

# Combinatorial Sequence Testing Using Behavioral Programming and Generalized Coverage Criteria

Achiya Elyasaf<sup>a</sup>, Eitan Farchi<sup>b</sup>, Oded Margalit<sup>a</sup>,  
Gera Weiss<sup>a</sup>, Yeshayahu Weiss<sup>a</sup>

<sup>a</sup>*Ben-Gurion University of the Negev*

<sup>b</sup>*IBM Haifa Research Lab*

---

## Abstract

This paper tackles three main issues regarding test design: (1) it proposes a new way to model what to test; (2) it offers a framework for specifying coverage criteria that generalizes previous types of coverage; (3) it outlines a Bayesian approach to an informed exploitation-exploration balance in the context of testing. In addition to the theoretical contribution, we present an empirical evaluation with a proof-of-concept tool that we have developed to support the conceptual advantages and to illustrate practical benefits.

*Keywords:* Sequence Testing, Behavioral Programming, Model-Based Testing, Test Optimization, Test Generation, Combinatorial Test Design, General Coverage Criteria, Bayesian Risk-Reduction

---

## 1. Introduction

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

---

*Email addresses:* [achiya@bgu.ac.il](mailto:achiya@bgu.ac.il) (Achiya Elyasaf), [farchi@il.ibm.com](mailto:farchi@il.ibm.com) (Eitan Farchi), [odedm@post.bgu.ac.il](mailto:odedm@post.bgu.ac.il) (Oded Margalit), [geraw@cs.bgu.ac.il](mailto:geraw@cs.bgu.ac.il) (Gera Weiss), [weissye@post.bgu.ac.il](mailto:weissye@post.bgu.ac.il) (Yeshayahu Weiss)

We propose in the paper to address the above challenge by tackling the following issues: (1) how to describe the space that needs 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 in order to optimally reduce bug risks.

Our answer to the first challenge is described using a proof-of-concept tool that we have implemented whose input language is founded on a scenario-based modeling paradigm, called Behavioral Programming (BP) [1, 2]. We show how this modeling language allows us to directly and compositionally specify aspects of the behavior we want to test [3]. Specifically, we demonstrate the automatic creation of composite tests using BP’s scenario integration mechanism. We demonstrate the advantages of this approach over traditional approaches where each usage scenario generates a single test.

Our answer to the second challenge of describing test coverage criteria is a generalization of the known Kuhn’s  $t$ -way sequence coverage approach [4]. We demonstrate, with examples from various domains, how our generalizations can naturally express useful coverage criteria that extend Kuhn’s vision to more types of applications. We also show how the proposed extension allows to include standard combinatorial test design (CTD) methods and Kuhn’s criterion under a unified formal vocabulary. Our approach to specifying coverage criteria is automata based. Each automaton in our framework represents a set of tests that are considered equivalent by testers, i.e., two tests in the same set are considered likely to both fail or both pass, presumably because of the probability that the system under test (SUT) handles them differently is low. We present various examples from different kinds of systems to demonstrate the naturalness of this approach and show that coverage requirements can be constructed to suit different systems irrespective of the test program specific to the system.

Our answer to the third challenge 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 probability of finding new bugs.

A proof-of-concept tool that we developed demonstrates how our approach to generalized sequence coverage can be applied in practice. Test plans are entered using the scenario-based BP approach, as described above. Additionally, testers can provide their own 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. A state-of-the-art optimization technique based on Genetic Algorithms (GA) is then applied to produce high-ranking test suites. Through this tool, we demonstrate a methodology we envision: use BP to specify what to test, automata to specify a library of coverage criteria, and apply optimization techniques to extract test suites. Our proposal is that practitioners can select criteria in the library based on, for example, the type of implementation at hand, the phase of the project, the code libraries in use, etc.

The paper is organized as follows. We begin, in Section 2, with the contribution number (2) of proposing a generalized coverage. The choice for starting with the second contribution is to stress the fact that it is independent of the first one — one can use the generalized coverage framework with any testing framework. Examples of how this generalization can be used in various domains are outlined in Section 3. Then, in Section 4, we outline contribution number (3) a Bayesian approach to decide when to exploit previous knowledge and when to explore new possibilities in the context of testing. Contribution number (1), the modeling language that we are proposing for testing, is presented in Section 5 together with a brief description of a proof-of-concept tool that we have developed. An evaluation of the proposed approach is presented in Section 6.

## 2. Generalized coverage criteria

We start by proposing a new type of coverage criteria that generalizes the  $t$ -way combinatorial sequence coverage criterion [5] and the classical  $t$ -way coverage [4]. The generalized framework gives testers tools for infusing domain knowledge into the test design. For example, a tester, based on previous experience with the system or on understanding of its implementation, may want to focus on testing some race conditions on the access of one shared resource only when another shared resource is held. We propose tools for specifying such focus. We first define our generalized notion of coverage and then motivate it by a series of examples.

In this section, we assume that tests can be modeled using words from  $\Sigma^a$  where  $\Sigma$  is some alphabet and  $a \in \mathbb{N} \cup \{*\}$ . The alphabet represent actions that can be applied to the system under test (SUT) using an external interface available for testing. These can include, as we will show later, validations and control of internal mechanisms made available as mock-ups (using, e.g, <https://site.mockito.org/>).

As described in the introduction, we propose to apply a test design process that starts with the definition of a test model  $P \subseteq \Sigma^a$ . In this section we abstractly consider  $P$  to be the set of all possible test scenarios. Once  $P$  is defined, ideally, we would 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. Coverage is defined using an indexed set of languages  $\{C(i)\}_{i \in I}, C(i) \subseteq \Sigma^a$ . The next definition captures what constitutes covering of  $P$  using  $\{C(i)\}_{i \in I}$ .

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

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

**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 for that, we always implicitly assume that in practice there is another coverage requirement  $\Sigma^a - \bigcup_{i \in I} C(i)$ , 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 definitely meet. For example, 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 freely define their test concerns. 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. In addition, experiments in applying coverage to software testing indicate that partition of the coverage requirements is not natural. Consider condition coverage, for example, where each condition in the SUT should evaluate to true by some test and to 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 formal-

izes a common practice used by testers 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 for achieving the relaxed coverage criterion. Consider code coverage, for example: when full coverage of all lines of code is not possible, people usually settle for covering at least one line of code of each function. In our context, this translates to a partition of  $I$  as defined above.

### 3. Coverage criteria examples

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 testing using *Combinatorial Test Design* [6]. We capture that as our first example.

**Example 1** (classic  $t$ -way coverage). *Classic  $t$ -way coverage is on finite spaces, i.e., when  $P \subseteq \Sigma^n$  for  $n \in \mathbb{N}$ . In this case, for a small  $t \leq n$  a coverage criterion is defined for each choice of  $t$  indexes  $i_1, \dots, i_t, 1 \leq i_l \leq n$  and  $t$  letters  $\sigma_1, \dots, \sigma_t \in \Sigma^t$  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 the design of tests in which the order of occurrences of events is important. We outline how the approach generalizes Kuhn's sequence coverage criterion [5]. The example below shows how to express Kuhn's coverage criterion using our framework.

**Example 2** (Kuhn's coverage criterion). *Under the above definitions, Kuhn's definition of  $t$ -sequence coverage [5] is obtained by taking  $I = \Sigma^t$  and*

$$C(\sigma_1 \cdots \sigma_t) = \mathcal{L}(\Sigma^* \sigma_1 \Sigma^* \cdots \Sigma^* \sigma_t \Sigma^*).$$

In addition to Kuhn’s coverage criterion, our framework allows for defining richer criteria. Indeed, we can consider criteria that are not exactly  $t$ -sequence coverage. We can, for example, require that  $\sigma_i$  does not appear unnecessarily, but only once. This leads to the following customized definition.

**Example 3** (Like Kuhn’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  $\Sigma' = \Sigma \setminus \{\sigma_1, \dots, \sigma_t\}$ .*

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

**Example 4** (message order). *Consider a language  $\Sigma$  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.*

Such customized test coverage requirements can be derived from known concurrent bug patterns [7] or other specific test concerns.

A more concrete SUT will lead to further customization of coverage requirements. Consider the case in which a bank’s customer has two accounts. A transaction may reduce the number of dollars in the first account by  $m$  dollars and another transaction may increase the number of dollars in the second account by  $m$  dollars. We are concerned that some transactions may calculate the overall money the customer has while the transfer of money from one account to the other has started but not completed. This leads to the following customized coverage requirement.

**Example 5** (transaction safety). *If  $\Sigma$  represents the set of possible transactions, and  $D \subseteq \Sigma$  is the set of transactions that deduce money from a customer account, and  $A \subseteq \Sigma$  is the set of transactions that add money to a customer account. Set  $\Sigma' = \Sigma - (D \cup A)$ . We capture our test concern by the following coverage requirements.*

$$C(d, a) = \mathcal{L}(d \Sigma' a), d \in D, a \in A$$

Many times coverage is motivated by symmetries. When we are required to see event  $\sigma_1$  and then event  $\sigma_2$ , we actually implicitly claim that all permutations of events occurring before  $\sigma_1$  are equivalent for the purpose of revealing a problem which is basically a symmetry claim (we also claim that all permutations between  $\sigma_1$  and  $\sigma_2$  as well as all permutations after  $\sigma_2$  are equivalent). Our framework supports the utilization of expected SUT symmetries in ways other than symmetries on executions order. This is illustrated in the next example.

Suppose we would like to check that a Tic-Tac-Toe<sup>1</sup> game-playing application works. There are 255,168 ways to play a game, i.e., put alternate Xs and Os on the  $3 \times 3$  board (depicted in Figure 1), where the game ends when one of the players wins or when the board is full (tie). Nevertheless, if we assume that the algorithm should work the same for symmetric boards (under eight rotations and reflections), we can reduce the test space considerably 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 of 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. relation. and  $[w]$  is the equivalence class of  $w$ .*

*In the TTT case,  $n = 9$  but this approach applies to any board game for which we believe that the playing algorithm has the same behavior over the possible rotations and reflections.*

*To illustrate the construction of  $C([w])$  we refocus on Tic-Tac-Toe with  $n = 9$ . We start by noticing that the first move has only three different possibilities (up to symmetry): center, corner, or center of an edge. If we number the places as  $1, 2, \dots, 9$  as shown in Figure 1, then the first move can only be from the symmetry equivalence classes  $[1]$ ,  $[2]$ , or  $[5]$ . Concretely, a move from  $[1]$  could*

---

<sup>1</sup>Wikipedia's definition of the game of Tic-Tac-Toe — <https://en.wikipedia.org/wiki/Tic-tac-toe>



1	2	3
4	5	6
7	8	9

Figure 1: A Tic-Tac-Toe board

be 1 but it also could be 3, 7 or 9. As we assume that the algorithm either makes a mistake or not on each equivalence set, then it is enough under that assumption to test, say 7, in order to test a game with the first movement coming from  $[1] = \{1, 3, 7, 9\}$ . We will relax this assumption by introducing a Bayesian approach that represents our confidence that an equivalence class does not contain a bug in Section 4.

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

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

We proceed in this way to define an equivalence set over the sequence of entire game plays which define the equivalent classes  $C([w])$ . Note 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  $w$ .

Many times additional tester domain knowledge may have to do with the partition of the set of possible events. For example, when the program statements are the set of events, tester knowledge may indicate that problems are more likely along error paths. This partitions the program statements (under an appropriate programming paradigm) to program statements that belong to error paths and program statements that do not. Our framework allows for the infusion of such tester knowledge into the test design. The following example illustrates one way of doing that that formally.

**Example 7** (disjoint sets of events). *For a partition  $\bigsqcup_{i=1}^n \Sigma_i = \Sigma$ , take  $I = \Sigma_1^* \times \cdots \times \Sigma_n^*$  and  $C(w_1, \dots, w_n) = \bigcap_{i=1}^n \mathcal{L}(\hat{\Sigma}_i^* w_i [1] \hat{\Sigma}_i^* \cdots \hat{\Sigma}_i^* w_i [|w_i|] \hat{\Sigma}_i^*)$  where  $\hat{\Sigma}_i = \Sigma \setminus \Sigma_i$ .*

#### 4. Testing methodology: 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 upper left corner and then center while not exposing it when playing upper right corner and then center.
2. In practice we do not control all of the systems under test inputs. Specifically, it is harder to control or even specify the environment configuration and state. For example, if the test opens a file and writes to it, the test

may control whether or not the file exists and that we have write permission to it. 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 the risk measure to guide the generation of additional tests to best decrease the risk of having a bug in the SUT.

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  $t$ . We execute  $t$  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 (the  $\alpha$  and  $\beta$  parameters are 1,  $\alpha$  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  $\alpha$  by one each time the test succeeds and increase  $\beta$  by one if it fails. When the SUT is corrected based on the detected bugs,  $\alpha$  and  $\beta$  are reset to 1 for each  $C(i)$  and the process repeats.

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 that we have a bug in the system. 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  $P(Y|X_{C_i} = 1)$ . 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  $\sum_{i \in I} l(C(i))P(Y|X_{C_i} = 1)P(X_{C(i)} = 1)$ .

To handle the case in which  $C(i)_{i \in I}$  is not a partition, we focus on the same random variables above and apply Bayesian-network discovery techniques [8] to learn the edges in the network. The network is learned from a profile of the system usage.

The above risk and loss functions can be used to 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 in order 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 to find a bug in  $C(i)$  is low.

## 5. Sampling and Optimization Based Tool

We have developed a tool that demonstrates how our approach can be used in practice. The tool is available at <https://github.com/bThink-BGU/TestSuiteGenerator>, including the code used for the experiments section below. In this section, we describe the design of this tool and how it can be used.

### 5.1. Input Language: Behavioral Programming

Since we are interested in covering test requirements not just in code coverage, we need a model of the requirements that tell us what to test. We chose to use the behavioral programming (BP) modeling paradigm to this end. The reason for this choice is that BP is designed to model the requirements in an

incremental way that allows specifying test scenarios without relation to the implementation. The paradigm is particularly designed to allow users to specify models in a natural and intuitive manner, that is aligned with how humans perceive the requirements of a system.

In 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 normally 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 is a model of what needs to be tested called a behavioral program (*b-program*). 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. [1] proposed a simple b-thread integration protocol. The protocol consists of each b-thread submitting a statement before the selection of 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 that 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 all 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 [9, 1]. In this paper, we use and slightly extend the implementation of BP, called BPjs, developed by Bar-Sinai and others [10]. In this implementation, the b-threads are specified using simple JavaScript code snippets (hence the name), and the integration mechanism is developed in Java using the Rhino library (see <https://github>.

com/mozilla/rhino). We use the tool for generating a labeled transition system that models the sequences of events that test the system. Since the set is usually huge, we need to sample it wisely, as described next.

### 5.2. A ranking-based sampling of the test model

To test the performance of a generalized coverage criteria in the real world, we added a mechanism to BPjs that allows users to generate small test suites (sets of tests) from their programs. The new mechanism uses a ranking function provided by the user and applies a genetic algorithm (GA) to construct test suites with high ranks.

Our experimental tool takes as input a procedure that defines the ranking function. This function takes a test-suite, represented as a set of arrays of events, and returns a number that models how well the suite covers the criteria that the user is interested in. In Section 6 we measure different coverage criteria using this tool. We used, for example, this mechanism to express our interest in counting how many different regular expressions of the form  $\Sigma^*\sigma_1\Sigma^*\sigma_2\Sigma^*$ , where  $(\sigma_1, \sigma_2)$  is a pair of events, are matched by at least one test in the suite. This is, of course, Kuhn’s 2-way sequence coverage criterion.

GA was used to extract small test suites (sets of tests) out of a large set of sampled tests. The individuals are candidate test suites and the fitness function is the ranking function that the user provided. We applied standard mutations and crossover operators (for set individuals) using the Jenetics (`jenetics.io`) Java library with the parameters: `CROSSOVER_OPERATOR=SinglePointCrossover`; `CROSSOVER_PROBABILITY=70%`; `MUTATION_PROBABILITY=30%`; `SELECTION_METHOD=Tournament` with  $k = 3$ ; `NUMBER_OF_GENERATIONS=300`; and `MAXIMAL_PHENOTYPE_AGE=10`.

The proposed tools are designed to support the methodology described in Section 2 above, in general, and Definition 2 specifically. While the definition only specifies a Boolean condition of coverage (i.e., a criterion is either covered or not), the tool, for practical reasons, allows for a quantitative measure of coverage. Specifically, we propose to count the number of  $C(i)$ s that a test suite covered and try to maximize it as much as possible.

```

int rankTestSuiteNext(Set<List<String>> testSuite) {
    var newTestSuite = new ArrayList<>();
    for (var test : testSuite) {
        var events = test.stream().collect(Collectors.toList());

        for (int x = 0; x < events.size()-1; x++)
            newTestSuite.add("(" + events.get(x) + "," + events.get(x+1) + ")");
    }
    return newTestSuite.stream()
        .collect(Collectors.groupingBy(e->e, Collectors.counting()))
        .size();
}

```

Listing 1: An example code (in Java) of a ranking function (2-way). The function receives a test suite and returns its rank, the number of sequences of two events covered.

The terms map as follows:

- The test model  $P$  is the set of executions of the b-program (the set of b-threads provided by the user).
- The test suite  $S$  is a set of tests generated by the tool. The size of this set is a parameter provided by the user, i.e., 5, 10, or 20, according to the resources available for testing.
- The coverage count  $\Gamma$  is given using a ranking function that also encapsulates the coverage criterion  $C$  implicitly. Specifically, we require users to provide us with a procedure, called ranking function, that takes a test suite  $S$  and returns  $\Gamma_C(S, P)$ . This procedure can count, for example, the number of different 2-way (or 3-way) sequences that appear in  $S$  as shown in Listing 1.

Our proof of concept tool first samples a large set of test cases from the model (50,000 in our experiments) and then extracts a test suite by looking for a subset of the required size (given by the user) with a high rank. As said above, we apply GA for this purpose, though other optimization techniques are applicable as well. 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.

## 6. Evaluation

For evaluation, we describe how our modeling language serves for describing test models for different types of systems and how generalized coverage criteria cater for extracting test suites from these models.

The first section of our evaluation is a qualitative, demonstration the applicability of the proposed modeling approach for testing. The second section is focused on a quantitative evaluation of our generalization of coverage criteria and optimization techniques.

### 6.1. Sequence-testing modeling with behavioral programming

We now turn to present some examples of sequence-testing (ST) models specified using the BP paradigm. All the examples in this paper can be found at [github.com/bThink-BGU/Papers-2021-Sequence-Testing](https://github.com/bThink-BGU/Papers-2021-Sequence-Testing).

We begin with one of the benchmark examples of Bombarda [5]. We show how the regular expression proposed by Bombarda can be modeled using BP offering better modularity and readability.

*The vault example.* This is an example of a vault that can be unlocked only by the combination “12345”. The code in Listing 2 specifies two b-threads — one for pressing the keys and one for checking the code correctness. 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 each of the two testing requirements. The first b-thread is aligned with the requirement that the safe shall have a keypad with nine digits that can be pressed at any order. It continuously requests to press 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).

#### 6.1.1. More complex, non-event-based, systems

We turn to demonstrate how BP can be used for modeling systems that are not usually perceived as event based. We will demonstrate it on Moodle



```

let AnyKey = [ Event("1"), ..., Event("9") ]

bthread('press keys', function() {
  while(true)
    // At each iteration, one event is selected randomly.
    sync({ request: AnyKey })
})

bthread('check code', function() {
  for(let i=0; i < 5; i++) {
    let key = sync({ waitFor: AnyKey }).name
    if (code[i] != key)
      stop() // stop the execution
  }
  // Code is correct.
  sync({ request: Event("Open"),
        block: Event("Open").negate() })
  state.accepting()
  stop()
})

```

Listing 2: The vault example

— a popular, open-source learning management system, used by educators to create private websites with online courses to achieve learning goals [11]. While Moodle is not considered an event-based system, the interaction between the users and the system can be modeled as an event-based system. Our BP-based Moodle model (Listing 3) has three b-threads, each aligned to a different aspect of the system behavior, handled by a different type of user. The first b-thread specifies a behavior of an administrator that creates a course and enrolls users to 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.

According to Moodle documentation<sup>2</sup>, a teacher cannot add questions to a quiz once a student attempts the quiz. 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 during the time that the teacher is adding another question, then the UI misbehave and displays an exception

---

<sup>2</sup>Moodle Quiz FAQ — [https://docs.moodle.org/39/en/Quiz\\_FAQ](https://docs.moodle.org/39/en/Quiz_FAQ)

States	Events	Edges	Traces
117	22	228	12200

Table 1: Statistics on the automatically generated LTS of the Moodle model

(tested on Moodle v3.9). 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. We used Selenium WebDriver<sup>3</sup> to automate the script in the browser.

The statistics on the automatically generated LTS are presented in Table 1. There are 22 distinct events in this behavioral program, one for each synchronization point with a `request`, plus two additional for the two `AnswerQuestion` events since there are two questions. These 22 events generate 12,200 distinct traces! While the specification of each b-thread is easy to understand, the generated LTS of these three b-threads (depicted in Figure 2) is too complex to manually specify, or even understand.

### 6.2. The alternating-bit protocol: using different coverage criteria

The alternating-bit protocol (ABP) [12] is a classical full-duplex protocol that sends data from *sender* to *receiver* on an unreliable communication channel. It is a standard benchmark for formal verification and modeling [13]. The ABP protocol starts by sending a data packet (the ‘send’ event in our model) and repeats sending this packet until receiving an acknowledgment (‘ackOK’). After the acknowledgment, the sender starts sending the next data item, if it exists. Each data packet is attached with metadata of a single bit that indicates the correct order of the messages. This bit alternates in each data item. As long as the sender does not receive any positive or negative acknowledgment (‘ackNok’), the same data packet is resent again and again. The receiver is waiting for data packets. When a packet is received, the receiver checks whether the message that has been received is indeed the message that the receiver ex-

---

<sup>3</sup>Selenium WebDriver — <https://www.selenium.dev/>

```

bthread('Admin adds a course', function () {
  sync({ request: Event('Session.Start', {s: 'admin'}) })

  sync({ request: Event('AddCourse.Begin', {name: 'course 1'}) })
  sync({ request: Event('AddCourse.Submit', {name: 'course 1'}) })
  sync({ request: Event('EnrollUser.Begin',
    {course: 'course 1', user: 'Terri Teacher', role: 'Teacher'}) })
  sync({ request: Event('EnrollUser.Submit',
    {course: 'course 1', user: 'Terri Teacher', role: 'Teacher'}) })
  sync({ request: Event('EnrollUser.Begin',
    {course: 'course 1', user: 'Sam Student', role: 'Student'}) })
  sync({ request: Event('EnrollUser.Submit',
    {course: 'course 1', user: 'Sam Student', role: 'Student'}) })

  sync({ request: Event('Session.End', {s: 'admin'}) })
})

bthread('Teacher adds a quiz with questions', function () {
  let c = sync({ waitFor: Any('EnrollUser.Submit',
    {role: 'Teacher'}) }).data
  sync({ request: Event('Session.Start', {s: 'teacher'}) })

  sync({ request: Event('AddQuiz.Start',
    {s: 'teacher', course: c.course, name: 'quiz 1'}) })
  sync({ request: Event('AddQuestion.Start',
    {s: 'teacher', quiz: 'quiz 1', name: 'Question 1'}) })
  sync({ request: Event('AddQuestion.Submit',
    {s: 'teacher', quiz: 'quiz 1', name: 'Question 1'}) })
  sync({ request: Event('AddQuestion.Start',
    {s: 'teacher', quiz: 'quiz 1', name: 'Question 2'}) })
  sync({ request: Event('AddQuestion.Submit',
    {s: 'teacher', quiz: 'quiz 1', name: 'Question 2'}) })
  sync({ request: Event('AddQuiz.Submit',
    {s: 'teacher', course: c.course, name: 'quiz 1'}) })

  sync({ request: Event('Session.End', {s: 'teacher'}) })
})

bthread('Student answers false to questions', function () {
  sync({ waitFor: Any('EnrollUser.Submit', {role: 'Student'}) }).data
  sync({ request: Event('Session.Start', {s: 'student'}) })

  let q = sync({ waitFor: Any('AddQuestion.Submit') }).data
  sync({ request: Event('AnswerQuestion.Start',
    {quiz: q.quiz, name: q.name, answer: 'False'}) })
  sync({ request: Event('AnswerQuestion.Submit',
    {quiz: q.quiz, name: q.name, answer: 'False'}) })

  sync({ request: Event('Session.End', {s: 'student'}) })
})

```

Listing 3: The Moodle example

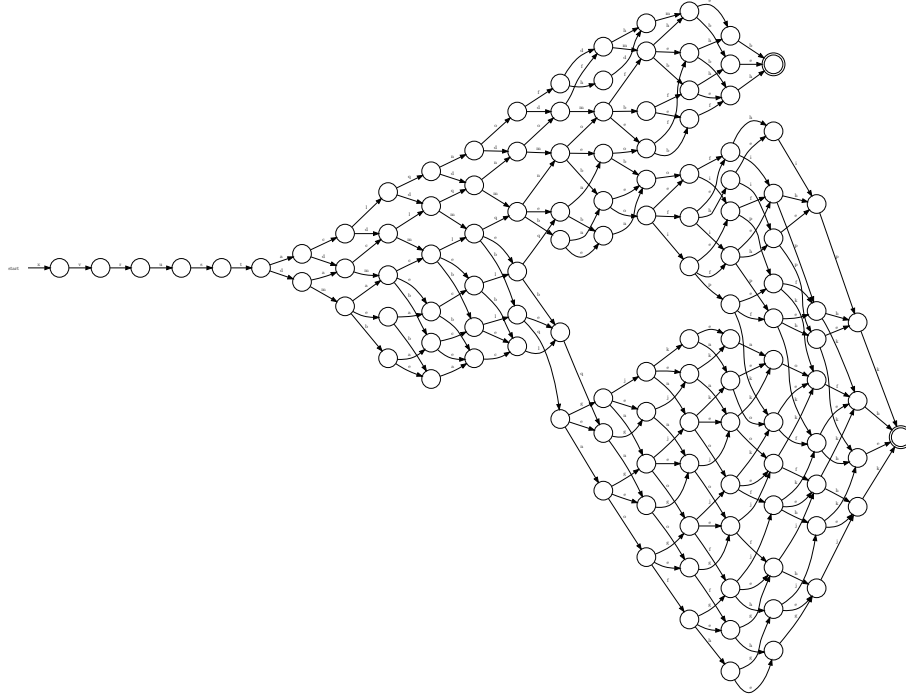


Figure 2: The generated LTS of the Moodle example.

pected to be received by examining the bit of the message order attached to the message. If it does match, it sends a message to the sender indicating that the message is correct ('recAck'); if the attached bit indicating the order of the messages does not match, it sends a failure message ('recNak') and waits to the following message. Since ABP allows to deal with communication in unreliable channels, our model has realized two types of noises that can occur. One is the loss of a message on the channel ('r2tLoss', 't2rLoss') and the other is the change of the order in which the messages arrive on the channel, an event that simulates the arrival of messages on a multi-routing channel ('r2tReorder', 't2rReorder'). These interruptions on the channels can occur in both directions from the sender to the receiver or vice versa.

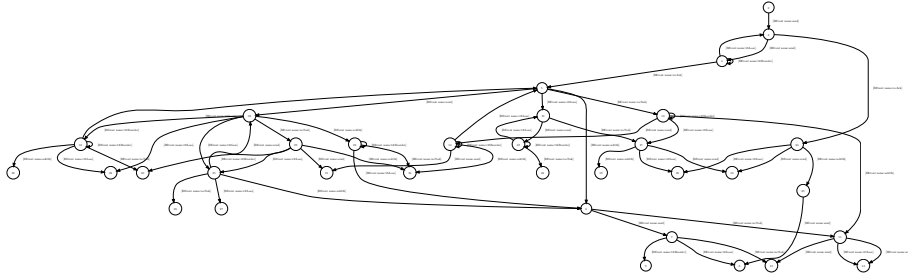


Figure 3: Part of generated LTS model of the alternating-bit protocol. The model presents up to a depth of seven.

### 6.2.1. The alternating-bit protocol: Events generator

Our demonstration of ABP models it as BP executable model; we can run the model and generate, for each data item as input, sequences of events that start with 'send' event and end with 'success' event if all data items are received correctly or 'fail' if not. Figure 3 presents part (depth of seven) of our ABP model. Each test case is a path in the graph, starting in a root and ending in an accepting node. Even though the number of different events in ABP is relatively small, it is very hard to generate coverage manually.

We borrowed our ABP model from an example proposed by Antti [14] included in the distribution of Intel's model-based testing named fMBT [15]. The model proposed by Antti [14] is based on rules for triggering events and transformations that happen when an event is triggered. This maps easily to COBP [9] by defining corresponding queries and context-dependent queries. The model is used to generate sequences of events that follow the protocol. Each sequence of events represents a test case for the system under test (SUT). We apply a 'white box' approach where the model is used to drive the system under test and to verify that it follows the protocol. Each test case (list of events) triggered the SUT to act and execute the event action (e.g., the 'send' event trigger the SUT to send a data packet). An instrumentation layer checks if the condition to execute the event are met and then the action is done and the ABP internal data of the SUT updates accordingly. If the conditions for triggering the event

```

send,recAck,r2tLoss,send,send,t2rReorder,recNak,send,ackOk,recNak,recNak,
ackNok,send,ackNok,recAck,r2tLoss,send,recNak,r2tLoss,send,send,recNak,
ackOk,recNak,r2tLoss,send,recAck,r2tLoss,send,recNak,send,recNak,r2tLoss,
ackOk,send,recAck,send,send,recNak,send,r2tReorder,t2rReorder,t2rReorder,
r2tReorder,t2rReorder,ackOk,recNak,r2tReorder,send,r2tReorder,r2tLoss,
recNak,ackNok,ackNok,send,recAck,ackOk,recNak,ackNok,send,send,t2rLoss,
send,recAck,send,t2rLoss,send,t2rReorder,recNak,ackOk,r2tLoss,success,
recNak,ackNok

```

Listing 4: An example of one test case that was generated automatically by the ABP model. This test case was designed to send six data items to the receiver. A test suite consists of several test cases such as this.

are not met by the SUT, an error occurs and the test fails. If all events execute successfully then the test succeeds. Listing 4 shows an example of one test case that was generated by our ABP model. Each test suite was constructed by grouping several test cases. We used the mechanism described in Section 5 for generating suites that maximize different ranking functions as elaborated below.

### 6.2.2. The alternating-bit protocol: POC Methods

We planted bugs in the SUT and checked, for each test case or test suite, whether it catches the planted bug or not. The bug we planted in the SUT is triggered when certain combinations of consecutive events occur, e.g., ‘recOk’ followed by ‘ackOk’ (“2-way” bug) or ‘recOk’ comes three times in a sequence (“3-way” bug). This is a common bug pattern that appears in many situations. Our goal was to use the BP-based executable model described above to generate many test cases, collect them into test suites, and examine how often they find the planted bugs in the SUT. We ran the model 50,000 times and generated 50,000 valid test cases. Out of these test 50,000 cases, we built test suites, each test suite holds ten test cases. Our challenge was to find the best test suites that have the highest probability to catch the planted bugs. To this end, we defined two ranking functions, as follows. One ranking function counts all  $n$  consecutive events in a test suite. The second ranking function is based on Kuhn’s method that counts all sequences of  $n$  events but not necessarily consecutive  $n$  events. Our thesis was that maximizing the consecutive events

maximizes the probability to catch the bug. i.e., that the first 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’s criterion, has real usage.

*6.2.3. The alternating-bit protocol: Catch the bug*

We tried three methods of optimizing the test suite and measured the probability of catching the planted bug. The results are summarized in Table 2. The first method, titled ‘Random’, is simply peeking ten test cases at random. The second, titled ‘Kuhn’, is maximizing the number of elements of  $\{\Sigma^* \sigma_1 \Sigma^* \sigma_2 \Sigma^* : \sigma_1, \sigma_2 \in \Sigma\}$  (or its equivalent for three letters bugs) that are matched by at least one test case. The third method, titled  $\Sigma^* w \Sigma^*$ , is maximizing the number of elements of  $\{\Sigma^* w \Sigma^* : w \in \Sigma^n\}$  (where  $n$  is two or three, depending on the length of the bug under examination) that is matched by at least one test case. In the two last methods, as described in the previous subsection, we first sampled 50,000 random tests and then used GA to find a subset with as any matched of the above regular expressions as possible.

We checked the probability of each of these methods to find four different bugs, each triggered by a specific sequence of events. The column titled ‘ $w$ ’ in the table species the sequence of events that triggered the planted bug. Some sequences are of length two and some of length three, some are more frequent and some happen only in rare corner cases. We ran this process 1,000 times for each of the planted bugs. In every iteration, we checked the three methods and counted how many test cached the bug.

The results in Table 2 show that the third method, titled  $\Sigma^* w \Sigma^*$  is always better and that it is much better than the method titled ‘Kuhn’ when the method titled ‘Random’ gives low results, i.e., when the bug is rare. The moral of this experiment is that there are cases where generalized coverage criteria are significantly more effective in catching certain types of bugs than random sampling and than Kuhn’s coverage criterion.

We tested with 3-way and with 2-way bugs, with different sequences of

events. The tables show that some sequences of events produced bugs that are detected by all three methods and some were detected with high enough probability only by a generalized coverage criterion. This is because some sequences are prevalent enough to be found randomly and some need to be directly targeted because the probability to get them in random is too low.

#### 6.2.4. *The alternating-bit protocol: Best test-suite generation method*

We have compared three methods for generating a test suite for different coverage criteria and different suite sizes. To produce the test suites, we first created a base of 50,000 valid test cases extracted from our model. We then used three test suite generation methods: 1) the ‘unranked’ method of peeking random  $n$  test cases from the base and using them as a test suite; 2) brute force method of repeating the ‘unranked’ process 1000 times and choosing the suite with the highest rank; 3) the GA method is described in Section 5.

We used two types of ranking functions for methods 2 and 3 above: 1) Kuhn’s sequence coverage criterion: the ranking function counts the number of pairs  $c_1, c_2 \in \Sigma$  such that the regular expression  $\Sigma^*c_1\Sigma^*c_2\Sigma^*$  is matched by at least one test case in a suite; 2) A generalized criterion: the ranking function counts the number of pairs  $c_1, c_2 \in \Sigma$  such that the regular expression  $\Sigma^*c_1c_2\Sigma^*$  is matched by at least one test case in a suite; The generalized criterion that we use models a heuristic, that applies to many systems, that many bugs are triggered by a specific sequence of events that come directly one after another. We call it a ‘generalized’ criterion because it demonstrates how developers can generalize Kuhn’s type of thinking for modeling new heuristics that target new types of bugs. The numbers in Table 2 show that this improves on using Kuhn’s criterion in some cases. Note, that we are not claiming that our criterion is better than Kuhn’s criterion in general, only that generalizing allows us to better target specific types of bugs.

We checked each method with both ranking functions for three sizes of a test suite: 5, 10, and 20. In each test, we measured the value of a ranking function (called ‘rank’) and runtime (‘time’) in seconds.



$w$	Unranked (Random)	Kuhn	$\Sigma^*w\Sigma^*$
$ackOk \cdot ackOk$	0.219	0.198	0.709
$recNak \cdot recAck$	0.732	0.643	0.988
$ackNok \cdot ackNok \cdot recAck$	0.067	0.091	0.216
$send \cdot send \cdot ackOk$	0.821	0.806	0.978

Table 2: The table shows the probability to catch a planted bug. Each row represents the probability to catch a specific bug: two cases when the bug was caused by two consecutive events and two cases by three consecutive events. Each column represents a test suite of ten test cases generated out of 50,000 test cases. The unranked method is peeking randomly test suites; Kuhn method finds the best test suite using Kuhn ranking and checking 1,000 test suites out of 50,000 test cases;  $\Sigma^*w\Sigma^*$  method finds the best test suite using 'consecutive events' ranking and GA algorithm.

Table 3 and Figure 4 shows that the GA technique works best both in terms of time and in terms of ranking.

## 7. Conclusion

“I always thought something was fundamentally wrong with the universe.” (the Hitchhiker’s Guide to the Galaxy series [16]). Nevertheless, we try to do the best we can.

System and software testing move along time as a pendulum between two opposite ends. On the one hand, there is a need to test the system in the best quality and deliver bug-less software. On the other hand, the need to reduce the testing effort in both resources and scheduling or time to market.

To tackle this dilemma, identified three related issues in testing and presented three major contributions, one for each:

1. **How to specify the space that needs to be covered:** We defined a generalized, automata-based approach for specifying coverage criteria.
2. **Finding a finite, relatively small, test suite that covers this space:**

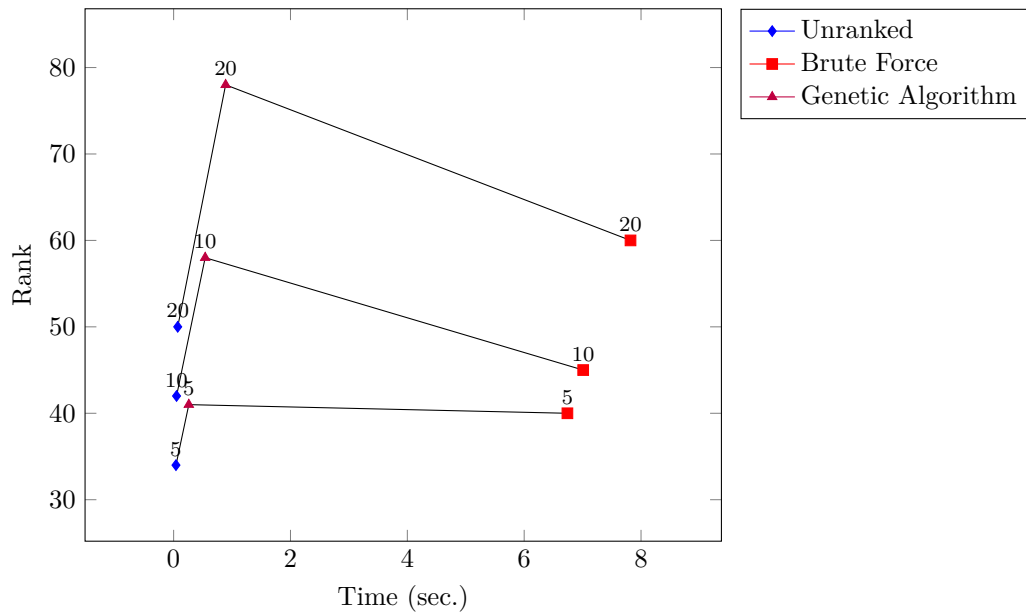


Figure 4: Efficiency graph - this graph displays the calculation time-to-quality ratio for the ranking function that we have developed for the  $\Sigma^*w\Sigma^*$  coverage criterion. The number above each data point represents the size of the test suite (number of test cases in a test suite). The graph displays the data of the second row in Table 3. This visualization clearly shows that the genetic programming heuristic search approach is better than brute force both in time and in the obtained rank in all the three test sizes that we examined. The advantage grows with the size of the test.

Opt. method		GA			Brute Force			Unranked (Random)		
Suite size		5	10	20	5	10	20	5	10	20
Kuhn	Time (s)	.27	.46	.69	7.18	8.85	9.4	.005	.005	.005
	Rank	121	121	121	121	121	121	119	119	119
$\Sigma^*w\Sigma^*$	Time (s)	.26	.54	.89	6.74	7.01	7.82	.004	.005	.007
	Rank	41	58	78	39	55	62	34	42	50

Table 3: The table introduces the effective method for finding a test suite in terms of ranking and running time. The table shows three methods for generating a test suite out of 50,000 possible valid test cases. One is an intelligent search using GA for search efficiency, the other is based on computing power and examines 1,000 cases from the space of possibilities, and the third is random, without ranking at all. For each of the methods, three possible sizes for test suites were examined: 5, 10, and 20. We also examined two scoring functions, one based on the Kuhn criterion and the other the criterion we defined, of ‘consecutive events’. In each case, we examined both the result obtained from the ranking function and the running time.

We proposed to model the testing software using the behavioral-programming (BP) paradigm, to allow the automatic creation of composite tests.

- 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.

In addition to the theoretical contributions, we presented the following practical contributions:

- Coverage criteria:** We provided a set of valuable coverage criteria that may be applied to various domains.
- Testing software modeling:** We provided BP-based models for a testing software for classic benchmark example of sequence testing. In addition, we demonstrated how the approach can be applied for modeling systems that are not usually perceived as event-based, specifically, the widely used, Moodle learning platform.

- **Proof-of-concept tool:** We developed a tool that allowed us to present the entire process. The tool starts with the system modeling phase in BP, later designing different test scenarios from the testing model using different ranking functions, and finally analyzing the results from various aspects.

From what we have presented in this article, it is possible to expand to other research areas in different directions in the space of testing. All of which are focused on achieving the goal of efficiently testing process, 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 [3]. This modeling is a prerequisite to the approach presented in this article. Another way to expand is using statistical tools in the stem of an efficient and focused test suites generation process to testing coverage, increasing the probability of identifying the faults.

## References

- [1] D. Harel, A. Marron, G. Weiss, Programming Coordinated Behavior in Java, in: ECOOP 2010 – Object-Oriented Programming, Springer Berlin Heidelberg, 2010, pp. 250–274. doi:10.1007/978-3-642-14107-2\_12.
- [2] D. Harel, A. Marron, G. Weiss, Behavioral Programming, Communications of the ACM 55 (7) (2012) 90. doi:10.1145/2209249.2209270.  
URL <http://dl.acm.org/citation.cfm?doid=2209249.2209270>
- [3] Y. Weiss, Testing reactive systems using behavioural programming, a model centric approach (2021). arXiv:2112.01538.
- [4] D. R. Kuhn, J. M. Higdon, J. F. Lawrence, R. N. Kacker, Y. Lei, Combinatorial methods for event sequence testing, in: 2012 IEEE Fifth International Conference on Software Testing, Verification and Validation, IEEE, 2012, pp. 601–609.
- [5] A. Bombarda, A. Gargantini, An automata-based generation method for combinatorial sequence testing of finite state machines, in: 2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE, 2020, pp. 157–166.
- [6] D. Kuhn, R. Bryce, F. Duan, L. Ghandehari, y. Lei, R. Kacker, Combinatorial testing. theory and practice, Advances in Computers 99 (2015) 1–66. doi:10.1016/bs.adcom.2015.05.003.
- [7] E. Farchi, Y. Nir, S. Ur, Concurrent bug patterns and how to test them, in: Proceedings International Parallel and Distributed Processing Symposium, 2003, pp. 7 pp.–. doi:10.1109/IPDPS.2003.1213511.
- [8] M. Koivisto, K. Sood, Exact bayesian structure discovery in bayesian networks, J. Mach. Learn. Res. 5 (2004) 549–573.
- [9] A. Elyasaf, Context-Oriented Behavioral Programming, Information and Software Technology 133 (2021) 106504. doi:https:

[//doi.org/10.1016/j.infsof.2020.106504](https://doi.org/10.1016/j.infsof.2020.106504).

URL <http://www.sciencedirect.com/science/article/pii/S095058492030094X>

- [10] M. Bar-Sinai, G. Weiss, R. Shmuel, BPjs - An Extensible, Open Infrastructure for Behavioral Programming Research, in: 21st ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings, MODELS-Companion 2018, 2018. doi: 10.1145/3270112.3270126.
- [11] Moodle — open-source learning platform, <https://moodle.org>, accessed: 2021-03-10.
- [12] K. A. Bartlett, R. A. Scantlebury, P. T. Wilkinson, A note on reliable full-duplex transmission over half-duplex links, *Communications of the ACM* 12 (5) (1969) 260–261.
- [13] K. Chukharev, D. Suvorov, D. Chivilikhin, V. Vyatkin, Sat-based counterexample-guided inductive synthesis of distributed controllers, *IEEE Access* 8 (2020) 207485–207498.
- [14] fmbt-alternating-bit protocol, <https://github.com/intel/fmbt/tree/master/examples/link-layer-protocol> (2019).
- [15] Intel, fmbt, <https://01.org/fmbt/>; <https://github.com/intel/fmbt> (2012).
- [16] D. Adams, *The restaurant at the end of the universe*, Harmony Books, New York, 1981.