

Generalized Coverage Criteria for Combinatorial Sequence Testing

Achiya Elyasaf¹, Eitan Farchi², Oded Margalit², Gera Weiss², and Yeshayahu Weiss²

Abstract—We present a new model-based approach for testing systems that use sequences of actions and assertions as test vectors. Our solution includes a method for quantifying testing quality, a tool for generating high-quality test suites based on the coverage criteria we propose, and a framework for assessing risks. For testing quality, we propose a method that specifies generalized coverage criteria over sequences of actions, which extends previous approaches. Our publicly available tool demonstrates how to extract effective test suites from test plans based on these criteria. We also present a Bayesian approach for measuring the probabilities of bugs or risks, and show how this quantification can help achieve an informed balance between exploitation and exploration in testing. Finally, we provide an empirical evaluation demonstrating the effectiveness of our tool in finding bugs, assessing risks, and achieving coverage.

Index Terms—Bayesian risk-Reduction, behavioral programming, combinatorial test design, model-based testing, sequence testing, test coverage, test generation, test optimization.

I. INTRODUCTION

TESTING faces one of its biggest challenges is knowing when to stop. How many tests are necessary? Should we aim to test every possible input or focus on inputs that are more likely to cause errors? Is it better to allocate our resources uniformly throughout the application or invest our efforts in the most critical paths?

To address these problems in black-box testing, several techniques have been proposed for providing a set of input vectors for testing the system. These techniques define coverage criteria on the possible input vectors, thus dramatically reducing the nearly infinite possibilities. These techniques include, for example, decision table testing, domain analysis, all-pairs testing, boundary-value analysis, and more [17].

The aforementioned techniques test the possible system behaviors by analyzing the effect of the input vector on the system

behavior and then clustering the input-vector space. Another type of technique clusters the system-behavior space, either by clustering the behaviors manually (e.g., by use cases or user stories) or systematically. Kuhn and Higdon¹ proposed a coverage criterion that is based on sequences of events of a system, called *t-way sequence coverage*, where sequences of any t events cluster the system behavior [11], [25], [28].

In this paper, we continue and extend Kuhn and Higdon's approach for a direct clustering of the system behavior best defined as a sequence of events. More generally, we propose to address the above testing challenge by tackling the following issues: (1) how to cluster the space of event sequences that need to be covered; (2) how to cover this space with a finite, relatively small number of tests; and (3) how to methodically explore and exploit knowledge from previous tests to optimally reduce bug risks.

Our answer to the first challenge is to define generalized coverage criteria for sequences of events. With examples from various domains, we demonstrate how our generalizations can naturally express useful coverage criteria that extend Kuhn and Higdon's vision by allowing more types of applications. We also show how the proposed extension enables the inclusion of standard combinatorial-test-design (CTD) methods and Kuhn and Higdon's criterion under a unified formal vocabulary [24]. Our approach to specifying coverage criteria is automata-based. Each automaton in our framework represents a set of tests considered equivalent by testers. For example, two tests in the same set are likely to both fail or both pass, presumably because the test designer expects that the probability that the system under test handles them differently is low. We present various examples from different systems to demonstrate the naturalness of this approach and show that testers can construct coverage requirements to suit different domains irrespective of a specific test model.

Our answer to the second challenge is a publicly available tool for test-suites generation that demonstrates how our generalized criteria can be used in practice. Our sequence coverage paradigm includes the definition of a set of automata. Testers can provide their test coverage criteria by providing a ranking function that counts the number of automata (in their coverage criteria) that pass at least one of the tests in a given test suite.

Manuscript received 3 May 2022; revised 26 April 2023; accepted 13 May 2023. Date of publication 24 May 2023; date of current version 15 August 2023. This work was supported by Israeli Science Foundation under Grant 2714/19. Recommended for acceptance by S. Chandra. (Corresponding author: Achiya Elyasaf.)

Achiya Elyasaf is with the Software and Information Systems Engineering Department, Ben-Gurion University of the Negev, Beersheba 84105, Israel (e-mail: achiya@bgu.ac.il).

Eitan Farchi is with the IBM Haifa Research Lab, Haifa 3498825, Israel (e-mail: farchi@il.ibm.com).

Oded Margalit, Gera Weiss, and Yeshayahu Weiss are with the Computer Science Department, Ben-Gurion University of the Negev, Beersheba 84105, Israel (e-mail: odedm@post.bgu.ac.il; geraw@cs.bgu.ac.il; weissye@post.bgu.ac.il).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TSE.2023.3279570>, provided by the authors.

Digital Object Identifier 10.1109/TSE.2023.3279570

¹We attribute this contribution to Kuhn and Higdon since the book "Practical Combinatorial Testing" by Kuhn, Kacker, and Lei [25] references an earlier work from 2009 by D.R. Kuhn and J.M. Higdon entitled "Testing Event Sequences". However, we could not obtain a copy of this paper.

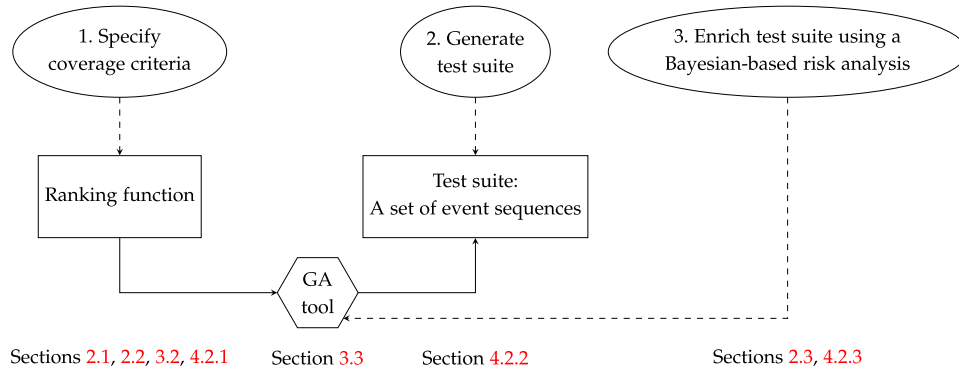


Fig. 1. A schema of our approach. It consists of three phases; each has an activity (oval) that generates an artifact (box). The test suite is generated using the GA-based, test-suite-generation tool (hexagon) that takes the system-behavior model and the ranking function and returns a set of sequences of events that achieve a high coverage rank. Below the tool and each phase are their relevant section numbers.

A genetic-algorithms (GA) based technique is then applied to produce high-ranking test suites.

Our answer to the third challenge of balancing exploitation and exploration is as follows. We acknowledge that the specification of the coverage criteria is not dichotomic, i.e., two tests may belong to the language of the same automaton, but one may exhibit a problem while the other does not. We thus apply a Bayesian approach. The Bayesian framing of the testing process yields a mathematical formula for balancing exploration and exploitation of the knowledge obtained by tests. Specifically, we consider each class of equivalent tests as a Bernoulli random variable with a certain probability of hitting a bug. We then show how to measure and use these probabilities to maximize the likelihood of finding new bugs.

We propose a methodology with three activities (depicted in Fig. 1) and demonstrate it through our test-suite-generation tool. Given a model of the system behavior (specified as, e.g., automata), it starts by specifying the ranking function. The two artifacts are then used to generate a test suite (i.e., a set of event sequences) that achieves a high coverage ranking. Finally, a Bayesian-based approach is applied to balance the need to exploit knowledge from previous tests and explore new areas. We envision practitioners defining their domain-specific coverage criteria and selecting predefined generalized criteria from a library. The selection can be based on, for example, the type of implementation at hand, the project phase, the code libraries in use, etc.

The paper is organized as follows. We begin with a theoretical presentation of our approach in Section II. Section III presents our publicly available tool and demonstrates our approach and methodology through this tool. This section also provides a qualitative evaluation of our approach, demonstrating its applicability for testing. We conclude with a quantitative evaluation of our approach in Section IV.

II. A MODEL-BASED APPROACH TO SEQUENCE TESTING

In this section, we present the conceptual approach proposed in this paper. We begin with our new coverage criteria and continue with a formulation of a Bayesian approach to risk

TABLE I
A SUMMARY OF THE MAIN MATHEMATICAL NOTATIONS

Symbol	Explanation
Σ	Alphabet of test actions like ‘ClickLoginButton’.
a	Length of the test sequences (a number or *).
$P \subseteq \Sigma^a$	A regular language of possible test sequences.
I	An index set for naming the coverage sets.
$C(i)$, for $i \in I$	A set of tests that cover some aspect.
$C = \{C(i)\}_{i \in I}$	A coverage criterion specifying the aspects.

assessment. Note that these contributions are independent and can be used separately. To simplify the reading of this section, the main mathematical notations we use are summarized in Table I.

A. Generalized Coverage Criteria

We propose a new type of coverage criteria that generalizes the t -way combinatorial sequence coverage criterion [25] and the classical t -way coverage. The generalized framework gives testers tools for infusing domain knowledge into the test design. For example, based on previous experience with the system or an understanding of its implementation, a tester may want to focus on testing some race conditions on the access of one shared resource only when another shared resource is held. We define our generalized coverage notion and then motivate it with examples.

We consider the test model as a set of all possible tests. Each is a sequence of actions that may be applied to the system under test (SUT) using an external interface available for testing. More formally, it is an abstract regular set $P \subseteq \Sigma^a$ where Σ is an alphabet that represents the possible actions, and $a \in \mathbb{N} \cup \{*\}$, meaning that each test may be of any length.

Once P is defined, we would ideally have liked to execute each $t \in P$ against the SUT, but this may be impossible to realize, as P may be huge or even infinite. We, therefore, propose to use coverage criteria to overcome that. We define coverage criteria C using an indexed set of languages $C = \{C(i)\}_{i \in I}$, $C(i) \subseteq \Sigma^a$.

The following definition captures what constitutes covering P using C .

Definition 1 (Coverage): A set of tests $S \subseteq P$ is said to cover a test-model $P \subseteq \Sigma^a$ under the coverage criteria $C = \{C(i)\}_{i \in I}$, $C(i) \subseteq \Sigma^a$ if $\forall i \in I : (C(i) \cap P \neq \emptyset \Rightarrow C(i) \cap S \neq \emptyset)$ and $\bigcup_{i \in I} C(i) = \Sigma^a$.

As explained above, a tester may want to focus on testing some race conditions on shared resources. In such a case, Σ^a will represent sequences of reads and writes to the shared resources of interest. The $C(i)$ s will specify specific races of interest, e.g., two consecutive writes to a shared table by one process, with a read occurring after the first write and before the second by another process. To describe such requirements more easily by the tester, domain-specific languages (DSL) may be used.

Coverage is used to analyze and identify missing tests and report on the progress of the test effort. To allow for the latter, we also define the percentage of coverage obtained.

Definition 2 (Coverage Ratio): Given a finite I , a set of tests $S \subseteq P$ for a test-model $P \subseteq \Sigma^a$, and a coverage criterion $C = \{C(i)\}_{i \in I}$, $C(i) \subseteq \Sigma^a$, we define the coverage ratio as

$$\Gamma_C(S, P) = \frac{|\{i \in I : C(i) \cap S \neq \emptyset\}|}{|\{i \in I : C(i) \cap P \neq \emptyset\}|}.$$

A few comments on the above definitions are in order: Based on test concerns, the budget allocated for testing, and the expected usage of the system, it is natural in some applications to require that each test falls within some $C(i)$ and that $P \subseteq \bigcup_{i \in I} C(i) \subset \Sigma^a$. To accommodate that, we always implicitly assume that in practice, there is another coverage requirement $\Sigma^a \setminus \bigcup_{i \in I} C(i)$, even if it has an empty intersection with P .

Note that a coverage requirement may be a single test. Typically, that may represent a happy path scenario the system should meet. For example, if a single user opens a file that exists, the user has read permission and reads one byte successfully from the opened file.

Another natural expectation is that $\{C(i)\}_{i \in I}$ is a partition. We chose not to enforce that because this requirement may limit the flexibility of testers to define their test concerns freely. Once $\{C(i)\}_{i \in I}$ is defined, a partition of $\bigcup_{i \in I} C(i)$ can be easily achieved if desired by considering coverage requirements of the form $\tilde{C}(i) = C(i) \setminus \bigcup_{i' < i} C(i')$, assuming an (arbitrary) order over the index set. As the following example shows, partitioning the coverage requirements is sometimes unnatural. Consider condition coverage, for example, where each condition in the SUT should be evaluated as true by some test and false by another. Given n conditions in the software, we will get $2n$ coverage requirements, namely, that the first condition is true, etc. If the conditions are implemented as nested if-statements we will have tests that simultaneously have the first and the second condition as, say, true. Thus, condition coverage requirements do not result in a partition.

Practicality, covering the criterion $\{C(i)\}_{i \in I}$, $C(i) \subseteq \Sigma^a$ may require excessive resources that are not available to testers. The following definition formalizes a common practice testers use

to overcome this situation. Specifically, it formalizes the notion of relaxing coverage requirements:

Definition 3 (Coverage Relaxation): A coverage criterion $\{C'(i)\}_{i \in J}$, $C'(i) \subseteq \Sigma^a$ is a relaxation of a coverage criterion $\{C(i)\}_{i \in I}$, $C(i) \subseteq \Sigma^a$ if J is a partition of I and if for $j \in J$, $j = \{i_1, \dots, i_k | i_l \in I, l = 1, \dots, k\}$ then $C'(j) = C(i_1) \cup \dots \cup C(i_k)$.

Clearly, $|J| \leq |I|$ in the above definition, thus reducing the resources required to achieve the relaxed coverage criterion. Consider code coverage, for example: When full coverage of all lines of code is impossible, people usually settle for covering at least one line of code for each function. In our context, this translates to a partition of I as defined above. Yet another relaxation example is found in synchronization coverage [6]. In synchronization coverage, there are coverage requirements $C(l, p_h^1, p_h^2)$ for each lock l , and each program location p_h^1 and p_h^2 that attempts to hold the lock l . The coverage requirement $C(l, p_h^1, p_h^2)$ requires a sequence of events in which lock l is held by p_h^1 and then program location p_h^2 attempts to hold the lock. Due to the availability of testing resources, the tester may want to relax this coverage requirement by only requiring some sequence of events in which any program location holds lock l . Then another program location attempts to obtain the lock l . This will result in a new coverage requirement $C(l)$ in which $C(l) = \bigcup_{p_h^1, p_h^2} C(l, p_h^1, p_h^2)$, which is a relaxation of the original synchronization coverage. In that way, coverage relaxation lets the testers control the testing effort.

B. Coverage Criteria Examples

In this section, we motivate our definitions by showing that they naturally extend existing coverage notions and how they allow for formal specifications of best practices. We will later examine specific use cases in Section IV.

We first show how our definitions generalize the classic t -way coverage on finite test spaces. Such coverage models have gained wide usage in the industry and are generally known as *Combinatorial Test Design* [24].

Example 1 (Classic t -Way Coverage): The classic t -way coverage is defined on finite spaces, i.e., when $P \subseteq \Sigma^n$ for $n \in \mathbb{N}$. In this case, for a small $t \leq n$ and for each choice of t indexes $i_1, \dots, i_t$, $1 \leq i_l \leq n$, and t letters $\sigma_1, \dots, \sigma_t \in \Sigma^t$, the coverage criterion is defined by $C(i_1, \dots, i_t, \sigma_1, \dots, \sigma_t) = \{T \in \Sigma^n : T[i_k] = \sigma_k \text{ for } k = 1, \dots, t\}$.

The following examples motivate the above definitions, focusing on test design in which the order of occurrences of events is essential. We outline how the approach generalizes Kuhn and Higdon's sequence coverage criterion [25], starting by demonstrating how to express Kuhn and Higdon's criterion using our framework.

Example 2 (Kuhn and Higdon's Coverage Criterion): Under Definition 1, Kuhn and Higdon's definition of t -sequence coverage [25] is obtained by taking $I = \Sigma^t$ and $C(\sigma_1 \dots \sigma_t) = \mathcal{L}(\Sigma^* \sigma_1 \Sigma^* \dots \Sigma^* \sigma_t \Sigma^*)$, where $\mathcal{L}$ is the set of words (language) defined by the regular expression.

In addition to Kuhn and Higdon's coverage criterion, our framework allows for defining richer criteria. Indeed, we can

consider criteria that are not precisely t -sequence coverage. We can, for example, require that σ_i does not appear unnecessarily, but only once. This leads to the following customized definition.

Example 3 (Like Kuhn and Higdon's Coverage Criterion): Take $I = \Sigma^t$ and $C(\sigma_1 \cdots \sigma_t) = \mathcal{L}(\Sigma'^* \sigma_1 \Sigma'^* \cdots \Sigma'^* \sigma_t \Sigma'^*)$ where $\Sigma' = \Sigma \setminus \{\sigma_1, \dots, \sigma_t\}$.

For example, with a SUT domain in mind, we would like to test all permutations of events in which a message is sent before receiving it.

Example 4 (Message Order): Consider a language Σ that contains subsets of send messages events $S \subseteq \Sigma$ and subsets of receive messages events $R \subseteq \Sigma$. We then consider $C(s, r) = \{T \in \Sigma^n : T[i] = s, T[j] = r \text{ for some } 1 \leq i < j \leq n\}, s \in S, r \in R$ as our set of coverage requirements. Note, of course, that an event can be either “send”, “receive,” and maybe even “other,” but $R \cap S = \emptyset$.

Such customized coverage criteria can be derived from known concurrent bug patterns or specific test concerns. Practice shows that this improves the quality of testing [16].

A more concrete SUT will lead to further customization of coverage requirements. Consider for example a bank customer with two accounts. A transaction may reduce m dollars from the first, and another may increase the second account by m dollars. We are concerned that inconsistency may arise if the transaction is stopped in the middle. This leads to the following customized coverage requirement.

Example 5 (Transaction Safety): We define Σ as the set of possible transactions, $D \subseteq \Sigma$ as the set of transactions that reduce money, $A \subseteq \Sigma$ as the set of transactions that add money, and $\Sigma' = \Sigma - (D \cup A)$. We capture our test concern by $C(d, a) = \mathcal{L}(\Sigma' d \Sigma' a \Sigma'), d \in D, a \in A$.

Many times coverage is motivated by symmetries. When we are required to see event σ_1 and then event σ_2 , we implicitly claim that all permutations of events occurring before σ_1 are equivalent for the purpose of revealing a problem which is basically a symmetry claim (we also argue that all permutations between σ_1 and σ_2 as well as all permutations after σ_2 are equivalent). Our framework supports the utilization of expected SUT symmetries in ways other than symmetries on execution order, as illustrated in the following example.

Suppose we would like to check that a tic-tac-toe game-playing application works. There are 255,168 ways to play a game, i.e., put alternate Xs and Os on free squares in the 3×3 board, where the game ends when a player wins or all cells are marked (tie). Nevertheless, if we assume the algorithm should work the same for symmetric boards (under eight rotations and reflections), we can considerably reduce the test space to 26,830.

Example 6 (Using Symmetries to test an $n \times n$ Board Game): Formally, the set, P , of equivalence classes of sequences of game-plays in an $n \times n$ game is $P = n^{2!}/\sim$ where $\sim$ is an equivalence relation over the eight rotations and reflections. In our setting, a coverage criterion is an equivalence class: $C([w]) = [w]$. We use the common notations where w is a game sequence and $[w]$ is the equivalence class that w belongs to.

To illustrate the construction of $C([w])$, we refocus on tic-tac-toe with $n = 3$, which means $n^2 = 9$ squares. We notice that the first move has only three possibilities (up to symmetry):

1	2	3
4	5	6
7	8	9

Fig. 2. A Tic-Tac-Toe board.

center, corner, or center of an edge. If we number the places as 1, 2, ..., 9 as shown in Fig. 2, then the first move can only be from the symmetry equivalence classes [1], [2], or [5]. Concretely, a move from [1] could be 1, but it also could be 3, 7, or 9. As we assume that on each equivalence set, the algorithm either makes a mistake or not, then it is enough to test, say 7, to test a game with the first movement coming from $[1] = \{1, 3, 7, 9\}$. We will relax this assumption in Section II-C by introducing a Bayesian approach representing our confidence that an equivalence class does not contain a bug.

To construct the equivalence classes $C([w])$, we apply the symmetries inductively to the game's evolution. The above symmetry application on the first move of the game reduces the space to be tested from $9!$ to $3 \cdot 8!$, which is only $1/3$ of the space. Furthermore, we can reduce it even further to $8 \cdot 7! + 16 \cdot 6!$ (which is only $1/7$ of the space) by noticing that:

- 1) after playing 1, the next move can only be 2,3,5,6, or 9.
 - a) after playing 1,5 only 2,3,6, or 9 can be played.
 - b) after playing 1,9 only 2,3,5, or 6 can be played.
- 2) after playing 2, the next move can only be 1,4,5,7, or 8.
 - a) after playing 2,5 only 2,3,6, or 9 can be played.
- 3) after playing 5, the next move can only be 1 or 2.
 - a) after playing 5,1 only 2,3,6, or 9 can be played.

We proceed this way to define an equivalence set over the sequence of entire game plays, which define the equivalent classes $C([w])$. Note that the permutation requirement can be encoded by having a finite automaton with $9!$ states. Adding the constraints above is straightforward.

Take $I = n^{2!}/\sim$ and $C([w]) = [w]$ where $\sim$ is an equivalence relation of the eight rotations and reflections, $n^{2!}/\sim$ denotes the set of equivalence classes of this relation and $[w]$ is the equivalence class of w .

C. A Bayesian Risk-Reduction Approach

Exhaustive execution of P is hard or even impossible as P may be infinite. If I is finite, executing representatives from $C(i)_{i \in I}$ is a feasible alternative to exhaustive execution of P , though even that may be too hard due to the size of I and difficulties in generating elements of $C(i)_{i \in I}$. Another challenge is that even if we execute a test in $C(i)$, we are not guaranteed, in general, that there are no bugs that can be exposed by executing another test in $C(i)$. There are several possible reasons for that:

- 1) The definition of $C(i)_{i \in I}$ is based on the tester's domain knowledge that may mistake. For example, the assumption in the tic-tac-toe example (Example 6), namely, that mistakes are invariant under the eight possible symmetries, may be wrong. Thus, we may expose a problem by playing the upper left corner and then the center while not revealing it when playing the upper right corner and then the center.

- 2) In practice, we do not control all of the SUT's inputs. Specifically, it is difficult to control and specify the environment configuration and state. For example, a test that writes to a file may control whether or not the file exists and that we have written permission. However, the level of the operating system or the state of the memory garbage collector may not be a parameter controlled by the test. As a result, our confidence of not having a problem due to the execution of an element in $C(i)$ only increases when repeatedly running elements in $C(i)$.

We thus apply a Bayesian approach that quantifies the risk of having a bug in the SUT given the tests that were executed so far. We also use a risk measure to guide the generation of additional tests to best decrease the risk of having a bug in the SUT. In the description below, we do not incorporate bug prediction information as was done in [33]. If available, the method can be extended to incorporate bug prediction information by modifying the priors.

We define an indicator random variable $X_{C(i)}$. For a given test t , $X_{C(i)}(t) = 1$ if $t \in C(i)$ and 0 otherwise. We first discuss the case in which $\{C(i)\}_{i \in I}$ is a partition. We randomly choose a test t , execute it, and determine if the test succeeded and to which $C(i)$ t belongs. Given that Y is the indicator variable of not having a bug, we model $P(Y|X_{C(i)} = 1)$ as a Bernoulli distribution with a beta conjugate prior initialized to the uniform beta distribution (both α and β parameters are set to one, where α is the weight representing the average success of the test t when executed on $C(i)$). Bayes rule is used to update the beta prior for each $C(i)$ as we gather evidence that $C(i)$ does not contain a bug. The update rule is simple; increase α by one each time the test succeeds and increase β by one if it fails. When the SUT is corrected based on the detected bugs, α and β are reset to 1 for each $C(i)$, and the process repeats. If $\{C(i)\}_{i \in I}$ is not a partition an evidence of a test t execution result may apply to more than one $C(i)$ and we can update the conditional probabilities estimation for each one of them.

To produce an overall estimation of the likelihood of a bug, $P(X_{C(i)} = 1)$ can be estimated using some profile of the software usage (either collected empirically or estimated). After running k chosen tests, $\sum_{i \in I} P(Y|X_{C(i)} = 1)P(X_{C(i)} = 1)$ is used to update the likelihood of a bug. Here we use the assumption that $C(i)$ is a partition and we use the current beta prior associated with each $C(i)$ to estimate $P(Y|X_{C(i)} = 1)$. We can always obtain a partition by refining the set of coverage criteria, $\{C(i)\}_{i \in I}$, using intersections. In addition, if $\{C(i)\}_{i \in I}$ is not a partition, the likelihood function above is still relevant but no longer has a probability interpretation. Instead, it serves to measure more than once evidence that applies to intersections of $C(i)$ s. If bugs in each $C(i)$ are associated with a different loss $l(C(i))$, we can represent the average expected loss of the system as $\sum_{i \in I} l(C(i))P(Y|X_{C(i)} = 1)P(X_{C(i)} = 1)$.

Another way of handling the case in which $C(i)_{i \in I}$ is not a partition is to focus on the same random variables as above and apply Bayesian-network discovery techniques [23] to learn the edges in the network. The network is learned from a profile of the system usage. More specifically, we consider the joint distribution of the variables $X_{C(i)}$, $i \in I$ and obtain N samples of

the joint distribution from the SUT profile. Assuming here that I is small, one may use the Bayesian scoring criterion described in chapter eight of [32] to choose the Bayesian network that best fits the N samples.

The above risk and loss functions can guide the test generation process. In general, we prefer to take steps in the test generation process that decrease the risk of finding the bug or the expected average loss from finding it. The process of generating a test t reaching $C(i)$ may include the realization of several conditions. For example, we need to read and then write to some shared resource to cover $C(i)$. Facing several generation alternatives of the tests in P , we may prefer to realize conditions that are needed in order to reach $C(i)$ over conditions that are needed to reach $C(j)$ if $P(Y|X_{C(i)} = 1) > P(Y|X_{C(j)} = 1)$. This includes the special case in which $C(i)$ was never visited, which will be preferred over $C(j)$ s that were visited many times. In addition, if the loss function is given and the loss associated with a defect in $C(i)$ is high, we may prefer revisiting $C(i)$, even if the probability of finding a bug in $C(i)$ is low.

Note that the case in which no bugs are found in the above procedure only indicates that the current abstraction level imposed by the set of coverage criteria $C(i)$, $i \in I$ is no longer effective in finding bugs in the SUT. It does not mean the system has no bugs. Indeed, as a tester, if my expert opinion is that the system still has defects, I should probably attempt to design, following our methodology, additional coverage criteria that need to be covered. On the other hand, if my expert opinion as a tester is that the set of coverage criteria $C(i)$, $i \in I$ is sufficient, then it is consistent on my part to stop the testing if the above search procedure no longer yields bugs.

III. A TOOL-DRIVEN METHODOLOGY

This section demonstrates how our approach can be applied in practice, demonstrating the coverage criteria specification and the test suite generation. To support this process and test the performance of generalized coverage criteria in the real world, we added a tool that allows users to generate small test suites (sets of tests) from their programs. This tool uses a ranking function provided by the user and applies a genetic algorithm (GA) [12] to construct test suites with high ranks (i.e., individuals represent test suites). The full implementation details are presented in Section IV-B2. The tool and code examples are at github.com/bThink-BGU/Papers-2023-TSE-Sequence-Testing.

A. Mapping From Theoretical to Implementation Terms

The proposed tools are designed to support the methodology described in Section II. While Definition 1 only specifies a Boolean condition of coverage (i.e., a criterion is either covered or not), our tool, for practical reasons, allows for a quantitative measure of coverage.

Specifically, we propose counting the number of $C(i)$ s that a test suite covers and maximizing it as much as possible. We note that our tool only gives a sub-optimal solution. The optimization process can be controlled by tweaking the parameters to get the required balance between computation resources and quality.

The terms map as follows:

- The test model P is a set of 50 k possible executions of the system-behavior model (i.e., paths).
- The test suite S is generated by the tool. Its size can be changed according to the available resources.
- The coverage criteria $C = \{C(i)\}_{i \in I}$, $C(i) \subseteq \Sigma^a$ is specified as a ranking function that takes a test suite S and returns the number of $C(i)$ s that are covered.
- The coverage ratio $\Gamma_C(S, P)$ (Definition 2) is computed by dividing the result of the ranking function by all possible sequences in C . For example, if there are 11 possible events, a 2-way ranking function of P is bounded in $11^2 = 121$.

B. Coverage-Criteria Specification

Our test-suite-generation tool uses a user-defined ranking function to determine the quality of the test suites. This function takes a test suite, represented as a set of lists of events, and returns a number that models how well the suite covers the criteria that the user is interested in. In Section IV, we measure different coverage criteria using this tool. We used, for example, this tool to express our interest in counting how many different regular expressions of the form $\Sigma^* \sigma_1 \Sigma^* \sigma_2 \Sigma^*$ are matched by at least one test in the suite, where (σ_1, σ_2) is a pair of events. This is, of course, Kuhn and Higdon's 2-way sequence coverage criterion, as presented in Listing 1. The ranking function is an implementation of coverage criteria. Searching for top test suites according to the coverage criteria means that the chosen test suite maximizes the coverage of the test-case space. The ranking function attaches value to each candidate test suite according to the suite's opportunity for system coverage.

Listing 1: A Ranking Function in Java That Implements Kuhn and Higdon's 2-Way Coverage Criterion.

```
private long twoWayRank(Set<List<String>> suite) {
    var allPairs = new ArrayList<>();
    for (var test : suite)
        for (var i = 0; i < test.size() - 1; i++)
            for (var j = i + 1; j < test.size(); j++)
                allPairs.add("(" + test.get(i) + ", " +
                    test.get(j) + ")");
    return allPairs.stream().distinct().count();
}
```

C. Test-Suite Generation

To extract test suites with a high rank from the BP model, we developed a tool based on a genetic algorithm (GA) [12] that evolves test suites. The tool enables the testing engineer the freedom to tailor the parameters required to build good test suites for the system. The tester may adopt the following characteristics: the size of the pool of valid test cases, the test-suite size, the ranking function, and the search algorithm. Given these parameters, the tool creates a vast pool of valid test cases and finds test suites with a high ranking.

Specifically, for the GA and the genetic operators, we use Jenetics—a Java library for evolutionary algorithms [40]. Our tool creates a pool of valid test cases by performing 50 K random

TABLE II
EVOLUTIONARY HYPERPARAMETERS

Representation	Set<List<String>>
Mutation	Uniform test replacement
Recombination	Partially Matched Crossover
Mutation probability	0.05
Recombination probability	0.7
Parent selection	Tournament with $k = 3$
Survivor selection	Generational replacement
Population size	100
Termination criteria	300 generations or best-fitness convergence ($\epsilon \leq 0.001$)

walks on the given system-behavior model. It then utilizes GA to search for highly ranked test suites. Each individual, i.e., a test suite, is represented as a set of n tests, where different n values can be used. The initial population is randomly created from the pool of valid test cases. The fitness function is the ranking function provided by the user. We apply a standard mutation operator that replaces each test in each individual with another random test, with a probability of 0.05. We used a partially matched crossover [12] with a distribution of 0.7 to avoid repetitions of tests within a test suite. The complete evolutionary hyperparameters are summarized in Table II. Notably, while the hyperparameters may be further optimized, depending on the model and coverage criteria, they were selected as they proved robust to many models and coverage criteria.

In Section IV, we evaluate the sensitivity of the various parameters to the quality of the obtained results.

IV. EVALUATION

Section III provided a qualitative assessment of our modeling approach, demonstrating its applicability for testing. For example, we showed how coverage criteria could be specified as ranking functions in JavaScript.

Here, we have focused on a quantitative evaluation of our approach. Specifically, we evaluate here: 1) the effectiveness of different coverage criteria in detecting bugs; 2) the efficiency of our GA-based tool, and 3) the validity of our Bayesian risk-reduction approach. To this end, we begin with the *alternating-bit protocol*, which is a standard benchmark for formal verification and modeling [10]. Next, we evaluate our approach on a web application called *Moodle*, which is a popular, open-source learning management system. The use case of Moodle demonstrates the applicability of the method to real-life testing.

For specifying the system-behavior model, we used the *Behavioral Programming* modeling paradigm [21], designed for specifying behavior in a natural and intuitive manner that is aligned with how users perceive the requirements of a system. We begin with a short description of this paradigm and a demonstration of its usefulness for specifying system-behavior models.

The systems under test, the testing model, and execution instructions are in our repository at github.com/bThink-BGU/Papers-2023-TSE-Sequence-Testing.

A. Behavioral Programming

In behavioral programming (BP), a user specifies a set of scenarios that may, must, or must not happen. Each scenario

is a simple sequential thread of execution and is thus called a *b-thread*. B-threads are typically aligned with system requirements, such as “user must log in before using the system,” or “every file-read action must be preceded by a file-open action,” etc. The set of b-threads, called behavioral program (*b-program*), specifies the overall system behavior and, in our case, what needs to be tested. At run-time, all b-threads participating in a b-program are combined, yielding a complex behavior that is consistent with all the b-threads.

To synchronize the b-threads behaviors, Harel et al. [21] proposed a simple b-thread integration protocol. The protocol consists of each b-thread submitting a statement before selecting an action to perform, where actions are represented as events. The statement declares which events the b-thread requests, which events it waits for (but does not request), and which events it blocks (forbids from happening). After submitting the statement, the b-thread is paused. When all b-threads have submitted their statements, we say the b-program has reached a *synchronization point*. Then, a central event arbiter selects a single event that was requested and was not blocked. Having selected an event, the arbiter resumes b-threads that requested or waited for that event. The rest of the b-threads remain paused, and their current statements are used in the next synchronization point.

From a formal point of view, BP semantics are typically defined in terms of transition systems, where each b-thread is a labeled transition system (LTS), and the execution engine generates a cohesive LTS on the fly [13], [20]. In this paper, we use an implementation of BP, called BPjs [3], where b-threads are specified using simple JavaScript code snippets (hence the name), and the integration mechanism is developed in Java using the Rhino library [31]. We use BPjs to specify the test model, generate the cohesive LTS that models all possible sequences of events that test the system and execute the test model (which is equivalent to performing a random walk on the cohesive LTS). As described next, since all of these sequences are usually huge, we need to sample them wisely.

BP has been used for different development phases, such as requirement specification and solicitation [14], [15], [34], implementation [20], and verification [19], and also for testing [2], [18], [38], [39]. To understand the concepts of behavioral programming and how it can be used for specifying system-behavior model, we present a benchmark example of Bombarda and Gargantini [5], showing how the regular expression they proposed can be modeled using BP. This is an example of a vault that can be unlocked by the combination “12345”. The code in Listing 2 specifies two b-threads, one for pressing the keys and one for checking the combination. While this small example can be modeled using a single b-thread, breaking the specification into two modules allows us to align each b-thread to a different testing requirement. The first b-thread is aligned with the requirement that the safe has a keypad with nine digits that can be pressed in any order. It continuously requests pressing any of these keys. The second b-thread is aligned with the requirement that the safe is opened only when the correct code is dialed. While it does not request keys, it waits for the correct sequence and then requests the Open event while blocking all

other events (i.e., keys). At run-time, the b-threads are executed simultaneously and synchronized using the protocol described before.

Listing 2: The Vault Example [5] Implemented With BP. The Program has Two b-Threads — the First Continuously Requests to Press Any of These Keys. The Second b-Thread Waits for the Correct Key Sequence and Then Opens the Safe While Blocking Any Other Action.

```
const code = '12345'
const AnyKey = [ Event('1'), ..., Event('9') ]

bthread('press keys', function() {
  while(true)
    sync({ request: AnyKey }) //press a random key
})

bthread('check code', function() {
  for(let i=0; i<5; i++) {
    let key = sync({ waitfor: AnyKey })
    if (code[i] != key) //wrong vault code
      sync({ block: bp.all })
  }
  //correct vault code
  sync({ request: Event('Open'),
        block: Event('Open').negate() })
  sync({ block: bp.all })
})
```

B. Alternating-Bit Protocol

The alternating-bit protocol (ABP) [4] is a full-duplex communication protocol using an unreliable communication channel. Each packet in this protocol is repeatedly sent until the sender receives an acknowledgment from the client. Each packet is attached with a single bit of metadata that indicates the correct order of the messages. This bit alternates in each packet. When a packet is received, the receiver verifies that the attached bit is correct and sends back an acknowledgment message with the same bit attached.

In order to assess our methodology, we implemented the ABP protocol as described in [22]. Our implementation involved translating the model to Java for convenience and utilizing it as the system under test. Additionally, we introduced a variety of defects to investigate different facets of our approach, which are explained in detail below.

In addition to the SUT implementation, we created a testing model using BP that also includes “rainy-day” scenarios. Specifically, the model specifies two types of noises that can occur in both communication directions (sender to receiver and vice versa). The noises are a loss of a message and a change in the order of the messages. Both the SUT and the testing model are in our repository.

We applied a “white-box” testing approach where the testing model is used to drive the SUT and check that it follows the protocol. Each test case (list of events generated from the b-program) triggers the SUT to act and execute the event action. For example, the send event triggers the SUT to send a data packet. An instrumentation layer checks if the conditions to execute the event are met, and then the action is carried out,

and the internal state of the SUT updates accordingly. If the conditions for triggering the event are not met by the SUT, an error occurs, and the test fails. The test succeeds if all events execute successfully. We represented some actions using two events to allow a high granularity in the test cases. For example, when a receiver sends an acknowledgment message, it may be either correct (event `rAck`) or incorrect (`rNak`). Similarly, the sender may receive this message correctly (`sAck`) or incorrectly (`sNak`).

1) *Coverage-Criteria Evaluation*: Section II-B showed how testers could specify coverage criteria that generalize the notion of coverage and help in focusing the test efforts. Here we evaluate the effectiveness of coverage criteria in detecting different bugs. To this end, we injected four bugs in the SUT as follows:

- `sAck,sAck`: When the sender receives two acknowledgments in a row, it disregards them both and re-sends the message.
- `rNak,rAck`: When the receiver has to send an acknowledgment after a not-acknowledges, it does not send an acknowledgment.
- `sNak,sNak,rAck`: When the receiver has to send an acknowledgment after the sender receives two following not-acknowledges, it does not send an acknowledgment.
- `send,send,sAck`: When the sender receives an acknowledgment after sending the following message twice, it disregards the acknowledgment and re-sends the last message.

Next, we used the BP-based test model to generate many test cases, collect them into test suites, and examine how often they find the injected bugs in the SUT. We ran the model 50 K times and generated 50 K valid test cases. We built test suites out of these test cases, each has ten test cases. Our challenge was finding the best test suites with the highest probability of catching the injected bugs. To this end, we defined two ranking functions and used our GA-based tool to find test suites with a high rank.

The first ranking function is based on Kuhn and Higdon's method that counts all sequences of n events but not necessarily consecutive n events (i.e., $\{\Sigma^* \sigma_1 \Sigma^* \sigma_2 \Sigma^* : \sigma_1, \sigma_2 \in \Sigma\}$ or its equivalent for three letters bugs). The second-ranking function counts all n consecutive events in a test suite (i.e., $\Sigma^* w \Sigma^*$). Our thesis is that maximizing the consecutive events maximizes the probability of catching the bug, i.e., that the second-ranking function is better for our purposes. The point that we are trying to make here is that there are situations where our generalization that allows ranking functions other than Kuhn and Higdon's criterion has real usage. For a baseline, we also used a 'random' method for the test-suite generation that simply peeks ten test cases at random.

For each method and an injected bug, we counted how many tests detected the bug. We repeated this process 1 K times and averaged the results. The results are summarized in Table III. The second column specifies the sequences of events that trigger the injected bugs. For example, the sequence `sNak, sNak, rAck` represents the case where the sender receives twice an acknowledgment message with the wrong bit, and then the receiver sends the correct acknowledgment message. As the random column

TABLE III
THE PROBABILITY OF CATCHING BUGS IN THE ALTERNATING-BIT PROTOCOL AND MOODLE LMS, FOR EACH TEST-SUITE GENERATION METHOD

Domain	w	Unranked (Random)	Kuhn- Higdon	$\Sigma^* w \Sigma^*$
Alt. Bit	<code>sAck, sAck</code>	0.219	0.198	0.709
	<code>rNak, rAck</code>	0.732	0.643	0.988
	<code>sNak, sNak, rAck</code>	0.067	0.091	0.216
	<code>send, send, sAck</code>	0.821	0.806	0.978
Moodle	<code>teacher.AddQuestion, student.SubmitExam, teacher.SubmitQuestion</code>	0.372	0.986	0.512

The unranked method randomly generates test suites. w represents a sequence that triggers the bug.

suggests, some bugs are frequent, and some occur in rare corner cases.

The results show that the generalized criterion $\Sigma^* w \Sigma^*$ is always better and that it is much better than Kuhn and Higdon's criterion when the bug is rare. This criterion model a heuristic that applies to many systems where specific sequences of consecutive events trigger bugs. We call it a 'generalized' criterion because it demonstrates how developers can generalize Kuhn and Higdon's approach for modeling new heuristics that target new types of bugs. Note that we are not claiming that our criterion is better than Kuhn and Higdon's criterion in general, only that generalizing allows us to target specific types of bugs better.

2) *Test-Suite Generation Evaluation*: To evaluate our GA-based tool for generating test suites, we compare it to two other methods for different coverage criteria and different suite sizes. To produce the test suites, we first created a pool of 50 K valid test cases extracted from our model. We then used three methods for generating test suites from the pool (n is the test suite size):

- 1) *Unranked*: Randomly peeking n test cases.
- 2) *Best of 1 K*: Repeating the 'unranked' process 1 K times and choosing the suite with the highest rank.
- 3) *GA*: Our GA-based tool described in Section III-C.

We check each method with the same coverage criteria as before (Kuhn and Higdon and $\Sigma^* w \Sigma^*$) and three test-suite sizes: 5, 10, and 20. For each test, we measure the ranking function value (called 'rank') and the run-time ('time') in seconds. The results presented in Table IV, show that GA works best in terms of time and ranking. Kuhn and Higdon's 2-way rank is constant since there are eleven possible events, and therefore the rank is bounded in $11^2 = 121$.

Fig. 3 depicts the average generation time for each method (i.e., the time row of the $\Sigma^* w \Sigma^*$ criterion in Table IV). This visualization clearly shows that GA outperforms other methods both in time and obtained rank in all three test sizes. The advantage grows with the size of the test.

3) *Bayesian Risk-Reduction Evaluation*: This section demonstrates how our Bayesian approach computes a stable estimation of the risk involved with each coverage criteria $C(i)$ when the system is continually tested and not fixed.

In our setting, there are 121 coverage criteria, namely 11^2 and we have run $t_1, \dots, t_n$ tests, $n = 50,000$. When a test, t_i

TABLE IV
THE PERFORMANCE OF TEST-SUITE GENERATION METHODS IN TERMS OF QUALITY AND TIME, EVALUATED ON THE ALTERNATING-BIT PROTOCOL AND MOODLE LMS

		Method Suite size	5	GA 10	20	5	Best of 1K 10	20	Unranked (Random)		
Alt. Bit	Kuhn-Higdon	Time (mSec)	78.1	88.7	125	622.4	859.4	810	0	0	0
		Rank (max)	121	121	121	121	121	121	121	121	121
		Rank (mean)	120.457	120.797	120.879	120.879	120.879	120.879	112.983	117.496	118.919
	$\Sigma^*w\Sigma^*$	Time (mSec)	31.2	38.9	38.9	519.3	528.9	561	0	0	11
		Rank (max)	41	58	78	39	55	62	34	42	50
		Rank (mean)	37.579	54.162	63.801	36.249	45.915	53.572	25.575	32.61	40.325
Moodle	Kuhn-Higdon	Time (Sec)	177.433	299.626	474.957	476.969	945.588	1757.526	0.863	1.987	3.68
		Rank (max)	700	700	700	696	700	700	673	696	699
		Rank (mean)	684	695	699	673	692	699	611	647	668
	$\Sigma^*w\Sigma^*$	Time (Sec)	14.671	15.353	22.282	21.158	31.242	49.068	0.151	0.135	0.113
		Rank (max)	76	111	150	71	104	139	68	97	133
		Rank (mean)	68	103	141	66	98	132	57	85	118

We examined the performance for each method (GA, “best of 1 k”, and unranked), possible test-suite size (5, 10, and 20), and scoring functions (kuhn and higdon and $\Sigma^*w\Sigma^*$). Numbers are averaged over 1 k runs.

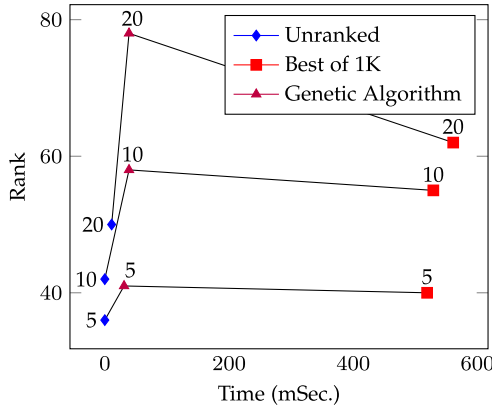


Fig. 3. Efficiency graph for the $\Sigma^*w\Sigma^*$ criterion, evaluated on the alternating-bit protocol. The graph displays each method’s ratio between calculation and time to quality. The number of data points represents the test-suite size.

is executed, it either hits a coverage criterion $C(i)$ or does not. Assuming that $C(i)$ is hit when t_i is executed, we either find a defect or not, and the test either passes or not. Our focus is on the statistical dependency under a Bayesian setting of these two random variables. Discovering or failing to discover a defect during a test that targets some $C(i)$ triggers an update of the prior probability of finding a defect associated with $C(i)$. As the prior is a Beta function, it is controlled by two parameters α_i, β_i in which $E_i = \frac{\alpha_i}{\alpha_i + \beta_i}$ represents the estimated average of finding a defect if $C(i)$ is hit. The counters α_i and β_i are continually updated using the Beta Bernoulli conjugate prior rule. Namely, if $C(i)$ was covered during the execution of t_i , we increment α_i by one if a bug was found and β_i otherwise. We want to establish that the prior stabilizes over time, i.e., that E_i converges when $n \rightarrow \infty$ and that the probability mass of the prior concentrates. The second is established if the variance of the Beta distribution, namely, $V_i = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ vanishes. We have run 50 k tests and updated the priors for each coverage criterion, as explained above. We then removed all the $C(i)$ s not covered by at least 1 k tests and computed the maximal

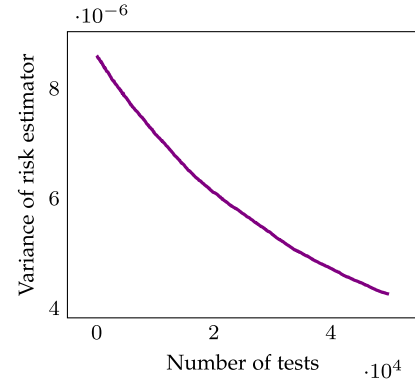


Fig. 4. The stabilization of our Bayesian-based risk estimation, evaluated on the alternating-bit protocol. The estimation of the risk of each coverage criteria $C(i)$ is stabilized when the system is continually tested and is not fixed. The graph shows that the variance of these parameters gets very small very quickly, indicating that our risk estimation gets very accurate after a reasonable number of tests.

variance of the remaining $C(i)$ s over time. Note that the current testing efforts are insufficient for the removed $C(i)$ s, requiring additional test design and implementation. As depicted in Fig. 4, the variance converges to zero as expected, indicating that our risk estimation gets very accurate after a reasonable number of tests.

C. Moodle LMS

BP can be used for modeling systems that are not usually perceived as event-based. We demonstrate it on Moodle LMS — a popular, open-source learning management system used by educators to create private websites with online courses to achieve learning goals [29]. While Moodle is developed as an object-oriented system, use it as a black box and model the interaction between the users and the system as an event-based system.

In the supplemented material, we provide the specification of a BP-based model with three b-threads, each aligned to a different

TABLE V
STATISTICS ON MOODLE'S B-PROGRAM STATE SPACE

B-threads	Events	States	Edges	Traces (possible tests)
3	22	103	195	33,489

aspect of the system behavior, handled by a different type of user. The first b-thread specifies an administrator behavior that creates a course and enrolls users in it. The second b-thread specifies how an enrolled teacher adds a quiz with two questions to the course. Finally, the last b-thread specifies how an enrolled student waits for a question to be added and then answers the question. We used the Selenium WebDriver [36] for performing actions on Moodle UI. Using this full stack of b-program, actuators, Selenium WebDriver, and Moodle v3.9, we were able to run a few tests and detect the following bug.

According to Moodle documentation [30], a teacher cannot add questions to a quiz once a student attempts it. This behavior is enforced by the user interface (UI), as the “Add question” button disappears once a student attempts the quiz. Yet, if a student attempts the quiz while the teacher adds another question, the UI misbehaves and displays an exception. These are the exact steps that reveal the bug: (1) a teacher starts adding a second question to a quiz; (2) a student attempts to answer the first question; (3) the teacher submits the second question.

The Moodle example is far more complicated than the alternating-bit protocol. For the purpose of evaluating its complexity, we generated an automaton that represents the entire state space of our model. Table V shows that although our model has only three intuitive b-threads and 22 distinct events, the state space consists of 103 states, 195 edges, and 33,489 distinct traces. Notably, the generated model includes only “high-level” events, such as `login(username, password)` or `enrollUser(course-id, username)`. To actuate the website using Selenium, we added “low-level” events to the model, such as `writeText(xpath, text)` and `click(xpath)` that receive an element unique path on the UI and perform actions on it. This extended model is too large to compute, though it does not raise a problem since it is executed directly without requiring the generation of the state space.

1) *Coverage-Criteria Evaluation*: Here we present how we evaluated the effectiveness of coverage criteria in detecting different bugs. While evaluating our approach on Moodle, we detected the aforementioned bug using Kuhn and Higdon’s “3-way” criterion, emphasizing the usefulness of this criterion and its advantage. Since the SUT is a real system with real bugs, we did not have to inject bugs. Other than that, the evaluation method is the same as presented in Section IV-B1.

The results are summarized in Table III. The second column specifies the sequences of events that trigger the injected bugs. Specifically, the sequence `teacher.AddQuestion, student.SubmitExam, teacher.SubmitQuestion` represents the detected bug, as explained above. According to the findings, Kuhn and Higdon’s 3-way criterion is more effective in this case, which makes sense considering that the bug in question is dependent on three independent events, and the occurrence of

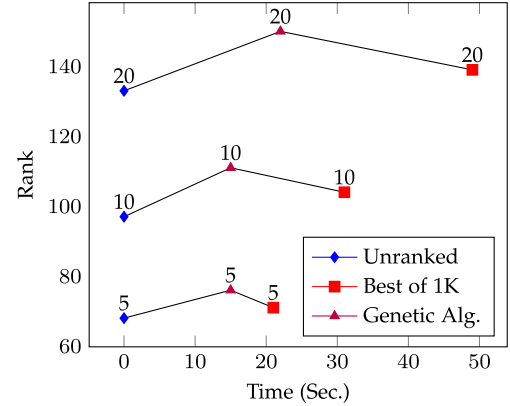


Fig. 5. Efficiency graph for the $\Sigma^*w\Sigma^*$ criterion, evaluated on Moodle LMS. The graph displays the ratio between calculation and time to quality for each method. The number of data points represents the test-suite size.

other events has no impact on it. This indicates the necessity of a universal approach that does not replace current state-of-the-art methods but rather complements them with more generalized versions.

2) *Test-Suite Generation Evaluation*: Similar to Section IV-B2, we compared our GA-based tool with two baselines, unranked and “best of 1 K”, measuring the value of the ranking function (called ‘rank’) and the run-time (‘time’) in seconds over three test-suite sizes: 5, 10, and 20. The results are presented in Table V, showing that GA works best in terms of time and ranking.

Fig. 5 depicts the average generation time for each method (i.e., the time row of the $\Sigma^*w\Sigma^*$ criterion in Table V). Like in the alternating-bit protocol, GA is better than “best of 1 K” both in time and obtained rank in all three test sizes. The advantage grows with the size of the test.

3) *Bayesian Risk-Reduction Evaluation*: We performed the same analysis as in Section II-A and similar to Fig. 4, the variance converges to zero as expected, indicating that our risk estimation gets very accurate after a reasonable number of tests.

There are a total of 1,728 theoretical coverage criteria, which is equivalent to 12 cubed. However, when examining the Moodle example and accounting for our testing model’s fixed order and position of four events, only 512 (or eight cubed) of the C_i s intersect with P . We generated 30,000 test cases, which gave us 1,568 unique test cases. We repeated the experiment shown in Fig. 4 and received a similar graph only with fewer tests and the best variance of risk estimator approaching 10^{-5} .

V. RELATED WORK

A. Coverage-Criterion Specification

Producing all possible paths (e.g., [35]) does not seem to be practical in large systems. Thus, one of the critical issues is the definition of the equivalence classes of tests according to the coverage criterion.

Lübke et al. [27] defined equivalence classes by dividing the tested system parameters into two types — parameters with a limited, small number of options (e.g., a list of payment methods)

and parameters with many options (e.g., a bank account number or first name). It is necessary for this approach to test each parameter of the first type and at least one of the second type. Thus, all models of the second type are in the same equivalence class.

In some cases, such as autonomous driving, coverage requirements are enforced by regulation. To address such large-scale systems, [26] realized the optimization of the testing process based on CTD. Nevertheless, they were still required to reduce the number of overall test scenarios. They achieved that by value quantization and by preventing illogical scenarios. For example, they defined medium traffic load in the range of three to six vehicles and detected illegal scenarios like making a left turn on a straight road.

B. Test-Suite Generation

Kuhn and Higdon proposed the first sequence-testing variant, known as t -way sequence testing, that allows only one event triggering in a test case [25]. Later versions allowed multiple triggered events and extended the algorithms to support additional features [5], [111], [37]. Current sequence testing methods for valid test suite generation consist of two steps. The first is to generate a list of all relevant sequences of a length t (called “target sequences”), and the second step is to generate test cases that cover the target sequences (called “test sequences”) [25]. The latter uses a greedy algorithm that handles constraints between two events. Then, a labeled transition system is proposed to model the SUTs’ requirements, and graph path methods are used to find the optimal valid test cases. Based on this work, additional work has been done to expand the language of constraints by, e.g., adding the possibility of contiguous values [37] and allowing more complex relationships between more than two factors [11].

Two problems remain open with these sequence-testing approaches. First, these approaches rely on the existing test model (not part of their work). Second, the solution must be effective in run time and size. To address these challenges, Bombarda and Gargantini [5] proposed to model the SUT by a finite state machine and to create the test cases using automata theory.

Some of the aforementioned papers have shown that generating test cases to cover every possible combination of input parameters is not necessary. This is due to the principle of t -way testing, which seeks to identify the combinations of input parameters that could result in failures. By selecting a small value of t , which represents the number of parameters in each combination, a much smaller test set can be generated compared to exhaustive testing.

C. Utilizing Knowledge From Previous Runs

Our approach randomly generates test sequences in a black-box manner (we used white-box events to verify the test result in Section IV, however, the tests were generated in a black-box manner). Random black-box test generators, or fuzzes, as they are sometimes called, have evolved to include a feedback loop that facilitates the efficient discovery of bugs, reduction of risk, and increase of coverage. The feedback loop may consist of

information on newly achieved code-coverage objectives or the occurrences of desired events during execution, such as buffer overflow [7].

Other ideas for directing the test generation include: choosing inputs far away from the previous inputs [8], dividing the inputs into sub-domains, and using translation to obtain a new test in a different sub-domain [9]. Another approach defines exclusion zones around tests that were already executed and discard randomly chosen inputs if they are chosen from an excluded zone [1].

Other approaches attempt to generate inputs that are more likely to cause a failure based on failure models. For example, boundary inputs are preferred as it is well known that they are more likely to cause a failure [1]. Another example chooses test sequences that mutate the values of fields in the object under test [41].

VI. CONCLUSION

In conclusion, recall the famous quote: “I always thought something was fundamentally wrong with the universe.” (The Hitchhiker’s Guide to the Galaxy series [42]). This feeling is familiar to everyone in the software testing industry. In this paper, we addressed the dilemma of every testing team: what and how much to test before we declare our software as ready to launch?

To tackle this dilemma, we identified three related issues and presented theoretical and practical contributions:

- 1) *How to specify the test space that needs to be covered:* We defined a generalized, automata-based approach for specifying coverage criteria. We also provided a set of valuable coverage criteria that may be applied to various domains.
- 2) *Finding a finite, relatively small, test suite that covers this space:* We developed a tool that allows us to translate the coverage criteria as ranking functions, generate test suites for these functions, and analyze the results from various aspects.
- 3) *How to utilize knowledge from previous runs to optimally reduce bug risks:* We proposed a Bayesian-based formula for balancing exploration and exploitation of the knowledge obtained by tests.

The approach we presented relies on the existence of a system model. We used the behavioral-programming paradigm to create this model, though other approaches can be used. Evaluating the model construction effort and the paradigm’s applicability for modeling large-scale systems, are out of the scope of this paper. Nevertheless, as demonstrated in Section IV-C, three small and intuitive b-threads generate a large state space and are sufficient for detecting bugs. This is due to the ability to specify behavioral aspects using small b-threads, without explicitly specifying the cohesive behavior. Furthermore, our approach does not require state-space generation since BP models are executable. Finally, the agility of the paradigm allows for incrementally adding more testing requirements, without changing the existing model.

From what we have presented in this article, it is possible to expand to other research areas in different directions in the

testing space. All of these are focused on achieving the goal of efficiently testing processes, software, and system. A possible trend we have begun to explore is the system modeling process concerning the testing resulting from system requirements using BP tools [38]. Another way to expand is by using statistical tools in the stem of an efficient and focused test suites generation process to test coverage, increasing the probability of identifying the faults.

ACKNOWLEDGMENTS

The authors would like to thank Gal Amram for a thorough review of the draft of this paper.

REFERENCES

- [1] M. A. Ahmad, "New strategies for automated random testing," Ph.D. dissertation, Comput. Sci. Dept., University of York, 2014.
- [2] M. Bar-Sinai, A. Elyasaf, A. Sadon, and G. Weiss, "A scenario based on-board software and testing environment for satellites," in *Proc. 59th Isr. Annu. Conf. Aerosp. Sci.*, 2019, pp. 1407–1419.
- [3] M. Bar-Sinai, G. Weiss, and R. Shmuel, "BPjs - an extensible, open infrastructure for behavioral programming research," in *Proc. IEEE/ACM 21st Int. Conf. Model Driven Eng. Lang. Syst.: Companion Proc.*, 2018, pp. 59–60.
- [4] K. A. Bartlett, R. A. Scantlebury, and P. T. Wilkinson, "A note on reliable full-duplex transmission over half-duplex links," *Commun. ACM*, vol. 12, no. 5, pp. 260–261, 1969.
- [5] A. Bombarda and A. Gargantini, "An automata-based generation method for combinatorial sequence testing of finite state machines," in *Proc. IEEE Int. Conf. Softw. Testing, Verification Validation Workshops*, 2020, pp. 157–166.
- [6] A. Bron, E. Y. Farchi, Y. MagidNir, and S. Ur, "Applications of synchronization coverage," in *Proc. ACM SIGPLAN Symp. Princ. Pract. Parallel Program.*, Chicago, IL, USA, K. Pingali, K. A. Yelick, and A. S. Grimshaw, Eds., Jun. 15–17, 2005, pp. 206–212. [Online]. Available: <https://doi.org/10.1145/1065944.1065972>
- [7] G. Candea and P. Godefroid, *Automated Software Test Generation: Some Challenges, Solutions, and Recent Advances*. Cham, Switzerland: Springer, 2019, pp. 505–531.
- [8] T. Chen, F.-C. Kuo, R. Merkel, and T. Tse, "Adaptive random testing: The art of test case diversity," *J. Syst. Softw.*, vol. 83, pp. 60–66, 2010.
- [9] T. Y. Chen, F. C. Kuo, and H. Liu, "On test case distributions of adaptive random testing," in *Proc. 19th Int. Conf. Softw. Eng. Knowl. Eng.*, 2007, pp. 141–144.
- [10] K. Chukharev, D. Suvorov, D. Chivilikhin, and V. Vyatkin, "SAT-based counterexample-guided inductive synthesis of distributed controllers," *IEEE Access*, vol. 8, pp. 207485–207498, 2020.
- [11] F. Duan, Y. Lei, R. N. Kacker, and D. R. Kuhn, "An approach to T-way test sequence generation with constraints," in *Proc. IEEE Int. Conf. Softw. Testing, Verification Validation Workshops*, 2019, pp. 241–250.
- [12] A. E. Eiben et al., *Introduction to Evolutionary Computing*, vol. 53. Berlin, Germany: Springer, 2003.
- [13] A. Elyasaf, "Context-oriented behavioral programming," *Inf. Softw. Technol.*, vol. 133, May 2021, Art. no. 106504. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S095058492030094X>
- [14] A. Elyasaf, A. Marron, A. Sturm, and G. Weiss, "A context-based behavioral language for IoT," in *Proc. CEUR Workshop Proc.*, R. Hebig and T. Berger, Eds., Copenhagen, Denmark, 2018, pp. 485–494. [Online]. Available: <http://ceur-ws.org/Vol-2245>
- [15] A. Elyasaf, A. Sadon, G. Weiss, and T. Yaacov, "Using behavioral programming with solver, context, and deep reinforcement learning for playing a simplified RoboCup-Type game," in *Proc. IEEE 22nd Int. Conf. Model Driven Eng. Lang. Syst. Companion*, 2019, pp. 243–251.
- [16] E. Farchi, Y. Nir, and S. Ur, "Concurrent bug patterns and how to test them," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, 2003, pp. 286–293.
- [17] I. Forgács and A. Kovács, *Practical Test Design: Selection of Traditional and Automated Test Design Techniques*. London, U.K.: BCS Publishing, 2019.
- [18] J. Greenyer, "Scenario-based modeling and programming of distributed systems," in *Proc. Int. Conf. Appl. Theory Petri Nets Concurrency*, 2021, pp. 168–187.
- [19] D. Harel, R. Lampert, A. Marron, and G. Weiss, "Model-checking behavioral programs," in *Proc. 9th ACM Int. Conf. Embedded Softw.*, 2011, pp. 279–288.
- [20] D. Harel, A. Marron, and G. Weiss, "Programming coordinated behavior in Java," in *Proc. Eur. Conf. Object-Oriented Program.*, Berlin, Germany: Springer, 2010, pp. 250–274.
- [21] D. Harel, A. Marron, and G. Weiss, "Behavioral programming," *Commun. ACM*, vol. 55, no. 7, pp. 90–100, Jul. 2012.
- [22] Intel, fmbt, 2012. [Online]. Available: <https://01.org/fmbt/>; <https://github.com/intel/fmbt>
- [23] M. Koivisto and K. Sood, "Exact Bayesian structure discovery in Bayesian networks," *J. Mach. Learn. Res.*, vol. 5, pp. 549–573, Dec. 2004.
- [24] D. Kuhn, R. Bryce, F. Duan, L. Ghandehari, Y. Lei, and R. Kacker, "Combinatorial testing: Theory and practice," *Adv. Comput.*, vol. 99, pp. 1–66, 2015.
- [25] D. R. Kuhn et al., "Practical combinatorial testing," *NIST Special Pub.*, vol. 800, no. 142, 2010, Art. no. 142.
- [26] C. Li, C.-H. Cheng, T. Sun, Y. Chen, and R. Yan, "ComOpT: Combination and optimization for testing autonomous driving systems," 2021, *arXiv:2110.00761*.
- [27] D. Lübke, J. Greenyer, and D. Vatlin, "Effectiveness of combinatorial test design with executable business processes," in *Empirical Studies on the Development of Executable Business Processes*. Berlin, Germany: Springer, 2019, pp. 199–223.
- [28] J. R. Maximoff, D. R. Kuhn, M. D. Trela, and R. Kacker, "A method for analyzing system state-space coverage within a t-wise testing framework," in *Proc. IEEE Int. Syst. Conf.*, 2010, pp. 598–603.
- [29] Moodle, "Moodle – open-source learning platform," Accessed: Jan. 26, 2023. [Online]. Available: <https://moodle.org>
- [30] Moodle, "Moodle's quiz faq," Accessed: Jan. 26, 2023. [Online]. Available: https://docs.moodle.org/39/en/Quiz_FAQ
- [31] Mozilla, "Mozilla/rhino: Rhino is an open-source implementation of javascript written entirely in Java," Accessed: Jan. 26, 2023-. [Online]. Available: <https://github.com/mozilla/rhino>
- [32] R. E. Neapolitan, "Learning Bayesian networks," Pearson Prentice Hall Upper Saddle River, New York, NY, USA, 2004, vol. 38.
- [33] A. Perera, A. Aleti, B. Turhan, and M. Boehme, "An experimental assessment of using theoretical defect predictors to guide search-based software testing," *IEEE Trans. Softw. Eng.*, vol. 49, no. 1, pp. 131–146, Jan. 2023.
- [34] R. Poliansky, M. Sipper, and A. Elyasaf, "From requirements to source code: Evolution of behavioral programs," *Appl. Sci.*, vol. 12, no. 3, 2022, Art. no. 1587. [Online]. Available: <https://www.mdpi.com/2076-3417/12/3/1587>
- [35] M. Rocha, A. Simão, and T. Sousa, "Model-based test case generation from uml sequence diagrams using extended finite state machines," *Softw. Qual.*, vol. 29, no. 3, pp. 597–627, 2021.
- [36] Selenium, "Selenium – a suite of tools for automating web browsers," Accessed: Jan. 26, 2023. [Online]. Available: <http://www.selenium.dev/>
- [37] Y. Sheng et al., "Extended covering arrays for sequence coverage," *Symmetry*, vol. 10, no. 5, 2018, Art. no. 146.
- [38] Y. Weiss, "Testing reactive systems using behavioral programming, a model centric approach," 2021, *arXiv:2112.01538*.
- [39] C. Wiecher, J. Fischbadh, J. Greenyer, A. Vogelsang, C. Wolff, and R. Dumitrescu, "Integrated and iterative requirements analysis and test specification: A case study at Kostal," in *Proc. IEEE/ACM 24th Int. Conf. Model Driven Eng. Lang. Syst.*, 2021, pp. 112–122.
- [40] F. Wilhelmstötter, "Jenetics: Java genetic algorithm library," 2022. [Online]. Available: <http://jenetics.io>
- [41] W. Zheng, Q. Zhang, M. R. Lyu, and T. Xie, "Random unit-test generation with mut-aware sequence recommendation," in *Proc. IEEE/ACM 25th Int. Conf. Automated Softw. Eng.*, Antwerp, Belgium, C. Pecheur, J. Andrews, and E. D. Nitto, Eds., Sep. 20–24, 2010, pp. 293–296. [Online]. Available: <https://doi.org/10.1145/1858996.1859054>
- [42] D. Adams, *The Restaurant At The End of the Universe*. New York, NY, USA: Harmony Books, 1981.