Flexible Combinatorial Interaction Testing

Hanefi Mercan, Arsalan Javeed, and Cemal Yilmaz, Member, IEEE

Abstract—We present Flexible Combinatorial Interaction Testing (F-CIT), which aims to improve the flexibility of combinatorial interaction testing (CIT) by eliminating the necessity of developing specialized constructors for CIT problems that cannot be efficiently and effectively addressed by the existing CIT constructors. F-CIT expresses the entities to be covered and the space of valid test cases, from which the samples are drawn to obtain full coverage, as constraints. Computing an F-CIT object (i.e., a set of test cases obtaining full coverage under a given coverage criterion) then turns into an interesting constraint solving problem, which we call *cov-CSP*. cov-CSP aims to divide the constraints, each representing an entity to be covered, into a minimum number of satisfiable clusters, such that a solution for a cluster represents a test case and the collection of all the test cases generated (one per cluster) constitutes an F-CIT object, covering each required entity at least once. To solve the cov-CSP problem, thus to compute F-CIT objects, we first present two constructors. One of these constructors attempts to cover as many entities as possible in a cluster before generating a test case, whereas the other constructor generates a test case first and then marks all the entities accommodated by this test case as covered. We then use these constructors to evaluate F-CIT in three studies, each of which addresses a different CIT problem. In the first study, we develop structure-based F-CIT objects to obtain decision coverage-adequate test suites. In the second study, we develop order-based F-CIT objects, which enhance a number of existing order-based coverage criteria by taking the reachability constraints imposed by graph-based models directly into account when computing interaction test suites. In the third study, we develop usage-based F-CIT objects to address the scenarios, in which standard covering arrays are not desirable due to their sizes, by choosing the entities to be covered based on their usage statistics collected from the field. We also carry out user studies to further evaluate F-CIT. The results of these studies suggest that F-CIT is more flexible than the existing CIT approaches.

Keywords—Combinatorial interaction testing, covering arrays, sequence covering arrays, constraint solving, structural coverage, coverage criteria

1 Introduction

Exhaustively testing the input spaces of modern software systems in a timely manner (if not impossible at all) is generally far beyond the available resources for testing [1], such as time, computers, storage devices, network resources, and person-hour. Combinatorial interaction testing (CIT) approaches systematically sample the input space and test only the selected instances of the system's behavior [1], [2]. Note that the term "input" in CIT is used in the most general sense to refer to any factor, which can affect program executions, such as configuration options, input parameters, user events, etc.

CIT approaches typically model the software under test as a set of parameters, each of which takes its values from a discrete domain. As not all possible combinations of parameter values may be valid in practice, the model can also have a set of constraints, which invalidate certain combinations. Based on this model, CIT then generates a sample, i.e., a set of test cases, which from now on will be referred to as a *CIT object*, meeting a specified *coverage criterion*. That is, the sample contains some specified combinations of parameters and their values. For instance, *t-way covering arrays* – a well-known CIT approach, where t is called the *coverage strength* – requires that each valid combination of parameter values for every combination of t parameters appears at least *once* in the sample [3], aiming to reveal all the failures caused by the interactions of t or fewer parameters.

As an example, which will further be studied in detail in Section 2, Figure 1a presents a configurable system with 6 compile-time configuration options $(o_1, \ldots, o_6)$ implemented by using preprocessor directives. Each option has two levels of settings $\{$*(T)rue*, *(F)alse*$\}$ and there are no inter-option constraints (i.e., all combinations of option settings are valid). The set of test cases in Figure 1b represent a 2-way covering array, i.e., a CIT object, for this system. Since $t = 2$, all pairwise combinations of settings for these 6 configuration options can be found in at least one of the 7 test cases selected by this CIT object.

To reduce the cost of testing, CIT constructors, i.e., the tools to compute CIT objects, aim to obtain a full coverage under the given criterion by using the smallest number of test cases possible. CIT has indeed been successfully used in many application domains, including systematic testing of network protocols [4], input parameters [5], software configurations [6], software product lines [7], multi-threaded applications [8], and graphical user interfaces [9].

• *H. Mercan, A. Javeed, and C. Yilmaz are with the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey. E-mail: {hanefimercan, ajaveed, cyilmaz}@sabanciuniv.edu*

```
1 #ifdef (o1 && o2)
2  #ifdef (o3 || o4)
3   ...
4  #endif
5 #endif

6 #ifdef (o5)
7  #ifdef (o6)
8   ...
9  #endif
10 #endif
```

(a)

test cases						decision outcomes			
o_1	o_2	o_3	o_4	o_5	o_6	$o_1 \wedge o_2$	$o_3 \vee o_4$	o_5	o_6
T	F	T	F	F	F	*F*	-	F	-
F	T	F	T	T	F	*F*	-	T	F
T	T	T	T	F	T	*T*	*T*	F	-
T	F	F	F	T	F	*F*	-	T	F
F	F	T	F	T	T	*F*	-	T	T
F	F	F	T	F	T	*F*	-	F	-
T	T	T	F	F	T	*T*	*T*	F	-

(b)

entities to be covered
$e_1 : (o_1 \wedge o_2)$
$e_2 : \neg(o_1 \wedge o_2)$
$e_3 : (o_1 \wedge o_2) \wedge (o_3 \vee o_4)$
$e_4 : (o_1 \wedge o_2) \wedge \neg(o_3 \vee o_4)$
$e_5 : (o_5)$
$e_6 : (\neg o_5)$
$e_7 : (o_5 \wedge o_6)$
$e_8 : (o_5 \wedge \neg o_6)$

(c)

test cases						decision outcomes			
o_1	o_2	o_3	o_4	o_5	o_6	$o_1 \wedge o_2$	$o_3 \vee o_4$	o_5	o_6
T	T	T	T	T	T	*T*	*T*	*T*	*T*
F	F	T	F	F	T	*F*	-	*F*	*T*
T	T	F	F	T	F	*T*	*F*	*T*	*F*

(d)

Fig. 1: (a) An example set of preprocessor directives for a system with 6 compile-time configuration options, (b) an example 2-way standard covering array created for the system, (c) entities to be covered to obtain full coverage under the decision coverage criterion, and (d) an example test suite obtaining full coverage under the decision coverage criterion.

We, however, observe that when the actual CIT problems differ from the ones addressed by the existing CIT approaches, it can be difficult to use these approaches in an efficient and effective manner [1], [10], [11]. Note that, in this context, changes in CIT problems refer to changes in the coverage criteria or in the properties of the test spaces from which the samples are drawn, such that existing CIT constructors cannot be used as they are (i.e., requiring modifications, if at all possible) or demand excessive number of test cases to guarantee full coverage.

For example, if the coverage criterion in our running example was changed from t-way coverage to decision coverage [12], where the goal is to cover every outcome of a decision in Figure 1a at least once, then, to guarantee full coverage, the strength of the standard covering array to be used would be at least 4 (i.e., $t \geq 4$). This is because the outcome of the decision in line 2 (Figure 1a) depends on the interactions among 4 options, namely o_1, o_2, o_3, and o_4. This, however, requires to have at least 16 test cases, while a full decision coverage in this scenario can be achieved by using as little as 3 test cases, such as the ones given in Figure 1d.

Different CIT problems typically necessitate the development of specialized constructors. Taking a brief look at the historical perspective of covering arrays can help understand this trend: The very first variants of covering array constructors supported only pairwise testing of binary parameters, where $t = 2$ and each parameter had exactly two levels of values [13]. When these strict conditions were not met, the aforementioned objects were of little worth. Consequently, new CIT constructors were developed to handle the CIT problems, in which the parameters could take on a different number of values and the covering arrays could be computed for $t \geq 2$ [3]. However, as these objects assumed that all possible combinations of parameter values were valid, they were not appropriate in the presence of system-wide inter-parameter constraints, causing wasted resources in testing [14], [15]. Thus, new CIT constructors were developed to handle system-wide constraints [16], [17]. However, these objects then became inappropriate in the presence of test case-specific constraints, which led to the development of test case-aware covering arrays and their specialized constructors [18].

Developing specialized constructors can, however, be quite challenging and time-consuming, which is also apparent from more than 50 papers published in the literature, the sole purpose of which is to compute standard covering arrays [1], [2].

In this work, we introduce *Flexible Combinatorial Interaction Testing* (F-CIT) to improve the flexibility of CIT by eliminating the necessity of developing specialized constructors for every distinct CIT problem. In F-CIT,

both the entities to be covered and the space of test cases, from which the samples will be drawn, are expressed as constraints. The problem of computing an F-CIT object to cover all the requested entities then turns into an interesting constraint solving problem, which we call *cov-CSP* [19], [20], [21]. Given a set of constraints, each of which represents an entity to be covered, cov-CSP aims to divide the constraints into a minimum number of satisfiable clusters, such that each cluster depicts a subset of the entities, which can be tested together in a single test case. A solution for a cluster then represents a test case, covering all the entities included in the cluster. Consequently, the collection of all the test cases generated (one per cluster) constitutes an F-CIT object that covers each required entity at least once. In the remainder of the paper, we use the terms "CIT object" and "F-CIT object" interchangeably to refer to a set of test cases, which obtain full coverage under a given coverage criterion.

Going back to our running example (Figure 1), a decision coverage-adequate F-CIT object can be computed by representing each configuration option as a Boolean variable. Then, each entity to be covered corresponds to a distinct outcome of a decision, represented as a constraint in Boolean logic. Figure 1c presents all the entities that need to be covered to obtain full decision coverage for the system given in Figure 1a. These entities can be divided into 3 satisfiable clusters: $\{e_1, e_3, e_5, e_7\}$, $\{e_2, e_6\}$, and $\{e_4, e_8\}$. A solution for each cluster represents a test case. For example, the three test cases in Figure 1d, each of which corresponds to a solution computed for a distinct cluster, represent an F-CIT object, achieving full decision coverage.

Note that we use the term "constraint" in the general sense in F-CIT. That is, any restriction, independent of the logic in which it is specified, is considered to be a constraint. Consequently, an F-CIT constructor can be used as long as the entities to be covered are expressed as constraints and an appropriate procedure (i.e., a "solver") is provided to determine if a given set of entities can be tested together in a single test case, i.e., if the respective constraints can be satisfied together. In our running example (Figure 1), for instance, we can use an ordinary SAT or CSP solver [22] to figure out whether the constraints included in the clusters are satisfiable or not. We, therefore, believe that F-CIT can be used in a wide spectrum of domains, including software product lines, system of systems, and cyber-physical systems, in addition to the domains, which we used for evaluating F-CIT in this work, i.e., highly-configurable systems and event-driven systems.

Note further that using constraint solving techniques for combinatorial interaction testing is not a new idea [6], [10], [15], [16], [17], [23], [24]. However, the constraints in F-CIT are interpreted quite differently than the ones used in existing CIT approaches. More specifically, while the constraints in existing CIT approaches are typically used to specify combinations of parameter values that should be avoided, they are used in F-CIT to specify both the combinations (i.e., the entities) to be covered and the space of valid test cases, from which the samples are drawn. Therefore, the scope of a constraint in existing CIT approaches is all the test cases included in a covering array. That is, all of the selected test cases should satisfy all the constraints. On the other hand, the scope of a constraint representing an entity to be covered in F-CIT is limited to a single test case. That is, such a constraint needs to be satisfied by at least one test case, rather than by all the test cases selected, allowing a considerable amount of flexibility.

For instance, in our running example (Figure 1), expressing o_5 and $\neg o_5$ (i.e., the outcomes of the decision in line 6) as constraints to selectively determine what to cover in standard covering arrays, prevents the generation of any covering arrays as these conflicting constraints are enforced to be satisfied by all of the selected test cases. In F-CIT, however, these constraints are required to be satisfied by different test cases. For example, in Figure 1d, while the former constraint is satisfied by the first and third test cases, the latter one is satisfied by the second test case.

F-CIT is not a methodology for deciding what needs to be tested. It, in fact, takes as input a set of entities to be covered and aims to cover them in a minimum number of test cases by accommodating as many entities as possible in a single test case. Note that for a given CIT problem, regardless of whether an F-CIT constructor is to be used or a specialized constructor is to be developed, entities to be covered need to be enumerated and a procedure needs to be devised to determine whether a given set of entities can be covered together in a single test case or not. Once these are given, though, F-CIT provides a constructor right away.

Furthermore, F-CIT does not aim to replace existing CIT approaches. We, indeed, don't see much value in using F-CIT to compute the same CIT objects that the existing CIT constructors compute, as the generalized F-CIT constructors may not be as efficient and as effective as their specialized counterparts. We rather aim to reduce the barriers to applying CIT to other domains and testing problems by generalizing the construction of CIT objects as much as possible, so that the collective effort spent for developing F-CIT constructors can be leveraged to address a wider spectrum of CIT problems.

In this work, we present two F-CIT constructors, namely *cover-and-generate* and *generate-and-cover*. While the former aims to cover as many entities as possible in a cluster first and then generates a test case for the cluster, the latter generates a test case first and then

marks all the entities accommodated by the test case as covered.

To evaluate F-CIT, we then carry out three case studies, each of which focuses on a different CIT problem. In the first study, we use F-CIT to compute structural code coverage-based test suites. In the second study, we use F-CIT to improve a number of existing order-based covering arrays for testing event-driven systems by taking the reachability constraints imposed by graph-based models directly into account during the construction of CIT objects. In the last study, we use F-CIT to compute usage-based CIT objects, where the entities to be covered are determined according to their usage statistics in the field – an approach which is of importance especially when standard covering arrays are not desirable due to their sizes.

In these studies, we observed that it was either unclear how to use the existing constructors (if at all possible) to compute the requested CIT objects; or the existing constructors required non-trivial modifications or excessive number of test cases to guarantee a full coverage. F-CIT, on the other hand, used the same constructor to compute all the requested CIT objects without requiring any modifications, demonstrating the flexibility of the proposed approach.

We also carry out user studies to further evaluate the proposed approach. More specifically, we observe human subjects working on the smaller instances of the very same CIT problems we study in this work and report the results we obtained together with the insights we gained.

In previous work [25], we presented an initial set of definitions for F-CIT and provided an algorithm for computing F-CIT objects. And, we did this without providing any implementations or empirical evaluations. In this work, however, we present a simplified set of more formal definitions, an additional F-CIT constructor, a tool implementing the F-CIT constructors, and three case studies together with user studies, in which F-CIT is evaluated.

The contributions of this work can be summarized as follows:

- A flexible approach, F-CIT, for computing combinatorial objects for testing,
- Two constructors together with a tool implementing these constructors to compute F-CIT objects,
- Definition and construction of structure-based F-CIT objects,
- Definition and construction of order-based F-CIT objects,
- Definition and construction of usage-based F-CIT objects,
- A series of experiments demonstrating the flexibility of F-CIT,
- User studies demonstrating the usability of F-CIT.

The remainder of the paper is organized as follows: Section 2 provides a motivating example; Section 3 introduces F-CIT; Section 4 develops two constructors for computing F-CIT objects; Section 5 presents three case studies, demonstrating the drawbacks of the existing CIT approaches and how F-CIT addresses these drawbacks; Section 6 presents the user studies; Section 7 provides a general discussion of the applicability of F-CIT; Section 8 discusses threats to validity; Section 9 discusses related work; and Section 10 presents concluding remarks and possible directions for future work.

2 Motivating Example

In this section, we provide more details on our running example used in Section 1. In this example, we are concerned with compile-time configuration options implemented in the form of preprocessor directives, such as `#ifdef` and `#ifndef` directives found in C and C++. Figure 1a presents a hypothetical system with 6 compile-time configuration options, namely $o_1, \ldots, o_6$, each of which happens to have two levels of settings *(T)rue* and *(F)alse*. In the remainder of the paper, we use the term "if-then-else directive" to refer to an `#ifdef`, `#ifndef`, or a similar conditional branch directive, the conditions of which are comprised of only compile-time configuration options and/or constants. Note that such directives allow the decision outcomes to be directly controlled from outside the system by modifying the settings of the compile-time options as a part of the build process.

An if-then-else directive essentially describes how configuration options interact with each other. That is, the outcome of a decision (thus the behavior of the system) may change due to these interactions. Consequently, these interactions may need to be tested. To this end, one structural test adequacy criterion that the developers can use is the decision coverage (DC) criterion. A full coverage under DC is obtained when every decision, such as $(o_1 \wedge o_2)$ and $(o_3 \vee o_4)$ in Figure 1a, is evaluated to both $true$ and $false$.

Consider a scenario where the goal is to create a DC-adequate test suite for the system given in Figure 1a. Note that since a single configuration can cover multiple decision outcomes, the number of configurations required to obtain full coverage under DC can be reduced by covering as many outcomes as possible in each of the selected configurations. This is, indeed, the main motivation behind CIT. Therefore, CIT should be of help.

2.1 Applying standard CIT

It, however, turns out that standard covering arrays are infeasible to achieve the aforementioned coverage criterion in an efficient and effective manner.

As an in initial attempt, a standard 2-way covering given in Figure 1b is created. The first 6 columns in this figure present the 2-way covering array and the last 4 columns depict the outcomes of the decisions: 'T' for $true$, 'F' for $false$, and '$-$' for decisions that are not exercised due to some unsatisfied guard conditions. For example, the first row indicates that the decision $(o_3 \vee o_4)$ is not exercised by the configuration $(o_1 = T, o_2 = F, o_3 = T, o_4 = F, o_5 = F, o_6 = F)$, because the guard condition $(o_1 \wedge o_2)$ evaluates to F.

This covering array while obtaining a full coverage for the if-then-else directive between the lines 6 and 10 in Figure 1a, obtains only 75% DC coverage for the if-then-else directive between the lines 1 and 5, covering 3 out of 4 decision outcomes required for full coverage. More specifically, out of the decision outcomes $\{(o_1 \wedge o_2), \neg(o_1 \wedge o_2), (o_1 \wedge o_2) \wedge (o_3 \vee o_4), (o_1 \wedge o_2) \wedge \neg(o_3 \vee o_4)\}$, the last one where the inner decision $(o_3 \vee o_4)$ needs to be evaluated to F, is not covered. Note that this outcome can only be achieved with a single 4-way combination, in which o_1=T, o_2=T, o_3=F, and o_4=F.

One solution approach to overcome this issue is to increase the strength of the covering array, i.e., to use a larger t. This, however, can excessively increase the number of configurations to be tested. For example, since the missing combination in our example is a 4-way combination, to guarantee the coverage of this combination, a 4-way covering array needs to be created at the very least. However, a 4-way covering array for this scenario can have as many as 28 configurations.

An alternative approach is to use a variable-strength covering array, requiring a 4-way coverage only for the options $\{o_1, \ldots, o_4\}$. However, since what is actually being requested is the exhaustive testing of all possible combinations of settings for these 4 binary options, at least 16 configurations are required by this alternative.

Note that decision outcomes that need to be covered cannot be expressed as constraints in standard covering arrays in an attempt to selectively determine what to cover. This is because constraints in standard covering arrays are globally enforced. That is, the constraints should be satisfied by each and every configuration included in the covering array. Therefore, expressing the decision outcomes as constraints in standard covering arrays prevents the creation of any CIT objects because the alternative outcomes of a decision are guaranteed to conflict with each other. For example, since the outcomes of the decision at line 6 in Figure 1a, i.e., o_5 and $\neg o_5$, conflict with each other, no configuration satisfying both of these constraints can be generated; thus, no standard covering array can be constructed.

2.2 Applying F-CIT

F-CIT, on the other hand, can flexibly be used as follows to obtain DC-adequate test suites. Each entity to be covered corresponds to a distinct decision outcome. The entities are then expressed as constraints by using Boolean logic where each configuration option is represented by a Boolean variable. For our running example, Figure 1c presents all the entities required to be covered to obtain full coverage under the DC criterion. For instance, the first two entities (e_1 and e_2) represent the T and F outcomes of the decision at line 1 in Figure 1a, respectively.

Given the entities in Figure 1c, an F-CIT constructor divides them into 3 clusters: $\{e_1, e_3, e_5, e_7\}$, $\{e_2, e_6\}$, and $\{e_4, e_8\}$, such that all the constraints within a cluster can be satisfied together and that the number of clusters required to cover all the entities is minimized as much as possible.

Each cluster represents a set of decision outcomes that can be covered together in a single configuration. Therefore, a solution computed for a cluster represents a configuration, which covers all the decision outcomes included in the cluster. Consequently, the F-CIT constructor generates the three configurations (one for each cluster) given in Figure 1d, which obtain full coverage under the DC criterion; the first configuration covers the entities $\{e_1, e_3, e_5, e_7\}$, the second configuration covers $\{e_2, e_6\}$, and last configuration covers $\{e_4, e_8\}$.

Note that neither the clusters nor the configurations generated in this study are unique in the sense that there are other sets of configurations that an F-CIT constructor can generate to achieve full coverage. This is indeed similar to what we have in standard covering arrays as different t-way covering arrays can be computed for the same input space model.

Note further that although half of the constraints in Figure 1c conflict with the other half, it does not create an issue for F-CIT. This is because as each of these constraints represents an entity to be covered, F-CIT enforces them at the level of a test case. This, in turn, improves the flexibility of CIT, compared to enforcing the constraints at the level of a test suite as is the case with standard covering arrays where each and every test case included in a test suite should satisfy all the constraints. That is, F-CIT aims to satisfy each constraint representing an entity in at least one test case, rather than enforcing all the selected test cases to satisfy all of the entity constraints. For example, in the test suite given in Figure 1d, the constraint for entity $e_2 : \neg(o_1 \wedge o_2)$ is satisfied by the second configuration only. The other configurations included in this test suite, indeed, violate this constraint.

3 F-CIT

F-CIT takes as input a set of entities E to be covered and a model $M =< P, D, C >$, where $P = \{p_1, p_2, \ldots, p_k\}$ is a set of parameters, $D = \{D_1, D_2, \ldots, D_k\}$ is a set of respective domains

of values, and C is a constraint defined over P. While C defines the space of valid test cases, from which the samples are drawn, E specifies what needs to be covered by these samples.

Next, we make a number of definitions, starting from the "standard" definitions and going towards the F-CIT-specific ones:

Definition 1. *A* constraint *is a tuple $< P', R >$ where $P' \subseteq P$ is a subset of $l \leq k$ parameters and R is an l-ary relation on the corresponding domains.*

Definition 2. *An* evaluation *is a function from a subset of parameters to a particular set of values in the corresponding subset of domains.*

Definition 3. *An* evaluation *satisfies a constraint $< P', R >$, if the values assigned to the parameters in P', satisfies the relation R.*

Definition 4. *An evaluation is* consistent *with respect to a set of constraints, if it satisfies all the constraints.*

Definition 5. *An evaluation is* complete*, if it includes all the parameters in P.*

Definition 6. *An* F-CIT testable entity *is a constraint over a subset of P, which has at least one evaluation consistent with C, representing an entity to be covered in testing.*

Definition 7. *An* F-CIT test case *is a complete evaluation of P, which is consistent with C.*

Definition 8. *An F-CIT testable entity is said to be* covered *by an F-CIT test case, if and only if the test case is consistent with the testable entity.*

Definition 9. *Given an F-CIT model $M =< P, D, C >$ and a set of testable entities E to be covered, an* F-CIT object *is a set of F-CIT test cases, such that every F-CIT testable entity in E, is covered by at least one F-CIT test case.*

Going back to our running example in Section 2, the F-CIT model $M =< P, D, C >$ is defined as follows: $P = \{o_1, \ldots, o_6\}$, $D = \{\{T, F\}, \{T, F\}, \{T, F\}, \{T, F\}, \{T, F\}, \{T, F\}\}$, and $C : true$, indicating that all possible configurations are valid. An F-CIT testable entity corresponds to a distinct decision outcome expressed as a constraint in Boolean logic. The testable entities to be covered $E = \{e_1, \ldots, e_8\}$ are then defined as they are given in Figure 1c. For example, the testable entity e_1 is defined as $\neg(o_1 \wedge o_2)$, representing the F outcome of the first decision in Figure 1a. An F-CIT test case corresponds to a configuration, in which each configuration option assumes the value of either T or F, such as the second configuration in Figure 1d where $o_1 = F$, $o_2 = F$, $o_3 = T$, $o_4 = F$, $o_5 = F$, and $o_6 = T$. An F-CIT object then corresponds to a decision coverage-adequate set of F-CIT test cases, such as the ones given in Figure 1d.

4 Computing F-CIT Objects

It turns out that computing F-CIT objects requires us to solve an interesting constraint satisfaction problem, which we call *cov-CSP*, inspired from the theoretical concepts for "measuring" the level of consistency in paraconsistent logic (i.e., "inconsistency-tolerant" systems of logic) [19], [20], [21].

Given a set of constraints H, cov-CSP aims to divide H into a minimum number of *satisfiable* clusters. That is, *cov-CSP* seeks to satisfy the constraints, not necessarily as a whole, but in groups. We first define cov-CSP in the most general sense and then show how solving this problem helps compute F-CIT objects.

Definition 10. *Given a set of constraints $H = \{h_1, \ldots, h_m\}$, cov-CSP divides H into a minimum number of clusters $S = \{H_1, \ldots, H_n\}$, such that $\bigcup_{H_i \in S} H_i = H$ and that for each $H_i \in S$, $\bigwedge_{h \in H_i} h$ is satisfiable, i.e., all the constraints in a cluster are satisfiable together.*

Given a model $M =< P, D, C >$ and a set of F-CIT testable entities to be covered $E = \{e_1, \ldots, e_m\}$, each of which is represented as a constraint, computing an F-CIT object proceeds by first solving the cov-CSP problem, so that E is divided into a "minimum" number of satisfiable clusters $S = \{E_1, \ldots, E_n\}$ (as specified by Definition 10). Note that since computing the global minimum may not be computationally feasible (or desirable), F-CIT aims to compute an approximation to it.

Each cluster depicts a set of testable entities that can be tested together. Therefore, a solution for a cluster, represents an F-CIT test case covering all the F-CIT testable entities included in the cluster. Consequently, the collection of all the test cases generated (one per cluster), constitutes an F-CIT object covering each testable entity in E at least once.

The only remaining detail to ensure the generation of valid test cases, is to take the model constraint C into account. To this end, when checking the satisfiability of a cluster $E_i \in S$ or computing a solution for it, the constraint to satisfy simply becomes $C \wedge \bigwedge_{e \in E_i} e$.

Note that, in order to reduce the number of test cases required, it is desirable to avoid redundancy as much as possible by covering each testable entity in exactly one test case. However, a testable entity, in the process of covering other testable entities, may end up being covered by multiple test cases. This can happen unintentionally (i.e., by chance) or intentionally to satisfy the model constraint C.

Next, we present two constructors for computing F-CIT objects (thus, for solving the cov-CSP problem), namely *cover-and-generate* and *generate-and-cover.*

Algorithm 1 The cover-and-generate constructor for computing F-CIT objects

Input: A test space model $M =< P, D, C >$
Input: A set of testable entities E to be covered
Output: An F-CIT object T

1: $S \leftarrow \{\}$
2: **for each** testable entity $e \in E$ **do**
3: $\quad accommodated \leftarrow false$
4: $\quad$ **for each** $E' \in S$ **do**
5: $\quad\quad$ **if** $satisfiable(e \wedge \bigwedge_{e' \in E'} e' \wedge C)$ **then**
6: $\quad\quad\quad E' \leftarrow E' \cup \{e\}$
7: $\quad\quad\quad accommodated \leftarrow true$
8: $\quad\quad\quad$ **break**
9: $\quad\quad$ **end if**
10: $\quad$ **end for**
11: $\quad$ **if not** $accommodated$ **then**
12: $\quad\quad S \leftarrow S \cup \{\{e\}\}$
13: $\quad$ **end if**
14: **end for**
15:
16: $T \leftarrow \{\}$
17: **for each** $E' \in S$ **do**
18: $\quad T \leftarrow T \cup solve(C \wedge \bigwedge_{e' \in E'} e')$
19: **end for**
20: **return** T

Algorithm 2 The generate-and-cover constructor for computing F-CIT objects

Input: A test space model $M =< P, D, C >$
Input: A set of testable entities E to be covered
Output: An F-CIT object T

1: $T \leftarrow \{\}$
2: **for each** testable entity $e \in E$ **do**
3: $\quad accommodated \leftarrow false$
4: $\quad$ **for each** $t \in T$ **do**
5: $\quad\quad$ **if** $satisfiable(e \wedge t \wedge C)$ **then**
6: $\quad\quad\quad accommodated \leftarrow true$
7: $\quad\quad\quad$ **break**
8: $\quad\quad$ **end if**
9: $\quad$ **end for**
10: $\quad$ **if not** $accommodated$ **then**
11: $\quad\quad T \leftarrow T \cup solve(e \wedge C)$
12: $\quad$ **end if**
13: **end for**
14: **return** T

4.1 The Cover-and-Generate Constructor

The cover-and-generate constructor (Algorithm 1) maintains a pool S of clusters, each representing a set of testable entities that can be covered together. The pool is initially empty (line 1). Then, for each testable entity $e \in E$, we attempt to accommodate it in an existing cluster $E' \in S$ (lines 4-10). To this end, we check to see if e is satisfiable together with all the constraints in E' as well as with the model constraint C, i.e., whether $e \wedge \bigwedge_{e' \in E'} e' \wedge C$ is satisfiable (line 5). If so, e is included in E' (line 6), indicating that e can be accommodated together in a single test case with the other testable entities in E'. Otherwise (i.e., if no such cluster is found), we populate S with a new cluster initially having only e (line 12). Once all the testable entities are processed, for each cluster $E' \in S$, we generate a test case by solving $C \wedge \bigwedge_{e' \in E'} e'$ (line 18). The collection of all the test cases generated (T), is then returned as the F-CIT object computed, covering all the testable entities in E (lines 17-20).

4.2 The Generate-and-Cover Constructor

The generate-and-cover constructor associates a cluster with an F-CIT test case, rather than with a set of F-CIT testable entities. Conceptually, this constructor generates a test case first and then marks all the testable entities accommodated by the test case as covered. Therefore, it is different than the cover-and-generate constructor, which attempts to cover as many testable entities as possible in a cluster before generating a test case. Consequently, the set of clusters maintained through the iterations of the generate-and-cover constructor, simply represents the F-CIT test cases that have already been included in the F-CIT object being computed.

Given a model $M =< P, D, C >$ and a set of testable entities E to be covered, one way to generate a test case is to compute a solution for the model constraint C, regardless of E. However, generating test cases without taking the testable entities to be covered into account, may make it quite difficult to cover the entities that are hard to cover by chance. We, therefore, employ an alternative approach in this work, which guarantees that at least one previously uncovered testable entity is covered by every test case generated.

Algorithm 2 presents the generate-and-cover constructor. The F-CIT object T is initially empty (line 1). Then, for each testable entity $e \in E$, we check to see if e has already been covered by a test case $t \in T$ (lines 4-9), i.e., if there exists a test case $t \in T$, which is consistent with e (line 5). If no such test case is found, a new F-CIT test case covering e, is generated by solving the constraint $e \wedge C$ and T is populated with the newly generated test case (lines 10-12). Once all the testable entities in E have been processed, T is returned as the F-CIT object computed (line 14).

4.3 A Seeding Mechanism

Both of the constructors we have discussed so far can also take as input a *seed*, which in this context refers

to a set of F-CIT test cases. Given a seed, all the F-CIT testable entities in the seed, are considered to have already been covered and additional F-CIT test cases are generated only to cover the remaining entities.

To this end, the only change that needs to be made is to modify line 1 in Algorithms 1 and 2, such that instead of starting with an empty pool of clusters, we start with an initially populated pool of clusters, each of which is created to include a single F-CIT test case in the seed. Nothing else in the algorithms needs to be changed.

In Section 5.1, we use the seeding mechanism both to compute higher strength F-CIT objects from lower strength F-CIT objects (by using the lower strength objects as seeds) and to generate F-CIT objects that satisfy multiple coverage criteria (by using an object satisfying a coverage criterion as a seed to compute another object satisfying a different coverage criterion).

4.4 Example: Computing DC-Adequate Test Suites as F-CIT Objects

In this section, for illustrative purposes, we use the cover-and-generate constructor (Algorithm 1) to compute DC-Adequate test suites as F-CIT objects using our running example in Section 2. For the sake of the discussion, however, we introduce the following system-wide constraint to the problem: $(o_2 = F) \implies (o_6 = T)$, i.e., if o_2 is $false$, then o_6 must be $true$, invalidating the combination $(o_2 = F, o_6 = F)$.

Modeling. The F-CIT model is defined as $M =< P, D, C >$, where $P = \{o_1, \ldots, o_6\}$, $D = \{\{T, F\}, \ldots, \{T, F\}\}$, and $C : (\neg o_2 \implies o_6)$. Each F-CIT testable entity then naturally corresponds to a decision outcome to be covered. Figure 1c presents all the F-CIT testable entities that need to be covered to obtain full coverage under the decision coverage criterion.

Assuming that the testable entities in Figure 1c are processed in the order $e_1, \ldots, e_8$, the cover-and-generate constructor proceeds as follows: First, $e_1 : (o_1 \wedge o_2)$ is processed. Since the pool S is initially empty (line 1), a new cluster $E_1 = \{e_1\}$ is created and S is populated with E_1, i.e., $S = \{E_1\}$ (line 12). Then, $e_2 : \neg(o_1 \wedge o_2)$ is processed. Since $e_1 \wedge e_2 \wedge C$, i.e., $(o_1 \wedge o_2) \wedge \neg(o_1 \wedge o_2) \wedge (\neg o_2 \implies o_6)$, is not satisfiable (line 5), e_2 cannot be placed in E_1. So, a new cluster $E_2 = \{e_2\}$ is created and S is updated to $\{E_1, E_2\}$ (line 12). Next, $e_3 : (o_1 \wedge o_2) \wedge (o_3 \vee o_4)$ is processed. Since $e_1 \wedge e_3 \wedge C$, i.e., $(o_1 \wedge o_2) \wedge ((o_1 \wedge o_2) \wedge (o_3 \vee o_4)) \wedge (\neg o_2 \implies o_6)$, is satisfiable (line 5), e_3 is included in E_1 (line 6). After processing all the remaining testable entities in Figure 1c, we have the clusters given in the first column of Table 1.

For each cluster in $S = \{E_1, E_2, E_3\}$, we then generate an F-CIT test case by satisfying the constraints included in the cluster together with the model constraint C (lines 16-19). For example, for E_1, solving $e_1 \wedge e_3 \wedge e_5 \wedge e_7 \wedge C$ produces the test case $(o_1 = T, o_2 = T, o_3 = T, o_4 = T, o_5 = T, o_6 = T)$. Processing all the clusters would then generate the F-CIT object given in the second column of Table 1 (line 20), which is, indeed, DC-adequate.

TABLE 1: An F-CIT object (second column) created for the set of satisfiable clusters $S = \{E_1, E_2, E_3\}$ (first column) obtained for the testable entities in Figure 1c.

satisfiable clusters	DC-adequate F-CIT object					
$S = \{E_1, E_2, E_3\}$	o_1	o_2	o_3	o_4	o_5	o_6
$E_1 = \{e_1, e_3, e_5, e_7\}$	T	T	T	T	T	T
$E_2 = \{e_2, e_6\}$	F	F	T	F	F	T
$E_3 = \{e_4, e_8\}$	T	T	F	F	T	F

4.5 Discussion

Regarding constraints and solvers. The terms "constraint" and "solver" are used in the general sense in F-CIT. That is, any restriction, independent of the logic in which it is specified, is considered to be a constraint and a solver conceptually determines whether a given set of testable entities can be covered together in a single test case or not. Therefore, F-CIT expects that the underlying solver supports essentially a single computational primitive, namely $solve$. The other primitive used in Algorithms 1 and 2, namely $satisfiable$, can actually be implemented by using $solve$ as the absence of a solution indicates unsatisfiability.

Having a simple interface between F-CIT constructors and solvers further improves the flexibility of F-CIT. For example, all of the widely-used SAT and CSP solvers, in one form or another, provide a $solve$ primitive. Furthermore, this feature also allows application- and domain-specific solvers to be used with F-CIT constructors (Section 5.3).

This interface can indeed be further generalized by having $solve$ to take as input a set of constraints, each of which can represent a testable entity, a model constraint, or a test case. Since an F-CIT constructor does not then need to interpret these constraints, the testable entities, the model constraints, and the test cases can be expressed in any form desired, which may not even need to be formal.

Regarding constructors. We have presented two constructors in this section, namely the cover-and-generate constructor and the generate-and-cover constructor. We introduced the latter solely to mimic one of the simplest ways of generating F-CIT objects: Keep on generating valid test cases until all the required entities have been covered. As such, we use this constructor as a base line for comparisons in our experiments (Section 5), demonstrating that computing F-CIT objects in an efficient and effective manner is not trivial. Indeed,

the results of our experiments strongly suggest that the cover-and-generate constructor performed better than the generate-and-cover constructor in reducing both the sizes and the construction times of F-CIT objects (Section 5).

We, therefore, generally suggest to use the cover-and-generate constructor. However, the generate-and-cover constructor can still be of practical interest in scenarios especially when it is costly to determine whether multiple testable entities can be covered together or not (due to, for example, the complexity of the constraints to be solved) and when it is easy to cover the entities by chance in valid test cases. Note that the presence of these factors favors the generate-and-cover constructor as multiple testable entities can be covered by generating a valid test case. Furthermore, by making sure that each test case covers at least one previously uncovered testable entity, the generate-and-cover constructor guarantees the convergence into full coverage. Clearly, the end-users can always experiment with both constructors to determine the one to use in their projects.

With all these in mind, we have implemented the F-CIT constructors given in Algorithms 1 and 2 in Python in the form of an extensible tool that can work with any types of constraints and solvers. The tool can be downloaded at https://github.com/susoftgroup/UCIT/.

The efficiency and effectiveness of the F-CIT constructors we introduced in this work (i.e., the construction times and the sizes of the F-CIT objects computed), can be effected by the order, in which the testable entities are processed. In the presence of some knowledge regarding a favorable order (or a partial order), the testable entities can be sorted accordingly before they are fed to an F-CIT constructor. If not, a random order can be used by shuffling the entities. Furthermore, the construction process can be repeated multiple times in an attempt to compute smaller F-CIT objects at the cost of increased construction times. In Section 5.3.4, we carry out additional set of experiments to evaluate the sensitivity of the cover-and-generate constructor (which generally performed better than the generate-and-cover constructor) to the order the testable entities are processed.

Furthermore, F-CIT constructors may not be as efficient as their specialized counterparts. Our ultimate goal, however, is not to perform better than the existing constructors when F-CIT is used to compute the same CIT objects that these constructors are specifically designed to compute. As a matter of fact, we don't see much value in using F-CIT in such scenarios unless the F-CIT constructors perform better than the existing ones. Our goal is rather to improve the flexibility, thus the applicability, of CIT by eliminating the necessity of developing specialized constructors for every distinct CIT problem, which is not addressed by the existing constructors.

5 Experiments

F-CIT does not aim to replace existing CIT constructors, but rather to reduce the barriers to applying CIT to other domains and problems. Note that, in this context, changing the underlying CIT problem is not the same as simply changing the parameters of an existing problem, but rather changing the problem itself. For example, for standard covering arrays, we don't consider the changes in system-wide constraints and/or the changes in model parameters to be a change in the underlying CIT problem. This is because the only thing that changes in such situations is the problem parameters, while the original problem remains intact, which is to cover all valid t-tuples at least once.

To evaluate F-CIT, we, therefore, carry out three case studies, each of which focuses on a different CIT problem. In the first study (Section 5.1), we compute structure-based CIT objects to obtain decision coverage-adequate objects. In the second study (Section 5.2), we compute order-based CIT objects, where the reachability constraints imposed by an underlying graph-based model are taken into account to cover various sequences of events. In the third study (Section 5.3), we compute usage-based CIT objects by selecting the tuples to be covered based on their usage statistics in the field, which is especially useful when standard covering arrays are not desirable due to their sizes.

In each study, we first introduce the CIT problem of interest and discuss the motivation behind this problem. We then discuss and empirically demonstrate that to compute the requested CIT objects, the existing constructors (as they are) require excessive number of test cases to guarantee full coverage. Or, they require non-trivial modifications. Or, it is not clear (if at all possible) how to modify them. We finally express the CIT problems in F-CIT and show that the very same F-CIT constructor (thus, the same construction approach) can compute all of the requested CIT objects in all the studies without any modifications, demonstrating the flexibility of the proposed approach.

In the experiments, we integrate different "solvers" with F-CIT. This, however, is solely for the purpose of demonstrating that F-CIT can work with different solvers. The very same solver, such as the CSP solver we use in Section 5.1, can in deed be used in all the studies.

Note further that although the CIT problems in our studies are different than the ones addressed by existing CIT constructors, we opt to use existing constructors for comparisons in the experiments to justify the need for F-CIT. That is, in these studies, we are not claiming that F-CIT constructors perform better than

standard CIT constructors (because the underlying CIT problems are different), but rather demonstrating that a different CIT constructor is indeed needed to compute the requested CIT objects in an efficient and effective manner. Otherwise, i.e., had the existing constructors addressed the CIT problems presented in this paper in an efficient and effective manner, there would be no need for F-CIT.

We, furthermore, use our generate-and-cover F-CIT constructor as a base line to show that computing F-CIT objects is not trivial at all and that better construction approaches, such as the cover-and-generate approach, are needed.

The raw data we obtained from the experiments can be found at https://github.com/susoftgroup/UCIT/.

5.1 Study 1: Structure-Based CIT

In this study, we use the same CIT problem discussed in Section 2.

5.1.1 Coverage criterion

In [26], [12], we introduced a novel CIT object, which given a structural coverage criterion, such as decision coverage (DC), computes a "minimal" test suite to obtain full coverage under the criterion. In this work, we not only express the same coverage criterion using F-CIT, demonstrating the expressiveness of F-CIT, but also generalize the aforementioned coverage criterion to higher coverage strengths, demonstrating the flexibility of F-CIT. We call this *structure-based CIT*.

In a nutshell, structure-based CIT takes as input the source code of the system under test, a coverage strength t, and a structural code coverage criterion. First, for each outer-most if-then-else directive in the implementation, a *virtual configuration option* is defined. Then, for a given a virtual configuration option, conditions that must be satisfied to obtain a full coverage under the given structural coverage criterion for the respective if-then-else directive, are defined as *virtual settings*. Finally, a number of configurations are selected to cover all valid t-way combinations of virtual option settings. The smaller the number of configurations selected, the better the approach is.

Next, without losing generality, we provide more details by using DC as the structural code coverage criterion of interest. The proposed approach, on the other hand, is readily available to use with other structural coverage criteria, such as condition coverage [27].

Definition 11. *A* virtual configuration option *(or* virtual option*, in short) represents an outer-most if-then-else directive, which is not nested in another if-then-else directive.*

For example, the system in Figure 1a has two virtual options: vo_1 representing the outer-most if-then-else directive between lines 1 and 5 and vo_2 representing the outer-most if-then-else directive between lines 6 and 10.

Definition 12. *Given a virtual configuration option, each feasible outcome of every decision in the respective if-then-else directive, is defined as a* virtual setting *and expressed as a constraint, such that covering all of these virtual settings obtains a full coverage under DC.*

For instance, the virtual option vo_1 in our running example has four virtual settings: $\{o_1 \wedge o_2, \neg(o_1 \wedge o_2), (o_1 \wedge o_2) \wedge (o_3 \vee o_4), (o_1 \wedge o_2) \wedge \neg(o_3 \vee o_4)\}$. The first two settings are respectively for covering the $true$ and $false$ branches of the decision $o_1 \wedge o_2$ and the last two settings are respectively for covering the $true$ and $false$ branches of the decision $o_3 \vee o_4$ while taking the guard condition $o_1 \wedge o_2$ into account. Similarly, vo_2 has four virtual settings: $\{o_5, \neg o_5, o_5 \wedge o_6, o_5 \wedge \neg o_6\}$.

Not all virtual settings of a virtual option may be valid due to some conflicting settings required for the actual configuration options that appear multiple times in the same if-then-else directive. Since each virtual setting is expressed as a constraint, an invalid virtual setting can be marked and filtered out by determining whether or not the respective constraint is satisfiable. That is, a virtual setting is invalid, if the respective constraint is not satisfiable. Clearly, covering invalid virtual settings is not required to achieve full coverage. Consequently, in the remainder of the paper, the term "virtual setting" is used to refer to valid virtual settings.

Definition 13. *A* t-combination *is a combination of virtual settings for a combination of* t *distinct virtual options, which is expressed by joining the respective constraints with the AND logical operator.*

As was the case with virtual settings, a t-combination is invalid, if the respective constraint is not satisfiable. In the remainder of the paper, the term "t-combination" is used to refer to valid t-combinations.

Note that each t-combination represents an interaction that can be tested. Going back to our running example and considering that $t = 2$, some example 2-combinations for the virtual options vo_1 and vo_2 are: $(o_1 \wedge o_2) \wedge (o_5)$, testing the interaction between the $true$ branches of the decisions at lines 1 and 6; and $((o_1 \wedge o_2) \wedge \neg(o_3 \vee o_4)) \wedge (o_6)$, testing the interaction between the $false$ branch of the decision at line 2 and the $true$ branch of the decision at line 7.

Definition 14. *Given a set of virtual configuration options, their virtual settings, and a coverage strength* t*,* t-way structure-based coverage criterion K_{struct} *marks all valid* t*-combinations for coverage.*

Definition 15. *Given a set of virtual configuration options, their virtual settings, and a coverage strength* t*, a* t-way structure-based F-CIT object *is a set of actual system*

TABLE 2: Information about the subject applications used in Study 1.

sut	version	description	actual options	virtual options	valid 1-combins	valid 2-combins	valid 3-combins
mpsolve	2.2	Mathematical solver	14	4	30	296	1104
dia	0.96.1	Diagramming application	15	11	42	734	7170
irissi	0.8.13	IRC client	30	11	70	2102	36056
xterm	2.4.3	Terminal emulator	38	31	78	2871	66497
parrot	0.9.1	Virtual machine	51	29	152	10359	426194
gimp	3.2.5	Vector graphics editor	79	28	198	16438	794050
pidgin	2.4.0	IM	53	43	199	17857	986926
python	2.6.4	Programming language	68	49	210	21180	1368012
xfig	2.6.8	Graphics manipulator	79	48	237	26985	1969006
vim	7.3	Text editor	79	49	239	27442	2019176
sylpheed	2.6.0	E-mail client	84	48	258	31597	2451586
cherokee	1.0.2	Web server	97	28	272	32530	2318986

configurations, in which each t-combination selected by K_{struct} is covered by at least one configuration.

In this context, an actual system configuration is said to cover a t-combination, if the configuration is consistent with the respective constraint.

Note further that the coverage strength t in K_{struct} can be 1, which simply marks the virtual settings of all the virtual options for coverage. Therefore, covering all valid 1-combinations (i.e., all virtual settings) guarantees to obtain full coverage under DC. Consequently, 1-way structure-based F-CIT objects are the same/similar combinatorial objects we introduced in our short paper [12], but expressed in F-CIT, demonstrating the expressiveness of F-CIT.

One issue with the 1-way structure-based F-CIT objects, however, is that they don't take the interactions between structurally isolated if-then-else directives into account. Take the 1-way structure-based object given in Figure 1d as an example, although a DC-adequate test suite, it does not, for example, test the interaction between the $true$ branch of the decision $o_1 \wedge o_2$ (line 1) and the $false$ branch of the decision o_5 (line 6).

This issue, which was not addressed in our previous work [12], can now easily be handled in F-CIT by simply increasing the strength of K_{struct}, demonstrating the flexibility of F-CIT by generalizing the coverage criterion introduced in [12]. Going back to our running example in Figure 1 and considering that $t = 2$, K_{struct} selects $4 * 4 = 16$ 2-combinations for vo_1 and vo_2, covering all the pairwise interactions between the settings of these virtual options.

5.1.2 Study setup

For the evaluations, we used 12 subject applications. Each application had a number of binary compile-time configuration options implemented by using preprocessor directives. Table 2 provides information about these subject applications. The columns of this table respectively present the subject applications, their versions and descriptions, the numbers of actual compile-time options they have, the numbers of virtual options extracted, and the numbers of 1-, 2- and 3-combinations selected by our structure-based coverage criterion. Note that since we were not aware of any inter-option constraints for these subject applications, all possible combinations of option settings were considered to be valid. Furthermore, to give an idea about the structural complexities of the virtual options we extracted, Table 3 presents the percentages of the virtual options that are of cyclomatic complexities of 2, 3, 4, 5, and $\geq$ 6, respectively. Throughout the paper cyclomatic complexities are computed on a per virtual option basis by using Radon [28] – a tool to compute various code metrics.

TABLE 3: Percentages of the if-then-else directives (one per virtual option) that are of cyclomatic complexity 2, 3, 4, 5, and ≥ 6.

	cyclomatic complexity				
sut	2	3	4	5	≥ 6
mpsolve	0	50	0	0	50
dia	9.09	63.64	27.27	0	0
irissi	0	36.36	36.36	0	27.27
xterm	54.84	25.81	6.45	6.45	6.45
parrot	24.14	37.93	13.79	6.90	17.24
gimp	0	57.14	10.71	28.57	3.57
pidgin	2.33	53.49	25.58	9.30	9.30
python	8.16	63.27	16.33	4.08	8.16
xfig	2.08	50	20.83	14.58	12.50
vim	4.08	48.98	20.41	14.29	12.24
sylpheed	10.42	56.25	8.33	6.25	18.75
cherokee	3.57	32.14	14.29	7.14	42.86

All the experiments, unless otherwise stated, were repeated 5 times and carried out on Google Cloud using Intel Xeon CPU 2.30GHz machine with 4 GB of

RAM, running 64-bit Ubuntu 17.10 as the operating system.

5.1.3 Applying standard CIT

Modeling. The very first observation we make is that standard covering arrays cannot be used (as they are) with virtual options because the settings of virtual options are constraints, rather than discrete values as is the case with standard covering arrays. For example, one setting for vo_1 is $(o_1 \wedge o_2) \wedge (o_3 \vee o_4)$ and another is $(o_1 \wedge o_2) \wedge \neg(o_3 \vee o_4)$. To the best of our knowledge, there is no standard covering array constructor that can take constraints as settings. Note that these virtual settings cannot be expressed as constraints in standard constructors either, because such constraints are globally enforced and virtual settings can conflict with each other, which prevents the creation of any covering arrays (Section 2).

An alternative approach can be to create a standard covering array for the actual configuration options to obtain full coverage under K_{struct}. This, however, may unnecessarily increase the number of configurations required. For example, the standard 2-way covering array given in Figure 1b obtains only 38% coverage under the 2-way K_{struct} criterion (covering only 9 out of 24 2-combinations). Since the maximum number of actual configuration options involved in a 2-combination is 6 in this example, a 6-way covering array needs to be used to guarantee full coverage. This, however, is the same as exhaustive testing. Indeed, using variable strength covering arrays as an alternative, also suffers from the same issue.

Next, to demonstrate that the CIT problem defined in this study is indeed different than the ones addressed by standard covering arrays, which justifies the need for a different constructor to guarantee full coverage in an efficient and effective manner, we apply standard CIT on the subject applications in Table 2.

Evaluations. We first observed that since standard covering arrays do not necessarily take the complex interactions between configuration options into account, they, especially in the presence of tangled options, either fail to obtain full decision coverage or require excessive number of test cases [26], [12].

More specifically, we first created standard 2-way and 3-way covering arrays for our subject applications and measured the t-way structure-based coverage they provided for $t = 1$, 2, and 3. The experiments for $t = 1$ and 2 were repeated 30 times, whereas those for $t = 3$ were repeated 5 times as measuring the coverage for higher strengths was costly. The average sizes of the standard 2-way and 3-way covering arrays created were 13.74 and 36.78, respectively.

Standard covering arrays did not even guarantee DC adequacy, i.e., 1-way structure-based coverage (Table 4). More specifically, in about 58% (14 out of 24) of the experimental setups, standard covering arrays could not obtain full DC coverage. Overall, the DC coverages achieved were 97.58% and 99.08%, on average, for $t = 2$ and 3, respectively.

TABLE 4: Percentages of the 1-, 2-, and 3-combinations covered by standard 2- and 3-way covering arrays. The experiments were repeated 30 times.

sut	standard 2-way CA			standard 3-way CA		
	% of t-combinations covered			% of t-combinations covered		
	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$
mpsolve	100	55	23	100	83	56
dia	99	39	18	100	46	27
irissi	100	36	11	100	49	22
xterm	97	49	29	98	55	38
parrot	90	29	8	94	33	15
gimp	95	36	14	98	47	21
pidgin	99	23	11	100	25	17
python	98	31	12	99	36	18
xfig	99	31	12	100	35	18
vim	99	30	11	100	34	18
sylpheed	97	39	16	98	45	25
cherokee	99	21	5	100	28	10

TABLE 5: Percentages of valid 1-combinations of various cyclomatic complexities covered by standard t-way covering arrays.

cyclomatic complexity	standard t-way covering arrays	
	$t=2$	$t=3$
2	100.00	100.00
3	100.00	100.00
4	98.96	100.00
5	98.17	99.84
≥ 6	94.17	97.28

Furthermore, the higher the strength of the structure-based criterion, the more the required combinations were missing from the standard covering arrays (Table 4). Overall, the 2- and 3-way standard covering arrays, while respectively covering 34.92% and 43.00% of all the 2-combinations, achieved 14.17% and 23.75% coverage of the 3-combinations.

Similarly, the more the cyclomatic complexity of the virtual options, the more the required combinations were missing (Table 5). For example, standard 2-way covering arrays, on average, covered 100.00%, 100.00%, 98.96%, 98.17%, and 94.17% of the 1-combinations for the virtual options with cyclomatic complexities of 2, 3, 4, 5, and ≥ 6, respectively.

We have then created higher strength as well as variable strength covering arrays. For the former, we determined the maximum number of distinct configuration options that appear in a t-way virtual option combination and used it as the strength of the standard covering array. For the latter, we determined the number of distinct configuration options that appear

TABLE 6: Using standard covering arrays to guarantee full coverage under structure-based coverage criterion. The columns indicate the subject application, the coverage strength of the standard covering array computed together with the average construction time and size obtained by repeating the experiments 3 times for 1-, 2-, and 3-way structure-based CIT, respectively. The symbol '-' marks experimental setups, for which the standard constructor failed with an "out of memory" exception.

sut	t-way standard covering arrays created for structure-based CIT								
	1-way structure-based CIT			2-way structure-based CIT			3-way structure based CIT		
	t	time	size	t	time	size	t	time	size
mpsolve	2	0.34	10	4	0.33	54	6	0.56	272
dia	3	0.36	26	5	0.46	134	7	0.97	608
irissi	4	0.90	82	7	-	-	9	-	-
xterm	9	-	-	12	-	-	15	-	-
parrot	10	-	-	15	-	-	18	-	-
xfig	6	-	-	9	-	-	12	-	-
python	5	616.80	299	9	-	-	12	-	-
pidgin	8	-	-	11	-	-	14	-	-
gimp	5	-	-	10	-	-	15	-	-
vim	5	-	-	10	-	-	15	-	-
sylpheed	10	-	-	16	-	-	20	-	-
cherokee	4	73.77	130	7	-	-	10	-	-

TABLE 7: Using variable strength covering arrays to guarantee full coverage under structure-based coverage criterion. The columns indicate the subject application and the average construction time and size of the variable strength covering arrays computed for 1-, 2-, and 3-way structure-based CIT, respectively. The experiments were repeated 3 times. The symbol '-' marks experimental setups, for which the standard constructor failed with an "out of memory" exception.

sut	variable strength covering arrays created for structure-based CIT					
	1-way structure-based CIT		2-way structure-based CIT		3-way structure-based CIT	
	time	size	time	size	time	size
mpsolve	0.29	8	0.41	47	0.88	252
dia	0.32	8	0.42	48	0.79	202
irissi	0.33	16	0.99	323	554.99	3217
xterm	0.54	512	12.05	4187	-	-
parrot	5.49	3750	-	-	-	-
xfig	364.78	585	-	-	-	-
python	0.40	32	4.70	845	6319.56	13350
pidgin	0.44	256	15.90	3447	-	-
gimp	0.41	32	6.07	730	4317.48	8908
vim	0.41	36	4.82	718	43198.74	9037
sylpheed	19.62	5062	-	-	-	-
cherokee	0.43	18	-	-	-	-

in each t-way virtual option combination and used it as the coverage strength to be satisfied for these configuration options. All of the covering arrays in these experiments were computed by using `ACTS` [29] and the experiments were repeated 3 times.

Tables 6-7 present the results we obtained. In 75% (27 out of 36) of the experimental setups for computing fixed-strength covering arrays and in 28% (10 out of 36) of the experimental setups for computing variable strength covering arrays, the standard constructor (ACTS) failed with an "out of memory" exception. The tables, therefore, present only the experiments, in which we were able to compute a covering array using the standard constructor. Although the covering arrays we could compute achieved full coverage, they did so at the expense of excessive number of configurations. For comparisons, the interesting reader can refer to Table 8 to check the sizes of the F-CIT objects computed for the study.

5.1.4 Applying F-CIT

Modeling. We have defined the F-CIT model as $M =< P, D, C >$, where P is the set of variables representing the actual configuration options; D is their respective domains, i.e., the settings that the actual configuration options can take on; and C is the model constraint (if any) invalidating certain combinations of option settings. Each F-CIT testable entity then naturally corresponded to a valid t-combination to be covered (Definition 13) and each F-CIT test case naturally corresponded to a configuration, in which every actual configuration option has a valid setting.

We have also used the seeding mechanism of F-CIT (Section 4.3) in this study to combine multiple coverage criteria. In particular, to construct 1-way structure-based F-CIT objects in some experiments, we used standard 2-way or 3-way covering arrays computed for the actual configuration options, as seeds. By doing so, we effectively computed t-way DC-adequate covering arrays, which not only covered all t-way combinations of actual option settings, but also achieved DC adequacy.

To further demonstrate that the very same seeding mechanism can also be used to incrementally compute F-CIT objects – a well-known approach for computing standard covering arrays [30], we have used lower strength structure-based F-CIT objects as seeds to compute higher strength F-CIT objects.

Cost. To extract virtual options from source code, we used `cppstats`, which is a static analysis tool for analyzing C/C++ preprocessor-based variability in highly configurable systems [31]. The tool parsed the if-then-else directives into an XML-based tree representation. We then simply traversed the representation to identify the elements that corresponded to virtual options. An if-then-else directive, which was not structurally contained in another if-then-else directive simply became a virtual option. Once a virtual option was found, we traversed the respective tree to determine the virtual settings, i.e., visiting the decisions in the possibly nested if-then-else directive. For each decision d with a guard condition g, two virtual settings were created: $g \wedge d$ and $g \wedge \neg d$. All told, developing a generic script to carry out these steps took about 10 hours.

We have integrated our constructors given in Algorithms 1 and 2 with SATisPy [32], which is a Python library that interfaces with various SAT solvers, such as MiniSat [33]. Since the decisions in the source code

TABLE 8: Information about the structure-based F-CIT objects created. The symbol '*' marks the experimental setups, in which the generate-and-cover constructor timed out after six days. The experiments were repeated 5 times.

	1-way				2-way				3-way			
	generate-and-cover		cover-and-generate		generate-and-cover		cover-and-generate		generate-and-cover		cover-and-generate	
sut	time	size	time	size	time	size	time	size	time	size	time	size
mpsolve	0.37	3.00	0.31	3.00	17.61	15.20	2.07	14.00	221.54	93.40	11.99	39.80
dia	0.37	4.40	0.34	4.20	16.35	19.60	2.26	19.40	482.35	131.80	24.79	70.60
irissi	0.69	4.00	0.66	4.00	74.21	25.20	13.16	24.20	8461.64	316.40	139.32	109.20
xterm	0.61	4.20	0.58	4.20	50.54	19.80	5.74	21.20	7025.89	271.60	92.54	79.00
parrot	2.03	10.00	1.95	10.00	877.18	57.80	46.65	55.80	206682.44	841.33	1070.67	317.40
gimp	2.45	8.20	2.27	8.00	825.78	49.80	67.11	48.00	457184.81	998.50	1645.61	272.80
pidgin	2.26	4.40	2.29	4.40	788.98	34.00	31.82	33.40	*	*	628.75	172.00
python	2.16	4.80	2.07	4.40	743.89	36.00	28.68	34.60	*	*	932.46	187.00
xfig	2.81	5.80	2.74	6.00	1355.77	46.00	78.54	45.80	*	*	2311.84	270.00
vim	2.82	6.40	2.69	6.20	1357.64	48.60	56.47	48.60	*	*	1679.70	291.20
sylpheed	3.18	6.00	3.04	6.60	1737.00	49.20	78.20	47.40	*	*	2724.60	279.20
cherokee	3.59	5.00	3.53	5.00	2792.24	45.40	79.89	45.00	*	*	2095.94	252.40

were already expressed as Boolean expressions and since the virtual settings (thus, the testable entities) were simply obtained by joining these expressions (or their negations) with the AND logical operator, the integration step took about 1 hour. Most of this time was, indeed, spent for developing simple syntactic transformations to match the input format of the solver. Furthermore, since all the testable entities in this study are expressed in Boolean logic, the SATisPy solver, which we opted to use in the first place due to its ease-of-use, can easily be replaced with any other SAT or CSP solver.

Evaluations. The t-way structure-based F-CIT objects we computed in this study covered all the required t-combinations by construction. Furthermore, the cover-and-generate constructor generally performed better than the generate-and-cover constructor in reducing both the sizes and the construction times (Table 8). We, therefore, ran the generate-and-cover constructor with a time-out period of six days per construction. Overall, the cover-and-generate constructor reduced the sizes by an average of 2%, 77%, and 66%, while at the same time reducing the construction times by an average of 3.31%, 95.39%, and 99.56%, when $t = 1$, 2, and 3, respectively. Note further that in 16.67% (6 out of 36) of the experimental setups, te generate-and-cover constructor timed out (Table 8). We, therefore, focus on the results obtained from the cover-and-generate constructor in the remainder of this section.

As expected, the higher the coverage strength, the larger the size and the construction time of the structure-based F-CIT objects tended to be. More specifically, the average sizes were 5.50, 36.45, and 195.05 with the average constructions times of 1.87, 40.88, and 1113.18 seconds for 1-, 2-, and 3-way structure-based F-CIT objects, respectively.

TABLE 9: Information about the t-way DC-adequate covering arrays created by computing 1-way structure-based F-CIT objects using t-way standard covering arrays as seeds. The column '+cfgs.' reports the average numbers of additional configurations needed. The experiments were repeated 5 times.

sut	using 2-way standard CAs as seeds				using 3-way standard CAs as seeds			
	generate-and-cover constructor		cover-and-generate constructor		generate-and-cover constructor		cover-and-generate constructor	
	time	+cfgs.	time	+cfgs.	time	+cfgs	time	size
mpsolve	0.72	0.00	0.61	0.00	0.70	0.00	0.60	0.00
dia	0.47	0.00	0.42	0.00	0.45	0.00	0.40	0.00
irissi	1.07	1.00	0.83	1.00	1.05	0.00	0.83	0.00
xterm	0.74	3.80	0.86	1.00	0.77	0.00	0.92	0.00
parrot	3.84	12.40	3.53	7.00	4.25	6.00	4.14	5.00
gimp	5.68	12.20	3.98	3.00	6.28	4.60	4.63	2.00
pidgin	2.53	1.00	3.09	1.00	2.71	0.00	3.31	0.00
python	3.82	5.00	3.51	2.00	3.85	0.00	3.61	0.00
xfig	4.23	3.00	4.24	1.00	4.41	0.00	4.20	0.00
vim	3.72	3.40	4.12	3.00	3.64	0.00	4.14	0.00
sylpheed	5.08	3.40	4.57	2.00	5.71	1.00	5.02	1.00
cherokee	5.80	3.00	6.12	1.00	6.22	1.00	6.13	1.00

Computing t-way DC-adequate covering arrays. Note that as the ultimate goal of the structure-based F-CIT objects is to obtain full coverage under the K_{struct} coverage criterion, they may not cover all the standard t-tuples. For example, the 1-way structure-based F-CIT objects we generated covered 67.33% and 40.00% of all the 2- and 3-tuples, on average, respectively. The numbers were 94.33% and 86.33% for the 2-way structure-based and 95.17% and 91.75% for the 3-way structure-based F-CIT objects.

One good thing about having a seeding mechanism in F-CIT is that it can be leveraged to satisfy multiple coverage criteria. For example, one way to obtain t-way DC-adequate covering arrays, i.e., standard t-way covering arrays that guarantee full DC coverage, is

TABLE 10: Using structure-based F-CIT objects as seeds to cover the missing 2- and 3-tuples by computing standard covering arrays. The column '+cfgs.' reports the average numbers of additional configurations needed. The experiments were repeated 5 times.

sut	standard 2-way CA using t-way structure-based objects as seeds						standard 3-way CA using t-way structure-based objects as seeds					
	$t=1$		$t=2$		$t=3$		$t=1$		$t=2$		$t=3$	
	time	+cfgs.	time	+cfgs.	time	+cfgs.	time	+cfgs.	time	+cfgs.	time	+cfgs.
mpsolve	0.07	7.00	0.06	2.40	0.06	0.80	0.07	19.40	0.08	13.20	0.07	8.40
dia	0.06	6.80	0.07	2.00	0.07	2.00	0.08	18.80	0.08	11.60	0.08	9.20
irissi	0.07	8.80	0.08	2.00	0.10	2.00	0.16	27.80	0.16	16.60	0.23	34.80
xterm	0.08	9.00	0.09	7.00	0.13	7.00	0.24	31.20	0.24	25.20	0.31	24.40
parrot	0.10	8.40	0.14	4.40	0.24	4.00	0.41	34.40	0.48	21.20	0.66	18.00
gimp	0.15	10.80	0.22	8.20	0.35	7.80	0.97	39.80	1.19	30.40	1.47	28.20
pidgin	0.10	10.00	0.12	7.00	0.21	7.00	0.42	35.40	0.46	24.60	0.69	29.80
python	0.13	10.40	0.19	5.00	0.27	5.00	0.72	38.60	0.70	24.20	1.19	58.60
xfig	0.16	10.80	0.22	3.00	0.34	3.00	1.04	40.00	1.05	21.00	1.43	23.60
vim	0.15	11.00	0.20	3.00	0.34	3.00	0.97	40.20	1.10	20.80	1.40	24.00
sylpheed	0.16	10.60	0.22	4.80	0.39	4.80	1.25	41.00	1.48	24.00	1.65	26.00
cherokee	0.19	12.00	0.27	7.40	0.48	6.80	1.73	43.80	2.24	29.80	2.26	26.00

TABLE 11: Information about the 3-way structure-based F-CIT objects created by using 2-way structure-based F-CIT objects as seeds. The experiments were repeated 5 times.

sut	generate-and-cover constructor		cover-and-generate constructor	
	time	size	time	size
mpsolve	45.75	30.00	49.78	34.00
dia	152.62	58.20	65.38	59.60
irissi	2050.65	97.60	1029.66	92.00
xterm	594.90	73.20	167.30	71.60
parrot	18647.29	278.40	4103.58	279.60
gimp	12679.40	216.40	5563.53	215.80
pidgin	52514.79	158.80	30171.88	157.20
python	38510.67	170.40	16897.34	168.20
xfig	59537.75	230.40	14543.20	222.40
vim	67225.58	258.20	19227.10	247.80
sylpheed	117420.77	236.00	67550.43	243.80
cherokee	161779.40	211.20	57712.16	208.40

TABLE 12: Information about the 4-way structure-based F-CIT objects created. Due to the cost, the experiments were repeated only once.

sut	valid 4-combins	time	size
mpsolve	1344	16.24	62
dia	32346	111.32	197
xterm	615994	576.92	281
irissi	395504	2067.48	442
pidgin	15293336	16772.38	751
python	19856465	16958.23	869
gimp	14678226	42706.50	1293
parrot	7587625	19631.76	1482
cherokee	47087747	90360.34	2300
sylpheed	81732014	111090.57	3040
xfig	76405845	149335.77	3987
vim	76661558	96900.90	4340

to use standard t-way covering arrays as seeds to compute 1-way structure-based F-CIT objects.

To demonstrate the feasibility of this approach, we generated 2- and 3-way DC-adequate covering arrays (Table 9). We observed that 1-way structure-based F-CIT objects turned the standard covering arrays into DC-adequate test suites with little increases in both the sizes and the construction times. The average numbers of additional configurations required on top of the standard 2-way and 3-way covering arrays were 1.83 and 0.75, respectively, with the additional construction times of 2.99 and 3.16 seconds, on average.

Note that using structure-based F-CIT objects as seeds to compute standard covering arrays is also possible. To demonstrate the feasibility, we used, 1-, 2-, and 3-way structure-based F-CIT objects as seeds to compute 2- and 3-way standard covering arrays (Table 10). The average numbers of additional configurations required on top of the 1-, 2-, and 3-way structure-based F-CIT objects were 21.92, 13.28, and 15.18, respectively, with the additional construction times of 0.40, 0.46, and 0.60 seconds, on average.

Incrementally computing structure-based F-CIT objects. Another use of the seeding mechanism is to leverage lower strength F-CIT objects as seeds for computing higher strength F-CIT objects. To demonstrate the feasibility, we used 2-way structure-based F-CIT objects as seeds to compute 3-way structure-based F-CIT objects. The results of these experiments can be found in Table 11.

Computing 4-way structure-based F-CIT objects. Last but not least, we have run our cover-and-generate constructor for $t = 4$. Table 12 presents the results we obtained. Overall, the minimum, the average, and the maximum sizes of the 4-way structure-based F-CIT objects we computed were 62, 1587, and 4340 with

the construction times of 16.24, $45544,03$, and 96900.90 seconds respectively.

5.1.5 Discussion

Standard covering arrays and structure-based F-CIT objects clearly employ different coverage criteria. We, therefore, do not claim that the F-CIT constructors developed in this work performed better than the standard CIT constructor used in the study. We rather demonstrate that a different CIT constructor is indeed needed to obtain full coverage under the structure-based coverage criterion in an efficient and effective manner. Had the existing constructors addressed the structure-based CIT problem in an efficient and effective manner, there would be no need for F-CIT.

5.2 Study 2: Order-Based CIT

In this study, we use graphs to model the input spaces of software systems, which we believe can address many interesting test scenarios, such as the ones that arise during the systematic testing of event-driven systems as well as multi-threaded applications. We first define the model of the input space in an abstract manner and briefly discuss two scenarios in which the same or similar models have been used for testing, then present a number of coverage criteria for which CIT can be used to satisfy and discuss the shortcomings of the state-of-the-art CIT approaches, and finally present how F-CIT overcomes these shortcomings.

The model of interest in this study, in its simplest form, is a directed graph $G = (V, E, v_0, v_\perp)$, where V is a set of nodes; E is a set of ordered pairs of the form (v, w), representing a directed edge from node $v \in V$ to node $w \in V$; and $v_0 \in V$ and $v_\perp \in V$ are two distinguished nodes, namely the *entry* and the *exit* node. The entry node has an in-degree of 0 and the exit node has an out-degree of 0. Furthermore, all the nodes are reachable from the entry node and the exit node is reachable from all the nodes. Figure 2 presents some example models.

Given a graph-based model, one high-level testing objective is to generate test cases to satisfy some structural coverage criterion, such as exercising every node and/or edge at least once [34]. When graphs are used as a model, however, the coverage criterion is often concerned with the order of the entities (e.g., nodes and edges) to be tested. For Figure 2a, one such criterion for example, would be to generate a set of paths from the entry node to the exit node, such that every valid order of two (not necessarily distinct) nodes is covered (not necessarily in a consecutive manner) by at least one path. Given this criterion, some example orders to be covered for Figure 2a are: $[v_3, v_4]$, $[v_1, v_6]$, $[v_6, v_6]$, and $[v_6, v_5]$, which can all be covered (together with other orders) by the path $(v_0, v_1, v_3, v_4, v_5, v_6, v_5, v_6, v_\perp)$. On the other hand, $[v_2, v_3]$ and $[v_2, v_4]$ are not valid orders since no paths can include them.

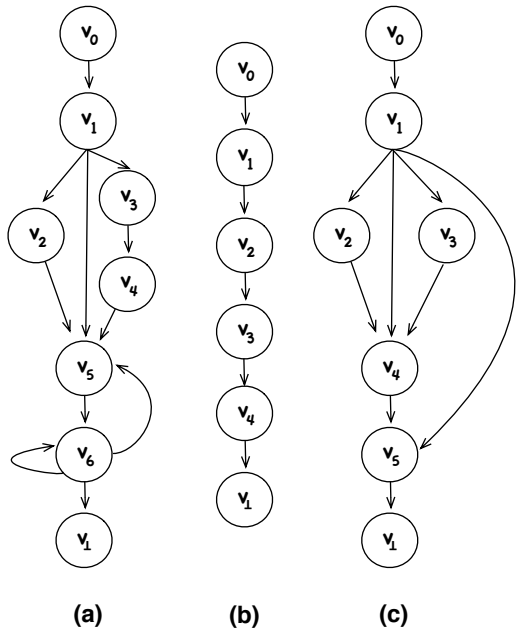


Fig. 2: Example graph-based models.

The same and similar graph-based models and coverage criteria have indeed been used for software testing. For example, in systematic testing of event-driven systems, such as graphical user interfaces (GUIs), graph-based models can capture the flow of events in the form of *event sequence graphs* [35] or *event-flow graphs* [36], [35], where each node represents an event and a directed edge from v_1 to v_2 indicates that event v_2 can follow event v_1. In this context, an event is considered as an environmental or a user stimulus that from the perspective of testing can be mimicked by a test case. Since the behavior of an event-driven system often depends on the order, in which the events are processed, testing such a system typically involves validating the system response under different event orders.

Another domain, in which graph-based models have been used for testing, concerns the systematic testing of multi-threaded applications. In this domain, the model of a thread captures all sequences of "atomic blocks" that might be traversed through the thread during its execution [37], [38]. In the remainder of the paper, the aforementioned models are referred to as *atomic block flow graphs* (AFGs).

Each node in an AFG represents an atomic block and the edges connecting the nodes represent the possible execution sequences of atomic blocks. For the programs adhering to a strict mutual-exclusion locking principle [38], an atomic block is defined as a code segment from one lock exit to the subsequent lock exit. And, a lock exit in this context corresponds to the release of a lock previously acquired

for a synchronized code segment [38], [39], [40]. For such programs, testing approaches aim to reveal non-deadlock errors, namely *atomicity-violation* and *order-violation* errors. Atomicity-violation errors occur when a sequence of operations that need to be carried in an atomic manner is erroneously divided into multiple atomic blocks, such that the atomicity of the entire operation cannot be guaranteed. Order-violation errors occur when an implicit execution order between two groups of atomic blocks is assumed, but not enforced, e.g., thread A is assumed to start before thread B. To detect these errors, different orders of atomic blocks need to tested.

Note that for these and similar scenarios, since a test case (e.g., a path from the entry node to the exit node) can cover more than one order, the number of test cases to obtain full coverage under a given coverage criterion can be reduced by carefully constructing the test cases. Consequently, CIT approaches can be of practical help.

With standard covering arrays, however, the order of parameter values in a test case is assumed to have no effect on the fault revealing ability of the test case. For example, given a software configuration, such as the ones studied in Section 5.1, any permutation of the option settings constituting the configuration, covers exactly the same set of option setting combinations, thus all these permutations should detect the same faulty interactions. For the scenarios we are interested in this study, however, the order matters. Consequently, the types of CIT objects we need in this study are quite different than the one we have computed in Section 5.1.

5.2.1 Coverage criterion

To take the orders into account, a different type of covering array, called a *sequence-covering* array, was defined in [41] and a number of interesting order-based coverage criteria were presented in [9]. In this study, we improve on these works by making both the coverage criteria and the construction approach take the reachability constraints imposed by a given graph-based model into account. Further discussion on this can be found in Section 5.2.3.

Definition 16. *Given $G = (V, E, v_0, v_\perp)$, a* path *is an ordered sequence of nodes $(v_{i_1}, \ldots, v_{i_n})$, such that $(v_{i_j}, v_{i_{j+1}}) \in E$ for $1 \leq j < n$.*

Definition 17. *Given $G = (V, E, v_0, v_\perp)$, a* test case *is a path from v_0 to $v_\perp$.*

For a given test case p of length n, let p_i, where $0 \leq i \leq n$, be the node located at position i in the test case, such that $p_0 = v_0$ and $p_n = v_\perp$.

Definition 18. *Given $G = (V, E, v_0, v_\perp)$, a t-order $[v_{i_1}, \ldots, v_{i_t}]$, where $v_{i_j} \in V$ for $1 \leq j \leq t$, is an ordered tuple of not-necessarily-distinct t nodes, such that there exists a test case p, in which the nodes $[v_{i_1}, \ldots, v_{i_t}]$ appear in the order they are given (not necessarily in a consecutive manner, though). A test case p of length n, such that $p_{m_j} = v_{i_j}$ and $0 < m_1 < m_2 < \cdots < m_t < n$ for $1 \leq j \leq t$, is said to cover the t-order $[v_{i_1}, \ldots, v_{i_t}]$.*

For instance, for the graph given in Figure 2a, $[v_1, v_2]$ and $[v_1, v_6]$, which are both covered by the test case $(v_0, v_1, v_2, v_5, v_6, v_\perp)$, are examples of 2-orders, whereas $[v_2, v_3]$ is not a 2-order, since there is no path from v_2 to v_3.

Definition 19. *Given $G = (V, E, v_0, v_\perp)$, a consecutive-t-order $[v_{i_1}, \ldots, v_{i_t}]$ is a t-order, such that there exists a test case p of length n, where $v_{i_j} \in V$ and $p_{m_j} = v_{i_j}$ for $1 \leq j \leq t$, and $m_{k+1} = m_k + 1$ for $1 \leq k < t$, i.e., $[v_{i_1}, \ldots, v_{i_t}]$ is a subpath in path p. Such a test case p is said to cover the consecutive-t-order $[v_{i_1}, \ldots, v_{i_t}]$.*

For instance, for the graph given in Figure 2a, $[v_1, v_2]$, $[v_6, v_5]$, and $[v_6, v_6]$, which all appear as subpaths in $(v_0, v_1, v_2, v_5, v_6, v_5, v_6, v_6, v_\perp)$, are examples of consecutive-2-orders, whereas $[v_1, v_6]$, although a 2-order, is not a consecutive-2-order, as there is no edge from v_1 to v_6.

Definition 20. *Given $G = (V, E, v_0, v_\perp)$, a non-consecutive-t-order $[v_{i_1}, \ldots, v_{i_t}]$ is a t-order, such that there exists a test case p of length n, where $v_{i_j} \in V$ and $p_{m_j} = v_{i_j}$ for $1 \leq j \leq t$, and $m_{k+1} - m_k > 1$ for at least one $1 \leq k < t$. Such a test case p is said to cover the non-consecutive-t-order $[v_{i_1}, \ldots, v_{i_t}]$.*

For Figure 2a, $[v_1, v_5]$ is an example of a non-consecutive-2-order, because there is at least one path, e.g., $(v_0, v_1, v_2, v_5, v_6, v_\perp)$, where the nodes constituting the order can appear in a non-consecutive manner. On the other hand, $[v_3, v_4]$, although a 2-order, is not a non-consecutive-2-order, because all the paths including this order have it in a consecutive manner.

Based on these definitions, we define the four coverage criteria given below (inspired from [9]). We call this *order-based CIT*.

Definition 21. *Given $G = (V, E, v_0, v_\perp)$, a set of test cases T is* t-order adequate, *if and only if for every t-order in G, there exists at least one test case in T, which covers it.*

Definition 22. *Given $G = (V, E, v_0, v_\perp)$, a set of test cases T is* t-cover adequate, *if and only if for every consecutive-t-order in G, there exists at least one test case in T, which covers it.*

Definition 23. *Given $G = (V, E, v_0, v_\perp)$, a set of test cases T is t^+-*cover adequate, *if and only if for every non-consecutive-t-order in G, there is at least one test case in T, which covers it.*

Definition 24. *Given $G = (V, E, v_0, v_\perp)$, a set of test cases T is t^*-*cover adequate, *if and only if T is both t-cover adequate and t^+-cover adequate.*

Note that to satisfy the t-order adequacy criterion, all possible t-orders need to be covered at least once regardless of whether they are covered in the form of a consecutive- or non-consecutive-t-order, whereas to satisfy the t-cover adequacy criterion all possible t-orders that can be covered in a consecutive manner need to be covered in the form of a consecutive-t-order. Similarly, to satisfy the t^+-cover adequacy criterion, all possible t-orders that can be covered in a non-consecutive manner need to be covered in the form of a non-consecutive-t-order. Finally, t^*-cover adequacy criterion is different than the t-order adequacy criterion, because when a t-order can be covered both in a consecutive and non-consecutive manner, the t^*-cover adequacy criterion guarantees that it is covered in the form of both consecutive- and non-consecutive-t-order, whereas for the t-order adequacy criterion covering it in either way is enough.

5.2.2 Study setup

In this study, we used 171 AFGs (atomic block flow graphs) obtained from Apache ActiveMQ v5.9.1 [42] – a high-performance, open source message oriented middleware – to evaluate the proposed approach. We unrolled the cycles in these graphs once to get acyclic graphs, which is a frequently used approach in bounded model checking [43] (see Section 5.2.4 for more details). After being unrolled, these graphs had an average of 312.82 nodes ($min = 12$ and $max = 3604$) and an average of 493.23 edges ($min = 12$ and $max = 6566$). All the experiments were carried out on the same Google Cloud platform we used in Study 1 (Section 5.1).

Note that due to the volume of the data to be reported in this section, using tabular notations was simply out of the question. Therefore, we opted to present different views of the data as we see fit by using plots, such as Figure 3, or by using summary tables, such as Tables 13-14. The raw data can, however, be found at https://github.com/susoftgroup/UCIT/.

In the summary tables, we first divide the experiments into 4 almost equal-size partitions with increasing complexity either by using the number of settings each configuration option has (e.g., Table 15) or by using the number of F-CIT entities to be covered (e.g., Tables 13 and 14). For each partition, we then report the minimum, median, and maximum results obtained in the partition. For a better interpretation of the results, we also filter out the experimental setups, in which the number of testable entities to be covered is less than 10. Furthermore, the partitions are indicated in the summary tables by the unique ID numbers reported under the "part." column.

5.2.3 Applying standard CIT

Modeling. The coverage criteria we have defined in Section 5.2.1 are inspired from [9], which empirically demonstrates that these order-based criteria are effective in detecting faults in event-driven software systems, such as graphical user interfaces (GUIs).

On the other hand, although the aforementioned work presents an approach to generate order-based CIT objects for a given graph-based model, it does not provide a systematic way of taking the reachability constraints imposed by the underlying graph into account during the construction of these objects. Such constraints are rather attempted to be handled after a CIT object is constructed with the aim of converting the invalid test cases, which are erroneously selected due to the overlooked-for constraints, to valid ones. However, no systematic way of carrying this post-mortem analysis is provided in [9]. Therefore, this approach can generate many invalid test cases, which may not be trivially "fixed." For example, for the model given in Figure 2b, out of 24 possible permutations of 4 nodes (excluding v_0 and $v_\perp$), only one of them (4.2%) is a valid test case, which is difficult to generate by chance. Invalid test cases is an important issue in CIT, because they often result in wasted testing resources [14], [15].

More specifically, the proposed construction approach in [9] uses standard covering arrays to compute order-based CIT objects. It takes as input a set of e events, a coverage strength t, and a predetermined length l for the test cases to be generated (i.e., only fixed-length test cases can be generated) and as output computes a standard t-way covering array for l options, each of which can take on e settings (one distinct setting per event). For example, to compute a 2-cover-adequate CIT object for the model given in Figure 2a, the aforementioned approach would generate a standard 2-way covering array for 6 options (because the minimum length of a test case to guarantee the coverage of all consecutive-2-orders is 6), each of which has 6 settings (because the number of nodes except for v_0 and $v_\perp$ is 6).

Evaluations. We used the aforementioned construction approach [9] to obtain full coverage under the order-based coverage criteria for the graphs discussed in Section 5.2.2. To this end, given a graph with e nodes, we, in an attempt to make sure that every requested order can be covered, used the longest path length l in the graph as the fixed-length. Note that this approach requires us to fix the length of the test cases to be generated. Consequently, the problem of covering different types of t-orders, independent of the actual coverage criterion used, was turned into a problem of computing a standard t-way covering array for l options, each of which can take on e settings. We used `ACTS` [29] to compute the required standard covering arrays. The experiments were repeated 5 times.

Table 15 presents the results we obtained. As the graphs got larger, since the number of settings for each option (i.e., the number of nodes e in the graph)

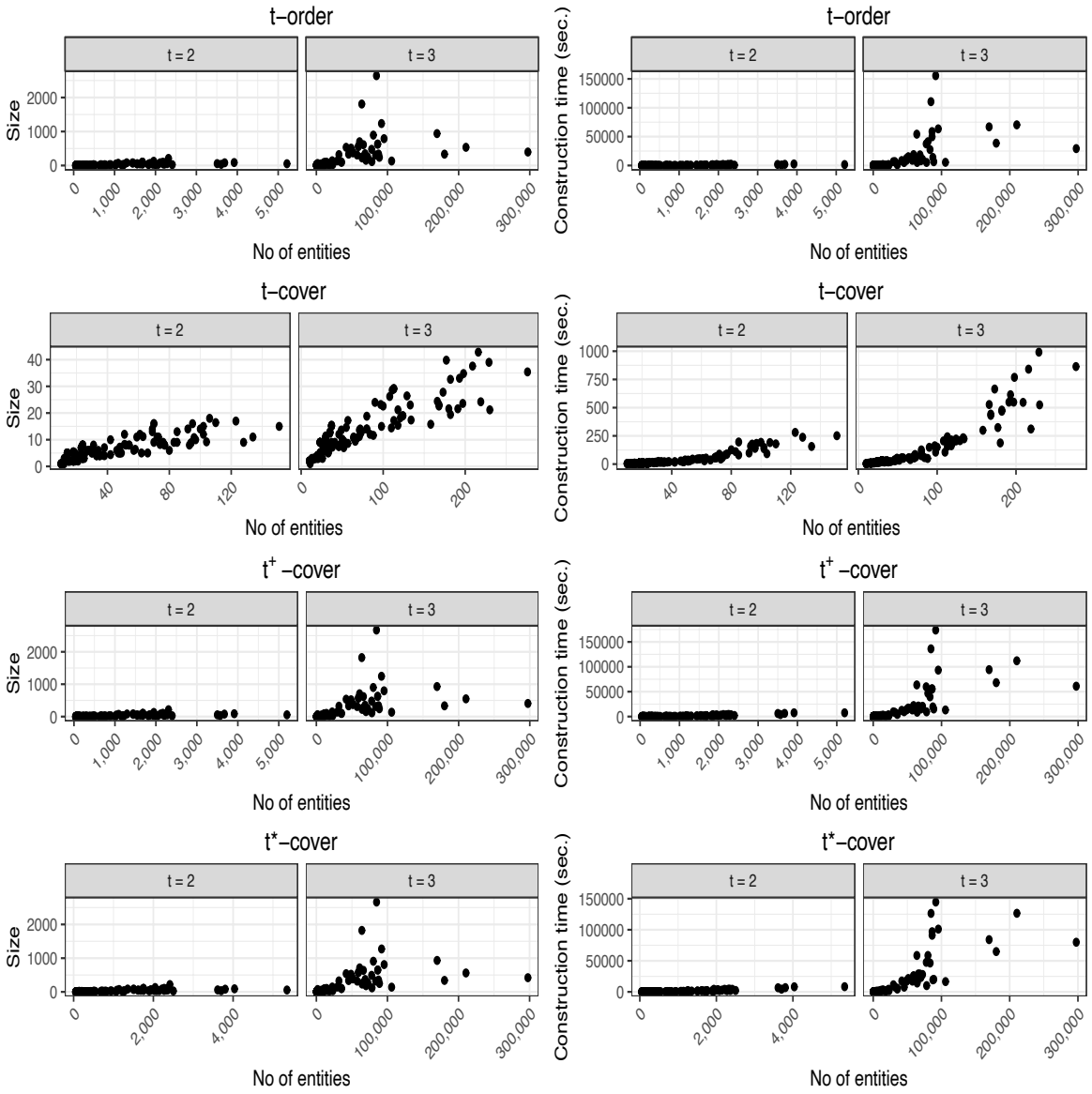


Fig. 3: Results obtained for the coverage criteria given in Definitions 21-24. The horizontal axes represent the numbers of F-CIT entities to be covered, whereas the vertical axes depict either the average sizes of the F-CIT objects constructed or the average construction times (in seconds), depending on the graph. The experiments were repeated up to 5 times.

increased, it took increasingly longer times to compute the required covering arrays. This was indeed the case even for relatively small option counts (i.e., small values of l). As it was not feasible for us to generate all the required covering arrays, we employed a threshold value of 70 when $t = 2$ and 30 when $t = 3$ on the number of settings an option can have. This enabled us to cover all of the experimental setups, in which $e < 70$ when $t = 2$ and $e < 30$ when $t = 3$. These thresholds were chosen, such that the standard covering array constructor had one day to compute the requested object. Within the allocated time limits, we were able to generate standard covering arrays for 96.49% (165 out of 171) of the models when $t = 2$ and for 66.67% (114 out of 171) of the models when $t = 3$.

None of the test cases in the generated covering arrays, on the other hand, were valid. As no systematic approach is presented in [9] to take the reachability constraints enforced by the underlying graphs into account or to fix the invalid test cases in a post-mortem

TABLE 13: Summary statistics for the construction times (in seconds) and the sizes of the t-way order-based F-CIT objects created for a) $t = 2$ and (b) $t = 3$ b. For each partition, the minimum, median, and maximum values encountered in the partition for the metrics in the columns are reported. The experiments were repeated up to 5 times.

	part.	stat.	nodes	edges	entities	time	size
t-order	1	min	12	12	31	2	1
		med	44	66	61	11	5
		max	46	67	71	18	6
	2	min	14	13	78	3	1
		med	34	38	123	13	3
		max	70	99	224	41	13
	3	min	35	40	227	18	2
		med	59	82	394	38	5
		max	172	248	1020	188	26
	4	min	118	150	1037	116	11
		med	580	871	1936	725	54
		max	3604	6566	5215	2608	210
t-cover	1	min	12	12	10	3	1
		med	28	37	15	5	5
		max	44	66	17	9	5
	2	min	19	23	18	6	2
		med	40	47	18	8	4
		max	54	73	21	11	5
	3	min	23	28	23	7	3
		med	69	98	30	15	5
		max	460	620	52	36	12
	4	min	112	150	54	29	5
		med	580	871	83	92	11
		max	3604	6566	151	399	21
t^+-cover	1	min	12	12	30	12	1
		med	44	66	59	27	5
		max	46	67	68	39	6
	2	min	16	18	73	27	1
		med	34	38	114	50	3
		max	71	111	222	141	13
	3	min	35	40	241	114	2
		med	59	82	397	215	5
		max	172	248	1032	722	47
	4	min	118	150	1089	695	11
		med	580	871	1936	2067	55
		max	3604	6566	5200	7559	213
t^*-cover	1	min	12	12	42	15	1
		med	44	66	76	35	8
		max	46	67	86	53	8
	2	min	16	18	88	35	2
		med	34	38	134	62	4
		max	70	99	240	151	14
	3	min	35	40	242	118	2
		med	59	82	409	232	6
		max	172	248	1065	807	27
	4	min	118	150	1084	644	14
		med	580	871	2029	2234	59
		max	3604	6566	5351	8383	218

(a)

	part.	stat.	nodes	edges	entities	time	size
t-order	1	min	12	12	121	3	1
		med	28	37	312	34	9
		max	44	66	340	49	12
	2	min	17	18	373	5	1
		med	46	67	526	38	8
		max	70	99	1023	81	19
	3	min	34	38	1329	36	3
		med	54	69	3742	90	8
		max	172	248	14056	1612	98
	4	min	99	128	15196	333	24
		med	580	864	65389	12517	347
		max	3604	6566	297308	155572	2643
t-cover	1	min	12	12	10	4	1
		med	28	37	20	11	4
		max	54	69	24	20	9
	2	min	20	24	25	15	3
		med	44	66	28	29	10
		max	50	72	30	33	11
	3	min	23	28	31	17	5
		med	69	98	46	29	10
		max	152	219	73	84	17
	4	min	112	150	80	49	11
		med	580	871	134	234	23
		max	3604	6566	276	1163	42
t^+-cover	1	min	12	12	121	18	1
		med	28	37	310	60	9
		max	44	66	330	82	12
	2	min	17	18	371	39	1
		med	46	67	521	74	9
		max	70	99	1007	156	19
	3	min	34	38	1327	160	3
		med	54	69	3735	277	8
		max	172	248	14034	2865	98
	4	min	99	128	15185	1169	25
		med	580	864	65389	17037	349
		max	3604	6566	297294	173798	2672
t^*-cover	1	min	12	12	136	22	1
		med	28	37	338	79	16
		max	44	66	346	142	18
	2	min	17	18	401	49	3
		med	46	67	544	119	14
		max	70	99	1038	175	26
	3	min	34	38	1373	182	4
		med	54	69	3763	394	15
		max	172	248	14144	3262	102
	4	min	99	128	15253	1359	32
		med	580	864	65507	23023	368
		max	3604	6566	297570	144652	2665

(b)

manner, it is was not clear at all how to avoid and/or fix these invalid test cases.

5.2.4 Applying F-CIT

Given a graph $G = (V, E, v_0, v_\perp)$, which models the input space of the system under test, an F-CIT testable entity corresponds to a t-order, a consecutive-t-order, or a non-consecutive-t-order to be covered, depending on the the coverage criterion (Definitions 21-24). Then, an F-CIT test case corresponds to a path from v_0 to $v_\perp$ (Definition 17). Finally, the graph G, as it restricts the orders to be covered as well as the test cases to be generated, is expressed as the F-CIT model constraint C.

To this end, we encode the problem of finding a path from a source node to a sink node as a single-source single-sink flow problem [44]. In particular, flow on an edge $(v_i, v_j) \in E$ (using the terminology of flow networks) is represented by a unique variable e_{ij}. From the perspective of finding a path, an edge $(v_i, v_j) \in E$ is either taken, i.e., $e_{ij} = 1$, indicating that there is a flow on the edge, or not taken, i.e., $e_{ij} = 0$, indicating that there is no flow on the edge:

TABLE 14: Summary statistics for the construction times (in seconds) and the sizes of the 4-way order-based F-CIT objects computed. For each partition, the minimum, median, and maximum values encountered in the partition for the metrics in the columns are reported. Due to the cost, the experiments were repeated only once.

	part.	stat.	nodes	edges	entities	time	size
t-order	1	min	12	12	274	1	1
		med	28	37	1059	34	9
		max	44	66	1432	61	10
	2	min	17	18	1456	3	1
		med	46	67	1735	35	6
		max	51	73	2866	59	15
	3	min	23	28	3189	6	1
		med	44	55	19424	51	6
		max	76	111	27601	786	191
	4	min	41	49	29070	35	4
		med	99	135	118962	559	64
		max	578	804	4360399	92109	5742
t-cover	1	min	12	12	10	2	1
		med	44	55	11	3	3
		max	2889	4956	14	23	7
	2	min	17	18	15	3	2
		med	51	70	18	6	7
		max	1703	3121	21	19	9
	3	min	19	23	23	9	5
		med	56	80	27	16	8
		max	2652	4147	39	67	15
	4	min	59	82	40	19	6
		med	456	702	51	40	14
		max	2748	4264	143	197	26
t^+-cover	1	min	12	12	264	15	1
		med	28	37	1044	81	9
		max	44	66	1426	129	9
	2	min	17	18	1438	43	1
		med	46	67	1724	103	6
		max	51	73	2833	140	15
	3	min	23	28	3174	107	1
		med	44	55	19410	290	6
		max	76	111	27572	1145	191
	4	min	41	49	29046	389	4
		med	112	154	149246	3071	70
		max	578	804	4360336	171050	5821
t^*-cover	1	min	12	12	274	20	1
		med	28	37	1059	93	11
		max	44	66	1432	123	13
	2	min	17	18	1456	47	2
		med	46	67	1735	112	11
		max	51	73	2561	172	17
	3	min	23	28	2866	119	4
		med	40	47	19424	259	7
		max	71	111	27214	1539	199
	4	min	41	49	27601	414	10
		med	99	137	117735	2381	53
		max	578	804	4360399	189773	5858

TABLE 15: Summary statistics for the construction times (in seconds) and the sizes of the standard covering arrays obtained by using the order-based construction approach presented in [9]. None of the test cases chosen by these standard covering arrays were valid. For each partition, the minimum, median, and maximum values encountered in the partition for the metrics in the columns are reported. The experiments were repeated 5 times.

	part.	stats.	options	settings per options	time	size
t=2	1	min	13	9	0.38	175
		med	28	12	0.64	336
		max	29	13	0.67	344
	2	min	15	14	0.47	289
		med	31	17	0.86	561
		max	47	18	2.57	812
	3	min	23	19	0.74	703
		med	39	22	3.22	1118
		max	129	37	80.44	3974
	4	min	54	38	14.99	3474
		med	180	47	377.5	6282
		max	1153	63	43879.56	12171
t=3	1	min	13	9	2.28	2594
		med	29	12	15.07	7014
		max	29	12	19.84	7014
	2	min	15	13	3.8	6496
		med	28	13	19.1	8776
		max	32	15	90.78	14156
	3	min	18	16	3.82	4096
		med	32	17	140.7	20501
		max	47	18	760.67	28240
	4	min	23	19	62.53	24659
		med	33	21	410.34	38834
		max	62	27	11241.67	103923

$$e_{ij} \in \{0, 1\}. \tag{1}$$

Leaving cyclic graphs aside for the moment (which will be discussed later on in this section), to generate a test case, i.e., to form a flow from the source node v_0 to the sink node $v_\perp$, one of the outgoing edges of v_0 and one of the incoming edges of $v_\perp$ must be taken:

$$\sum_{(v_0, v_i) \in E} e_{0i} = 1 \tag{2}$$

$$\sum_{(v_i, v_\perp) \in E} e_{i\perp} = 1. \tag{3}$$

Note that since the graph is acyclic, at most one of the incoming and at most one of the outgoing edges of a node can be taken, i.e., there can be a flow on at most one incoming and at most one outgoing edge.

The flow through node i is then expressed as a constraint indicating that the amount of outgoing flow from i is the same as the amount of incoming flow to i:

$$\sum_{(v_k, v_i) \in E} e_{ki} = \sum_{(v_i, v_l) \in E} e_{il} \;\leq\; 1. \tag{4}$$

Note that the source and the sink nodes are exempt from (4) since there is no flow into the source node and no flow out of the sink node.

As an example, Figure 4 presents an encoding to compute a test case, i.e., a path from v_0 to $v_\perp$, for the graph given in Figure 2c.

To make sure that a specific order is covered by a test case, additional constraints are needed. More

$$e_{01}, e_{12}, e_{13}, e_{14}, e_{15}, e_{24}, e_{34}, e_{45}, e_{5\perp} \in \{0, 1\}$$
$$e_{01} = 1$$
$$e_{5\perp} = 1$$
$$e_{01} = e_{12} + e_{13} + e_{14} + e_{15}$$
$$e_{12} = e_{24}$$
$$e_{13} = e_{34}$$
$$e_{24} + e_{14} + e_{34} = e_{45}$$
$$e_{45} + e_{15} = e_{5\perp}$$

Fig. 4: Single-source single-sink encoding to find a path from the entry node v_0 to the exit node $v_\perp$ in the graph given in Figure 2c.

formally, to cover a t-order $[v_{i_1}, \ldots, v_{i_t}]$ in a graph $G = (V, E, v_0, v_\perp)$, the following additional constraints are needed:

$$\sum_{(v_k, v_{i_s}) \in E} e_{ki_s} = 1 \; for \; 1 \leq s \leq t, \tag{5}$$

which indicate that all the nodes in the requested order must be visited. Since the graph is acyclic, the order of visit is guaranteed.

For example, to cover the 3-order $[v_1, v_4, v_5]$ in Figure 2c, the encoding in Figure 4 needs to be extended with:

$$\begin{aligned} e_{01} &= 1 \\ e_{24} + e_{14} + e_{34} &= 1 \\ e_{45} + e_{15} &= 1. \end{aligned} \tag{6}$$

To cover the same order $[v_{i_1}, \ldots, v_{i_t}]$ in a non-consecutive manner, however, the following constraint is required in addition to (5):

$$\sum_{1 < s \leq t} e_{i_{s-1} i_s} < t - 1, \tag{7}$$

which ensures that the length of the path from v_{i_1} to v_{i_t} is at least t.

Going back to our running example, to cover $[v_1, v_4, v_5]$ in a non-consecutive manner, the following additional constraint is required on top of (6):

$$e_{14} + e_{45} < 2. \tag{8}$$

If, on the other hand, the t-order $[v_{i_1}, \ldots, v_{i_t}]$ needs to be covered in a consecutive manner, then the following constraints are needed instead of (5) and (7):

$$e_{i_{s-1} i_s} = 1 \; for \; 1 < s \leq t, \tag{9}$$

making sure that the edges between all the consecutive pairs of nodes in the order are taken.

For our running example in Figure 4, we would need the following additional constraints to cover $[v_1, v_4, v_5]$ in a consecutive manner:

$$\begin{aligned} e_{14} &= 1 \\ e_{45} &= 1. \end{aligned} \tag{10}$$

For a given graph $G = (V, E, v_0, v_\perp)$, we have, therefore, defined the F-CIT model in this study as $M =< P, D, C >$, where P is the set of variables, each of which represents a distinct edge in the graph; D is a set of sets $\{0, 1\}$, one per edge, indicating whether there is flow on the edge or not (i.e., whether the edge is taken or not); and C is the model constraint capturing the reachability restrictions in the graph. More specifically, for a given graph, Equations 1-4 constitute the model constraint C. For example, Figure 4 presents the F-CIT model constraint created for the graph given in Figure 2c. Note that, given a graph, C stays the same regardless of the testable entities to be covered. Each F-CIT testable entity then corresponds to an order to be covered (Definitions 21-24). In particular, to cover a t-order, Equation 5; to cover a t-order in a non-consecutive manner, Equations 5 and 7; and to cover a t-order in a consecutive manner, Equation 9 needs to be used. As an example, Equation 6 presents the constraints to be used to cover the 3-order $[v_1, v_4, v_5]$ in the graph given in Figure 2c. Similarly, Equations 6 and 8 are needed to cover the same 3-order in a non-consecutive manner. And, Equation 10 is needed to cover it in a consecutive manner. Each F-CIT test case then corresponds to a path from the entry node v_0 to the exit node $v_\perp$ (Definition 17), covering a number of required orders.

Note that we have so far been concerned with directed acyclic graphs (DAGs). To work with cyclic graphs, we unroll the cycles k times (for this work, $k = 1$), which is a frequently used approach in bounded model checking [43]. To this end, we first convert a given graph to a regular expression [45], where all the Kleene plus operators are replaced by using the Kleene star operator, i.e., converting a^+ to aa^*. We then replace all the Kleene stars in the expression using the bounded repetition operator, such that the respective strings can be repeated at most k times, i.e., converting a^* to $a\{0, k\}$. Finally, the resulting regular expression is converted back to a graph.

For this work, we used the `Vcsn` tool [46] to carry out these steps. More specifically, converting a graph to a regular expression and a regular expression to a graph were carried out by using a single `Vcsn` shell command. And, replacing the unbounded Kleene star operators by bounded repetition operators was performed by another shell command using the `replace` string-replacement utility.

```
% graph
edge(v0, v1).
edge(v1, v2).
%...

% 'reaches' definitions
reaches(A, B) :- edge(A, B).
reaches(A, B) :- edge(A, C), reaches(C, B).

% 3-orders
order(A, B, C) :- reaches(A, B), reaches(B, C).

% consecutive-3-orders
consec(A, B, C) :- edge(A, B), edge(B, C).

% non-consecutive-3-orders
nonconsec(A, B, C) :- reaches(A, X), X!=B,
                      reaches(X, B),
                      reaches(B, C).
nonconsec(A, B, C) :- reaches(A, B),
                      reaches(B, X), X!=C,
                      reaches(X, C).
```

Fig. 5: An example ASP encoding for determining the valid consecutive, nonconsecutive, and regular 3-orders.

After having an acyclic graph, we used ASP (Answer Set Programming) [47], [48] to determine the different types of t-orders to be covered. Note that this step could also have been carried out by using reachability-based graph algorithms. We, however, chose to use ASP because, being a declarative logic programming paradigm, it was a perfect match for the task at hand. We were even able to provide whole code segments in the paper (e.g., Figure 5) to demonstrate the effort involved in the development

Figure 5 presents an example ASP encoding to determine the consecutive, non-consecutive, and regular 3-orders. Below, we explain the encoding in a nutshell with no intention to introduce ASP. For more details about ASP, the interested reader may refer to an introduction [49] or a book [50].

A DAG is expressed by using `edge(..)` facts. There is a path from node `A` to node `B`, i.e., `A` reaches `B` or `B` is reachable from `A`, if there is an edge from `A` to `B` (i.e., `edge(A, B)` holds) or there is an edge from `A` to `C` and `B` is reachable from `C`:

```
reaches(A, B) :- edge(A, B).
reaches(A, B) :- edge(A, C),
                 reaches(C, B).
```

Then, `[A, B, C]` is a valid 3-order, i.e., `order(A, B, C)`, if `A` reaches `B` and `B` reaches `C`:

```
order(A, B, C) :- reaches(A, B),
                  reaches(B, C).
```

For a 3-order `[A, B, C]` to be covered in a consecutive manner, there should be an edge from `A` to `B` and edge from `B` to `C`:

```
consec(A, B, C) :- edge(A, B),
                   edge(B, C).
```

And, for the same order to be covered in a non-consecutive manner, `B` should be reachable from `A` via another node or `C` should be reachable from `B` via another node:

```
nonconsec(A, B, C) :- reaches(A, X),
                      X!=B,
                      reaches(X, B),
                      reaches(B, C).
nonconsec(A, B, C) :- reaches(A, B),
                      reaches(B, X),
                      X!=C,
                      reaches(X, C).
```

Note that this encoding can trivially be extended to determine t-orders for any strength t.

Cost. All told, developing a generic Python script to unroll the cycles in a given graph using `Vcsn` took about 2 hours, which was mostly spent for writing procedures to match the input and output formats of `Vcsn`. Similarly, developing a generic Python script to automatically generate the ASP encodings for determining different types of t-orders to be covered, such as the one given in Figure 5, took about another 2 hours. Integrating a CSP solver, namely Sugar [51], with the constructors (as also discussed in Section 5.3.4) took less than 1 hour.

Evaluations. To evaluate the proposed approach, we first used our F-CIT constructors to compute t-way ($t = \{2, 3\}$) order-based F-CIT objects for the graphs discussed in Section 5.2.2. The experiments were repeated up to 5 times; 94% of the experiments with the best-performing F-CIT constructor (i.e., cover-and-generate) were repeated exactly 5 times. By construction, all the test cases selected by the F-CIT objects computed were valid and all these F-CIT objects achieved full coverage under the respective coverage criterion.

As was the case with the previous study (Section 5.1), the cover-and-generate constructor performed generally better than the generate-and-cover constructor. Therefore, we ran the generate-and-cover constructor with a time-out period of one day, while letting the cover-and-generate constructor run to completion. For 93.20% (1275 out of 1368) of the experimental setups, the generate-and-cover constructor computed the requested F-CIT objects within the allocated time limits. For these setups, the cover-and-generate constructor reduced the sizes by an average of 65.86% and 60.79%, while at the same time reducing the construction times by an average of 72.24% and 77.03% when $t = 2$ and 3, respectively. We, therefore, focus on the results obtained from the cover-and-generate constructor in the remainder of this section.

Figure 3 presents the results obtained from the cover-and-generate constructor and Table 13 provides some summary statistics. As expected, the coverage criteria listed in the order of increasing number of entities they required to cover, were: t-cover, t^+-cover, t-order,

and t^*-cover. These criteria respectively marked an average of 38.50, 665.87, 672.37, and 704.36 entities for coverage when $t = 2$; and an average of 61.73, 21678.98, 21685.99, and 21740.71 entities when $t = 3$.

The sizes of the order-based F-CIT objects as well their construction times tended to be correlated with the number of entities to be covered. Overall, the minimum, the average, and the maximum sizes of the F-CIT objects created were 1, 17.28, and 220, respectively, when $t = 2$; and 1, 54.27, and 2672 when $t = 3$. And the construction times for these objects respectively were 2.57, 461.32, and 4156.25 seconds when $t = 2$; and 3.24, 2568.12, and 136116.19 seconds when $t = 3$.

Another trend we observed was that although the numbers of entities to be covered by the t-order criterion were similar to those to be covered by the t^+-cover and t^*-cover criteria, covering the latter set of entities took longer than covering the former set of entities. The average constructions times were 251.29, 751.03, and 806.92 seconds for t-order, t^+-cover, and t^*-cover criteria, respectively, when $t = 2$; and 6885.59, 9508.46, and 10626.21 seconds when $t = 3$. We believe that this was because of the additional constraints to be satisfied to make sure that the requested orders are covered in a non-consecutive manner (i.e., need for solving the constraints in Equation 7 on top of Equation 5).

Computing 4-way order-based F-CIT objects. Last but not least, we ran our cover-and-generate constructor for $t = 4$ with a time-out period of 200 hours. For 88.01% (602 out of 684) of the experimental setups, the constructor was able to generate the requested F-CIT objects within the allocated time limits, whereas for the remaining 11.99% of the setups, it timed out.

Table 14 presents the results we obtained. Overall, the minimum, the average, and the maximum sizes of the 4-way order-based F-CIT objects computed were 1, 136.42, and 5858, respectively. And the construction times for these objects respectively were 1.87, 4522.53, and 189773.20 seconds.

5.2.5 Discussion

Note that given a graph, there are different approaches for solving the problem of finding a path covering certain sequences of nodes. In this study, however, our goal was to demonstrate that there is at least one solution, which can be expressed in F-CIT. For example, instead of using a constraint solver, one can use a model checker and formulate the same problem as a property stating that there is no path covering the requested orders. A counter example (if any) would then be a test case covering the orders. Similarly, one can even develop a special purpose constraint solver, which uses graph-based reachability algorithms, to determine whether a given set of orders can appear on a single path. These solutions would all work with F-CIT as long as the underlying solver supports the single primitive $solve$ as discussed in Section 4.

5.3 Study 3: Usage-Based CIT

An electronics company has approached us to improve their CIT-based testing practices. In particular, they were interested in testing the Internet connectivity feature of a consumer device, which they market in dozens of countries. The end-users of this device can customize the aforementioned feature by using 9 configuration options, which have 308, 280, 154, 82, 58, 41, 6, 3, and 2 settings, respectively. Since there is no system-wide constraint, all possible configurations (i.e., all possible combinations of option settings) are valid. All told, these options constitute a space of more than 90 billion valid configurations.

The company provided us with 526691 real configurations that they collected from the field during the month of May in 2016. Each configuration was obtained from a different consumer device and there were a total of 37503 distinct configurations, i.e., some configurations were used by multiple costumers.

Historically, configuration-related failures in this system have often been caused by the faulty interactions among the configuration options. However, exhaustive testing of neither the whole configuration space nor the distinct configurations seen in the field, is desirable for the company. Due to legal and privacy concerns, we are not able to provide further details.

5.3.1 Coverage criterion

We first attempted to create standard covering arrays for the scenario at hand (see Section 5.3.3 for more information). It turned out that the smallest covering array we could generate was a 2-way covering array of size 86241. It is, however, quite difficult to justify the use of all these configurations for testing when one knows that the total number of distinct configurations used in the field is 37503. Had the company had enough resources (i.e., time and computing platforms) to test all the distinct configurations in the field, they would have done it.

We, therefore, defined two novel coverage criteria, namely K_{seen} and $K_{weighted}$, based on the idea that when testing all t-tuples is not feasible, one should at the very least, consider testing the t-tuples appearing in the field. We call this *usage-based CIT*.

Definition 25. *The* seen-t-way coverage criterion K_{seen} *takes as input a set of configurations T, a coverage strength t, and a* cutoff frequency *in $[0, 1)$, and mark for coverage all the t'-tuples ($1 \leq t' \leq t$) appearing in T, the frequencies of which are greater than the cutoff frequency.*

The frequency of a tuple is computed as follows:

Definition 26. *Given a set of configurations T, the* frequency *of a tuple is the ratio of the number of configurations in T, in which the tuple appear, to the total number of configurations in T.*

Note that, when the frequency cutoff is 0, K_{seen} selects all the t'-tuples ($1 \leq t' \leq t$) appearing in T.

K_{seen} can further be extended to obtain variable strength coverage by using a weighted sum of the frequencies, where the weight of a tuple is defined as follows:

Definition 27. *Given a set of configurations T, the* weight *of a tuple is the ratio of the number of times the tuple appears in T to the total number of tuples in T.*

Note that computing the denominator in Definition 27 does not require to explicitly enumerate all possible tuples appearing in T. More specifically, since the number of tuples in a given configuration of k options is $2^k - 1$, the total number of tuples in T (thus the denominator) is $|T|(2^k - 1)$.

Definition 28. *The* weighted-t-way coverage criterion $K_{weighted}$ *takes as input a set of configurations T, a coverage strength t, and a* cutoff weight *in $(0, 1]$, and mark for coverage a minimal set of t'-tuples ($1 \leq t' \leq t$), the total weight of which is greater than or equal to the given cutoff weight.*

To determine the tuples to be covered by this criterion, all the t'-tuples ($1 \leq t' \leq t$) appearing in T are sorted by the descending order of their weights. Then, the minimum number tuples from the top of the list are selected, such that the total weight of the selected tuples is greater than or equal to the cutoff weight. Note that the $K_{weighted}$ criterion with the cutoff weight of 1 can be satisfied by selecting all the distinct configurations in T.

5.3.2 Study setup

For the evaluations, we used the aforementioned subject application with 9 configuration options, which had 308, 280, 154, 82, 58, 41, 6, 3, and 2 settings, respectively, together with the 526691 real configurations collected from the field, out of which 37503 were distinct.

All the experiments were carried out on the same Google Cloud platform with the previous two studies (Sections 5.1 and 5.2).

5.3.3 Applying standard CIT

Modeling. We first attempted to create standard covering arrays of various strengths by using a number of well-known covering array constructors, namely Jenny [52], PICT [53], and ACTS [29].

The very first thing we observed was that although we had a small number of configuration options (only 9), due to the large number of settings some of these options had, many of the existing covering array constructors failed to generate the requested covering arrays. For example, we were not even able to model the configuration space in Jenny, because it turned out Jenny employs the letters of the English alphabet to represent the settings of a configuration option, limiting the maximum number of settings that an option can have to 52 (the number of capital and lowercase letters in the English alphabet). On the other hand, PICT, which is specifically designed for scalability [53], was able to generate a 2-way covering array of size 86241 in 100 seconds. It, however, failed to generate a 3-way covering array in 10 days, after which we terminated the process. Whereas ACTS was able to generate a 2-way covering array of size 86255 in 16 seconds and a 3-way covering array of size 13283730 in 1887 seconds (about 32 minutes). However, when we attempted to generate 4-way covering arrays, ACTS crashed after a while with some memory-related errors.

Note that given a usage-based coverage criterion, neither the tuples to be covered nor the tuples not to be covered can be expressed as constraints in standard constructors in an attempt to selectively determine what to cover and what not to cover. This is because constraints in standard constructors are globally enforced, i.e., all of the test cases selected must satisfy all of the constraints. Therefore, expressing a tuple, which is selected by a given coverage criterion, as a constraint to indicate that the tuple needs to be covered, will enforce the same tuple to appear in all of the selected configurations. Since this can prevent conflicting tuples from being covered, no covering array can be created. Similarly, expressing a tuple, which is not selected by a given coverage criterion, as a constraint to indicate that the tuple needs to be avoided, can also prevent the creation of a covering array. It may not, for example, be possible to assign values to certain model parameters due to some invalidated tuple combinations.

An alternative approach might be to express tuples that are not needed to be covered as *soft constraints*, which mark combinations of parameter values that are permitted, but not desirable [16]. However, when the tuples to be covered is a small fraction of all the tuples, the number of soft constraints can get quite large, which can in turn cause performance and scalability issues. For example, in our experiments, 99.90% of all the tuples (of strength up to and including a given value of t), on average, did not need to be covered. In other words, had soft constraints been used to express these need-not-to-be-covered tuples, the number of constraints would have been as high as 4.7 trillion in some experiments. We couldn't experiment with this approach, because none of the standard constructors that we have access to, supported soft constraints.

Evaluations. All told, the size of the smallest standard covering array that we could generate was larger than the number of distinct configurations seen in the field, which rendered the use of standard covering arrays in this context hard to justify.

TABLE 16: Statistics about the K_{seen} coverage obtained by standard covering arrays. The columns, respectively, report the frequency cutoff values, the numbers of testable entities to be covered, and the numbers of testable entities covered by the standard 2-way and 3-way covering arrays created for the study.

	t=2			t=3			t=4			t=5			t=6		
	no of	% covered by standard CAs		no of	% covered by standard CAs		no of	% covered by standard CAs		no of	% covered by standard CAs		no of	% covered by standard CAs	
cutoff	entities	2-way	3-way	entities	2-way	3-way	entities	2-way	3-way	entities	2-way	3-way	entities	2-way	3-way
0.5	4	100	100	4	100	100	4	100	100	4	100	100	4	100	100
0.25	20	100	100	25	100	100	26	100	100	26	100	100	26	100	100
0.2	34	100	100	44	100	100	49	100	100	50	98	100	50	98	100
0.15	54	100	100	89	97	100	100	94	100	101	93	100	101	93	100
0.1	80	100	100	164	90	100	235	78	100	264	70	97	269	69	96
0.05	200	100	100	474	85	100	734	67	96	900	56	87	971	51	81
0.04	240	100	100	601	84	100	964	65	95	1204	53	85	1318	48	79
0.03	299	100	100	811	80	100	1395	58	94	1811	46	82	2001	42	75
0.02	422	100	100	1281	76	100	2382	53	92	3256	39	79	3705	35	71
0.01	669	100	100	2201	74	100	4475	47	90	6585	33	74	7825	28	63
0.005	1056	100	100	3751	71	100	7949	44	89	12028	30	71	14634	25	59
0.001	2652	100	100	11604	67	100	27890	39	85	45599	25	65	57741	19	53
0	22554	100	100	182952	42	100	658825	11	67	1240182	1	27	1321685	0	4

TABLE 17: Statistics about the $K_{weighted}$ coverage obtained by standard covering arrays. The columns, respectively, report the weight cutoff values, the numbers of testable entities to be covered, and the numbers of testable entities covered by the standard 2-way and 3-way covering arrays created for the study.

	t=2			t=3			t=4			t=5			t=6		
	no of	% covered by standard CAs		no of	% covered by standard CAs		no of	% covered by standard CAs		no of	% covered by standard CAs		no of	% covered by standard CAs	
cutoff	entities	2-way	3-way	entities	2-way	3-way	entities	2-way	3-way	entities	2-way	3-way	entities	2-way	3-way
0.70	332	100	100	2171	74	100	8037	44	89	18518	28	69	29272	22	56
0.75	426	100	100	2951	72	100	11240	42	87	26467	27	68	42694	21	54
0.80	566	100	100	4117	70	100	16252	41	87	39309	25	66	63498	19	52
0.85	793	100	100	6051	69	100	24798	39	86	60222	23	64	98886	17	50
0.90	1185	100	100	9675	68	100	40330	37	85	101015	21	62	169254	15	47
0.95	2073	100	100	17826	64	100	79330	33	83	211007	17	58	368650	12	43

To further demonstrate that obtaining full coverage in an efficient and effective manner under the usage-based coverage criteria is a non-trivial task, Tables 16-17 report the coverage percentages obtained by the standard covering arrays generated in this study. In particular, when $t > 3$ with K_{seen}, the standard 2- and 3-way covering arrays did not guarantee to cover all the requested tuples. For example, when cutoff=0.001, only 39% (85%), 25% (65%), and 19% (53%) of all the required tuples for $t = 4$, 5, and 6 under K_{seen}, were covered by the standard 2-way (3-way) covering arrays (Table 16). Similarly, the standard covering arrays did not guarantee to cover all the tuples requested by $K_{weighted}$ either, especially for large values of coverage strength and weight cutoff values. For example, when $t = 6$ and cutoff=0.95 only 12% and 43% of all the required tuples were covered by the standard 2-way and 3-way covering arrays, respectively (Table 17).

5.3.4 Applying F-CIT

Modeling. We have defined the F-CIT model as $M =< P, D, C >$, where $P = \{o_1, \ldots, o_9\}$, $D = \{\{1..308\}, \{1..280\}, \{1..154\}, \{1..82\}, \{1..58\}, \{1..41\}, \{1..6\}, \{1..3\}, \{1, 2\}\}$, and $C : true$, indicating that all possible configurations were valid.

Each F-CIT testable entity then naturally corresponded to a tuple selected by the coverage criterion K_{seen} or $K_{weighted}$, which was expressed as a constraint over finite sets. For example, the 3-tuple $(o_1 = 204, o_5 = 12, o_9 = 1)$ was expressed as $o_1 = 204 \wedge o_5 = 12 \wedge o_9 = 1$. Note that the very same approach can readily be used to define and compute standard t-way covering arrays as F-CIT objects by expressing all valid t-tuples as F-CIT testable entities.

Consequently, any solver that works with logical operators, such as $\wedge$ (AND), and equality constraints over finite sets, such as $o_1 = 204$, including the commonplace SAT and CSP solvers [54], [55], can be used with the F-CIT constructors compute the F-CIT objects satisfying the K_{seen} and $K_{weighted}$ criteria.

Indeed, being able to work with any type of constraints as long as an appropriate solver is provided, improves the flexibility of F-CIT. To demonstrate that this feature also enables the use of domain- and/or application-specific solvers, we have implemented a quite simple solver for this study, instead of trivially using an existing SAT or CSP solver.

Algorithm 3 presents the aforementioned solver. It simply determines whether a given set of tuples E can be accommodated together in a single configuration or not. In particular, it marks E as satisfiable as long as the option settings appearing in E do not contradict with each other (lines 7-8).

Cost. All told, developing a generic script to determine the tuples (i.e., the testable entities) selected by the K_{seen} and $K_{weighted}$ coverage criteria for any configuration space model, coverage strength, and cutoff value, took less than 2 hours. And, implementing the solver in Algorithm 3 and integrating it with the F-CIT constructors took less than 1 hour. To further demon-

Algorithm 3 Determine if a given set of tuples can be accommodated together in a configuration cfg.

Input: Set of tuples E

```
1: cfg ← undef
2: for each tuple e in E do
3:     for each option o in e do
4:         Let e[o] is the value of o in e
5:         Let cfg[o] is the value of o in cfg,
6:             which is initially undef
7:         if defined cfg[o] and cfg[o] ≠ e[o] then
8:             return False
9:         else
10:            cfg[o] = e[o]
11:        end if
12:    end for
13: end for
14: return True
```

strate the flexibility of F-CIT, we have also integrated our constructors with a CSP solver (namely, Sugar [51]) to solve exactly the same set of constraints. Interestingly enough, it took about the same time (less than 1 hour) for us to do that as we needed to implement a simple procedure to match the input format of the solver. The implementation was done in Python.

Evaluations. To evaluate the proposed approach, we carried out a series of experiments. In these experiments, we used the cover-and-generate and generate-and-cover constructors given in Algorithms 1 and 2 to compute F-CIT objects of various strengths. Since the cover-and-generate constructor performed generally better than the generate-and-cover constructor, the experiments with the latter constructor were repeated up to 3 times and with a time-out period of one day for each repetition to keep the cost of the experiments under control. The experiments with the former constructor, on the other hand, were repeated 100 times to evaluate the sensitivity of the proposed approach to the order, in which the testable entities are processed, except for the experimental setups where the frequency cutoff was 0 and $t > 2$, which were repeated only once, to keep the cost under further control. In all the experiments, the orders were randomly generated by shuffling the testable entities to be covered.

Evaluating the K_{seen} coverage criterion. Table 18 summarizes the results we obtained for the K_{seen} coverage criterion. We first observed that the F-CIT constructors, especially the cover-and-generate constructor, were scaled to obtain full coverage under K_{seen} for various values of t up to and including 6, even when the frequency cutoff was 0. As a matter of fact, we chose to stop at the strength of 6, because, in the presence of 9 options, increasing the strength any further was quickly becoming exhaustive testing, which, in this context, is the same as testing all the distinct configurations seen in the field.

We then observed that the cover-and-generate constructor performed generally better than the generate-and-cover constructor in reducing both the covering array sizes and construction times. More specifically, the cover-and-generate constructor reduced the sizes by an average of 65.62%, 82.55%, 86.35%, 88.25%, and 88.16% while at the same time reducing the construction times by an average of 96.35%, 97.20%, 97.88%, 97.72%, and 97.32% when $t = 2$, 3, 4, 5, and 6, respectively. We, therefore, focus on the results obtained from the cover-and-generate constructor in the remainder of this section.

When $t \leq 3$ and cutoff=0, i.e., when all the t-tuples seen in the field are required to be covered, the sizes of the F-CIT objects generated by the cover-and-generate constructor, were smaller than the number of distinct configurations seen in the field, i.e., 37503. More specifically, the average sizes were 3625.24 and 24971.00 with the average construction times of 205.55 and 7216.62 seconds for $t = 2$ and 3, respectively (Table 18). When $t > 3$ and cutoff=0, however, the F-CIT objects had more than 37503 configurations, on average (Table 18).

In reality, when testing all the t-tuples seen in the field is still not practical due to the cost, the cutoff parameters of the usage-based coverage criteria can be utilized to select a weighted fraction of the tuples for testing. For example, when the frequency cutoff was set to 0.001 with K_{seen}, i.e., when the tuples that appeared in at least one thousandth of the configurations seen in the field were to be covered, the average sizes of the F-CIT objects became 327.31, 817.66, 1124.70, 1123.06, and 1015.85 when $t = 2$, 3, 4, 5, and 6, respectively.

All the results we obtained under different coverage strengths and cutoff values can be found in Table 18. For a fixed strength, as the cutoff increased, the number of testable entities as well as the size of the F-CIT objects tended to decrease. For example, when $t = 6$, the average sizes of the F-CIT objects were 1015.85, 206.63, 95.96, 46.23, and 5.84 for cutoff=0.001, 0.005, 0.01, 0.02, and 0.1, respectively. For a fixed cutoff, as the strength increased, on the other hand, although the number of testable entities to be covered increased, this did not necessarily cause an increase in the sizes of the F-CIT objects computed. For example, when cutoff=0.005, the average size of the F-CIT objects was 228.94 for $t = 5$, but 206.63 for $t = 6$. We believe that this was because covering a frequently appearing t-tuple covers multiple frequently appearing t'-tuples, where $t' < t$. Thus, covering higher strength tuples may help reduce the number of test cases needed by covering more required tuples per test case. Regarding the construction times, except for the experimental setups, in which cutoff=0, all the constructions times

TABLE 18: Statistics about the F-CIT objects created for the K_{seen} coverage criterion, where the columns, respectively, report the coverage strengths, the frequency cutoff values, the numbers of testable entities to be covered, and the average construction times (in seconds) as well as the average sizes of the F-CIT objects computed by the generate-and-cover and cover-and-generate constructors together with the minimum, maximum, standard deviation, and coefficient of variation statistics for the results obtained from the latter constructor. The character '*' marks the experimental setups, in which the generate-and-cover constructor timed out after one day. Furthermore, the number of times the experiments were repeated are given in the column "repeat count."

			generate-and-cover constructor			cover-and-generate constructor										
						time					size					
t	cutoff	no of entities	avg. time	avg. size	repeat count	min.	avg.	max.	sd.	cv.	min.	avg.	max.	sd.	cv.	repeat count
2	0.5	4	0.02	2.00	3	0.00	0.03	0.06	0.01	57.86	1	1.00	1	0.00	0.00	100
2	0.25	20	0.05	8.67	3	0.01	0.08	0.19	0.05	56.35	2	2.31	3	0.46	20.02	100
2	0.2	34	0.13	16.00	3	0.01	0.13	0.27	0.07	56.09	3	3.89	5	0.61	15.80	100
2	0.15	54	0.25	24.00	3	0.00	0.18	0.38	0.11	61.40	4	4.70	7	0.62	13.29	100
2	0.1	80	0.60	38.33	3	0.02	0.29	0.60	0.16	56.09	5	6.50	9	0.83	12.78	100
2	0.05	200	3.02	94.33	3	0.06	0.73	1.54	0.41	55.85	13	16.00	19	1.33	8.34	100
2	0.04	240	4.51	124.00	3	0.06	0.93	1.95	0.52	55.76	19	21.72	25	1.46	6.74	100
2	0.03	299	6.05	152.33	3	0.04	1.14	2.40	0.66	58.07	21	27.46	31	1.77	6.46	100
2	0.02	422	10.47	211.33	3	0.14	1.74	3.61	0.97	55.53	35	40.53	45	1.93	4.75	100
2	0.01	669	23.79	333.00	3	0.32	3.07	6.07	1.67	54.36	61	69.34	75	2.82	4.06	100
2	0.005	1056	50.97	518.33	3	0.30	4.92	9.60	2.75	55.91	113	118.83	126	2.91	2.44	100
2	0.001	2652	265.32	1275.00	3	2.29	14.85	27.08	7.51	50.58	311	327.31	342	5.89	1.80	100
2	0	22554	10389.32	9606.33	3	228.72	364.05	493.08	75.38	20.71	3589	3625.24	3666	17.06	0.47	100
3	0.5	4	0.01	2.33	3	0.00	0.02	0.05	0.01	56.45	1	1.00	1	0.00	0.00	100
3	0.25	25	0.10	8.33	3	0.01	0.08	0.18	0.04	58.47	2	2.26	3	0.44	19.41	100
3	0.2	44	0.12	15.33	3	0.01	0.12	0.29	0.07	58.07	3	3.82	6	0.77	20.07	100
3	0.15	89	0.61	32.00	3	0.01	0.16	0.38	0.10	58.08	4	4.96	8	0.95	19.11	100
3	0.1	164	1.20	53.67	3	0.03	0.28	0.68	0.17	57.97	5	6.86	10	1.33	19.34	100
3	0.05	474	8.16	170.00	3	0.06	0.79	2.66	0.50	63.02	14	18.54	24	2.22	12.00	100
3	0.04	601	10.22	194.00	3	0.13	1.08	2.09	0.57	52.87	20	26.24	32	2.32	8.84	100
3	0.03	811	20.99	285.33	3	0.12	1.37	2.82	0.74	54.17	27	34.60	42	3.14	9.08	100
3	0.02	1281	40.73	456.00	3	0.19	2.30	4.48	1.26	54.66	45	57.47	68	3.95	6.87	100
3	0.01	2201	106.93	781.67	3	0.58	4.66	9.26	2.48	53.10	103	113.27	129	4.67	4.13	100
3	0.005	3751	251.85	1340.67	3	1.26	9.05	17.85	4.59	50.71	203	217.86	238	6.87	3.15	100
3	0.001	11604	1813.94	4137.00	3	14.21	43.31	71.85	16.66	38.46	790	817.66	853	12.96	1.59	100
3	0	182952	197208.59	72642.00	2	8266.04	8266.04	8266.04	n/a	n/a	24971	24971.00	24971	n/a	n/a	1
4	0.5	4	0.02	2.33	3	0.00	0.02	0.06	0.01	60.87	1	1.00	1	0.00	0.00	100
4	0.25	26	0.05	7.67	3	0.01	0.08	0.18	0.04	58.93	2	2.31	3	0.46	20.02	100
4	0.2	49	0.15	16.00	3	0.01	0.11	0.28	0.07	59.04	3	3.65	5	0.73	19.90	100
4	0.15	100	0.43	24.33	3	0.01	0.15	0.37	0.09	57.23	4	4.91	8	0.91	18.46	100
4	0.1	235	1.69	65.33	3	0.01	0.24	0.55	0.14	58.37	5	6.32	10	1.22	19.23	100
4	0.05	734	10.29	188.33	3	0.07	0.66	1.29	0.34	51.88	11	16.82	22	2.09	12.40	100
4	0.04	964	16.06	235.00	3	0.12	0.90	1.99	0.49	54.20	16	23.26	31	3.06	13.15	100
4	0.03	1395	31.92	368.67	3	0.14	1.25	2.85	0.66	52.82	26	32.07	41	2.90	9.04	100
4	0.02	2382	74.45	606.67	3	0.30	2.17	4.51	1.10	50.58	44	54.46	66	3.93	7.21	100
4	0.01	4475	189.34	1136.00	3	1.10	5.07	10.46	2.43	47.96	99	118.57	138	7.23	6.10	100
4	0.005	7949	503.37	2073.33	3	2.97	11.06	19.49	4.85	43.84	223	248.91	270	9.51	3.82	100
4	0.001	27890	3793.76	7266.00	3	39.78	76.39	114.12	20.34	26.63	1089	1124.70	1165	17.37	1.54	100
4	0	658825	*	*	1	62011.15	62011.15	62011.15	n/a	n/a	59960	59960.00	59960	n/a	n/a	1
5	0.5	4	0.01	3.00	3	0.00	0.02	0.05	0.01	56.31	1	1.00	1	0.00	0.00	100
5	0.25	26	0.05	8.67	3	0.01	0.08	0.19	0.05	59.27	2	2.30	3	0.46	19.92	100
5	0.2	50	0.10	14.33	3	0.00	0.11	0.28	0.07	60.11	3	3.76	5	0.72	19.22	100
5	0.15	101	0.39	28.33	3	0.01	0.15	0.39	0.09	59.85	4	4.89	7	0.86	17.57	100
5	0.1	264	1.31	53.00	3	0.02	0.21	0.46	0.12	56.31	5	5.84	10	1.01	17.25	100
5	0.05	900	11.66	185.00	3	0.10	0.62	1.44	0.33	52.70	13	15.91	21	1.65	10.37	100
5	0.04	1204	18.24	236.33	3	0.14	0.85	1.76	0.45	52.63	16	21.74	28	2.70	12.42	100
5	0.03	1811	32.68	339.33	3	0.21	1.19	4.40	0.65	54.62	24	29.79	38	2.82	9.47	100
5	0.02	3256	68.51	598.67	3	0.48	2.07	4.04	0.94	45.63	40	48.82	57	3.92	8.02	100
5	0.01	6585	217.98	1188.67	3	1.20	4.64	7.92	1.95	41.96	89	106.41	126	7.00	6.58	100
5	0.005	12028	564.90	2256.67	3	4.33	11.50	19.43	4.18	36.39	206	228.94	244	8.89	3.88	100
5	0.001	45599	4400.78	8645.33	3	66.72	100.02	133.16	18.82	18.81	1056	1123.06	1168	20.17	1.80	100
5	0	1240182	*	*	1	171331.44	171331.44	171331.44	n/a	n/a	88314	88314.00	88314	n/a	n/a	1
6	0.5	4	0.01	2.33	3	0.00	0.03	0.05	0.01	56.15	1	1.00	1	0.00	0.00	100
6	0.25	26	0.03	5.00	3	0.01	0.07	0.18	0.04	56.12	2	2.22	3	0.41	18.66	100
6	0.2	50	0.14	14.33	3	0.00	0.11	0.64	0.08	75.56	3	3.73	5	0.72	19.28	100
6	0.15	101	0.36	27.00	3	0.01	0.15	0.41	0.09	57.95	4	4.73	7	0.86	18.15	100
6	0.1	269	1.18	43.33	3	0.02	0.20	0.59	0.12	57.36	5	5.84	10	0.97	16.55	100
6	0.05	971	10.15	155.67	3	0.07	0.57	1.27	0.30	52.93	13	15.45	20	1.66	10.77	100
6	0.04	1318	16.74	223.33	3	0.11	0.80	1.64	0.40	50.33	17	21.15	29	2.22	10.47	100
6	0.03	2001	25.10	280.67	3	0.24	1.14	2.17	0.53	46.50	24	28.33	35	2.45	8.63	100
6	0.02	3705	68.32	538.33	3	0.51	1.93	3.87	0.84	43.62	38	46.23	60	3.58	7.74	100
6	0.01	7825	196.90	1061.00	3	1.39	4.44	8.35	1.77	39.80	78	95.96	112	6.31	6.58	100
6	0.005	14634	493.02	2031.00	3	5.10	10.83	17.33	3.52	32.50	186	206.63	225	7.91	3.83	100
6	0.001	57741	3880.13	7842.67	3	74.39	105.56	139.61	16.58	15.71	963	1015.85	1065	20.88	2.06	100
6	0	1321685	*	*	1	179456.00	179456.00	179456.00	n/a	n/a	80350	80350.00	80350	n/a	n/a	1

TABLE 19: Statistics about the F-CIT objects created for the $K_{weighted}$ coverage criterion, where the columns, respectively, report the coverage strengths, the weight cutoff values, the numbers of testable entities to be covered, and the average construction times (in seconds) as well as the average sizes of the F-CIT objects computed by the generate-and-cover and cover-and-generate constructors together with the minimum, maximum, standard deviation, and coefficient of variation statistics for the results obtained from the latter constructor. Furthermore, the number of times the experiments were repeated are given in the column "repeat count."

			generate-and-cover constructor			cover-and-generate constructor										
						time					size					
t	cutoff	no of entities	avg. time	avg. size	repeat count	min.	avg.	max.	sd.	cv.	min.	avg.	max.	sd.	cv.	repeat count
2	0.70	332	7.16	167.67	3	0.07	0.76	1.73	0.48	62.75	27	31.74	37	1.97	6.22	100
2	0.75	426	10.23	209.67	3	0.09	0.99	2.16	0.63	63.43	36	41.10	47	2.20	5.36	100
2	0.80	566	17.90	279.33	3	0.13	1.38	3.03	0.88	63.99	49	55.73	61	2.26	4.05	100
2	0.85	793	34.33	400.33	3	0.24	2.11	4.64	1.30	61.60	77	85.40	94	3.00	3.51	100
2	0.90	1185	63.95	574.67	3	0.46	3.33	7.03	1.99	59.87	126	133.28	141	3.60	2.70	100
2	0.95	2073	162.53	990.33	3	1.10	6.42	14.11	3.73	58.04	235	245.66	256	4.46	1.82	100
3	0.70	2171	100.70	777.33	3	0.49	2.73	5.86	1.54	56.41	100	111.32	120	4.15	3.73	100
3	0.75	2951	181.72	1090.00	3	0.89	4.14	8.51	2.25	54.34	150	164.80	178	5.04	3.06	100
3	0.80	4117	303.40	1488.67	3	1.59	6.60	12.70	3.40	51.50	227	242.96	263	6.28	2.58	100
3	0.85	6051	582.78	2164.00	3	3.40	11.20	21.29	5.25	46.83	362	384.02	405	8.09	2.11	100
3	0.90	9675	1291.66	3447.33	3	9.26	22.47	38.32	8.79	39.12	639	666.56	696	10.16	1.52	100
3	0.95	17826	3671.09	6332.00	3	34.85	63.21	94.30	17.61	27.86	1320	1347.32	1387	14.03	1.04	100
4	0.70	8037	493.52	2095.67	3	3.07	7.74	13.70	3.07	39.64	218	251.25	275	8.98	3.57	100
4	0.75	11240	850.59	2905.67	3	6.23	13.25	25.50	4.72	35.59	357	382.40	414	10.73	2.80	100
4	0.80	16252	1608.39	4222.33	3	12.85	23.91	37.19	7.35	30.76	568	605.22	645	13.96	2.31	100
4	0.85	24798	3177.50	6454.67	3	31.82	50.10	71.79	11.96	23.88	947	984.22	1039	19.26	1.96	100
4	0.90	40330	6942.08	10454.67	3	90.62	127.55	200.33	23.77	18.64	1675	1734.26	1816	24.83	1.43	100
4	0.95	79330	21111.57	20702.00	3	403.89	459.80	500.68	19.50	4.24	3768	3868.14	3923	29.77	0.77	100
5	0.70	18518	1070.09	3516.67	3	10.01	16.76	24.58	4.39	26.17	362	386.81	413	11.02	2.85	100
5	0.75	26467	1896.90	5018.33	3	20.37	31.69	46.81	7.37	23.24	573	608.41	662	15.30	2.51	100
5	0.80	39309	3540.80	7484.33	3	47.71	66.26	109.47	12.47	18.83	911	956.32	1003	19.17	2.00	100
5	0.85	60222	6776.82	11397.67	3	115.87	132.83	148.14	7.46	5.61	1499	1555.03	1612	25.24	1.62	100
5	0.90	101015	15938.67	19338.33	3	362.41	411.02	439.87	14.26	3.47	2751	2849.14	2937	37.23	1.31	100
5	0.95	211007	50420.67	40552.50	2	1977.17	2081.92	2156.12	39.50	1.90	6738	6857.15	6974	48.97	0.71	100
6	0.70	29272	1380.18	4027.33	3	18.73	27.23	40.74	5.67	20.81	442	475.20	502	12.64	2.66	100
6	0.75	42694	2511.12	5857.00	3	40.92	54.86	90.64	9.36	17.06	708	743.64	789	16.65	2.24	100
6	0.80	63498	4650.73	8831.33	3	86.44	102.17	113.87	5.70	5.58	1065	1113.34	1156	20.65	1.85	100
6	0.85	98886	9152.57	13601.33	3	227.07	260.64	288.15	10.86	4.17	1765	1853.89	1936	32.88	1.77	100
6	0.90	169254	22546.59	23611.67	3	775.99	846.07	896.63	25.20	2.98	3359	3449.67	3540	37.66	1.09	100
6	0.95	368650	75936.20	52238.00	1	4482.59	4611.18	4785.40	61.42	1.33	8495	8627.77	8794	57.54	0.67	100

were under 106 seconds, with a majority of them being under 12 seconds, on average.

Evaluating the $K_{weighted}$ coverage criterion. Table 19 summarizes the results we obtained from the experiments, in which we used the $K_{weighted}$ coverage criterion.

As was the case with K_{seen}, the cover-and-generate constructor, compared to the generate-and-cover constructor, computed smaller covering arrays at a fraction of the cost. More specifically, the cover-and-generate constructor reduced the sizes by an average of 77.39%, 80.93%, 83.29%, 83.09%, and 80.29% while at the same time reducing the construction times by an average of 94.94%, 98.20%, 98.00%, 95.88%, and 92.00%, when $t = 2$, 3, 4, 5, and 6, respectively. We, therefore, focus on the results obtained from the cover-and-generate constructor in the remainder of this section.

We observed that the sizes of all the F-CIT objects we computed for the study, were profoundly smaller than the number of distinct configurations observed in the field (Table 19). More specifically, the maximum average size was 8627.77, which occurred when $t = 6$ and the weighted cutoff was 0.95. That is, to cover 95% of the most frequently appearing t-tuples for all $1 \leq t \leq 6$ in a weighted manner as described in Definition 28, an F-CIT object of size 8627.77 was needed, on average.

For a fixed strength, as the cutoff decreased, the number of testable entities to be covered as well as the size of the F-CIT objects tended to decrease. For example, when $t = 6$, the sizes of the F-CIT objects for cutoff=0.95, 0.90, 0.85, 0.80, 0.75, and 0.70, were, respectively, 8627.77, 3449.67, 1853.89, 1113.34, 743.64, and 475.20 (Table 19). Similarly, for a fixed cutoff, as the coverage strength decreased, both the number of testable entities to be covered as well as the size of the F-CIT objects tended to decrease. For example, when cutoff=0.95, the average sizes were 8627.77, 6857.15, 3868.14, 1347.32, and 245.66 for $t = 6$, 5, 4, 3, and 2, respectively. Last but not least, the maximum average construction time in all the experiments was 4611.18 seconds, which happened when $t = 6$ and $cutoff = 0.95$. A majority of the construction times (79.4%) were, on the other hand, under 150 seconds (Table 19).

Evaluating sensitivity to the order of processing. Tables 18 and 19 report the minimum, maximum, standard deviation, and coefficient of variation (i.e., the ratio of the standard variation to the mean, in short

CV) results obtained from the cover-and-generate constructor by repeating the experiments 100 times (except for the experimental setups where the frequency cutoff was 0 and $t > 2$, which were repeated only once due to their costs). Clearly, the performance of the cover-and-generate constructor can be affected by the order, in which the testable entities are processed. Consequently, in the absence of any knowledge regarding a favorable order (or a partial order), a random order can be used by shuffling the entities to be covered before they are fed to the constructor. This process can further be repeated multiple times in an attempt to generate smaller CIT objects at the cost of increased construction times.

5.3.5 Discussion

Note that the maximum coverage strength that can be used with the $K_{weighted}$ coverage criterion is the number of configuration options that the system under test has. Therefore, $K_{weighted}$, in a sense, offers a solution to an important, but still an open question of how to determine the coverage strength in CIT, by automatically determining strength based on usage statistics. That is, the strength of a tuple to be covered by $K_{weighted}$, essentially depends on how frequently the tuple appears in the field. Consequently, the strength may vary across the test space. This is different than variable strength covering arrays [56], because in variable strength covering arrays, the strengths are determined a priori and they vary at the level of option combinations. In the $K_{weighted}$ coverage criterion, on the other hand, the strengths vary at the level of option setting combinations and they are determined based on usage statistics. Therefore, no strength needs to be determined beforehand.

Note that F-CIT does not aim to replace standard covering array constructors. We, indeed, don't see much value in using F-CIT to compute the same CIT objects that the existing CIT constructors compute, as the generalized F-CIT constructors may not be as efficient and as effective as their specialized counterparts. For example, when we used the cover-and-generate constructor to compute standard covering arrays for the configuration space models used in the experiments, the aforementioned F-CIT constructor generated a 2-way standard covering array of size 87586 in 32263.90 seconds (vs. a 2-way covering array of size 86241 generated in 100 seconds by ACTS) and failed to generate a 3-way standard covering array within a day, after which we stopped the constructor (vs. a 3-way covering array of size 13283730 generated in 32 minutes by ACTS).

The point we want to emphasize, however, is that even if F-CIT was able to reduce the sizes by half and did so in seconds, it would still not be feasible at all (for the consumer company, for which we carried out the study) to run all the test cases selected. Therefore, the coverage criteria needed to be changed. However, existing constructors, as they are, could not take advantage of these new criteria, which required fewer tuples to be covered.

Last but not least, it seems that for the usage-based CIT problem, it may actually be possible to modify an existing constructor. This, however, requires that the source code of the constructor is available, the code is reversed engineered, and a modification strategy is implemented, tested, and maintained. Note, however, that even if this was possible, these modifications would be of little help (or of no help at all) to compute the structure-based and order-based CIT objects we discussed in Section 5.1 and Section 5.2, respectively. Consequently, another set of modifications would be required to compute the structure-based CIT objects, such that the values of model parameters can be expressed as arbitrarily complex Boolean expressions. Similarly, different set of modifications would be required to compute the order-based CIT objects, such that the reachability restrictions imposed by a graph-based model can be expressed as constraints to cover various orders of nodes. As a matter of fact, we don't know how these modifications can be made without the solution quickly converging to F-CIT. This is exactly why F-CIT aims to eliminate the need of modifying existing constructors or developing specialized constructors, by generalizing the construction of CIT objects as much as possible.

6 User Studies

To further evaluate the proposed approach, we have also carried out user studies.

6.1 Study Setup

We asked the Junior, Senior, and graduate-level computer science students studying at Sabanci University whether they would take part in the study. A total of 13 graduate-level and 7 undergraduate-level students agreed to participate on a voluntary basis. Table 20 summarizes the demographic information about the participants. Note that students at Sabanci University study standard combinatorial interaction approaches at different levels and/or for different purposes in the Software Engineering (undergraduate level), Software Verification and Validation (undergraduate/graduate level), and Automated Debugging (graduate level) courses, which explains the participants knowledgable of CIT in Table 20.

The participants were first given a 1-hour lecture. In this lecture, after a brief introduction of how the study would be carried out, the basic concepts in constraint solving, such as Boolean logic and satisfiability, were discussed. Then, F-CIT was introduced. To this end,

TABLE 20: Demographic information about the participants.

	undergraduate				graduate			
no of participants	7				13			
knowledgeable of CIT	**yes**		**no**		**yes**		**no**	
	1		6		10		3	
experience in programming (in years)	$\leq \mathbf{2}$	**3**	**4**	$\geq \mathbf{5}$	$\leq \mathbf{2}$	**3**	**4**	$\geq \mathbf{5}$
	2	3	2	0	0	2	5	6
experience in software testing (in years)	$\leq \mathbf{2}$	**3**	**4**	$\geq \mathbf{5}$	$\leq \mathbf{2}$	**3**	**4**	$\geq \mathbf{5}$
	7	0	0	0	11	1	0	1

the definitions and the algorithms given in Sections 3-4 were studied. Finally, a short tutorial on the F-CIT tool, which we had developed for the study was given (see below for more information about this tool).

The participants, after taking the lecture, took part in the study at their spare times. To gain better insight, each participant carried out the study with one of the authors playing the role of an observer, sitting by the participant and taking notes. The participants were asked to think out loud as much as possible. When it was not clear for the observer what the participant was doing, the observer prompted the participant with questions, such as what do you want to do now? Is the output what you were expecting? What do you think what went wrong? etc.

Each participant was given with the same three problems. These problems were, indeed, the smaller instances of the very same problems we studied in Section 5.1, Section 5.2, and Section 5.3, respectively. For each problem, the participants were first asked to develop an F-CIT model $M =< P, D, C >$, then to express a number of F-CIT entities as constraints using M, and finally to generate an F-CIT object (by using the F-CIT constructor provided) to cover all of the given entities. The problems as well as the entities used in the study were given in Table 21.

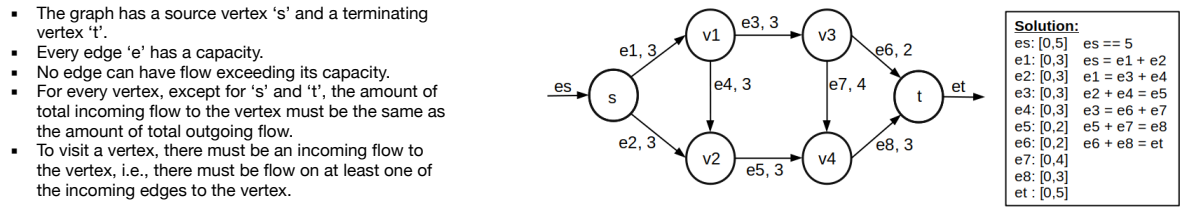
Note that F-CIT is not a methodology for choosing the entities to be tested. It rather takes as input a set of entities to be tested. Therefore, the participants in the study were given with a set of entities to cover. For each problem, the entities were presented starting from the easier ones progressing to the more challenging ones. Furthermore, the number of entities was kept small not to tire the participants.

We designed the studies such that if a participant working on a problem got stuck after the first 10 minutes, the observer would remind the participant of the basic concepts that 1) the F-CIT model $M =< P, D, C >$ should define a set of parameters P and their domains D; 2) the model constraint C is a constraint that should be satisfied by all of the F-CIT test cases generated; and 3) the entities should be expressed as constraints over P.

Furthermore, for the second problem (Table 21), if the participant got stuck after the first 15 minutes, he/she was provided with a description of the network flow problem [57] given in Figure 6a. If the participant got stuck again 15 minutes after reading the description, a solution for the example flow problem given in Figure 6b was presented to the participant. Note that both the description in Figure 6a and the example in Figure 6b are general enough that they can be found in

TABLE 21: Problems used in the user studies.

Problem 1
description
The same problem in Section 5.1 with 6 Boolean configuration options: $p_1, \ldots, p_6$.
entities
p_3
$p_1 \wedge p_2$
$\neg p_6 \wedge p4$
$p_2 \wedge \neg p_3 \wedge \neg p4$
$\neg(p_5 \vee \neg p_6)$
Problem 2
description
The same problem in Section 5.2 with the graph given in Figure 7.
entities
– t-orders –
$[a_1, a_5]$
$[a_1, a_3, a_4]$
$[a_2, a_4, a_6]$
– consecutive-t-orders –
$[a_4, a_6]$
$[a_1, a_3, a_4]$
$[a_3, a_4, a_6]$
– non-consecutive-t-orders –
$[a_1, a_2]$
$[a_1, a_3, a_4]$
$[a_4, a_6]$
Problem 3
description
The same problem in Section 5.3 with 5 parameters: $p_1, p_2 : [1, 3]$, $p_3 : [1, 4]$, and $p_4, p_5 : [1, 5]$.
entities
$(p_5 = 4)$
$(p_1 = 1, p_2 = 3)$
$(p2 = 1, p5 = 2)$
$(p3 = 3, p4 = 3, p5 = 2)$
$(p1 = 2, p2 = 2, p3 = 2, p4 = 5, p5 = 2)$



(a) (b)

Fig. 6: Explanations used for the second problem in the user study: (a) the description of the network flow problem and (b) an example network flow with incoming flow as 5 (i.e., $es = 5$) and its solution, where each edge has a label in the form of ex, c, indicating that c (except es and et) is the capacity of the edge ex.

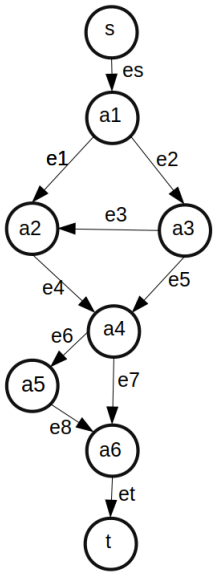


Fig. 7: The graph-based model used in the user studies.

any textbook on the subject [57]. Given these artifacts, the participants were still required to figure out how to express reachability in a graph as a flow problem and how to express different types of order-based entities (Section 5.1) as flow constraints.

We did this because solving the aforementioned problem requires specific knowledge of network flow problems and not all participants might have had the right background. Therefore, by providing a general description of the network flow problems together with an example, we aim to answer the following question: Had the participants had a basic background information on network flow problems, could they have leveraged it in F-CIT to obtain full coverage under various order-based coverage criteria?

Last but not least, we have developed an F-CIT tool for the practitioners to use in the study. Figure 8 presents a screen dump taken from this tool. At a very high level, the tool has three frames. A description frame on the left, presenting the problem to be solved. An F-CIT frame in the middle where the participant expresses a solution to the given problem in F-CIT. An output frame on the right, which (among other things, see below for more information) presents the results obtained from the cover-and-generate constructor (Section 4.1) for the F-CIT formulation given in the middle frame.

TABLE 22: The exit survey used in the user studies. All the questions, except for the last two, were Likert scale questions. Questions 1-2 and 6-8 had the following answer options: *1 - strongly disagree*, *2 - disagree*, *3 - neutral*, *4 - agree*, and *5 - strongly agree*. And, questions 3-5 had the following answer options: *1 - very difficult*, *2 - difficult*, *3 - normal*, *4 - easy*, and *5 - very easy*. The last three questions (5-7) were open-ended questions.

no	question
Q1	I understand the following concepts:
a	constraints
b	satisfiability
c	unsatisfiability
Q2	I understand the following concepts:
a	F-CIT models
b	F-CIT entities
c	F-CIT test cases
d	F-CIT objects
Q3	For problem 1 – Difficulty of encoding:
a	F-CIT model
b	F-CIT entities
Q4	For problem 2 – Difficulty of encoding:
a	F-CIT model
b	F-CIT entities
Q5	For problem 3 – Difficulty of encoding:
a	F-CIT model
b	F-CIT entities
Q6	I found F-CIT useful.
Q7	I would use F-CIT in projects.
Q8	I would recommend F-CIT to others.
Q9	What was the most challenging part in the study?
Q10	Any suggestions to improve F-CIT?

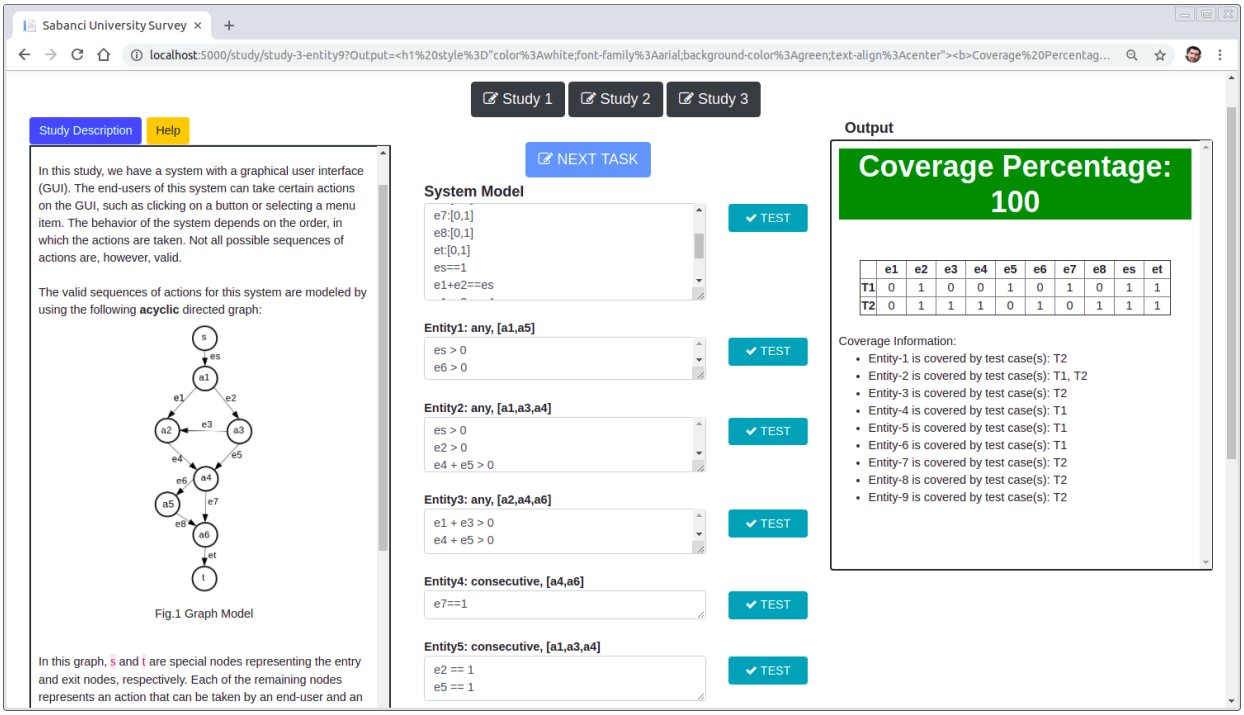


Fig. 8: A screen dump of the tool we have developed for the user studies.

The middle frame had a multi-line text field (model field) to express the F-CIT model and a multi-line text field (entity field) for expressing each entity to be covered. Each field had a "Test" button associated with it. When the Test button of the model field was clicked, the constraints entered for this field were fed to a constraint solver and the result was displayed in the output frame, allowing the participant to check whether the F-CIT model is capable of generating valid test cases. When the Test button of an entity field was clicked, on the other hand, the constraints entered in the respective entity field and those in the model field were combined and fed to the constraint solver. The result was then displayed in the output frame, allowing the participant to check whether the entity can be covered in a valid test case. In both cases, when the constraints were satisfiable, the output frame presented a solution where each parameter defined in the F-CIT model took on a valid value. Otherwise, a warning message indicating that the constraints were not satisfiable, was emitted.

In addition to the Test buttons, we also had a "Generate" button, which fed all the constraints entered (the ones entered in the model and entity fields) to the cover-and-generate constructor (Section 4.1) to compute an F-CIT object. When an F-CIT object was created (which is, indeed, a set of test cases), it was displayed in the output frame in the form of a table, where rows represented the test cases generated and columns depicted the parameters defined in the F-CIT model (Figure 8). Furthermore, the entities covered by each test case are reported.

After completing all the studies, participants filled out an exit survey. Table 22 presents this survey.

6.2 Evaluation Framework

To evaluate the proposed approach, we first counted the number of successful formulations. For a given problem, we define a *successful formulation* as a formulation where both the F-CIT model and the entities to be covered are correctly expressed in a *generalizable manner*, such that an F-CIT object obtaining full coverage can be computed. Note that we also take the generalizability of the formulation into account because we observed that (solely for the second problem) some participants came up with formulations that are too specific for the problem instance given in the study and that, therefore, are non-trivial to generalize for other instances of the same problem. These formulations were often obtained by introducing additional constraints in the F-CIT model in an ad hoc manner just to avoid some undesirable results. More discussion on this can be found in Section 6.3.

We also measured the time it took for the participants to complete the study. More specifically, for a given problem, we measured the *completion time* as the difference between the time the description of the problem was presented to the participant and the time the participant completed the study. Note that a study was completed whenever the participant chose to finish the study. In all but two cases, this

TABLE 23: Demographic information about the participants categorized based on their performances in addressing the second problem.

		degree level		knowledgeable of CIT		experience (in years) in programming		experience (in years) in software testing	
		cat.	count	cat.	count	cat.	count	cat.	count
successful formulations	after seeing only the description	undergrad.	2	yes	4	$\leq \mathbf{2}$	2	$\leq \mathbf{2}$	6
						3	1	3	0
		grad.	4	no	2	4	1	4	0
						$\geq \mathbf{5}$	2	$\geq \mathbf{5}$	0
	after seeing both the description and example	undergrad.	2	yes	1	$\leq \mathbf{2}$	0	$\leq \mathbf{2}$	4
						3	1	3	0
		grad.	2	no	3	4	2	4	0
						$\geq \mathbf{5}$	1	$\geq \mathbf{5}$	0
formulations with issues	missing constraints	undergrad.	0	yes	2	$\leq \mathbf{2}$	0	$\leq \mathbf{2}$	2
						3	1	3	0
		grad.	3	no	1	4	1	4	0
						$\geq \mathbf{5}$	1	$\geq \mathbf{5}$	1
	non-trivial generalization	undergrad.	3	yes	2	$\leq \mathbf{2}$	0	$\leq \mathbf{2}$	5
						3	2	3	0
		grad.	2	no	3	4	3	4	0
						$\geq \mathbf{5}$	0	$\geq \mathbf{5}$	0
give-ups		undergrad.	0	yes	2	$\leq \mathbf{2}$	0	$\leq \mathbf{2}$	1
						3	0	3	1
		grad.	2	no	0	4	0	4	0
						$\geq \mathbf{5}$	2	$\geq \mathbf{5}$	0

happened after computing an F-CIT object achieving full coverage. In two cases, however, the participants chose to stop working on the current problem in the middle of the study as they found the problem "very difficult" (see Section 6.3 for more information). Each participant worked on the problems one after another.

We, furthermore, counted the *number of errors* made by the participants. To this end, we counted the number of times the Test and the Generate buttons were clicked (Section 6.1) and the result obtained did not meet the expectation of the participant. More specifically, if a participant, after clicking on a Test button or a Generate button, made some changes and clicked on the same button, we assumed that the participant made an error before the first click (as the expectation of the participant after the first click did not seem to be met). Note that this metric provides an approximation of the number of errors made because on numerous occasions, we observed that the participants intentionally developed incorrect or missing constraints to test their hypotheses or to gain insight into the problem. We still opted to use this metric because attempting to figure out the actual intention of the participant after every click of a button would have introduced a great deal of intervention.

Furthermore, the percentage of the participants, who "agreed" or "strongly agreed" with a question group in the exit survey, was computed as the average percentage of the participants, who "agreed" or "strongly agreed" with the questions in the group. The percentage of the participants, who found the problems "difficult" or "very difficult," is computed in the same manner.

6.3 Data and Analysis

We first observed that all of the participants understood how F-CIT works. In particular, none of the participants were reminded of the basic F-CIT concepts during the study.

We then observed that the participants could also formulate previously unseen problems in F-CIT. More specifically, all of the participants successfully formulated the first and the third problems. That is, for each of these problems, all of the participants correctly expressed the F-CIT model as well as all of the F-CIT entities to be covered in a generalizable manner and generated an F-CIT object obtaining full coverage. And, they did so in a relatively fast manner. The average time it took for the participants to complete these studies was 4.88 minutes ($min = 1.47$ and $max = 13.89$) for the first problem and 5.38 minutes ($min = 2.29$ and $max = 9.02$) for the third problem.

As expected, the participants found the second problem more difficult than the other two problems, which was also reflected on the outcomes of the exit survey. While 65% of the participants found the second problem "difficult" or "very difficult," none of the participants thought the same thing for the first and third problems. As a matter of the fact, based on the answers given to the open-ended Q9 (Table 22), the most challenging part in the entire study turned out to be expressing an F-CIT model for the second problem. Two participants, indeed, chose to terminate this study in the middle of it after spending 42.56 minutes on average as they found the problem "very difficult" (see the row marked with "give-ups" in Table 23 for the demographic information of these participants).

Half (10 out of 20) of the participants, however, successfully formulated the problem in a generalizable manner by using the same (or similar) approach introduced in Section 5.2.4 and obtained full coverage. 6 of them did so after seeing the description in Figure 6a and 4 after seeing both the description and the example flow problem in Figure 6b (see the rows marked with "after seeing only the description" and "after seeing both the description and example" in Table 23, respectively, for the demographic information of these participants). None of the participants, who came up with a generalizable solution for this problem, did so without seeing the description or the example. The average completion time was 47.14 minutes ($min = 25.79$ and $max = 69.49$).

The remaining 40% (8 out of 20) of the participants, although generated F-CIT objects obtaining full coverage, either came up with a formulation, the generalization of which was non-trivial, or covered some of the entities by chance. More specifically, 3 participants developed generalizable formulations by representing edges using Boolean variables (rather than using integer variables), which were quite similar to the formulation we developed in Section 5.2.4. However, the constraints, which should have invalidated the presence of multiple independent flows, were missing from the F-CIT models (see the row marked with "missing constraints" in Table 23 for the demographic information of these participants). The participants failed to identify the issue because their formulations happened to obtain full coverage by generating valid test cases for the graph given in the study. Had they worked on larger graphs, however, they might have pinpointed and fixed the issue. The average completion time for this category of participants was 37.36 minutes ($min = 33.91$ and $max = 43.01$).

The remaining 5 participants developed formulations, the generalizations of which were non-trivial (see the row marked with "non-trivial generalization" in Table 23 for the demographic information of these participants). In particular, all of these participants chose to represent each vertex (rather than each edge) in the graph by using an integer variable, the value of which represents the order in which the vertex is visited. For a given variable, the set of possible values were determined manually by considering all possible paths that could be traversed. The invalid combinations of variable values (i.e., invalid paths) were then prevented by introducing model constraints in a rather ad hoc manner every time the participant observed that some of the generated test cases were invalid and/or some of the entities could not be covered. 4 (out of 5) of these participants correctly expressed all the constraints as well as the entities for the graph given in the study and obtained full coverage. The remaining participant, although had some missing and/or faulty constraints in the F-CIT model, happened to obtain full coverage by chance. The average completion time for this category of participants was 46.42 minutes ($min = 23.26$ and $max = 64.54$).

We did not find any correlations between the performances of the participants and their levels of degree, knowledge of CIT, or experiences in programming and testing. We, however, observed that the knowledge of the domain was influential in successfully completing the studies. More specifically, for the first and third problems, which required basic knowledge of programming and testing, all of the participants successfully formulated the problems in F-CIT. For the second problem, which required basic knowledge of network flow problems, all of the participants, who successfully formulated the problem in F-CIT, did so either after seeing a definition of the network flow problem or after seeing both the definition and an example flow problem. Table 23 provides demographic information about the participants categorized based on their performances in addressing the second problem.

Regarding the mistakes made during the study, we first observed that (as expected) the participants made more mistakes when solving the second problem, compared to the other two problems. While the average number of mistakes made for the second problem was 5.06, those for the first and third problems were 0.50 and 0.67, respectively.

We then observed a debugging pattern. The participants, solely for the second problem, found expressing the entities as constraints easier than expressing the F-CIT models. They, therefore, used the entity constraints to debug the models. More specifically, to gain insight as well as to test their hypotheses, they tended to click on the Test buttons associated with the entity fields. When the results obtained were not expected, they modified and fixed the models.

Another interesting observation we made is that more than half of the participants (especially for the first and third problems) encoded the F-CIT models as they were reading the study descriptions. That is, as they discovered new system constraints (e.g., parameters and their domains), they updated the models, suggesting that they knew what to look for in the requirements to develop the F-CIT models.

Last but not least, Table 24 presents the outcome of the exit survey. Regarding the questions Q1-Q2 (i.e., Q1.a-Q1.c and Q2.a-Q2.d) and Q6-Q8, all of the participants "agreed" or "strongly agreed" that 1) they understood the basic concepts both in constraint solving and in F-CIT, 2) they found F-CIT useful, 3) they would use F-CIT in a project, and 4) they would recommend F-CIT to others.

Regarding Q10, one suggestion was to improve the syntax of the language we used for expressing the

TABLE 24: Responses to the exit survey given in Table 22.

	strongly disagree	disagree	neutral	agree	strongly agree
Q1a	0	0	0	2	18
Q1b	0	0	0	3	17
Q1c	0	0	0	4	16
Q2a	0	0	0	3	17
Q2b	0	0	0	6	14
Q2c	0	0	0	4	16
Q2d	0	0	0	7	13
Q6	0	0	0	7	13
Q7	0	0	0	7	13
Q8	0	0	0	4	16
	very difficult	difficult	normal	easy	very easy
Q3a	0	0	1	3	16
Q3b	0	0	1	4	15
Q4a	1	12	7	0	0
Q4b	1	6	9	4	0
Q5a	0	0	0	2	18
Q5b	0	0	0	2	18

constraints in a way that closely resembles the Boolean expressions used in main stream programming languages. Another suggestion was to develop a means of expressing "long and repetitive" constraints in a more efficient and effective manner, which, in turn, can further simplify the processes of developing the constraints.

6.4 Discussion

We observed that some problems are more difficult to formulate in F-CIT than others. This is, indeed, to be expected. After all, solving some problems may require specific background knowledge and not everybody may possess it. Note, however, that the proposed approach still allows experts to formulate such problems in F-CIT and others to use the existing formulations to compute F-CIT objects. For example, the F-CIT formulation we developed to express reachability in graph-based models, can be used to obtain full coverage under other reachability-based coverage criteria (other than the ones we studied in this work) by changing the entities to be covered.

7 General Discussion

In this section, we informally discuss the proposed approach in an attempt to 1) address the additional questions that the reader may have, 2) discuss the big picture, in which we envision F-CIT to be an integral part, and 3) present possible avenues for future research. To further support the discussion, we also provide simple examples and code segments as we see fit.

F-CIT is not a methodology for choosing the entities to be tested. It rather takes as input a set of testable entities and aims to find a minimum number of test cases, such that every required entity is covered by at least one test case.

In the absence of a methodology or a tool that can automatically determine what needs to be tested, such as the existing structural code coverage criterion we used in Section 5.1, identifying the set of entities to be tested may require some effort. Note, however, that if the entities at question should really be tested, then they, in one way or another, must be defined and enumerated regardless of whether F-CIT or a specialized constructor is used.

Once the entities are determined, one may consider developing a specialized CIT constructor. To do that, however, a procedure, which determines whether a given set of entities can be covered together in a test case or not, needs to be devised. But, then, the very same procedure can be used as the "solver" in F-CIT, which in turn offers a constructor for free. Note that, as we have discussed in Section 4.5, given such a procedure, the entities can be represented in any form desired (e.g., not necessarily in formal logic), since F-CIT does not need to interpret them.

After all, developing specialized CIT constructors may not be easy. As a matter of fact, we introduced our generate-and-cover constructor (Algorithm 2) to mimic one of the simplest ways of generating CIT objects: Keep on randomly generating valid test cases until all the required entities have been covered. However, developing a high-performing specialized constructor is quite challenging, which is also apparent from more than 50 papers published in the literature, the sole purpose of which is to compute standard covering arrays [2]. Therefore, our goal is to generalize the construction as much as possible, so that the collective effort spent for developing F-CIT constructors can be leveraged in a wider spectrum of test scenarios, which in turn increases the flexibility of CIT.

We believe that the performance (i.e., the construction times) of F-CIT constructors can further be improved by using "hints." The idea behind using hints stems from a simple observation of ours: Testable entities to be covered are typically composed of the same set of sub-entities, e.g., the same conjuncts appear in multiple testable entities. Therefore, in the processes of computing F-CIT objects, the same constraints are often solved multiple times. Consequently, capturing the relationships between these recurring constraints (i.e., sub-entities) in the form of hints can improve the efficiency of F-CIT constructors by reducing the number of times the solver is called and/or by calling the solver with simpler constraints. For example, if it is known that the 2-orders $[v_1, v_2]$ and $[v_2, v_3]$ are covered by a test case for a DAG, then it can be inferred without even using a solver that the 2-order $[v_1, v_3]$ as well as the 3-order $[v_1, v_2, v_3]$ are also covered by the same test case. We have indeed been working on expressing hints

in a way (e.g., as constraints), such that they can be leveraged regardless of the F-CIT constructor in use. Although initial feasibility studies show promise, the aforementioned work is still at its early stages.

We also believe that tools (e.g., front-ends) that can provide various means for defining the coverage criteria as well as the input spaces, such that the testable entities selected by the coverage criteria are automatically generated, can be of further practical use. The important point to note, however, is that regardless of the way these tools operate, the cov-CSP problem, thus, the F-CIT constructors we have developed in this work, will stay intact.

To this end, we make two observations. First, declarative modeling approaches, such as the Answer Set Programming (ASP) approach we used in Section 5.2, is a good fit for F-CIT, as they express the logic of a computation without describing its control flow. Second, these approaches can further leverage the commonalities in the coverage criteria and the test spaces that may be present in different application domains. For example, the test spaces we have used in all of our three studies (Sections 5.1-5.3) can indeed be expressed by using a single DAG-based model. In such a model, a node can be associated with a possibly empty set of configuration options, each of which takes its value from a discrete domain. And, an edge can be associated with a condition, which needs to be satisfied before the edge can be taken.

With regard to Study 2. We, indeed, used such DAG-based models in Study 2 (Section 5.2). Since the only thing that mattered in that study was the possible orderings of nodes, no configuration option was associated with the nodes and all the edge conditions were `true`.

With regard to Study 1. The virtual options in Study 1 (Section 5.1) can also be modeled by using DAGs. More specifically, each virtual option can be expressed by using the control flow graph (CFG) of the respective if-then-else directive, in which an edge originating from a decision node is associated with a condition that specifies the respective outcome of the decision. For example, one of the edges that originates from the decision `if (c) then{...} else{...}` would be associated with `c`, representing the `then` branch, and the other edge would be associated with `¬c`, representing the `else` branch. Then, covering the virtual settings of a virtual option is the same as making sure that all the edges originating from the decision nodes in the CFG are covered. That is, each edge originating from a decision node corresponds to a virtual setting, which is expressed as the conjunction of all the conditions on the path from the entry node to the destination node of the edge.

To demonstrate the proposed approach, we have extended the ASP encoding given in Figure 5. Figure 9 presents the extended encoding. The first thing we did was to add another parameter to `edge(..)` to express the conditions associated with edges. More specifically, a fact `edge(A, B, Condition)` now indicates that there is an edge from `A` to `B`, which can only be taken if the condition `Condition` holds. Note that the encoding in Figure 5 can be made compatible with the new encoding by replacing all the `edge(A, B)` facts by `edge(A, B, true)`, indicating that all edge conditions are `true`.

```
% graph
edge(v0, v1, c1).
edge(v1, v2, c2).
%...

visit(v0, true).
visit(A, Condition) :- visit(X, Guard),
       edge(X, A, Decision),
       Condition=@conjunct(Guard, Decision).

decisionNode(A) :- edge(A, X, _),
       edge(A, Y, _),
       X != Y.

setting(Setting) :- decisionNode(A),
       visit(A, Guard),
       edge(A, B, Decision),
       Setting=@conjunct(Guard, Decision).
```

Fig. 9: An ASP encoding for determining the virtual settings of a virtual option expressed as a control flow graph, in which the edges originating from a decision are associated with the conditions required to take these edges.

Given the control flow graph of a virtual option, in which the entry node is `v0` and the unconditional edges are associated with `true`, a decision node `A` is a node, from which at least two edges originate (i.e., a node with an out-degree of at least 2):

```
decisionNode(A) :- edge(A, X, _),
  edge(A, Y, _),
  X!=Y.
```

The condition `Condition` required to visit node `A` starting from `v0` is the conjunct of the condition `Guard`, which is required to visit the preceding node `X`, and the condition `Decision`, which is associated with the edge from `X` to `A`:

```
visit(v0, true).
visit(A, Condition) :- visit(X, Guard),
  edge(X, A, Decision),
  Condition=@conjunct(Guard, Decision).
```

For every edge originating from a decision node, the condition required to take the edge then becomes a virtual setting:

```
setting(Setting) :- decisionNode(A),
  visit(A, Guard),
  edge(A, B, Decision),
```

```
% graph
node(v0, o1, 0).
node(v0, o1, 1).
node(v0, o2, 0).
node(v0, o2, 1).
%...

tuple(O1, S1, O2, S2)  :- node(v0, O1, S1),
                          node(v0, O2, S2),
                          O1 < O2.
```

Fig. 10: An ASP encoding enumerating all 2-tuples.

```
Setting=@conjunct(Guard, Decision).
```

Note that this encoding is not concerned with solving the conditions, which will indeed be carried out by the F-CIT constructor. It rather forms the virtual settings by joining the individual edge conditions, each of which is expressed as a Boolean expression. Therefore, the edge conditions can be as complex as possible and be processed in any format desired, e.g., as strings. Furthermore, since virtual settings will be Boolean expressions by construction, they can directly be used by our F-CIT constructors without any further processing.

With regard to Study 3. Last but not least, the space of all valid configurations (thus, the space of all valid t-tuples) in Study 3 can also be expressed by using the same DAG-based modeling approach. To this end, all we need is a single node, with which all the configuration options of the subject application under test are associated. Since all possible combinations of option settings were valid in Study 3, there is no need to leverage the reachability constraints that can be provided by a graph-based model. The space of all valid t-tuples can then be expressed as all possible combinations of settings for t distinct options.

To demonstrate the proposed approach, we have extended the ASP encoding given in Figure 5, such that each node can be associated with a set of configuration options, each of which can take its value from a discrete domain. All the configuration options of the system under test can then be associated with a single node, e.g., the entry node `v0`.

Figure 10 presents an example encoding for enumerating 2-tuples. In this encoding, the node `v0` has two binary options, namely `o1` and `o2`, each of which can take on either 0 or 1. All possible 2-tuples for this model can then be enumerated by using:

```
tuple(O1,S1,O2,S2)  :- node(v0,O1,S1),
 node(v0,O2,S2),
 O1<O2.
```

Note that the constraint `O1 < O2` above is used to put the configuration options in an order, so that only distinct tuples are generated. The tuple weights can then be computed as described in Section 5.3.

Note further that both of the encodings given in Figures 9 and 10 can trivially be extended for different coverage strengths.

8 Threats to Validity

All empirical studies suffer from threats to their internal and external validity. For this work, we were primarily concerned with threats to external validity since they limit our ability to generalize the results of our studies to industrial practice.

One threat concerns the representativeness of the case studies as well as the subject applications used in the experiments. To alleviate this issue, we addressed a different CIT problem in each case study.

In the first study (Section 5.1), we enhanced standard CIT with a well-known structural code coverage metric, namely decision coverage, and conducted comparative studies on 12 well-known software systems, including `Python`, `vim`, and `xterm`. In the second study (Section 5.2), we enhanced a number of existing order-based coverage criteria. In the third study (Section 5.3), we developed solutions for a problem faced by a successful consumer electronics company and evaluated the proposed approach by using the data collected from the field. Furthermore, not only the CIT problems we have addressed, but also the solution approaches we have developed were diverse. In the first study (Section 5.1), the values of the parameters were Boolean constraints, rather than discrete values, and we used F-CIT with a SAT solver. In the second study (Section 5.2), we expressed the reachability problem in DAGs as a constraint satisfaction problem (CSP) and used a CSP solver. In the third study (Section 5.3), we worked with parameters, each of which takes on a value from a discrete set of values and used F-CIT with a simple, application-specific constraint solver.

We have, however, not directly studied the fault-detection abilities of the F-CIT objects we computed. In the first study (Section 5.1), we developed F-CIT objects to obtain full decision coverage. The decision coverage criterion is, indeed, a well-known structural code coverage criterion for measuring the adequacy of test suites. Therefore, its fault-detection abilities are well-studied [58]. In the second study (Section 5.2), we enhanced a number of existing order-based coverage criteria, which have already been shown to be effective in testing event-driven systems [9]. In the third study (Section 5.3), we developed usage-based CIT objects to reduce the size of the interaction test suites by covering only the t-tuples (or a fraction of them) seen in the field. Consequently, for the test scenarios, in which standard covering arrays are infeasible (or undesirable) due to their sizes, usage-based CIT objects would offer the same (or similar) fault revealing abilities for the faults caused by the t-tuples seen in the field.

The number of times we repeated the experiments in the paper varied depending on the cost of the respective experiments. We, however, opted to work on larger CIT problems with smaller repetition counts, rather than working on smaller formulations with larger repetition counts. Furthermore, for each study, we have added a discussion for the costs involved in the study. The actual costs, however, may vary depending on the experience of the tester.

Regarding the user studies, all the participants in these studies were students. We, however, had both undergraduate- and graduate-level students with some background on software testing. Furthermore, more than half of these students had taken at least one course where standard CIT approaches were studied. A related threat concerns the representativeness of the problems used in the user studies. We, however, used the smaller instances of the very same problems we studied in this work (Sections 5.1-5.3). Furthermore, these problems were not known to the participants before taking part in the study. The participants were asked to finish working on one problem before moving to the subsequent problem. They were also required to finish all the studies in a single session. Had they been given more time and/or more instances of the same problems, more participants might have successfully formulated them in a generalizable manner and/or identified and fixed the issues with their formulations.

9 Related Work

The basic justification for using standard t-way covering arrays is that they can (under certain assumptions) reveal the failures caused by the interactions of t or fewer parameters. As a matter of fact, the results of many empirical studies suggest that a majority of parameter-related failures in practice are often caused by the interactions of only a small number of parameters [59], [2], [60], [61]. That is, t is generally much smaller than the number of parameters, typically $2 \leq t \leq 6$ with t=2 (i.e., pairwise testing) being the most common case [3]. Furthermore, for a fixed t, as the input space grows (e.g., as the number of parameters increases), the size of a covering array represents an increasingly smaller proportion of the whole input space. Thus, covering arrays can efficiently handle very large input spaces.

Computing standard covering arrays is an NP-hard problem [62]. Standard covering array constructors can be divided into four broad categories based on the construction approach they employ: random search-based constructors [60], [63], greedy constructors [3], [64], [65], metaheuristic search-based constructors [66], [67], [68], [69], [70], [71], and mathematical constructors [72], [73]. Our work is different in that, while these constructors compute standard covering arrays, the constructors we have presented in this paper compute F-CIT objects, of which the standard covering arrays are a special instance, by solving the cov-CSP problem.

Constraints in combinatorial interaction testing have also been of practical [15] and theoretical interest [23]. Based on their strictness, CIT constraints can be divided into *hard constraints* and *soft constraints*. Hard constraints mark combinations of option settings that are not permitted [17]. Soft constraints, on the other hand, mark combinations of option settings that are permitted, but not desirable [16]. In this work, we define two types of constraints: model constraints and constraints representing entities to be covered. The former types of constraints are system-wide hard constraints, whereas the latter types are hard constraints, the scope of which is limited to the test case covering the entities, i.e., only the test case that covers an entity must satisfy the respective constraint.

A relatively recent work [24] discovers the GUI widgets and Android Activities that interact with each other, by using static analysis and uses this information together with a constraint solver to reduce the number of combinations (thus the number of test cases) required for testing mobile applications. Our work is different in that the aforementioned work uses constraint solving (in a flexible manner by using Alloy [74]) solely to enumerate the testable entities to be covered, which in this context correspond to sequences of Android Activities that can be visited (e.g., prime paths in the activity transition graph of the application under test [24]) and/or the combinations of values that can be tested for the interacting GUI widgets. Once these entities are enumerated, the actual test cases are generated by first computing the Cartesian product of the values for the interacting GUI widgets in each Activity and then by computing the Cartesian product of the test cases generated for the interacting Activities. F-CIT, on the other hand, takes the entities to be covered as input and uses constraint solving for generating test cases to cover all the required entities. And, it does so by generalizing the construction of CIT objects in the form of the cov-CSP problem. From this perspective, F-CIT complements the aforementioned work (rather than replacing it) because once the testable entities are enumerated in [24], F-CIT can be used to generate the test cases. Furthermore, the aforementioned work [24] does not concern about generalizing the construction of CIT objects, which is, indeed, the main concern addressed in our work. For example, it is not clear at all whether the construction approach in [24] can be used to compute the different types of F-CIT objects introduced in this work (Sections 5.1-5.3).

Seeding has also been frequently used in combinatorial interaction testing [2]. Some example uses can be summarized as follows: 1) to guarantee the inclusion of certain configurations by having them in the seed; 2) to reduce the cost of testing by including already

tested configurations in the seed; and 3) for incremental construction of covering arrays by using lower strength covering arrays as seeds to compute higher strength covering arrays [30], [75]. In this work, we have developed a seeding mechanism for F-CIT and used it in two different ways: 1) to combine multiple coverage criteria and 2) to incrementally construct F-CIT objects.

10 Concluding Remarks

In this work, we have first presented F-CIT to make combinatorial interaction testing more flexible. In F-CIT, both the testable entities to be covered and the space of test cases, from which the samples will be drawn, are expressed as constraints. Consequently, the problem of computing F-CIT objects, turns into an interesting constraint solving problem, which we call *cov-CSP*. Given a set of constraints, each representing a testable entity to be covered, cov-CSP aims to divide the set into a minimum number of satisfiable clusters, such that a solution for a cluster represents a test case, covering the testable entities included in the cluster. The collection of all the test cases computed for the clusters constitute the F-CIT object, covering each required testable entity at least once.

We have then developed two constructors, namely cover-and-generate and generate-and-cover, to solve the cov-CSP problem, thus to compute F-CIT objects. These constructors can work with any types of constraints as long as an appropriate solver, the purpose of which is to determine whether a given set of entities can be covered in a single test case or not, is provided.

To evaluate F-CIT, we have first carried out three case studies, each of which focused on a different CIT problem, demonstrating that F-CIT is more flexible than the existing CIT approaches. We have arrived at this conclusion by noting that, in these studies, it was either unclear how to use the existing constructors (if at all possible) to compute the requested CIT objects, or the existing constructors required non-trivial modifications or excessive number of test cases to guarantee full coverage. F-CIT, on the other hand, used the same F-CIT constructor to compute all the requested CIT objects these studies without requiring any modifications.

We have also carried out user studies to further evaluate F-CIT, demonstrating the usability of the proposed approach. One thing we observed in these studies is that some problems are more difficult to formulate in F-CIT than others as they require some specific background knowledge, which may not be possessed by everybody. To alleviate these issues, we are developing declarative modeling-based front-ends to flexibly define both the coverage criteria to be used and the test spaces, from which the samples are drawn.

Another avenue for future work is to develop alternative approaches for solving the cov-CSP problem, i.e., developing better constructors. Yet another avenue is to capture the domain knowledge in the form of "hints", which are also expressed as constraints, to develop even better constructors. We are also interested in demonstrating the flexibility of F-CIT in different application domains.

11 Acknowledgments

This research was supported by the Scientific and Technological Research Council of Turkey (118E204).

References

[1] C. Yilmaz, S. Fouche, M. B. Cohen, A. Porter, G. Demiroz, and U. Koc, "Moving forward with combinatorial interaction testing," *Computer*, vol. 47, no. 2, pp. 37–45, 2014.

[2] C. Nie and H. Leung, "A survey of combinatorial testing," *ACM Computing Surveys (CSUR)*, vol. 43, no. 2, p. 11, 2011.

[3] D. M. Cohen, S. R. Dalal, M. L. Fredman, and G. C. Patton, "The AETG system: an approach to testing based on combinatorial design," *IEEE Trans. on Soft. Eng.*, vol. 23, no. 7, pp. 437–44, 1997.

[4] A. W. Williams and R. L. Probert, "A practical strategy for testing pair-wise coverage of network interfaces," in *Proceedings of Seventh International Symposium on Software Reliability Engineering*. IEEE, 1996, pp. 246–254.

[5] P. J. Schroeder, P. Faherty, and B. Korel, "Generating expected results for automated black-box testing," in *In Proceedings of the 17th IEEE International Conference on Automated Software Engineering, ASE 2002*. IEEE, 2002, pp. 139–148.

[6] C. Yilmaz, M. B. Cohen, and A. A. Porter, "Covering arrays for efficient fault characterization in complex configuration spaces," *IEEE Trans. Software Eng.*, vol. 32, no. 1, pp. 20–34, 2006.

[7] M. F. Johansen, Ø. Haugen, and F. Fleurey, "An algorithm for generating t-wise covering arrays from large feature models," in *Proceedings of the 16th International Software Product Line Conference-Volume 1*. ACM, 2012, pp. 46–55.

[8] Y. Lei, R. H. Carver, R. Kacker, and D. Kung, "A combinatorial testing strategy for concurrent programs," *Software Testing, Verification and Reliability*, vol. 17, no. 4, pp. 207–225, 2007.

[9] X. Yuan, M. B. Cohen, and A. M. Memon, "GUI interaction testing: Incorporating event context," *IEEE Transactions on Software Engineering*, vol. 37, no. 4, pp. 559–574, 2011.

[10] C. Yilmaz, "Test case-aware combinatorial interaction testing," *IEEE Transactions on Software Engineering*, vol. 39, no. 5, pp. 684–706, 2013.

[11] G. Demiroz and C. Yilmaz, "Cost-aware combinatorial interaction testing," in *Proceedings of the Internatinoal Conference on Advances in System Testing and Validation Lifecycles*, 2012, pp. 9–16.

[12] A. Javeed and C. Yilmaz, "Combinatorial interaction testing of tangled configuration options," in *2015 IEEE Eight International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2015, pp. 1–4.

[13] J. Lawrence, R. N. Kacker, Y. Lei, D. R. Kuhn, and M. Forbes, "A survey of binary covering arrays," *the electronic journal of combinatorics*, vol. 18, no. 1, p. P84, 2011.

[14] E. Dumlu, C. Yilmaz, M. B. Cohen, and A. Porter, "Feedback driven adaptive combinatorial testing," in *Proceedings of the International Symposium on Software Testing and Analysis*. ACM, 2011, pp. 243–253.

[15] M. B. Cohen, M. B. Dwyer, and J. Shi, "Interaction testing of highly-configurable systems in the presence of constraints," in *Proceedings of the 2007 International Symposium on Software Testing and Analysis*, 2007, pp. 129–139.

[16] R. C. Bryce and C. J. Colbourn, "Prioritized interaction testing for pair-wise coverage with seeding and constraints," *Information and Software Technology*, vol. 48, no. 10, pp. 960 – 970, 2006, advances in Model-based Testing.

[17] G. Mats, O. Jeff, and M. Jonas, "Handling constraints in the input space when using combination strategies for software testing," University of Skövde, School of Humanities and Informatics, Tech. Rep. HS- IKI -TR-06-001, 2006.

[18] C. Yilmaz, "Test case-aware combinatorial interaction testing," *IEEE Transactions on Software Engineering*, vol. 39, no. 5, pp. 684–706, 2013.

[19] M. Makaś, "A certain version of preservationism," *Logic and Logical Philosophy*, vol. 26, no. 1, pp. 63–77, 2016.

[20] P. K. Schotch and R. E. Jennings, "Inference and necessity," *Journal of Philosophical Logic*, vol. 9, no. 3, pp. 327–340, 1980.

[21] N. Rescher and R. Manor, "On inference from inconsistent premisses," *Theory and decision*, vol. 1, no. 2, pp. 179–217, 1970.

[22] A. Biere, M. Heule, and H. van Maaren, *Handbook of satisfiability*. IOS press, 2009, vol. 185.

[23] P. Danziger, E. Mendelsohn, L. Moura, and B. Stevens, "Covering arrays avoiding forbidden edges," *Theoretical Computer Science*, vol. 410, no. 52, pp. 5403–5414, 2009.

[24] N. Mirzaei, J. Garcia, H. Bagheri, A. Sadeghi, and S. Malek, "Reducing combinatorics in gui testing of android applications," in *Proceedings of the 38th International Conference on Software Engineering*, ser. ICSE '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 559–570.

[25] H. Mercan and C. Yilmaz, "A constraint solving problem towards unified combinatorial interaction testing," in *Proceedings of the 7th Workshop on Constraint Solvers in Testing, Verification, and Analysis*, vol. 1639, no. 5. CEUR, 2016, pp. 24–30.

[26] A. Javeed, "Gray-box combinatorial interaction testing," Master's thesis, 2015.

[27] Y. T. Yu and M. F. Lau, "A comparison of mc/dc, mumcut and several other coverage criteria for logical decisions," *Journal of Systems and Software*, vol. 79, no. 5, pp. 577–590, 2006.

[28] "Radon." [Online]. Available: https://radon.readthedocs.io/en/latest/

[29] L. Yu, Y. Lei, R. N. Kacker, and D. R. Kuhn, "ACTS: A combinatorial test generation tool," in *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 2013, pp. 370–375.

[30] S. Fouché, M. B. Cohen, and A. Porter, "Incremental covering array failure characterization in large configuration spaces," in *Proceedings of the eighteenth international symposium on Software testing and analysis*. ACM, 2009, pp. 177–188.

[31] J. Liebig, S. Apel, C. Lengauer, C. Kästner, and M. Schulze, "An analysis of the variability in forty preprocessor-based software product lines," in *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 2010, pp. 105–114.

[32] "SATisPy." [Online]. Available: https://github.com/netom/satispy

[33] N. Eén and N. Sörensson, "An extensible SAT-solver," in *International conference on theory and applications of satisfiability testing*. Springer, 2003, pp. 502–518.

[34] P. Samuel and A. T. Joseph, "Test sequence generation from uml sequence diagrams," in *Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD'08*. IEEE, 2008, pp. 879–887.

[35] F. Belli, M. Beyazit, and N. Güler, "Event-Based GUI testing and reliability assessment techniques–an experimental insight and preliminary results," in *IEEE Fourth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2011, pp. 212–221.

[36] A. M. Memon, "An event-flow model of GUI-based applications for testing," *Software testing, verification and reliability*, vol. 17, no. 3, pp. 137–157, 2007.

[37] M. Ç. Çalpur, "Interleaving coverage criteria oriented testing of multithreaded applications," Master's thesis, 2012.

[38] D. L. Bruening, "Systematic testing of multithreaded Java programs," Ph.D. dissertation, Massachusetts Institute of Technology, 1999.

[39] M. Musuvathi, S. Qadeer, T. Ball, M. Musuvathi, S. Qadeer, and T. Ball, "Chess: A systematic testing tool for concurrent software," Technical Report MSR-TR-2007-149, Microsoft Research, Tech. Rep., 2007.

[40] S. Lu, W. Jiang, and Y. Zhou, "A study of interleaving coverage criteria," in *the 6th joint meeting on European software engineering conference and the ACM SIGSOFT symposium on the foundations of software engineering: companion papers*. ACM, 2007, pp. 533–536.

[41] D. R. Kuhn, J. M. Higdon, J. F. Lawrence, R. N. Kacker, and Y. Lei, "Combinatorial methods for event sequence testing," in *Proceedings of the 5th IEEE International Conference on Software Testing, Verification and Validation*, 2012, pp. 601–609.

[42] "Apache ActiveMQ." [Online]. Available: http://activemq.apache.org/

[43] A. Biere, A. Cimatti, E. M. Clarke, O. Strichman, Y. Zhu *et al.*, "Bounded model checking." *Advances in computers*, vol. 58, no. 11, pp. 117–148, 2003.

[44] R. Ahlswede, N. Cai, S.-Y. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on information theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[45] J. E. Hopcroft, *Introduction to Automata Theory, Languages and Computation: For VTU, 3/e*. Pearson Education India, 2013.

[46] S. Lombardy, Y. Régis-Gianas, and J. Sakarovitch, "Introducing vaucanson," *Theoretical Computer Science*, vol. 328, no. 1-2, pp. 77–96, 2004.

[47] V. W. Marek and M. Truszczyński, "Stable models and an alternative logic programming paradigm," in *The Logic Programming Paradigm*. Springer, 1999, pp. 375–398.

[48] I. Niemelä, "Logic programs with stable model semantics as a constraint programming paradigm," *Annals of Mathematics and Artificial Intelligence*, vol. 25, no. 3-4, pp. 241–273, 1999.

[49] T. Eiter, G. Ianni, and T. Krennwallner, "Answer set programming: A primer," in *Reasoning Web International Summer School*. Springer, 2009, pp. 40–110.

[50] C. Baral, *Knowledge representation, reasoning and declarative problem solving*. Cambridge university press, 2003.

[51] N. Tamura and M. Banbara, "Sugar: A CSP to SAT translator based on order encoding," *Proceedings of the Second International CSP Solver Competition*, pp. 65–69, 2008.

[52] B. Jenkins, "jenny: A pairwise testing tool," http://www.burtleburtle.net/bob/index.html, 2005.

[53] J. Czerwonka, "Pairwise testing in the real world: Practical extensions to test-case scenarios," *Microsoft Corporation, Software Testing Technical Articles*, 2008.

[54] H. Katebi, K. Sakallah, and J. Marques-Silva, "Empirical study of the anatomy of modern sat solvers," *Theory and Applications of Satisfiability Testing-SAT 2011*, pp. 343–356, 2011.

[55] L. M. de Moura and N. Bjørner, "Satisfiability modulo theories: An appetizer," *SBMF*, vol. 5902, pp. 23–36, 2009.

[56] M. B. Cohen, P. B. Gibbons, W. B. Mugridge, C. J. Colbourn, and J. S. Collofello, "A variable strength interaction testing of components," in *Proceedings of 27th IEEE Annual International Computer Software and Applications Conference, (COMPSAC)*. IEEE, 2003, pp. 413–418.

[57] M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali, *Linear programming and network flows*. John Wiley & Sons, 2011.

[58] X. Cai and M. R. Lyu, "The effect of code coverage on fault detection under different testing profiles," *ACM SIGSOFT software engineering notes*, vol. 30, no. 4, pp. 1–7, 2005.

[59] R. Kuhn, Y. Lei, and R. Kacker, "Practical combinatorial testing: Beyond pairwise," *It Professional*, vol. 10, no. 3, 2008.

[60] P. J. Schroeder, P. Bolaki, and V. Gopu, "Comparing the fault detection effectiveness of n-way and random test suites," in *Proc. of the 2004 Int'l Symp. on Empirical Software Engineering*, 2004, pp. 49–59.

[61] D. R. Kuhn and V. Okum, "Pseudo-exhaustive testing for software," in *30th Annual IEEE/NASA Software Engineering Workshop. SEW'06*. IEEE, 2006, pp. 153–158.

[62] E. Maltais and L. Moura, "Hardness results for covering arrays avoiding forbidden edges and error-locating arrays," *Theoretical Computer Science*, vol. 412, no. 46, pp. 6517–6530, 2011.

[63] R. Huang, X. Xie, T. Y. Chen, and Y. Lu, "Adaptive random test case generation for combinatorial testing," in *Proceedings of 36th IEEE Annual International Computer Software and Applications Conference, (COMPSAC)*. IEEE, 2012, pp. 52–61.

[64] Y. Lei, R. Kacker, D. R. Kuhn, V. Okun, and J. Lawrence, "IPOG/IPOG-D: efficient test generation for multi-way combinatorial testing," *Softw. Test. Verif. Reliab.*, vol. 18, pp. 125–148, September 2008.

[65] Z. Wang and H. He, "Generating variable strength covering array for combinatorial software testing with greedy strategy." *JSW*, vol. 8, no. 12, pp. 3173–3181, 2013.
[66] M. B. Cohen, P. B. Gibbons, W. B. Mugridge, and C. J. Colbourn, "Constructing test suites for interaction testing," in *Proceedings of the 25th International Conference on Software Engineering*, 2003, pp. 38–48.
[67] R. C. Bryce and C. J. Colbourn, "One-test-at-a-time heuristic search for interaction test suites," in *Proceedings of the 9th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2007, pp. 1082–1089.
[68] H. Wu, C. Nie, F.-C. Kuo, H. Leung, and C. J. Colbourn, "A discrete particle swarm optimization for covering array generation," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 4, pp. 575–591, 2015.
[69] P. Galinier, S. Kpodjedo, and G. Antoniol, "A penalty-based Tabu search for constrained covering arrays," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, 2017, pp. 1288–1294.
[70] Y. Jia, M. B. Cohen, M. Harman, and J. Petke, "Learning combinatorial interaction test generation strategies using hyperheuristic search," in *Proceedings of the 37th IEEE International Conference on Software Engineering (ICSE)*, vol. 1. IEEE, 2015, pp. 540–550.
[71] J. Torres-Jimenez and E. Rodriguez-Tello, "New bounds for binary covering arrays using simulated annealing," *Information Sciences*, vol. 185, no. 1, pp. 137–152, 2012.
[72] N. Kobayashi, "Design and evaluation of automatic test generation strategies for functional testing of software," Ph.D. dissertation, Osaka University, Osaka, Japan, 2002.
[73] C. J. Colbourn, "Combinatorial aspects of covering arrays," *Le Matematiche (Catania)*, vol. 58, no. 121-167, pp. 0–10, 2004.
[74] D. Jackson, "Alloy: a lightweight object modelling notation," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 11, no. 2, pp. 256–290, 2002.
[75] C. Yilmaz, E. Dumlu, M. B. Cohen, and A. Porter, "Reducing masking effects in combinatorial interaction testing: A feedback driven adaptive approach," *IEEE Transactions on Software Engineering*, vol. 40, no. 1, pp. 43–66, 2014.

Hanefi Mercan received the BS degree in mathematics and the MS degree in computer science in 2012 and 2015, respectively. He is currently a PhD student in Sabanci University. His research interests include software testing and parallel computing.

Arsalan Javeed received BS degree in Telecommunication Engineering and MS in Computer Science and Engineering in 2011 and 2015 respectively. He is currently a PhD student at Sabanci University, Turkey. His research interests include Software Engineering, Testing and Security.

Cemal Yilmaz received the BS and MS degrees in computer engineering and information science from Bilkent University, Ankara, Turkey, in 1997 and 1999, respectively. In 2005, he received the PhD degree in computer science from the University of Maryland at College Park. Between 2005 and 2008, he worked as a post-doctoral researcher at IBM Thomas J. Watson Research Center, Hawthorne, New York. He is currently an assistant professor of computer science in the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey. His current research interests include software engineering and software quality assurance.