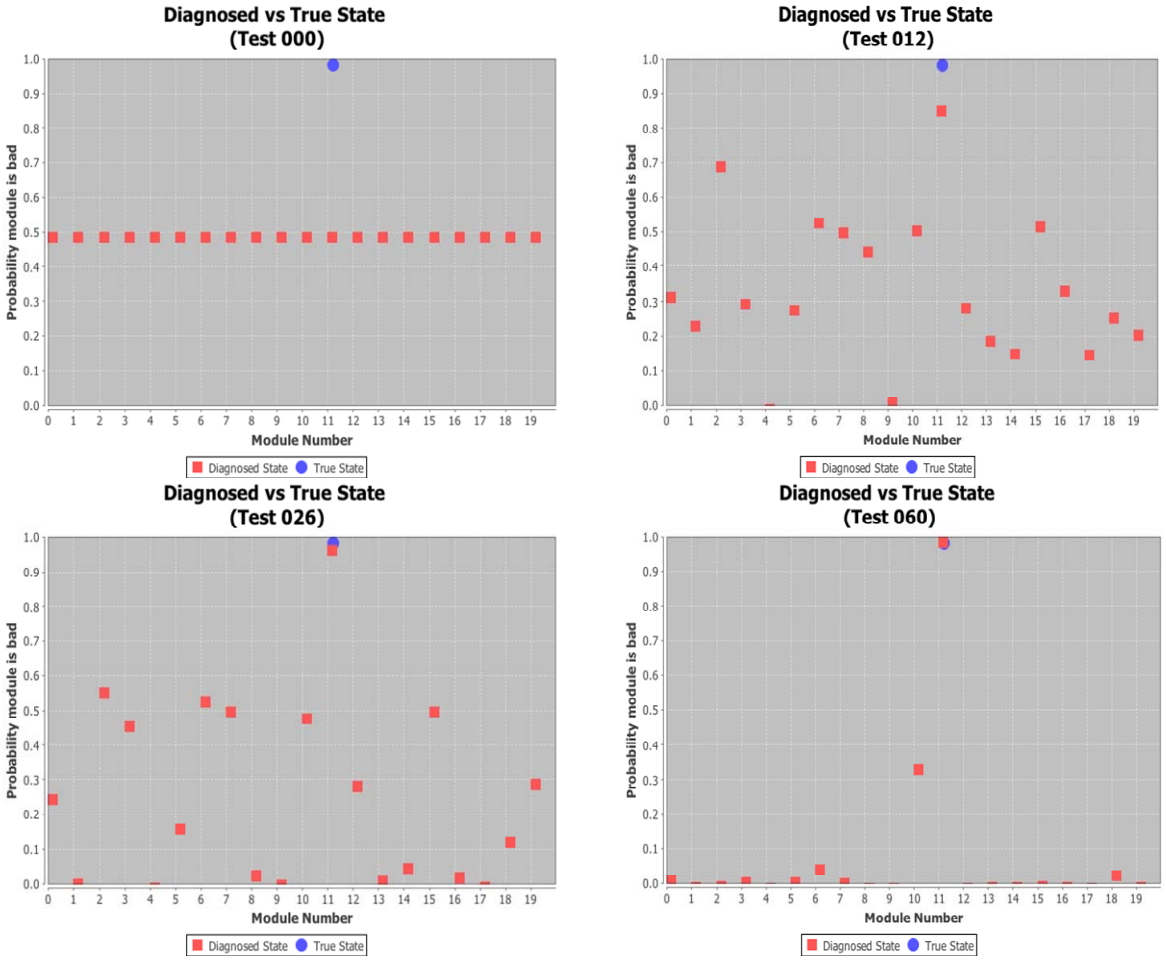

Figure 4. Diagnostic Sequence From an Idealized Simulation

Note. From simulation results in the decision support tool, this is a diagnostic sequence or trial from the Monte Carlo simulation of testing. In this trial, a single defect has been planted in Module 11 (blue ellipse), and testing improves the knowledge of the state of the system in the probability vector $\{b_i\}$ (red squares). Although we appear to have a good diagnosis by Test 026 (lower left), with Module 11 identified as bad, we can see that further testing more clearly eliminates all other modules as suspect. This would be most important in a regression test-retest scenario.

The decision support tool developed in Pfeiffer et al. (2009) simulates the testing of a complex system using minimal descriptions of tests, modules, and their connecting coverages. For idealized simulations, a range of coverages is specified between tests and modules, coupled with a target number of tests per module and modules per test. Simulations may be run with zero or more defects. The zero-defect case is particularly important for investigating test-retest or regression cases.

Each simulation typically involves a large number of trials (notionally 100 to 1000) and this facilitates examining the bounds of the idealized assumptions. The diagnostic sequence from a single trial is depicted in Figure 4. The reduction in residual entropy across the system is apparent as testing progresses from the initial state (Figure 4, upper left) to a usable diagnosis (Figure 4, lower right) with the defective module correctly

identified. The system entropy is computed as the aggregate of residual entropy associated with each module:

$$H = \sum_{i=1}^{n} -b_i \log_2 b_i - (1-b_i)\log_2(1-b_i) \tag{2}$$

After execution of a test, we update the prior probability b_i of each module M_i to the new probability b_i' based on the test outcome (PASS or FAIL):

$$b_i' = \begin{cases} \mathrm{P}(B_i \mid P_x) & \text{if } T_x \text{ passes} \\ \mathrm{P}(B_i \mid F_x) & \text{if } T_x \text{ fails} \end{cases} \tag{3}$$

These conditional probabilities are connected to test coverages through the Bayesian relations:

$$\mathrm{P}(B_i \mid P_x) = \frac{\mathrm{P}(P_x \mid B_i)\mathrm{P}(B_i)}{\mathrm{P}(P_x)} = \left(\frac{\mathrm{P}(P_x \mid B_i)}{\mathrm{P}(P_x)}\right) b_i \tag{4}$$

$$\mathrm{P}(B_i \mid F_x) = \frac{\mathrm{P}(F_x \mid B_i)\mathrm{P}(B_i)}{\mathrm{P}(F_x)} = \left(\frac{\mathrm{P}(F_x \mid B_i)}{\mathrm{P}(F_x)}\right) b_i \tag{5}$$

And these probabilities are computed with the following:

$$\mathrm{P}(P_x) = \prod_{i=1}^{n} \left[1 - \alpha_{ix} b_i\right] \tag{6}$$

$$\mathrm{P}(F_x) = 1 - \prod_{i=1}^{n} \left[1 - \alpha_{ix} b_i\right] \tag{7}$$

Knowledge of the coverage dyad $\{\alpha_{ix}\}$ is thus intrinsic to minimizing system entropy (Equation 2) and developing a cost-effective strategy for system testing. How precisely must these coverages be specified to be useful, though?

Simulation Results

In previous work (Pfeiffer et al., 2009a, 2009b), we have investigated the relative performance improvement in testing strategies using a best next test (one-test look ahead) and best next two tests (two-test look ahead). The coverages for these investigations were constructed by sampling a uniform distribution on [0.1,0.9] for each connected test and module pair.

Results from Pfeiffer et al. (2010) suggest that the best-next-two-tests strategy offers some improvement over a one-test look ahead, though the time required developing the two-test look ahead is on the order of 2.5 times the one-test strategy. A random test selection strategy was also used in this work as a baseline or no-strategy approach. Both best and best-two strategies clearly outperformed this random approach.

In this work, we also introduced an equivalent metric to residual information entropy (Equation 1) using instead the maximum probability q_i related to b_i by:

$$q_i = \max(b_i, 1 - b_i) \tag{8}$$

This measure is more intuitive than Equation 2, and represents an expected value of a replacement (or maintenance) decision with respect to a particular module. If, for example, a particular module has a $b_i = 0.70$, we may replace it knowing that this informed guess should be correct 70% of the time. This also means that in 30% of these cases we will unnecessarily replace or perform more granular debugging on this module. Our number of correct diagnoses across the system will increase as each b_i is adjusted, by testing, away from b_i=0.5 towards either 0 or 1 (Figure 4). In Pfeiffer et al. (2009a), we have shown that minimizing system entropy is approximately equivalent to maximizing the number of correct diagnoses.

In evaluating the best next test (or best next two tests), the measure (Equation 8) is aggregated as a system measure for a particular test T_x:

$$Q(T_x) = \sum_{i=1}^{n} q_{ix} \tag{9}$$

At any point in diagnostic testing, all available T_x are evaluated with Equation 9 and the largest $Q(T_x)$ indicates the next best T_x. The conditional probabilities dependent upon the specification of coverage (Equations 3–7) are intrinsic to this computation.

In simulation work using the decision support tool for complex testing, we examined the sensitivity of test strategies to the specification of test coverage within the model. Specifically, we examined both random and best-next test strategies with different specifications of coverage about a mean coverage per module of 0.7 or 70%. All $\{b_i\}$ were initialized with a maximum entropy value of $b_i = 0.5$, consistent with our assertion that the iterative simulation is relatively insensitive to the initial $\{b_i\}$ (Pfeiffer et al., 2009). All runs were made with zero defects present, to emphasize the utility of this work for test-retest or regression scenarios.

The fundamental connection of coverage to information (Equations 2, 6, and 7) suggests that, in general, more coverage per test should yield more information. For this investigation, four specifications were used: a uniformly distributed coverage among tests and modules from 50% to 90%, or [0.5,0.9]; and fixed coverages of 50%, 70% and 90%. A nominal 300 trials were used for this work, though the model output statistics were examined for 1000 trials without significant difference.

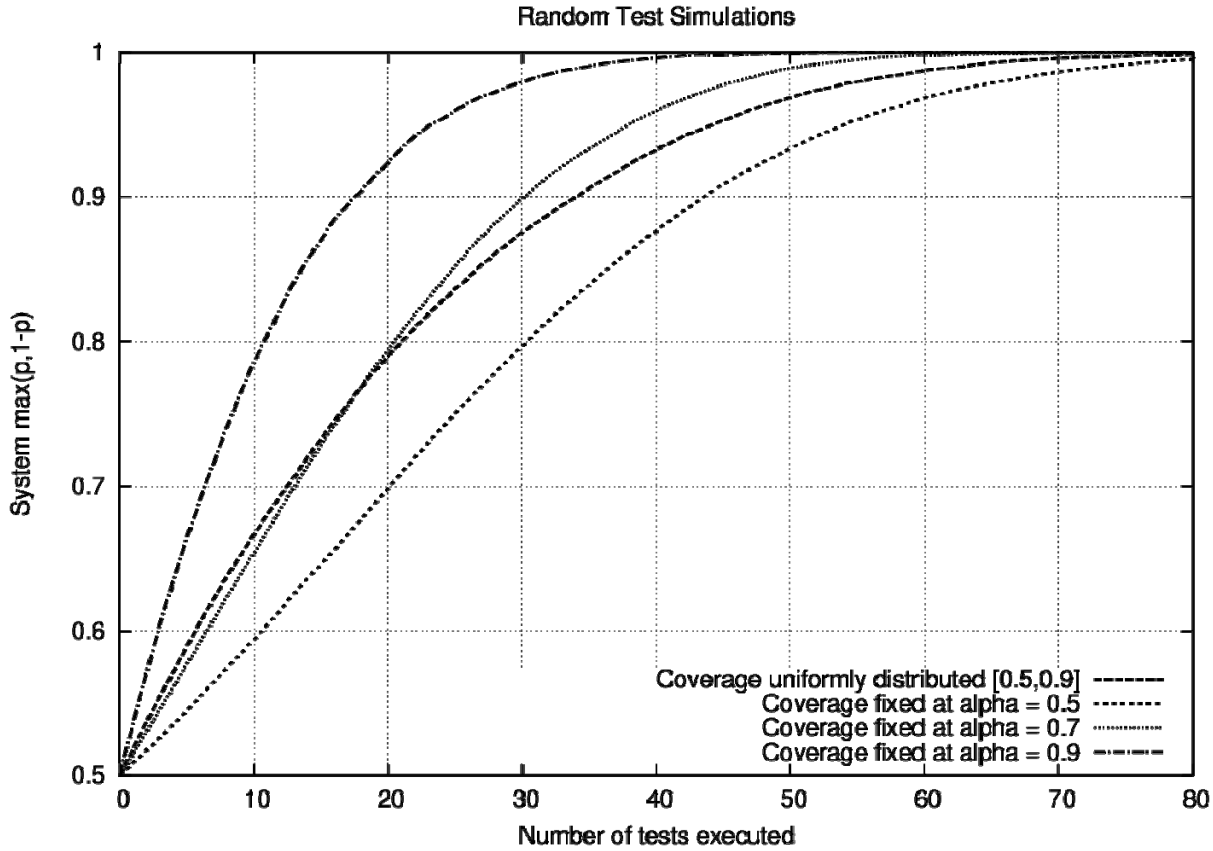

Figure 5. No-Strategy (Random) Testing Simulations

Note. Simulation results using different coverage specifications for a random test selection or a no-strategy approach show, in general, more information (smaller *H* or larger *Q*) when coverage per test increases.

The random strategy simulations (Figure 5) do indeed show more information returned when the coverage is fixed at 90%, and significantly less information returned when the coverage is fixed at 50%. Perhaps more interesting is the comparison of fixed coverage at 70% with a random coverage on the interval [0.5, 0.9], which has a mean of 70%. These runs for the random strategy appear quite similar up to about the first 20 tests (Figure 5). At this point, the fixed coverage at 70% appears to outperform the random coverage on [0.5, 0.9].

In contrast to the no-strategy approach, the best next test simulations (Figure 6) show pronounced differences among coverage specifications. The fixed 90% coverage run appears somewhat better in information returned per test execution, though interestingly the 70% and 50% coverage runs appear to underperform compared to the random simulation (Figure 5). These differences are not consistent over the test execution profile, however. This is particularly interesting because both the random and best next simulations were run with random number generators seeded identically; thus, the differences highlighted between Figures 5 and 6 are solely a function of the differences in the rate of information returned by the two strategies.

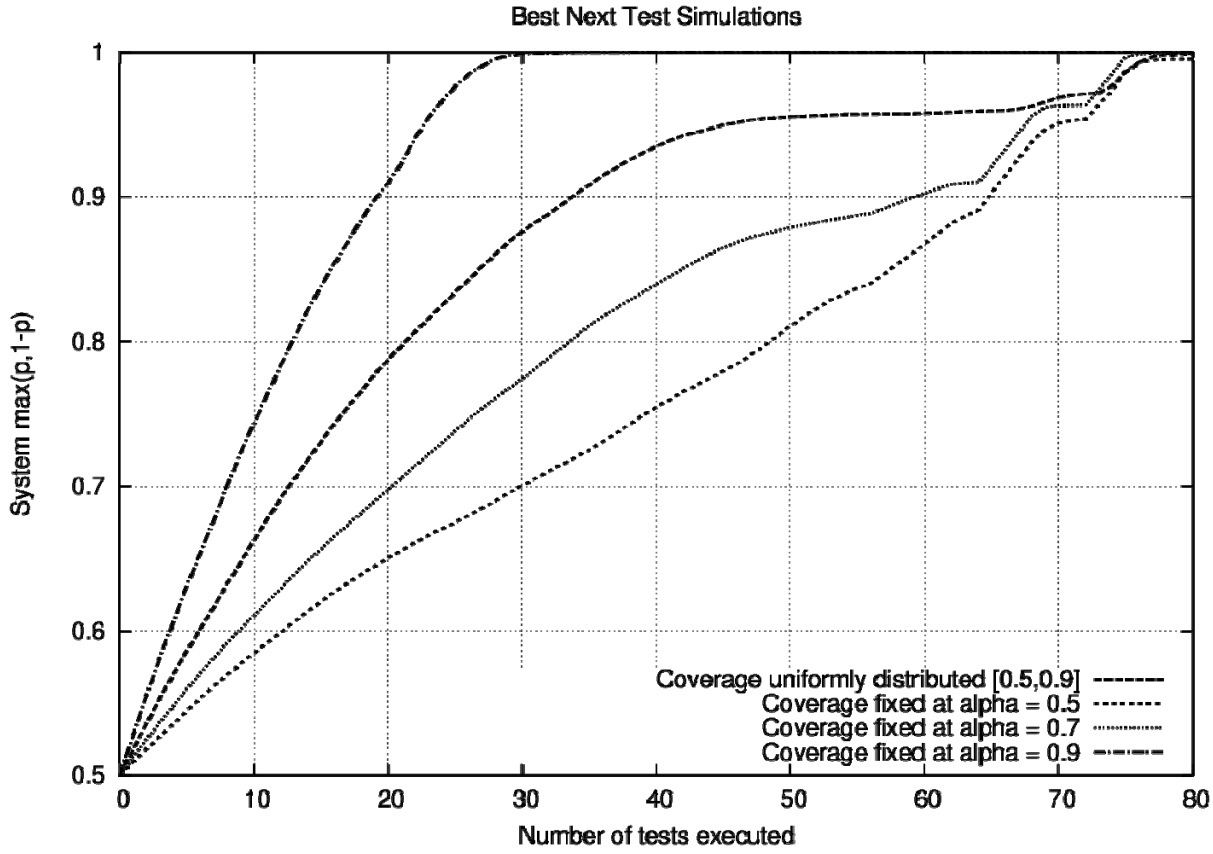

Figure 6. Best-Next Testing Simulations

Note. Simulation results using different coverage specifications for a best-next test selection show marked differences among runs. While it is still true, in general, that more coverage yields more information, the uniform random coverage [0.5, 0.9] and low-coverage (fixed at 0.5) runs appear to underperform when compared to the random test selection (Figure 5).

All of these simulations were conducted with no defects present, and identification of a defect tends to sharply alter the information profile in a run; intuitively, this is because the first FAIL result in test execution should sharply reduce the number of suspect modules across the system. In the absence of defects, it is possible, particularly as the testing progresses and alters the vector $\{b_i\}$, that the differences among tests in information returned (Equation 9) may vary widely on a one-test look ahead.

Consistent with our previous studies (e.g., Pfeiffer et al., 2010), we made a two-test look-ahead simulation to better assess the sensitivity of coverage specification to test selection strategy. Overall results (Figure 7) show little improvement from the one-test strategy (Figure 6), though the best performer (fixed coverage at 90%) does show some early improvement over the first ten tests executed. These results do suggest that the effectiveness of a test selection strategy is connected to the precision with which the test coverages are specified.

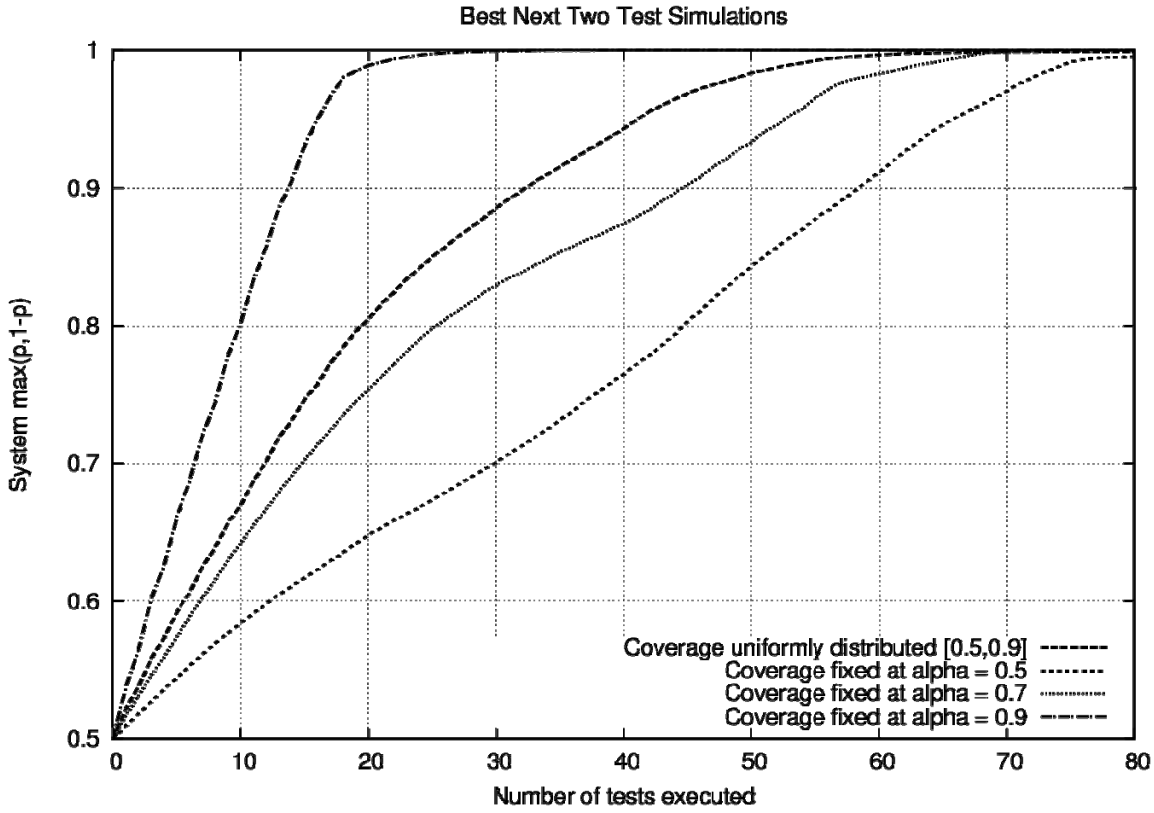

Figure 7. Best-Next-Two Test Simulations

Note. Simulation results using different coverage specifications for a best next two-test selection also show marked differences among runs, similar to the next-best-test simulations (Figure 6). Note that the best coverage specification (fixed at 90%) does significantly improve with a two-test strategy, though the other fixed runs and uniform random run show little improvement.

We should keep in mind that these idealized simulations place no constraint on the overlap among coverages on tests. We verified in simulation log files that significant overlap among tests (e.g., Figure 3) increased as the fixed coverage progressed from 50% to 90%. The impact of this overlap on test selection appears most dramatic in the non-random simulations (Figures 6 and 7) as best next and best-next-two strategies make better use of the information returned by each test. This overlap also means that many or most of the modules in the idealized system were completely covered by some number of tests in the test suite, leading to the near perfect information after about 30 tests have been executed (Figures 6 and 7). We expect real-world systems would rarely achieve perfect coverage regardless of the number of tests available, because of the nature of complex systems. For anything but a trivial component or module, we are unlikely to construct a set of tests that cover *all* branch paths or all of the input and output space.

An obvious conclusion from these results is that more coverage per test appears to improve the testing process. While this result may be somewhat intuitive, an equally useful and interesting result is that better specification of coverages increases the benefits from a rigorous test selection strategy. Further investigation with both real coverage data from

operational systems and idealized simulations with more complex distributions of coverage should yield additional insights into this problem.

Conclusions and Future Work

Effective, cost-efficient testing and re-testing is critical to the long-term success of open architecture. Using the framework for complex system testing developed in Pfeiffer et al. (2009), we have conducted additional simulation work to examine the sensitivity of test selection strategies to the specification of test coverages. Characterization of test coverages, particularly for software-intensive systems, remains a difficult challenge (Zhu et al., 1997), though in this work we did not address this problem directly. Rather, in the framework of our existing model, we have examined the impact of precision in specifying coverage on the information returned per test.

Not surprisingly, the test selection strategies we have investigated are quite sensitive to various specifications of test coverage. In these idealized simulations, less precision in coverage specification appears to flatten the information returned per test. Incorporation of more real test data from operational systems should help with further investigation of this point. The idealized work permitted complete (100%) or near-complete coverage of modules with overlapping tests (particularly for fixed coverages of 70% to 90%) that would be unlikely in real-world testing. In fact, we speculate that in simulating real-world systems, we will likely encounter test scenarios where no module is completely covered by testing, and real coverages are at best 95% to 98% with all overlapping coverages considered.

The decision support tool used in these simulations could also be further refined to permit specifying test-to-module coverages in terms of a collection of triangle or uniform distributions. This should better capture subject matter expertise in a quantitative manner. For example, a quasi-idealized simulation of a Garage Door Opening System could work with a specification that the *Object Detection Test* exercised at least 30% of the *Remote Control Module*, though no more than 50%, with a mode or mean of 40%. While these numbers may still be speculative or notional on the part of the subject matter expert, these confidence bounds would represent useful input to the overall test selection strategy.

References

Athans, M. (1987). Command and control (C2) theory: A challenge to control science. *IEEE Trans. Autom. Control, 32*(4), 286–293.

Barford, L., Kanevsky, V., & Kamas, L. (2004). Bayesian fault diagnosis in large-scale measurement systems. In *IMTC 2004: Instrumentation and Measurement Technology Conference* (pp. 1234–1239). Como, Italy: IEEE.

Brazet, M. D. (1993). AEGIS ORTS—The first and future ultimate integrated diagnostics system. *Aerospace and Electronic Systems Magazine, 9*(2), 40–45.

Butterworth, R. (1972). Some reliability fault-testing models. *Operations Research, 20*(2), 335–343.

Caruso, J. (1995). The challenge of the increased use of COTS: A developer's perspective. In *Proceedings of the Third Workshop on Parallel and Distributed Real-Time Systems* (pp. 155–159). Santa Barbara, CA: IEEE.

Dalcher, D. (2000). Smooth seas—rough sailing: The case of the lame ship. In *Seventh International Conference on Engineering of Computer Based Systems (EBCS 2000)* (pp. 393–395). Edinburgh, Scotland: IEEE.

Dijkstra, E. (1972). Notes on structured programming. In O.-J. Dahl, E. Dijkstra, & C. Hoare, *Structured Programming* (pp. 1–72). London, England: Academic Press.

Fishman, G. S. (1990). How errors in component reliability affect system reliability. *Operations Research, 38*(4), 728–732.

Garey, M. (1972). Optimal binary identification procedures. *SIAM Journal on Applied Mathematics, 23*(2), 173–186.

Leung, H., & White, L. (1991). A cost model to compare regression test strategies. In *Proceedings of the Conference on Software Maintenance* (pp. 201–208). Sorrento, Italy: IEEE.

Mao, C., & Lu, Y. (2005). Regression testing for component-based software systems by enhancing change information. In *Proceedings of the 12th Asia–Pacific Software Engineering Conference (APSEC'05)* (pp. 1–8). IEEE.

Moore, E., & Shannon, C. (1956a). Reliable circuits using less reliable relays, Part I. *Journal of the Franklin Institute*, 191–208.

Moore, E., & Shannon, C. (1956b). Reliable circuits using less reliable relays, Part II. *Journal of the Franklin Institute*, 281–298.

Pfeiffer, K. D., Kanevsky, V. A., & Housel, T. J. (2009a). *Reducing the cost of risk-based testing: Management of testing options to manage risk in test and evaluation.* Naval Postgraduate School. Monterey, CA: Acquisition Research Program.

Pfeiffer, K. D., Kanevsky, V. A., & Housel, T. J. (2009b). Testing of complex systems. *INFORMS Annual Meeting* (pp. 130–135). San Diego, CA: INFORMS.

Pfeiffer, K. D., Kanevsky, V. A., & Housel, T. J. (2010). An information–theoretic approach to software test-retest problems. *Acquisition Research Symposium* (pp. 100–120). Monterey, CA: Acquistion Research Program.

Rothermel, G., Untch, R. H., & Harrold, M. J. (2001). Prioritizing test cases for regression testing. *IEEE Transactions on Software Engineering, 27*(10), 929–948.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423; 623–656.

Sobel, M., & Groll, P. (1966). Binomial group-testing with an unknown proportion of defectives. *Technometrics, 8*(4), 631–656.

Weyuker, E. (1998). Testing component-based software: A cautionary tale. *IEEE Software, 15*(5), 54–59.

White, L., & Leung, H. (1992). A firewall concept for both control-flow and data-flow in regression integration testing. In *Proceedings of the Conference on Software Maintenance* (pp. 262–270). IEEE.

Zhu, H., Hall, P. A., & May, J. H. (1997). Software Unit Test Coverage and Adequacy. *ACM Computing Surveys, 29*(4), 366–427.