

Learning-Guided Fuzzing for Testing Stateful SDN Controllers

RAPHAËL OLLANDO, University of Luxembourg, Luxembourg

SEUNG YEOB SHIN, University of Luxembourg, Luxembourg

LIONEL C. BRIAND*, Lero SFI Centre for Software Research, University of Limerick, Ireland and University of Ottawa, Canada

Controllers for software-defined networks (SDNs) are centralised software components that enable advanced network functionalities, such as dynamic traffic engineering and network virtualisation. However, these functionalities increase the complexity of SDN controllers, making thorough testing crucial. SDN controllers are stateful, interacting with multiple network devices through sequences of control messages. Identifying stateful failures in an SDN controller is challenging due to the infinite possible sequences of control messages, which result in an unbounded number of stateful interactions between the controller and network devices. In this article, we propose SeqFuzzSDN, a learning-guided fuzzing method for testing stateful SDN controllers. SeqFuzzSDN aims to (1) efficiently explore the state space of the SDN controller under test, (2) generate effective and diverse tests (i.e., control message sequences) to uncover failures, and (3) infer accurate failure-inducing models that characterise the message sequences leading to failures. In addition, we compare SeqFuzzSDN with three extensions of state-of-the-art (SOTA) methods for fuzzing SDNs. Our findings show that, compared to the extended SOTA methods, SeqFuzzSDN (1) generates more diverse message sequences that lead to failures within the same time budget, and (2) produces more accurate failure-inducing models, significantly outperforming the other extended SOTA methods in terms of sensitivity.

CCS Concepts: • Networks → Programmable networks; • Software and its engineering → Software testing and debugging.

Additional Key Words and Phrases: Software-Defined Networks, Software Testing, Fuzzing

ACM Reference Format:

Raphaël Ollando, Seung Yeob Shin, and Lionel C. Briand. 2025. Learning-Guided Fuzzing for Testing Stateful SDN Controllers. 1, 1 (May 2025), 46 pages. <https://doi.org/10.1145/mnnnnnn.mnnnnnn>

1 Introduction

Software-defined networks (SDNs) [35], which have been applied in many domains such as data centres [28, 84], satellite communications [32, 54], and the Internet of Things [71, 75], have gained popularity due to the programmability of their controllers, enabling the deployment of network services through software. An SDN controller is a centralised software component in the SDN that enables the implementation of advanced network functionalities, such as dynamic traffic engineering [75] and network virtualisation [11]. However, implementing such functionalities

*Part of this work was done when he was affiliated with the Interdisciplinary Centre for Security, Reliability, and Trust (SnT) of the University of Luxembourg.

Authors' Contact Information: Raphaël Ollando, raphael.ollando@uni.lu, University of Luxembourg, Luxembourg, Luxembourg; Seung Yeob Shin, seungyeob.shin@uni.lu, University of Luxembourg, Luxembourg, Luxembourg; Lionel C. Briand, lionel.briand@lero.ie, Lero SFI Centre for Software Research, University of Limerick, Limerick, Ireland and University of Ottawa, Ottawa, Canada.



Please use nonacm option or ACM Engage class to enable CC license

This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM XXXX-XXXX/2025/5-ART

<https://doi.org/10.1145/mnnnnnn.mnnnnnn>

increases the complexity of SDN controllers. Furthermore, having a centralised controller introduces new attack surfaces (e.g., ARP spoofing for SDN [3, 81]) that can be exploited by malicious users to manipulate SDNs. Hence, testing SDN controllers becomes even more important and poses specific challenges compared to testing traditional networks, which typically lack software controllers and provide static network operations.

An SDN controller is a stateful software component that maintains a holistic view of the SDN, capturing the state information of network devices, links, and the controller itself. This enables the controller to provide dynamic network operations in an efficient and effective manner. However, testing stateful SDN controllers is challenging. An SDN controller interacts with multiple network devices through sequences of inbound and outbound control messages defined in the underlying SDN communication protocol (e.g., OpenFlow [63]). If a failure can occur only in a certain state of an SDN controller, discovering such a stateful failure requires engineers to identify message sequences that bring the controller into that state. However, discovering such stateful failures is a hard problem due to the potentially infinite number of possible sequences of control messages. This is because the size of sequences is unbounded, and there are various types of control messages with different sizes. In addition, even if engineers obtain sequences of control messages that cause failures, manual inspection of these sequences is time-consuming and error-prone. This may result in misunderstandings of the causes of failures and hence the application of unreliable fixes.

Fuzzing techniques have been widely applied for testing various network systems [40, 50, 60, 66]. Among these, state-aware fuzzing techniques that do not depend on protocol specifications could be considered for testing SDN controllers, as, to our knowledge, no existing state-aware fuzzing techniques account for the specificities (e.g., architecture and protocol) of SDNs. For example, AFLNet [66] constructs finite state machines (FSMs) based on the response codes generated by the server under test and uses these FSMs to guide the fuzzing process. AFLNet employs common byte-level fuzz operators, such as bit flipping as well as the insertion, deletion, and substitution of byte blocks. However, AFLNet operates under the working assumption that communication protocols embed special codes in response messages, which is not always the case, as in our SDN context. StateAFL [60] infers FSMs based on the in-memory states of the server, leveraging compile-time instrumentation and fuzzy hashing techniques; hence, it does not require response codes. During the fuzzing process, StateAFL guides the generation of new inputs to the server based on the inferred FSMs. It employs both byte-level and message-level fuzz operators, which do not rely on protocol specifications. NSFuzz [69] uses a combination of static analysis and manual annotation on the server's source code to identify states based on program variables and construct FSMs that capture the transitions between these states. It then performs FSM-guided fuzzing using fuzz operators similar to those in AFLNet. However, the state-aware fuzzing techniques introduced in this research strand are applicable to the server-client architecture by replacing a client with a fuzzer. The fuzzer replays captured message sequences and modifies them during the fuzzing process. In contrast, the SDN architecture differs significantly from the server-client architecture. For example, in the SDN architecture, communication is initiated between an SDN controller and switches, whereas in the server-client architecture, clients typically initiate requests to the server. Additionally, SDN switches also communicate with one another to enable network communication and services. Therefore, replacing an SDN switch with a fuzzer for testing an SDN controller is challenging. Furthermore, the working assumptions of these techniques, such as response codes, compile-time instrumentation, and source-code analysis and annotation, make them difficult to apply when testing an SDN controller. SDN operators are more concerned with potential failures that can occur in realistic scenarios, such as when a malicious user intercepts messages and disrupts the SDN during its operation [26, 40, 50, 58, 62].

There are some prior studies [40, 50, 62] that test SDN controllers by taking into account the architecture and protocols of SDNs. For example, DELTA [50] is a security assessment framework for SDNs. It reproduces existing SDN-related attack scenarios and uncovers new ones through fuzzing. Specifically, in fuzzing, DELTA modifies control messages by treating them as byte streams and randomising them. BEADS [40] is an automated attack discovery tool for SDNs. In contrast to DELTA, BEADS fuzzes control messages while adhering to the SDN protocol (i.e., OpenFlow), aiming to create test scenarios that can exercise components beyond the protocol parsers of SDN controllers. FuzzSDN [62] also adheres to the SDN protocol in its fuzzing process to test components beyond the protocol parsers of an SDN controller. In addition, FuzzSDN employs machine learning techniques to infer failure-inducing models that characterise the conditions under which failures occur, and uses them to guide the fuzzing. These techniques position their fuzzers between the SDN controller and the SDN switches to sniff and modify control messages, leveraging the man-in-the-middle attack strategy [21]. Hence, they do not require any modifications, replacements, annotations, or instrumentation of the components (i.e., switches and controllers) in SDNs, enabling the testing of SDN controllers in a realistic setting. However, these techniques, which account for the architecture and protocols of SDNs, do not consider the stateful nature of SDN controllers.

Contributions. In this article, we propose SeqFuzzSDN, a learning-guided fuzzing method for testing stateful SDN controllers. SeqFuzzSDN aligns with the aforementioned research strand that leverages the architecture and protocols of SDNs. Hence, SeqFuzzSDN tests SDN controllers in a realistic operational setting without requiring any compile-time instrumentation, manual annotation of source code, and replacing an SDN switch with a fuzzer. Instead, SeqFuzzSDN sniffs and fuzzes control messages exchanged between the SDN controller and switches by being aware of the stateful behaviours of the controller. SeqFuzzSDN employs a fuzzing strategy guided by Extended Finite State machines (EFSMs) in order to (1) efficiently explore the space of states of the SDN controller under test, (2) generate effective and diverse tests (i.e., message sequences) to uncover failures, and (3) infer accurate EFSMs that characterise the sequences of control messages leading to failures. Note that since the SDN communication protocol specifies various message fields, their values, and relations, guard conditions on state transitions in EFSMs are well-suited to capture state changes associated with such message fields, values, and relations.

We evaluated SeqFuzzSDN by applying it to two well-known open-source SDN controllers: ONOS [7] and RYU [72]. Additionally, we compared SeqFuzzSDN against our extensions of three state-of-the-art (SOTA) methods—DELTA [50], BEADS [40], and FuzzSDN [62]—which were used as baselines for generating tests for SDN controllers. We extended DELTA, BEADS, and FuzzSDN to produce EFSM models, since these SOTA methods were not originally designed to generate such models. It is important to note that, these three baselines are the best available options for evaluating SeqFuzzSDN when testing SDN controllers by fuzzing control messages. Our experiment results show that SeqFuzzSDN significantly outperforms the three baselines. Specifically, compared to the baselines, SeqFuzzSDN generates more diverse and effective tests (i.e., message sequences) that lead to failures, as well as more accurate EFSMs that characterise failure-inducing message sequences. In addition, SeqFuzzSDN can be applied to large SDNs since its performance is independent of the network size. Our complete evaluation results and the SeqFuzzSDN tool can be accessed online [61].

Organisation. The remainder of this article is structured as follows: Section 2 provides the background and defines the problem of testing stateful SDN controllers. Section 3 details the steps of SeqFuzzSDN. Section 4 presents the empirical evaluation of SeqFuzzSDN. Section 6 compares SeqFuzzSDN with related work. Finally, Section 7 concludes the article.

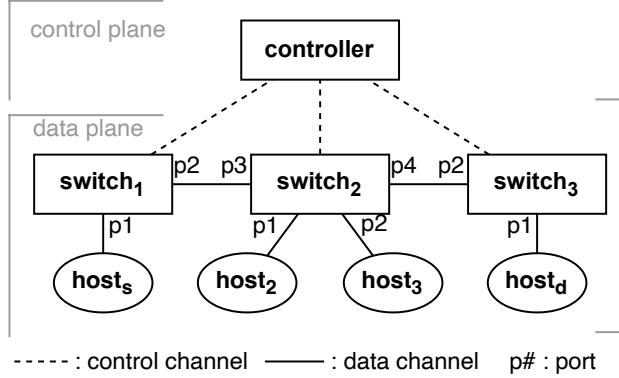


Fig. 1. An SDN topology example.

2 Background and problem description

This section introduces SDN concepts, including the SDN architecture and control messages. We then define the problem of testing stateful SDN controllers.

Architecture. The SDN architecture [35] separates the network into the control plane and the data plane, which is a key distinction from traditional networks that do not possess this separation. In the control plane, an SDN controller provides network administrators with a global view of the network, enabling centralised control over network operations. The centralised control allows administrators to optimally manage network resources and to effectively enforce network policies. Furthermore, an SDN controller provides engineers with APIs [35], enabling them to develop and install custom applications on the controller to address system-specific needs (e.g., dynamic adaptive traffic control application [75]). The data plane consists of SDN switches, which are responsible for forwarding data messages (i.e., data packets) based on the instructions provided by the controller. SDN switches are connected to hosts that generate and receive data messages. Such hosts can be servers, clients, and IoT devices in a networked system. In the SDN architecture, the SDN controller serves as the central software component that enables the provision of flexible and efficient network services.

Figure 1 depicts an SDN topology example that consists of a controller, three switches, and four hosts. The controller communicates with the three switches via control channels that carry control messages (e.g., OpenFlow messages [63]). The switches and hosts, on the other hand, are connected via data channels that carry data messages encapsulated by standard network protocols such as ARP [65, 67] and IP [65, 68].

Message sequences. In the control plane of an SDN, an SDN controller exchanges sequences of control messages with SDN switches to establish and manage communication among hosts, monitor network status, and enforce network policies. In the data plane of an SDN, hosts exchange sequences of data messages through SDN switches to transmit and receive various types of data, such as audio and video streams. For example, Table 1 presents an example sequence of messages aimed at discovering host locations (i.e., MAC addresses [65]) in the SDN network shown in Figure 1.

Regarding the example sequence listed in Table 1, we consider an SDN setup in which the controller in Figure 1 is unaware of a path across switches that enables the transmission of data messages from $host_s$ to $host_d$. The address resolution protocol (ARP) is typically used to map an IP address of a host to its physical (MAC) address [65, 67]. The first ARP message m_1 generated by $host_s$ is an ARP request aimed at obtaining the MAC address of $host_d$. The ARP request reaches

Table 1. An example sequence of messages for discovering host locations. The messages in this table are generated by the hosts, switches, and a controller depicted in Figure 1.

m_i	message	sender	receiver	channel
1	<u>arp_req(host_d)</u>	host _s	switch ₁	data
2	pkt_in(arp_req(host _d))	switch ₁	controller	control
3	pkt_out(arp_req(host _d),flood)	controller	switch ₁	control
4	arp_req(host _d)	switch ₁	switch ₂	data
5	pkt_in(arp_req(host _d))	switch ₂	controller	control
6	pkt_out(arp_req(host _d),flood)	controller	switch ₂	control
7	arp_req(host _d)	switch ₂	host ₂	data
8	arp_req(host _d)	switch ₂	host ₃	data
9	arp_req(host _d)	switch ₂	switch ₃	data
10	pkt_in(arp_req(host _d))	switch ₃	controller	control
11	pkt_out(arp_req(host _d),flood)	controller	switch ₃	control
12	<u>arp_req(host_d)</u>	switch ₃	host _d	data
13	<u>arp_rep(host_d)</u>	host _d	switch ₃	data
14	pkt_in(arp_rep(host _d))	switch ₃	controller	control
15	pkt_out(arp_rep(host _d),port2)	controller	switch ₃	control
16	arp_rep(host _d)	switch ₃	switch ₂	data
17	pkt_in(arp_rep(host _d))	switch ₂	controller	control
18	pkt_out(arp_rep(host _d),port3)	controller	switch ₂	control
19	arp_rep(host _d)	switch ₂	switch ₁	data
20	pkt_in(arp_rep(host _d))	switch ₁	controller	control
21	pkt_out(arp_rep(host _d),port1)	controller	switch ₁	control
22	<u>arp_rep(host_d)</u>	switch ₁	host _s	data

to switch₁ that is connected to host_s. The switch then sends the ARP request to the controller by encapsulating it through the packet-in control message m_2 . The controller is now aware of the information regarding the source of the ARP request, i.e., host_s. However, since the controller does not know the location of host_d, it instructs switch₁ to flood the ARP request to the connected switches using the packet-out message m_3 . The ARP request is then flooded in the network (via m_4 to m_{11}) until it reaches the destination host_d (via m_{12}). The destination host_d then sends the ARP reply m_{13} to switch₃ in order to inform the source host_s of its location (MAC). Note that, at this stage, since the controller knows the location of host_s, it directly instructs the three switches with the exact directions (i.e., port numbers) to forward the ARP reply (see m_{14} to m_{22}). After this procedure, the controller usually installs forwarding rules for both ARP and IP messages to the switches, resulting in different sequences of messages compared to the example sequence mentioned above.

Failures. Like any software component, SDN systems are susceptible to failures that can affect their functionality. These failures may result in service disruptions noticeable to users. Numerous studies have examined these failures in the context of SDN testing [40, 49, 50, 62, 77]. Furthermore, the centralisation of SDN system logic within its controller makes it a critical point of failure. A controller crash or loss of connection with the switches can disrupt the entire network. This vulnerability underscores the necessity for thorough testing to ensure the system's robustness and reliability. Such testing entails exploring the state space of the SDN system, including scenarios that are not easily reached. Unfortunately, no work has yet investigated how to automate such state space exploration in SDN systems.

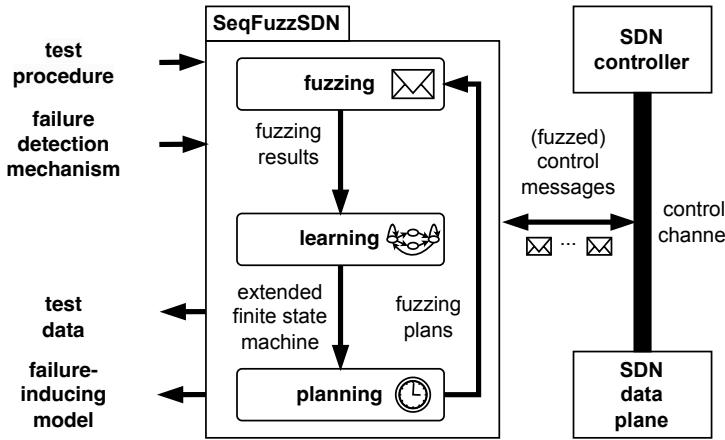


Fig. 2. Approach overview. To test a stateful SDN controller, SeqFuzzSDN fuzzes control message sequences guided by inferred extended finite state machines (EFSMs) that capture failure-inducing message sequences.

Problem. SDN controllers are inherently stateful. They manage complex states that encompass their internal states, the connected switches' states, and the overall network states. When developing and operating SDN systems, engineers must address system failures triggered by unexpected sequences of control messages. Specifically, they must ensure that the system behaves acceptably regardless of its current state. In SDN systems, a stateful controller is susceptible to entering incorrect states, sending unexpected messages, or triggering system failures. These failures may occur only when the controller, the connected switches, or the network reach specific states. For instance, an erroneous control message may be handled correctly under nominal conditions, but if the same message is transmitted during a state of 'recovery' of the system, a system failure may occur. When such a failure occurs, engineers must determine the state of the controller at the time of the failure and identify the sequence of messages that led to that state. Precisely identifying these conditions is crucial, as it enables engineers to diagnose the failure with a clear understanding of the conditions that caused it. Additionally, engineers can utilise this information to generate extended sets of control message sequences for testing the system after implementing fixes. Our work aims to efficiently and effectively test the controller of an SDN system by identifying control message sequences that cause failures, and then automatically derive an accurate model that characterises the sequences of messages leading to failures.

3 Approach

3.1 Overview

Figure 2 shows an overview of SeqFuzzSDN. SeqFuzzSDN takes as input a test procedure and a failure detection mechanism. The test procedure specifies the steps required to (1) initialise the controller under test, switches, and hosts in an SDN, (2) execute a use scenario, e.g., pair-wise ping test [12], to test the controller, and (3) properly tear down the SDN based on the given use scenario to test the controller again. Note that depending on the given use scenario, sequences of control messages exchanged between the controller and switches can vary. The failure detection mechanism, defined by engineers for the given test procedure, allows SeqFuzzSDN to determine whether the controller fails. For example, unexpected communication breakdowns and significant performance degradation can be considered as failures depending on the given test procedure.

Regarding the outputs of SeqFuzzSDN, it produces a test data set and a failure-inducing model. The former contains sequences of control messages that are fuzzed by SeqFuzzSDN and lead to failures detected by the failure detection mechanism. The failure-inducing model characterises sequences of control messages leading to either successes or failures. When the failure detection mechanism does not detect any failures, SeqFuzzSDN considers the corresponding message sequences as successful. In summary, SeqFuzzSDN aims at generating a test data set that contains diverse failure-inducing sequences of control messages and a failure-inducing model that accurately characterises them.

SeqFuzzSDN is an iterative fuzzing method consisting of three steps (see Figure 2), as follows: (1) The fuzzing step involves sniffing and modifying control messages that pass through the control channel between the SDN controller and the SDN switches. Hence, it does not require any changes to the SDN controller and switches. (2) The learning step takes as input the control message sequences and failure detection results obtained from the fuzzing step. The learning step then builds a model to characterise the message sequences. Specifically, the learning step infers an extended finite state machine (EFSM) [1] that captures the controller’s behaviour in terms of state transitions representing control messages received or sent by the controller. Unlike FSMs, EFSMs can capture state transitions associated with data variables, which are essential for modelling state changes caused by control messages. Indeed, control messages typically involve control operations that depend on data (e.g., flow tables and packet statistics). The inferred EFSM contains two types of final states representing success and failure, enabling SeqFuzzSDN to classify and predict which sequences of state transitions (i.e., control messages) induce either success or failure. (3) The planning step takes as input the EFSM inferred by the learning step and generates fuzzing plans. These fuzzing plans aim to guide the fuzzing step in efficiently exploring the possible space of control message sequences and discovering diverse failure-inducing sequences of control messages. In the following subsections, we provide detailed descriptions of the three steps in SeqFuzzSDN.

3.2 Fuzzing

The fuzzing step of SeqFuzzSDN relies on the man-in-the-middle attack (MITM) technique [21], which is widely used and studied in the network security domain. This technique enables SeqFuzzSDN to position itself between the controller under test and the SDN switches that are communicating with the controller. Using MITM, SeqFuzzSDN can intercept control messages and potentially fuzz them while ensuring that the controller and switches remain unaware of the presence of SeqFuzzSDN. Furthermore, employing this attack technique allows SeqFuzzSDN to generate realistic potential threats (i.e., unexpected sequences of control messages) that the controller may face in practice.

When fuzzing control messages, SeqFuzzSDN accounts for the syntax requirements (i.e., grammar) defined in an SDN protocol (e.g., OpenFlow) to ensure fuzzed control messages are syntactically valid. An SDN controller typically rejects syntactically invalid messages at its message parsing layer [40]. Hence, producing valid control messages is desirable in practice to test the controller’s behaviour beyond the parsing layer [40, 62].

SeqFuzzSDN employs a mutation-based fuzzing strategy [88] in which fuzz (i.e., mutation) operators introduce small changes to sniffed control messages while adhering to the syntax requirements of the messages. Below, we first describe five fuzz operators employed in SeqFuzzSDN that can modify control messages and their sequences. We then describe in detail how SeqFuzzSDN uses the fuzz operators.

3.2.1 Fuzz Operators. As shown in Figure 3, when SeqFuzzSDN sniffs a control message, it can apply one of the following fuzz operators: deletion, insertion, duplication, delay, and modification.

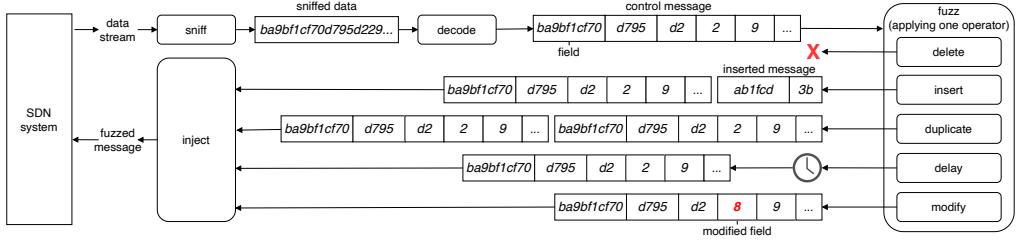


Fig. 3. A data flow example of fuzzing by applying either the deletion, insertion, duplication, delay, or modification operator.

These operators are based on those used in BEADS [40], with modifications tailored for the learning-guided fuzzing of SeqFuzzSDN. We describe further details of the fuzz operators below.

Deletion. The deletion operator drops an intercepted message. For example, when SeqFuzzSDN intercepts a packet-in message from the control channel, it can omit retransmitting the message to the channel, thereby deleting the packet-in message from the control channel.

Insertion. The insertion operator inserts a new control message into the control channel. For example, SeqFuzzSDN can insert a new packet-in message to the control channel while it sniffs messages passing through the channel. Note that, such a new message is either predefined by engineers or randomly generated, as configured in SeqFuzzSDN.

Duplication. The duplication operator duplicates a sniffed message. For example, when SeqFuzzSDN intercepts a packet-in message, it can copy the same message and resend both the original and copied messages to the control channel. Hence, the channel carries the duplicated packet-in messages.

Delay. The delay operator holds a control message for a certain amount of time. For example, SeqFuzzSDN can hold an intercepted packet-in message for 200ms and resend it after the delay time. When SeqFuzzSDN holds a synchronous message (e.g., barrier-request), the sender will also wait for a response from the receiver. However, if SeqFuzzSDN delays an asynchronous message (e.g., packet-in), the sender continues its processing without waiting for the receiver to respond. Note that the delay time can be configured in SeqFuzzSDN.

Modification. The modification operator modifies the content (i.e., fields) of an intercepted control message. For example, when SeqFuzzSDN intercepts a packet-in message, it can change the version field of the message and inject the fuzzed message into the control channel.

We note that the modification operator behaves differently for the initial fuzzing phase and the subsequent learning-guided fuzzing phases. Algorithm 1 shows how the modification operator functions during the initial fuzzing phase when there is no guidance available for fuzzing. Given an intercepted message msg , the modification operator parses the content of msg in terms of its fields (line 1). For each field f of msg and a given probability pf of fuzzing a field, the operator replaces its value with a random value within its syntactically valid value range (lines 2-7). For example, if msg contains ten fields and $pf = 0.2$, after applying the operator, the expected number of modified fields in msg is two. The operator then returns the fuzzed message msg' to transmit it through the control channel.

In the subsequent iterations of SeqFuzzSDN, the modification operator leverages planning outputs obtained from the learning and planning steps (see Figure 2). For readability, we describe the modification operator guided by learning in Section 3.5, after introducing the learning and planning steps.

Algorithm 1 Modification: Syntax-aware random**Input:**

msg : control message to be fuzzed
 pf : probability of fuzzing a field

Output:

msg' : control message after fuzzing

```

1:  $F \leftarrow \text{GET\_FIELDS}(msg)$ 
2:  $msg' \leftarrow msg$ 
3: for all  $f \in F$  do
4:   if  $\text{RAND}(0, 1) \leq pf$  then
5:      $msg' \leftarrow \text{REPLACE}(msg', f, \text{RAND\_VALID}(f))$ 
6:   end if
7: end for
8: return  $msg'$ 

```

3.2.2 Initial Fuzzing. During the initial fuzzing phase of SeqFuzzSDN, since no failure-inducing model has been inferred, SeqFuzzSDN applies the fuzz operators randomly, as described in Algorithm 2. The algorithm takes as input a probability pm of fuzzing a message. It then returns a sequence seq' of messages after fuzzing. While executing a given test procedure (see the repeat block on lines 2-30 in Algorithm 2), SeqFuzzSDN intercepts each of the control messages passing through the control channel between the SDN controller and switches (line 3). For each control message msg and the given probability pm , SeqFuzzSDN decides whether it fuzzes msg or not (line 5). When SeqFuzzSDN does not fuzz msg , it resends msg to the control channel and appends msg to seq' to record a processed message sequence (lines 6-7). For fuzzing the message msg , SeqFuzzSDN randomly selects one of the five fuzz operators (line 11). SeqFuzzSDN then applies the selected operator to msg and updates seq' accordingly (lines 12-29).

3.2.3 Data Collection. To generate failure-inducing models, SeqFuzzSDN relies on an inference technique that takes as input event traces and produces an extended finite state machine (EFSM), such as MINT [82]. This EFSM captures the event traces as state transitions with guard conditions. In our context, an event trace corresponds to a message sequence seq' obtained from the fuzzing step. Each event e in the trace is associated with the corresponding message msg listed in seq' . Specifically, an event e is a tuple (l, m, v) , where l denotes the type of msg , m denotes the fuzz operator applied to msg , v denotes the field values of msg . Note that m can be null if msg is not fuzzed in the given message sequence seq' . For example, consider a control message sequence in which a hello control message [63], used to discover and establish a connection between the controller and switches, was delayed by 200ms using the delay operator. SeqFuzzSDN encodes this hello message into an event e as follows: $(\text{hello}, \text{delay} : 200, < 0x5, 0x0, 0x10, 0xA34BF >)$, where the field values of the hello message are version = $0x5$, type = $0x0$, length = $0x10$, and xid = $0xA34BF$. The last event in the trace indicates either success or failure, determined by the failure detection mechanism for the given sequence seq' of messages. In addition, the event e is associated with both the sender and receiver of msg , enabling SeqFuzzSDN to track this information.

We note that, at each iteration i of SeqFuzzSDN, the fuzzing step executes the input test procedure (see Figure 2) n times, determined by a time budget. Hence, for each iteration i , the fuzzing step generates a dataset D_i that contains n event traces, i.e., $|D_i| = n$.

Algorithm 2 Initial fuzzing**Input:**

pm : probability of fuzzing a message

Output:

seq' : sequence of messages after fuzzing

```

1:  $seq' \leftarrow \langle \rangle$ 
2: repeat
3:    $msg \leftarrow \text{RECEIVE}()$ 
4:   // no fuzzing
5:   if  $\text{RAND}(0, 1) > pm$  then
6:      $\text{SEND}(msg)$ 
7:      $seq' \leftarrow \text{APPEND}(seq', msg)$ 
8:     continue
9:   end if
10:  // fuzzing
11:   $op \leftarrow \text{RAND\_SELECT\_FUZZ\_OPERATOR}()$ 
12:  if  $op$  is a deletion operator then
13:    // do nothing
14:  else if  $op$  is an insertion operator then
15:     $msg' \leftarrow \text{GET\_MESSAGE}(op)$ 
16:     $\text{SEND}(msg, msg')$ 
17:     $seq' \leftarrow \text{APPEND}(seq', msg, msg')$ 
18:  else if  $op$  is a duplication operator then
19:     $\text{SEND}(msg, msg)$ 
20:     $seq' \leftarrow \text{APPEND}(seq', msg, msg)$ 
21:  else if  $op$  is a delay operator then
22:     $t \leftarrow \text{GET\_DELAY}(op)$ 
23:     $\text{DELAY\_SEND}(msg, t)$ 
24:     $seq' \leftarrow \text{DELAY\_APPEND}(seq', msg, t)$ 
25:  else if  $op$  is a modification operator then
26:     $msg' \leftarrow \text{MODIFY}(msg, op)$ 
27:     $\text{SEND}(msg')$ 
28:     $seq' \leftarrow \text{APPEND}(seq', msg')$ 
29:  end if
30: until the test procedure has finished
31: return  $seq'$ 

```

3.3 Learning

At each iteration i of SeqFuzzSDN, the learning step takes as input a dataset D of event traces obtained from the fuzzing step through the 1st to i th iterations, i.e., $D = D_1 \cup \dots \cup D_i$. The learning step then outputs an EFSM inferred based on D . The inferred EFSM M is then used to guide the fuzzing process, which entails exploiting state transitions in M , exploring less-visited states in M , and discovering new states not captured in M . Furthermore, SeqFuzzSDN provides engineers with an accurate EFSM, achieved through iterative refinement of M . This EFSM serves as a failure-inducing model that characterises the generated failure-inducing message sequences (i.e.,

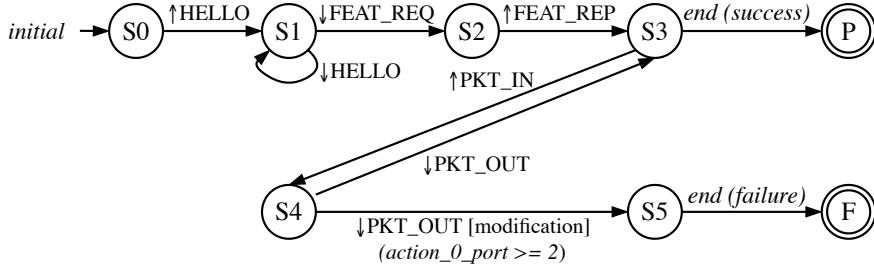


Fig. 4. A simplified EFSM example produced by SeqFuzzSDN. The \uparrow and \downarrow arrows indicate that the corresponding control messages are received and sent, respectively, by the controller under test.

event traces), enabling engineers to gain a more comprehensive understanding of failure-inducing sequences rather than individually inspecting each of them.

We note that, to infer an EFSM, SeqFuzzSDN relies on MINT [82], a state-of-the-art model inference tool that takes as input a dataset containing event traces and produces an EFSM. We opted to use MINT since it is one of the few tools available online and has been applied in many software engineering studies [30, 74]. In addition, the implementation of MINT is the most reliable among the tools available online, enabling us to focus on developing the main contributions of SeqFuzzSDN.

3.3.1 States and transitions. An SDN controller takes as input control messages and produces control messages in response, which are observable via MITM techniques [21]. In an EFSM inferred by SeqFuzzSDN, which captures sequences of observed control messages, a state is a placeholder for the transitions between different sequences of these messages, rather than representing a specific internal condition of the controller, which is not visible to SeqFuzzSDN. In this state, the controller is capable of processing a particular control message and generating a corresponding response message. A transition is defined as a tuple (s, l, c, m, d) , where s denotes a source state, l denotes the type of a control message, c denotes a guard condition on the fields of the message, m denotes a fuzz operator applied to the message, and d denotes a destination state. In an EFSM produced by SeqFuzzSDN, using the dataset D containing event traces, each transition (s, l, c, m, d) corresponds to an event (l, m, v) in D (see the event definition in Section 3.2.3).

For example, Figure 4 shows an EFSM produced by SeqFuzzSDN, simplified for clarity. The EFSM contains eight states in total. Among these states, state $S0$ represents the initial state of the EFSM, state P represents the success state, and state F represents the failure state. Additionally, the EFSM includes ten transitions. For instance, in the transition from state $S4$ to state $S5$, $S4$ serves as the source state (s), $PACKET_OUT$ as the label (l), $action_0_port \geq 2$ as the guard condition (c), $modification$ as the mutation operator (m), and $S5$ as the destination state (d). Note that the arrow \uparrow (resp. \downarrow) annotated before each label (e.g., \uparrow HELLO and \downarrow HELLO) indicates that a message is received by the controller (resp. sent by the controller).

3.3.2 Guard condition inference. SeqFuzzSDN aims at efficiently producing an accurate EFSM that correctly captures the event traces D . Since the overall accuracy of an EFSM highly depends on the accuracy of transitions' guard conditions, in this section, we first explain how SeqFuzzSDN uses MINT to infer guard conditions from D . For further details, such as merging states and removing non-determinism, we refer readers to the paper introducing MINT [82]. We then introduce how SeqFuzzSDN efficiently infers an EFSM from D in Section 3.3.3.

MINT employs a supervised machine learning algorithm [85] that requires labelled datasets to infer guard conditions of state transitions. To create labelled datasets, SeqFuzzSDN groups events in the event traces D into event groups E based on the event type defined by (l, m) , where the control message type l and the fuzz operator m are elements in an event $e = (l, m, v)$. This ensures that each group contains events with the same event type. For each event e in an event group, SeqFuzzSDN then labels e with the type of the next event following e in the corresponding event trace of e in the event traces D , as required by MINT. We note that one of the key reasons why MINT labels each event with the next event type is to enable machine learning classifiers to learn the guards that govern transitions between states, specifically identifying the conditions under which a state that allows a given event can transition to another state that allows an event of the next type. More precisely, given an event trace $e_1, \dots, e_i, e_{i+1}, \dots, e_n$, the event e_i in an event group is labelled with the type (l_{i+1}, m_{i+1}) of e_{i+1} . For each event group E , SeqFuzzSDN creates a dataset that contains pairs of the field values of an event $e \in E$ and the assigned label of e .

For example, Table 2 shows the creation of six datasets (see Table 2 (b)) from two event traces (see Table 2 (a)). The datasets could be used to infer the EFSM presented in Figure 4. As shown in Table 2 (a), Trace 1 and Trace 2 contain six distinct event types (l, m) : $(\text{HELLO}, \text{null})$, $(\text{FEAT_REQ}, \text{null})$, $(\text{FEAT_REP}, \text{null})$, $(\text{PKT_IN}, \text{null})$, $(\text{PKT_OUT}, \text{null})$, and $(\text{PKT_OUT}, \text{modification})$. For each event type, SeqFuzzSDN creates its corresponding dataset as presented in Table 2 (b). Note that the third and last columns in the table indicate the content of a labelled dataset, including field values of a control message and associated labels (i.e., the next event type).

Regarding supervised machine learning, SeqFuzzSDN relies on RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [20], an interpretable rule-based classification algorithm. RIPPER has shown successful applications in many software engineering problems involving classification and condition inference [14, 37, 56]. In particular, we select RIPPER because it generates pruned decision rules (i.e., if-conditions) that are more concise and, as a result, more interpretable than commonly used tree-based classification algorithms, such as C4.5 [70], which are susceptible to the replicated subtree problem [85].

3.3.3 Sampling Event Traces. Due to the computational complexity of the model inference problem, existing model inference techniques (including MINT) face scalability issues [74]. Among the works addressing the scalability problem, Shin et al. [74] recently introduced PRINS, which is the most relevant to MINT, the EFSM inference tool we selected for SeqFuzzSDN. PRINS aims to improve the scalability of EFSM inference for component-based systems. It employs a divide-and-conquer approach, first deriving individual models for each system component based on their respective logs. These component models are then systematically merged, incorporating the event flow across components as recorded in the logs. However, PRINS is not suitable for SeqFuzzSDN, as it requires prior knowledge of which system components generate logs.

In the context of SeqFuzzSDN, the number of event traces can continuously grow as SeqFuzzSDN iterates through the fuzzing, learning, and planning steps multiple times and fuzzes message sequences corresponding to those event traces. Hence, when the number of event traces and their events become large (e.g., 5000 event traces, containing 15000 events), MINT either crashes due to running out of memory or takes a prohibitively long time to complete its execution. To address the scalability problem in our context, when learning an EFSM at each iteration i , SeqFuzzSDN uses a subset D^s of event traces instead of all event traces D generated up to the current iteration.

Let M be an EFSM inferred at the $i-1$ th iteration of SeqFuzzSDN. At iteration i , to sample event traces from the event traces D_i obtained at i and those used in learning M , SeqFuzzSDN first separate D_i into *accepted* and *rejected* traces. Given an EFSM M , accepted traces are traces that are already explained by M , i.e., traces that follow paths (i.e., transition sequences) in M . Note that

Table 2. An example illustrating the creation of datasets based on event traces: (a) Two event traces (i.e., Trace 1 and Trace 2). (b) Six datasets created based on Trace 1 and Trace 2.

(a) event traces				
	event (e)	label (l)	mutation (m)	value (v)
Trace 1	e_{10}	HELLO	null	$v(e_{10})$
	e_{11}	HELLO	null	$v(e_{11})$
	e_{12}	FEAT_REQ	null	$v(e_{12})$
	e_{13}	FEAT REP	null	$v(e_{13})$
	e_{14}	PKT_IN	null	$v(e_{14})$
	e_{15}	PKT_OUT	null	$v(e_{15})$
	e_{16}	end(success)	null	null
Trace 2	e_{21}	HELLO	null	$v(e_{21})$
	e_{22}	HELLO	null	$v(e_{22})$
	e_{23}	FEAT_REQ	null	$v(e_{23})$
	e_{24}	FEAT REP	null	$v(e_{24})$
	e_{25}	PKT_IN	null	$v(e_{25})$
	e_{26}	PKT_OUT	modification	$v(e_{26})$
	e_{27}	end(failure)	null	null

(b) datasets				
event type (l, m)	event (e)	value (v)	next event type (l', m')	
(HELLO, null)	e_{10}	$v(e_{10})$	(HELLO, null)	
	e_{11}	$v(e_{11})$	(FEAT_REQ, null)	
	e_{21}	$v(e_{21})$	(HELLO, null)	
	e_{22}	$v(e_{22})$	(FEAT_REQ, null)	
(FEAT_REQ, null)	e_{12}	$v(e_{12})$	(FEAT REP, null)	
	e_{23}	$v(e_{23})$	(FEAT REP, null)	
(FEAT REP, null)	e_{13}	$v(e_{13})$	(PKT_IN, null)	
	e_{24}	$v(e_{24})$	(PKT_IN, null)	
(PKT_IN, null)	e_{14}	$v(e_{14})$	(PKT_OUT, null)	
	e_{25}	$v(e_{25})$	(PKT_OUT, modification)	
(PKT_OUT, null)	e_{15}	$v(e_{15})$	(end(success), null)	
(PKT_OUT, modification)	e_{26}	$v(e_{26})$	(end(failure), null)	

when guard evaluations are needed while SeqFuzzSDN walks over M with traces, it uses Z3 [22], a well-known and widely applied SMT solver. In contrast, rejected traces refer to traces that do not follow any path in M . Hence, to create a set of event traces for learning a new EFSM M' at iteration i , SeqFuzzSDN includes the rejected traces in the set in order to ensure that they are explained by M' . However, SeqFuzzSDN does not include the accepted traces in the learning process because they do not contribute to refining M into M' .

SeqFuzzSDN then further separates the rejected event traces obtained from iteration i of SeqFuzzSDN into success event traces and failure event traces. Drawing inspiration from the observation that balanced datasets often yield higher accuracy in ML [8, 9, 85], SeqFuzzSDN manages two

Algorithm 3 Sampling event traces. Note that the sets of event traces used in this algorithm contain only success or failure event traces.

Input:

M : EFSM

\mathcal{D}^M : set of success (resp. failure) event traces used to generate M

\mathcal{D}_i : set of success (resp. failure) event traces obtained from the i th iteration of SeqFuzzSDN

$n_{\mathcal{D}}$: maximum size of an output set of event traces

Output:

\mathcal{D} : set of success (resp. failure) event traces for learning a new EFSM

```

1: // Case: include all traces
2: if  $|\mathcal{D}^M \cup \mathcal{D}_i| \leq n_{\mathcal{D}}$  then
3:    $\mathcal{D} \leftarrow \mathcal{D}^M \cup \mathcal{D}_i$ 
4:   return  $\mathcal{D}$ 
5: end if
6:
7: // Case: replace traces
8:  $n_r \leftarrow |\mathcal{D}^M \cup \mathcal{D}_i| - n_{\mathcal{D}}$  // number of traces to replace
9: for  $n_r$  times do
10:    $\mathbb{G} \leftarrow \text{GROUP\_BY\_PATH}(\mathcal{D}^M, M)$  //  $\mathbb{G}$ : set of trace groups
11:    $G \leftarrow \text{SELECT\_MAX\_GROUP}(\mathbb{G})$ 
12:    $t \leftarrow \text{RAND\_SELECT\_TRACE}(G)$ 
13:    $\mathcal{D}^M \leftarrow \mathcal{D}^M \setminus \{t\}$ 
14: end for
15:  $\mathcal{D} \leftarrow \mathcal{D}^M \cup \mathcal{D}_i$  //  $|\mathcal{D}| = n_{\mathcal{D}}$ 
16: return  $\mathcal{D}$ 

```

distinct sets of event traces: one leading to success and the other to failure. These sets have the same maximum number of event traces and are used together to learn an EFSM.

Algorithm 3 presents our heuristic for sampling event traces. SeqFuzzSDN applies the algorithm separately to both success rejected event traces and failure rejected event traces. The algorithm takes as input an EFSM M inferred at iteration $i-1$, a set \mathcal{D}^M of success (resp. failure) event traces used to learn M , a set \mathcal{D}_i of success (resp. failure) rejected event traces obtained from iteration i , and the maximum size $n_{\mathcal{D}}$ of an output set \mathcal{D} . The algorithm then outputs a set \mathcal{D} of success (resp. failure) event traces for learning a new EFSM. As shown on lines 1–5 of the algorithm, when the size of $\mathcal{D}^M \cup \mathcal{D}_i$ does not exceed the maximum size $n_{\mathcal{D}}$, the algorithm returns $\mathcal{D}^M \cup \mathcal{D}_i$. Otherwise, on line 8, the algorithm computes the number n_r of event traces to remove from \mathcal{D}^M to ensure that the output set \mathcal{D} contains $n_{\mathcal{D}}$ event traces (see line 15). On lines 9–14, the algorithm removes n_r event traces from \mathcal{D}^M as follows: It first partitions \mathcal{D}^M into groups, each containing event traces that follow the same path in M . It then selects a group G that contains the largest number of event traces compared to the other groups. On lines 12–13, it randomly selects an event trace t and removes it from \mathcal{D}^M . On lines 15–16, the algorithm returns $\mathcal{D}^M \cup \mathcal{D}_i$, where $|\mathcal{D}| = n_{\mathcal{D}}$. Note that the selection mechanism on lines 10–11 aims at minimising information loss in \mathcal{D} with regard to learning an EFSM. Since the selection mechanism (lines 10–11) selects an event trace from group G containing the largest number of event traces and removes the selected trace from \mathcal{D}^M (lines 12–13), the remaining traces in G will still contribute to creating a new EFSM that contains the same path (i.e., no information loss after the removal) and accepts the remaining traces.

3.4 Planning

The planning step of SeqFuzzSDN takes as input an EFSM and outputs fuzzing plans to guide the subsequent fuzzing iteration. The fuzzing plans are defined as sequences of state transitions, i.e., paths in an EFSM, that guide the fuzzing step at the subsequent iteration. SeqFuzzSDN produces the fuzzing plans, aiming at (O1) exploring less-visited or new states of the controller under test, (O2) improving the accuracy of a failure-inducing model (i.e., EFSM) and (O3) increasing the diversity of message sequences (i.e., event traces) exercised for testing the controller. Hence, SeqFuzzSDN employs a multi-objective search algorithm [23] to address the planning problem. Below, we describe the multi-objective search-based planning approach in SeqFuzzSDN by defining the solution representation, the fitness functions, and the search algorithm.

3.4.1 Representation. Given an EFSM M obtained from the learning step, a candidate solution is a set C of sequences of state transitions (i.e., paths) in M where each transition sequence starts from the initial state s_1 of M and ends at a state s_o selected during search, representing a valid traversal of M . Depending on a fuzzing probability, each transition sequence in C can be associated with a fuzz operator m_o —deletion, insertion, duplication, delay, or modification described in Section 3.2.1—to be applied when the controller’s state is s_o in the subsequent iteration of SeqFuzzSDN.

3.4.2 Fitness Functions. SeqFuzzSDN aims at searching for candidate solutions with regard to the three objectives: (O1) coverage, (O2) accuracy, and (O3) diversity, described earlier. To quantify how a candidate solution fits these three objectives, below we define three fitness functions.

Coverage. SeqFuzzSDN relies on an EFSM M that models the state changes of the controller under test. To test various behaviours of the controller, SeqFuzzSDN aims at finding a candidate solution that ensures a similar (ideally equal) number of visits to each state in M . Hence, each state in M can be explored in different ways regarding how is reached and what happens after traversing it. Given an EFSM M at iteration i of SeqFuzzSDN and a set D of event traces obtained from the first to the i th iterations, SeqFuzzSDN counts the number of visits for each state in M by traversing M using each event trace in D .

To quantify the extent to which a candidate solution C satisfies the coverage objective regarding the state-visit numbers, SeqFuzzSDN leverages Shannon’s Entropy [73]. In general, entropy characterises the average level of uncertainty inherent to the stochastic variable’s possible outcomes. In our context, the entropy defines the level of uncertainty associated with visits to a state in an EFSM M . Intuitively, the higher the entropy, the more evenly the states in M are visited.

Let S be a set of all states in an EFSM M obtained at iteration i and D be a set of event traces obtained from the first to the i th iterations. For each state $s \in S$, we denote by $nv(s, C)$ the sum of the following: (1) the number of visits to s by the event traces in D , and (2) the number of visits to s expected by a candidate solution C . We denote by $nv(S, C)$ the total number of state visits for S and define $nv(S) = \sum_{s \in S} nv(s, C)$. Based on Shannon’s entropy equation, we formulate the following fitness function $fitcov(S, C)$ for the coverage objective as below. SeqFuzzSDN aims at maximising the fitness $fitcov(S, C)$.

$$fitcov(S, C) = - \sum_{s \in S} \frac{nv(s, C)}{nv(S, C)} \log_2 \frac{nv(s, C)}{nv(S, C)}$$

We note that, in practice, an EFSM inference technique is not always able to infer an EFSM M that allows the traversal of all event traces in D [82]. Hence, SeqFuzzSDN computes $nv(s, C)$ using those event traces in D that are traceable by M and a candidate solution C . To improve the accuracy of an inferred EFSM over iterations of SeqFuzzSDN, it accounts for an additional fitness function described below.

Accuracy. SeqFuzzSDN builds an EFSM M using MINT, which relies on supervised machine learning. Recall from Section 3.3 that MINT converts the event traces D into labelled training datasets (i.e., event groups) for building supervised classifiers. Hence, building accurate classifiers is beneficial to improve the overall accuracy of an EFSM M .

Note that the imbalance problem [85] is one of the main reasons that usually cause the low performance of supervised classification algorithms. In a labelled dataset, when the number of data instances of a class is significantly different from that of the other classes, classification algorithms tend to favour predicting the majority class, which is often not desirable in practice [85]. Hence, SeqFuzzSDN aims to address imbalance by planning to generate control message sequences that alleviate the problem.

To quantitatively assess the imbalance problem of each event group E (i.e., labelled training dataset) converted from the event traces D , SeqFuzzSDN uses the multi-class imbalance metric [55]. Given an event group E obtained from the event traces D , we denote by $nc(E)$ the number of classes in E , $ni(E)$ the number of data instances in E , and $ni(c)$ the number of data instances labelled with the class c . According to the multi-class imbalance metric, the imbalance ratio $ir(E)$ of E is computed as follows:

$$ir(E) = \frac{nc(E) - 1}{nc(E)} \sum_{c \text{ in } E} \frac{ni(c)}{ni(E) - ni(c)}$$

For example, consider an event group E that consists of three classes ($nc(E) = 3$)—namely c_1, c_2, c_3 —along with a total of 1200 data instances ($ni(E) = 1200$). In the case where the class distribution is balanced (i.e., $ni(c_1) = ni(c_2) = ni(c_3) = 400$), the imbalance ratio is $ir(E) = 1$. However, in a situation where the class distribution is imbalanced, such as $ni(c_1) = 5$, $ni(c_2) = 200$, $ni(c_3) = 995$, the imbalance ratio increases to $ir(E) \approx 3.37$.

To estimate the degree to which a candidate solution C (i.e., fuzzing plan) impacts the imbalance problem, SeqFuzzSDN augments each event group E obtained from the event traces D using C . Recall from Section 3.3 that each event group E contains events that have the same event type. The class assigned to an event e in E is determined by the event following e in the corresponding event trace (i.e., message sequence) containing e . Hence, we can estimate how many new events will be added to each event group when SeqFuzzSDN generates message sequences guided by a candidate solution C . Precisely, given a sequence $(s_1, l_1, m_1, c_1, d_1), \dots, (s_i, l_i, m_i, c_i, d_i), (s_{i+1}, l_{i+1}, m_{i+1}, c_{i+1}, d_{i+1}), \dots, (s_o, l_o, m_o, c_o, d_o)$ of state transitions in C , SeqFuzzSDN can, for example, augment an event group E that corresponds to the event type (l_i, m_i) with a new event that is labelled with (l_{i+1}, m_{i+1}) . We denote by $ir(E, C)$ the imbalance ratio of an event group that contains both the labelled events in the event group E and the augmented events from C . Below, we define the fitness function $fitacc(D, C)$ for the accuracy objective, where $ng(D)$ denotes the number of event groups in D . SeqFuzzSDN aims at maximising the fitness $fitacc(D, C)$.

$$fitacc(D, C) = \sum_{E \text{ in } D} ir(E, C) / ng(D)$$

Diversity. SeqFuzzSDN aims at testing the controller under test using diverse sequences of control messages. To this end, at each iteration i of SeqFuzzSDN, it plans to guide fuzzing in the $i+1$ th iteration to generate sequences of control messages that are different from the sequences exercised from the first to the i th iterations, which are captured in the event traces D . Given a candidate solution C , SeqFuzzSDN quantifies the difference between D and event traces (i.e., message sequences) that can be produced by C using the normalised compression distance (NCD) [18]. NCD measures the difference between two objects X and Y based on their compression, the Kolmogorov

complexity [43], and the information distance [6]. Precisely, $NCD(X, Y)$ is defined as follows:

$$NCD(X, Y) = \frac{Z(XY) - \min\{Z(X), Z(Y)\}}{\max\{Z(X), Z(Y)\}}$$

where $Z()$ is an actual compressor such as gzip [25], $Z(X)$ and $Z(Y)$ are the compressed sizes of the objects X and Y , and $Z(X, Y)$ is the compressed size of the concatenation of X and Y . Note that $NCD(X, Y) = 0$ indicates that the two objects are identical in terms of compressed information. In contrast, $NCD(X, Y) = 1 + \epsilon$ implies that they are distinct, where ϵ is a small positive value dependent on how closely the compressor Z approximates the Kolmogorov complexity. We opt to use NCD because it is applicable for comparing two sets of event traces, wherein individual events can have different message types, fuzz operators, and field values. Furthermore, the lengths of the event traces may differ from one another, and the two sets contain different numbers of event traces. Hence, applying simple sequence comparison methods is not straightforward in our context.

Given the event traces D , in order to use NCD as the diversity fitness for a candidate solution C , SeqFuzzSDN converts C into event traces T^C to be generated in the subsequent iteration. Specifically, for each transition sequence $p = (s_1, l_1, m_1, c_1, d_1), \dots, (s_i, l_i, m_i, c_i, d_i), (s_{i+1}, l_{i+1}, m_{i+1}, c_{i+1}, d_{i+1}), \dots, (s_o, l_o, m_o, c_o, d_o)$ in C , the sequence p is converted into an event trace $tr = (l_1, m_1, nil), \dots, (l_i, m_i, nil), (l_{i+1}, m_{i+1}, nil), \dots, (l_o, m_o, nil)$ by excluding the source and destination states s and d from the transitions while preserving their message type l and the fuzz operator m , along with their original order. Note that, in predicted event traces, field values are set to nil (i.e., $v_i = nil$) since transition sequences do not capture field values.

To quantify the degree to which a candidate solution C is different from the event traces D , we denote by T^C the predicted event traces when fuzzing is guided by C , and below, we define the fitness function $fitdiv(D, C)$ to address the diversity objective. SeqFuzzSDN aims at maximising the fitness $fitdiv(D, C)$.

$$fitdiv(D, C) = NCD(D, D \cup T^C)$$

3.4.3 Computational search. SeqFuzzSDN employs NSGA-II (Non-Dominated Sorting Genetic Algorithm II) [24], which has been applied in many software engineering studies [15, 45–47, 53, 76], to search for a near-optimal fuzzing plan (i.e., solution C). Algorithm 4 describes the search process. Briefly, the algorithm first generates an initial population P (lines 1-6), containing n_p candidate solutions. Subsequently, the algorithm evolves the population iteratively until finding the ideal Pareto front or the allocated time budget is exhausted (line 9-24). At each iteration, the algorithm evaluates each candidate solution $C \in P$ according to the fitness functions defined in section 3.4.2 (line 11-15). The algorithm then updates the archive P_α (lines 16-17). It then computes the Pareto ranking of the solutions in the archive P_α , along with their associated sparsity values (line 18-19). These sparsity values quantify the distribution of optimal solutions based on their fitness values. These ranks and sparsities are used to select the appropriate n_p solutions to be kept in the archive, as well as to identify the best Pareto front (line 20-21). The algorithm then creates a new population P by breeding the solutions in the archive (lines 22-23). After the search process (lines 9-24), the algorithm returns a selected solution (lines 25-26). Below, we describe in detail the initial population generation, breeding, and solution selection mechanisms that are specific to SeqFuzzSDN.

Initial population. Given an EFSM M , SeqFuzzSDN generates an initial population for the search, containing n candidate solutions. Algorithm 5 describes how SeqFuzzSDN creates a candidate solution at the beginning of the search process (see line 4 of Algorithm 4). Algorithm 5 takes as input an EFSM M , the number n of transition sequences in a candidate solution C , a probability μ of fuzzing a message, and the number k of (different) shortest paths. At each iteration of the repeat

Algorithm 4 Searching best candidate traces to be used in the EFSM-guided fuzzing step, based on NSGA-II.

Input:

M: An EFSM
D: A set of generated event traces
 n_p : size of the population and the archive
 n_s : size of a candidate solution
 n_k : number of shortest paths used during the generation of a candidate solution
 μ_f : candidate solution fuzzing probability
 μ_c : crossover probability
 μ_m : mutation probability

Output:

C_b : Best solution

```

1: // generate the initial population
2:  $P \leftarrow \emptyset$ 
3: repeat
4:    $C \leftarrow \text{GENERATECANDIDATE}(M, n_s, \mu_f, n_k)$ 
5:    $P \leftarrow P \cup C$ 
6: until  $|P| = n_p$ 
7: // create an empty archive
8:  $P_\alpha \leftarrow \emptyset$ 
9: repeat
10:   // assess the fitness of each individual
11:   for each  $C \in P$  do
12:      $f_1(C) = \text{fitcov}(\text{states}(M), C)$ 
13:      $f_2(C) = \text{fitacc}(D, C)$ 
14:      $f_3(C) = \text{fitdiv}(D, C)$ 
15:   end for
16:   // update the archive
17:    $P_\alpha \leftarrow P_\alpha \cup P$ 
18:   COMPUTEFRONTRANKS( $P_\alpha$ )
19:   COMPUTESPARSITIES( $P_\alpha$ )
20:    $P_\alpha \leftarrow \text{SELECTARCHIVE}(P_\alpha, n_p)$ 
21:    $BestFront \leftarrow \text{PARETOFRONT}(P_\alpha)$ 
22:   // create a new population
23:    $P \leftarrow \text{BREED}(P_\alpha, n_p, \mu_c, \mu_m)$ 
24: until  $BestFront$  is the ideal Pareto front or the algorithm run out of time
25:  $C_b \leftarrow \text{SELECTONE}(BestFront)$ 
26: return  $C_b$ 

```

block (lines 2-11), the algorithm finds a transition sequence p to be added into C , and this process is repeated n times. To find a transition sequence p , the algorithm first randomly selects a state s in M (line 3). It then finds the k shortest paths from the initial state of M to the selected state s using the k -shortest path algorithm [31] (line 4). SeqFuzzSDN uses the k -shortest path algorithm to obtain different transition sequences (paths) from the initial state to s . Given the fuzzing probability μ , the algorithm decides whether it applies a fuzz operator or not (line 6). If the algorithm decides to

Algorithm 5 Creating a candidate solution**Input:**

M : EFSM to generate a candidate solution (i.e., paths on M)
 n : size of a candidate solution
 μ : probability of fuzzing a message
 k : number of shortest paths to generate

Output:

C : candidate solution

```

1:  $C \leftarrow \emptyset$ 
2: repeat  $n$  times
3:    $s \leftarrow \text{RAND\_SELECT\_STATE}(M)$ 
4:    $P \leftarrow \text{FIND\_K\_SHORTEST\_PATHS}(M, s, k)$ 
5:    $p \leftarrow \text{RAND\_SELECT\_PATH}(M, P)$ 
6:   if  $\text{RAND}(0, 1) \leq \mu$  then
7:      $op \leftarrow \text{RAND\_SELECT\_FUZZ\_OPERATOR}()$ 
8:      $\text{ASSOCIATE\_FUZZ\_OPERATOR}(p, s, op)$ 
9:   end if
10:   $C \leftarrow C \cup \{p\}$ 
11: end
12: return  $C$ 

```

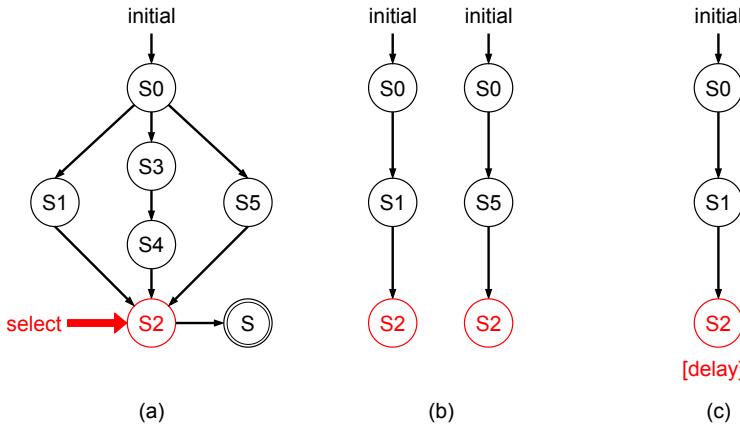


Fig. 5. An example illustration of generating a candidate solution from a simple EFSM: (a) a simple EFSM for clarity, (b) two shortest paths from S_0 to S_2 , and (c) a candidate solution and its associated fuzz operator, i.e., delay.

apply a fuzz operator, the algorithm randomly selects one of the five fuzz operators described in Section 3.2.1 (line 7). It then associates the transition sequence p with the selected fuzz operator m . This guides SeqFuzzSDN in the subsequent iteration to apply m when the controller reaches the selected state s following the transition sequence p . Since we do not know what will happen after applying m in the subsequent iteration of SeqFuzzSDN, it allows SeqFuzzSDN to potentially discover new states that are not captured in the current EFSM M .

For example, Figure 5 illustrates how SeqFuzzSDN generates initial candidate solutions using a simple EFSM M (Figure 5 (a)) for brevity. Given M , when Algorithm 5 selects state S2, it then finds two shortest paths (Figure 5 (b)). After that, the algorithm randomly selects a fuzz operator (e.g., delay) to apply to the selected candidate solution.

Breeding. The breeding mechanism uses the following genetic operators [24]: *selection*, *crossover*, and *mutation* operators. SeqFuzzSDN employs the binary tournament selection and the one-point crossover [24]. Specifically, given two parent solutions C^l and C^r , each containing transition sequences (paths) $\{p_1^l, \dots, p_i^l, \dots, p_j^l\}$ and $\{p_1^r, \dots, p_i^r, \dots, p_k^r\}$, respectively, the crossover operator randomly selects a crossover point i . It then generates two offspring solutions by swapping transition sub-sequences separated by i between the parents, resulting in $\{p_1^r, \dots, p_i^l, \dots, p_j^l\}$ and $\{p_1^l, \dots, p_i^r, \dots, p_k^r\}$. Further, SeqFuzzSDN relies on the uniform mutation operator [80]. Specifically, SeqFuzzSDN first randomly selects a transition sequence p in a candidate solution C . It then replaces p with a new transition sequence obtained with Algorithm 5, setting the parameter n to 1 to create a single transition sequence.

Selecting a near-optimal solution. Algorithm 4, which is based on NSGA-II, outputs a set of Pareto-optimal solutions, which are equally viable with respect to the three objectives regarding coverage, accuracy, and diversity (described in Section 3.4.2). However, SeqFuzzSDN requires selecting one of the solutions to guide fuzzing at the subsequent iteration. Various methods to select a near-optimal solution in a Pareto front have been proposed in the literature, such as selecting a *knee solution* [13], or selecting a *corner solution* [64] for a specific objective. In our context, SeqFuzzSDN uses a knee solution, which is often favoured in search-based software engineering studies [13, 16]. This preference is due to the observation that selecting other solutions on the front to achieve a slight improvement in one objective could result in a significant deterioration in at least one other objective [13]. Given the three objectives regarding coverage, accuracy, and diversity, SeqFuzzSDN favours a candidate solution that achieves a balanced optimisation across all these objectives.

Given a selected set of candidate solutions, containing planned paths (state transitions), SeqFuzzSDN finds transitions that are associated with the modification fuzz operator and a guard condition. It then solves the guard condition using Z3 in order to apply the modification fuzz operator, ensuring the guard condition is satisfied during our EFSM-guided fuzzing (described in Section 3.5). For example, given a state transition $(s_i, \uparrow \text{PACKET_IN}, f_k < 20 \wedge f_k > 8, \text{modification}, s_j)$, SeqFuzzSDN solves the guard condition, such as $f_k = 10$. When the transition is exploited during fuzzing, SeqFuzzSDN modifies a PACKET_IN message by assigning 10 to the field f_k of the message. Note that SeqFuzzSDN solves guard conditions during the (offline) planning step, rather than the (online) fuzzing step, in order to improve efficiency during fuzzing.

3.5 EFSM-Guided Fuzzing

After the initial fuzzing step, SeqFuzzSDN uses the learning and planning outputs to guide fuzzing sequences of control messages to test the SDN controller. This section first describes an EFSM-guided fuzzing method in SeqFuzzSDN, and then illustrates the method through a running example.

3.5.1 EFSM-guided Fuzzing Algorithm. Algorithm 6 describes the fuzzing procedure in SeqFuzzSDN once an EFSM M is available, after the initial fuzzing step. The algorithm takes as input an EFSM M and a set C of planning paths (i.e., sequences of state transitions) on M , and iterates lines 2-19 until the test procedure has finished executing. At the beginning of each iteration, on line 3, the algorithm first identifies the current state s in M according to the currently observed sequence seq' of control messages. On line 4, SeqFuzzSDN then finds a set P of applicable paths from the set C of planning paths to guide fuzzing. The applicable paths contain state transitions on M that

Algorithm 6 EFSM-Guided Fuzzing**Input:**

M : EFSM generated from the learning step

C : set of planned paths on M

Output:

seq' : sequences of messages after fuzzing

C' : set of planned paths after applying one of them

```

1:  $seq' \leftarrow \langle \rangle$ 
2: repeat
3:    $s \leftarrow \text{CURRENT\_STATE}(M, seq')$ 
4:    $P \leftarrow \text{FIND\_APPLICABLE\_PATHS}(C, M, seq')$ 
5:    $msg \leftarrow \text{RECEIVE}()$ 
6:    $TN \leftarrow \text{FIND\_APPLICABLE\_TRANSITIONS}(s, msg, P)$ 
7:    $op \leftarrow \emptyset$ 
8:   if  $TN \neq \emptyset$  then
9:      $tn \leftarrow \text{RAND\_SELECT}(TN)$ 
10:     $op \leftarrow \text{GET\_FUZZ\_OPERATOR}(tn)$ 
11:   end if
12:   if  $op$  is a fuzz operator then
13:      $msg \leftarrow \text{FUZZ}(msg, op)$ 
14:      $seq' \leftarrow \text{APPEND}(seq', msg, op)$ 
15:   else
16:      $seq' \leftarrow \text{APPEND}(seq', msg)$ 
17:   end if
18:    $\text{SEND}(msg)$ 
19: until the test procedure has finished
20:  $p \leftarrow \text{FIND\_USED\_PATH}(M, C, seq')$ 
21:  $C' \leftarrow C \setminus \{p\}$ 
22: return  $seq', C'$ 

```

start from the current state s . Below, Algorithm 7 further describes this procedure. On line 5, the algorithm receives a control message msg passing through the SDN control channel and then finds a set TN of applicable transitions from P . The applicable transitions start from the current state s , are triggered by an event type l corresponding to msg , and, if there are guards, the guards hold on the field values of msg . On lines 7-11, if some applicable transitions are found, the algorithm randomly selects a transition tn among the applicable transitions (line 9), and gets the fuzz operator op of tn , if tn has one (line 10). On lines 12-15, if the fuzz operator op is present, the algorithm applies it to msg (line 13) and appends it (line 14) to the output sequence seq' along with the applied fuzz operator (op). On lines 16-17, if no fuzz operator is present, the algorithm simply appends the originally received message msg to the output sequence seq' . On line 18, the algorithm sends back the msg , which could be fuzzed, into the control channel. Since, at each iteration of SeqFuzzSDN, the test procedure is run multiple times, on lines 20-21, the algorithm removes a planned path that has been applied, enabling the subsequent executions of the test procedure to be fuzzed, guided only by the remaining planned paths.

Algorithm 7 identifies a set C' of applicable paths on an EFSM M based on a given sequence seq of control messages and a set C of planned paths on M . On lines 1-3, the algorithm initialises a

Algorithm 7 Finding Applicable Paths**Input:**

C : set of planned paths
 M : EFSM
 seq : sequence of messages

Output:

C' : set of applicable paths

```

1:  $C' \leftarrow \emptyset$ 
2:  $p^s \leftarrow \text{WALK}(M, seq)$ 
3:  $s \leftarrow \text{CURRENT\_STATE}(M, seq')$ 
4: for each  $p \in C$ , where  $s$  is on  $p$  do
5:    $p' \leftarrow \text{SUBPATH}(p, s)$ 
6:   if for all  $s \in p'$ ,  $s \in p^s$ , and all  $s$  appear in the same order on both  $p'$  and  $p^s$  then
7:      $C' \leftarrow C' \cup \{p\}$ 
8:   end if
9: end for
10: return  $C'$ 

```

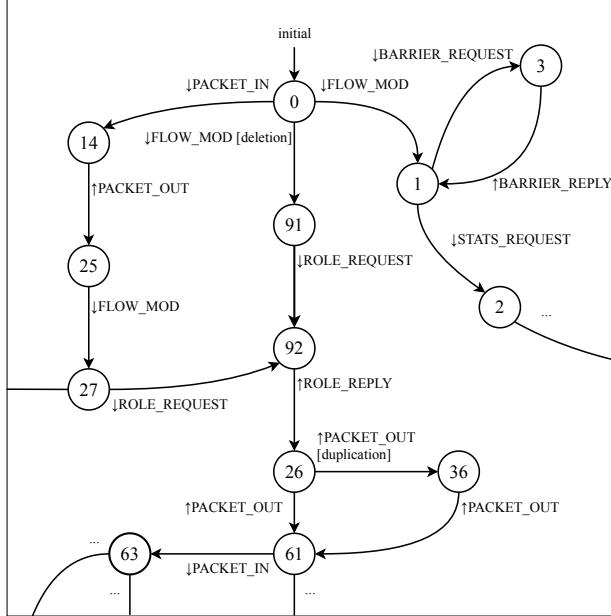
return set C' of applicable paths on M , converts seq into a path p^s on M , and identifies the current state s on M for the given seq . The algorithm then examines each path p in C to determine whether the current state s appears on p (line 4) and whether the sub-path p' of p from the start state to s in M is a derivative of p^s that corresponds to the sequence seq of control messages (lines 5-6). If p' is a derivative of p^s , p' can be derived from p^s by deleting some transitions without changing the order of the remaining transitions. Recall from Algorithm 5 that SeqFuzzSDN uses the k -shortest path algorithm to create planned paths. Hence, Algorithm 7 checks whether planned (sub-)paths on M are derivatives of the paths in M that correspond to sequences of control messages. The algorithm then returns a set C' of applicable paths that satisfy the conditions described above.

3.5.2 EFSM-Guided Fuzzing Example. Figure 6 presents a part of an EFSM M (Figure 6a) inferred from the learning step and three planned paths C (Figure 6b) in M created by the planning step. Given the EFSM M and the planned paths C , when SeqFuzzSDN begins executing the test procedure (e.g., ping test), Algorithm 6 starts with the initial state 0 in M to perform EFSM-guided fuzzing. Then all the three planned paths, p_1 , p_2 , and p_3 shown in Figure 6b, are identified as applicable paths. The algorithm then receives the first control message (generated by the test procedure), which, in this example, we assume to be “ \uparrow HELLO”. In this case, however, there are no planned paths that contain transitions starting from state 0 and taking the event (message) “ \uparrow HELLO”. Hence, the “ \uparrow HELLO” message is sent back into the control channel without any modification. The “ \uparrow HELLO” message is then appended to the output sequence seq' , as follows:

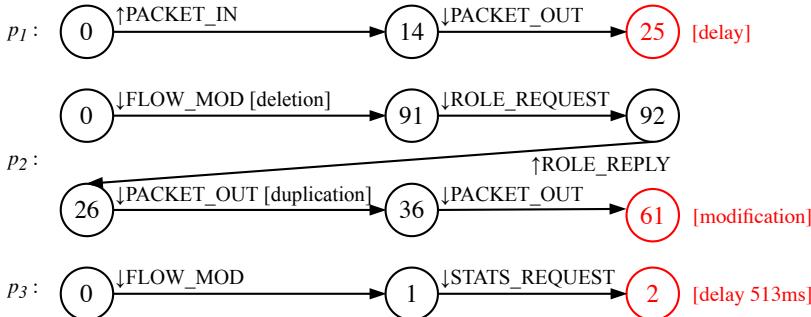
$$seq' = \langle \uparrow\text{HELLO} \rangle$$

After receiving the “ \uparrow HELLO” message, in this example, the algorithm receives three more control messages, as follows: “ \downarrow HELLO”, “ \downarrow FEATURES_REQUEST”, and “ \uparrow FEATURES_REPLY”. In these cases, the EFSM M remains in state 0 since there are no transitions from state 0 that can be taken by the three messages. In addition, there are no applicable paths. As a result, those messages are sent back to the control channel, and the output sequence seq' is as follows:

$$seq' = \langle \uparrow\text{HELLO}, \downarrow\text{HELLO}, \downarrow\text{FEATURES_REQUEST}, \uparrow\text{FEATURES_REPLY} \rangle$$



(a) A partial EFSM example inferred from the learning step of SeqFuzzSDN.



(b) Three planned paths to guide fuzzing, created by the planning step of SeqFuzzSDN.

Fig. 6. Output examples of the learning and planning steps: (a) a partial EFSM and (b) three planned paths on the EFSM.

Next, the algorithm receives the “ $\downarrow \text{FLOW_MOD}$ ” message, while the EFSM M is in state 0. Then, the algorithm identifies paths p_2 and p_3 as applicable paths since they contain transitions starting from state 0 and taking the event “ $\downarrow \text{FLOW_MOD}$ ”. Among the two transitions, i.e., $(0, \downarrow \text{FLOW_MOD}, \text{nil}, \text{deletion}, 91)$ on p_2 and $(0, \downarrow \text{FLOW_MOD}, \text{nil}, \text{nil}, 14)$ on p_3 , the algorithm randomly selects the second one on p_3 . Since no fuzz operator is associated to the transition, the algorithm simply sends the message back to the control channel, and appends the message to the output sequence seq' of control messages, as follows:

$$seq' = \langle \uparrow \text{HELLO}, \downarrow \text{HELLO}, \downarrow \text{FEATURES_REQUEST}, \uparrow \text{FEATURES_REPLY}, \downarrow \text{FLOW_MOD} \rangle$$

In the subsequent iteration of the algorithm, the current state of the EFSM M changes to state 1, as there is a transition from state 0 to 1 that takes the “ \downarrow FLOW_MOD” event. The algorithm then finds only p_3 as an applicable path since it has a transition starting from state 1. If the algorithm receives the “ \downarrow BARRIER_REQUEST” message, the message is forwarded to the control channel without applying any fuzz operators, as there are no applicable transitions on p_3 , and is appended to the output sequence seq' .

After this iteration, the algorithm changes the current state of the EFSM M to state 6, since there is a transition from state 1 to 6 taking the “ \downarrow BARRIER_REQUEST” event. In this case, there are no applicable paths. If the algorithm receives the “ \downarrow BARRIER_REPLY” message, it sends the message back to the control channel without any modification and updates the output sequence seq' .

Since there is a transition from state 6 to 1 in the EFSM M , in the next iteration of the algorithm, the current state is set to state 1. Path p_3 is applicable in this situation. If the algorithm receives the “ \downarrow STATS_REQUEST” message, the corresponding transition on p_3 is identified by the algorithm. However, since there is no fuzz operator associated to the transition, the algorithm sends the message back to the control channel and updates the output sequence seq' .

In the next iteration of the algorithm, the current state of the EFSM M is set to state 2 by taking the transition from state 1 to 2 due to the “ \downarrow STATS_REQUEST” event. In the state, path p_3 is applicable. Note that, however, state 2 is the end state of path p_3 and is associated with the delay fuzz operator, which holds a message for 513ms and then sends it back to the control channel. This indicates that for any receiving message, the algorithm applies the delay fuzz operator. If the algorithm receives the “ \uparrow STATS_REPLY”, it applies the delay fuzz operator. From the subsequent iterations of the algorithm, no planned paths are applicable as p_3 has been exploited in its entirety. Hence, the algorithm simply forwards the receiving messages to the control channel and updates the output sequence seq' until the end of the test procedure execution. After executing the test procedure, we can obtain the following sequence of control messages, which leads to the SDN controller failing:

$$\begin{aligned}
 seq' = & \langle \uparrow \text{HELLO}, \downarrow \text{HELLO}, \downarrow \text{FEATURES_REQUEST}, \\
 & \uparrow \text{FEATURES_REPLY}, \downarrow \text{FLOW_MOD}, \downarrow \text{BARRIER_REQUEST}, \\
 & \uparrow \text{BARRIER_REPLY}, \downarrow \text{STATS_REQUEST}, \\
 & \uparrow \text{STATS_REPLY}[\text{delay 513ms}], \downarrow \text{ERROR}, \text{FAILURE} \rangle
 \end{aligned}$$

4 Evaluation

In this section, we empirically evaluate SeqFuzzSDN. Our complete evaluation package is available online [61].

4.1 Research Questions

RQ1 (comparison): *How does SeqFuzzSDN compare against other state-of-the-art fuzzing techniques for SDNs?* We investigate whether SeqFuzzSDN can outperform state-of-the-art testing techniques for SDNs, including DELTA [50], BEADS [40], and FuzzSDN [62]. We choose these techniques as they rely on fuzzing to test SDN controllers and their implementations are available online.

RQ2 (ablation study) *How does the sampling technique employed by SeqFuzzSDN influence its performance?* We assess the impact of the sampling technique (defined in Algorithm 3), which is our heuristic for sampling event traces to learn EFSMs. Specifically, we assess the impact of the technique in terms of execution time, the accuracy of EFSMs, and the diversity and coverage of the fuzzing results. To achieve this, we compare SeqFuzzSDN with its variant SeqFuzzSDN^{NS}, which does not sample event traces, and subsequently analyse the impact of the sampling algorithm.

RQ3 (scalability): *Can SeqFuzzSDN fuzz sequences of control messages and learn stateful failure-inducing models in practical time?* We investigate the correlation between SeqFuzzSDN’s execution time and network size. To do so, we carry out experiments involving SDNs of different network sizes.

4.2 Simulation Platform

To conduct large-scale experiments, we employ a simulation platform that emulates the physical networks. Specifically, we utilise Mininet [44] to create virtual networks of various sizes. Mininet leverages real-world SDN switch programs, resulting in emulated networks that closely match real-world SDNs. Hence, Mininet has been widely adopted in numerous SDN studies [40, 50, 62, 75].

We note that SeqFuzzSDN can also be applied to actual physical SDNs. However, assessing SeqFuzzSDN on actual physical networks through large-scale experiments, such as the ones reported in this article, is prohibitively expensive in terms of both cost and time.

Our experiments were conducted on 10 virtual machines, each equipped with 4 CPUs and 10GB of RAM. Each experiment was conducted with a time budget of 5 days for ONOS and 3 days for RYU. We note that, within this budget, the sensitivity values of the EFSMs generated by SeqFuzzSDN reach their plateaus. Due to the randomness of SeqFuzzSDN, we repeated our experiments 10 times. These experiments took approximately 60 days of concurrent execution on the 10 virtual machines.

4.3 Study Subject

We evaluate SeqFuzzSDN by testing two open-source and actively maintained SDN controllers, ONOS [7] and RYU [72], both of which are still widely used in SDN studies [40, 49–51, 62, 77, 89]. Both controllers’ implementations are based on the OpenFlow SDN protocol specification. SeqFuzzSDN, which fuzzes OpenFlow control messages, is therefore capable of testing any SDN controller that adheres to the OpenFlow specification.

For our evaluation, we created five virtual networks with 1, 2, 4, 8, and 16 switches respectively. Each network is managed by either ONOS or RYU. In each network, the switches possess emulated physical connections with all the other switches, forming a fully connected topology. Each switch is connected to two hosts, simulating devices that transmit and receive data, such as video and audio streams.

We note that the study subjects, comprising of 5×2 synthetic systems built on the five networks managed by ONOS and RYU, are representative of both existing SDN studies and real-world SDNs. For instance, in prior SDN studies testing ONOS and RYU, DELTA was evaluated using an SDN with two switches, BEADS was evaluated using an SDN with three switches, and FuzzSDN was evaluated on SDNs with 1, 3, 5, 7, and 9 switches, due to the significant computational resources required for conducting experiments with SDNs.

4.4 Experimental Setup

EXP1. To answer RQ1, we conduct a comparative analysis of SeqFuzzSDN with three other SDN testing tools: FuzzSDN [62], DELTA [50], and BEADS [40]. FuzzSDN is a testing framework that generates rule-based failure-inducing models and test cases. FuzzSDN employs a grammar-based machine learning-guided fuzzing technique, which enables it to progressively refine the generated failure-inducing models, offering interpretable models that describe the conditions leading to a failure. DELTA is a security framework designed for SDNs that allows engineers to automatically replicate established attack scenarios associated with SDNs and uncover new attack scenarios through fuzzing. DELTA accomplishes this by changing control messages, employing a fuzzing technique that randomises the control message byte stream, regardless of the OpenFlow protocol specificities. Lastly, BEADS is an automated attack discovery technique that relies on a range of

mutation (fuzz) operators, with the aim of discovering attack scenarios. BEADS also fuzzes control messages but employs strategies such as message dropping, duplication, delay, and modification while adhering to the OpenFlow specification. This allows BEADS to generate fuzzed control messages that can pass beyond the message parsing layer of the system under test.

To compare SeqFuzzSDN with these three SDN testing tools, we create three baselines: FuzzSDN^E, DELTA^E and BEADS^E. These baselines extend FuzzSDN, DELTA and BEADS respectively, to infer EFSMs, as the original testing tools do not produce EFSMs as part of their test outputs. FuzzSDN^E (resp. DELTA^E and BEADS^E) encodes the fuzzed control messages and the test output (i.e., success and failure) as a dataset to infer EFSMs. The baselines then use MINT to generate EFSMs. Unlike SeqFuzzSDN, FuzzSDN^E, DELTA^E, and BEADS^E do not leverage the generated EFSM to guide their fuzzing operations.

We use two synthetic systems, each with a single switch, controlled by either ONOS or RYU. We leverage a test procedure (see Section 3) that specifies a pairwise ping test [12], which has been used in many SDN studies [26, 40, 50, 58, 62]. This test procedure is important as it enables practitioners to verify communication between hosts, measure latency, detect packet loss, and identify routing issues. For the failure detection mechanism, we identify spurious switch disconnections. In our experiments, we identify switch disconnections that lead to communication breakdowns as failures. These failures cannot be localised using stack traces to pinpoint the causes of the failures.

In our comparison, we count the number of failures observed during the execution of SeqFuzzSDN and the baselines. In addition, from the final EFSMs inferred by the four tools, we measure the number of unique loop-free paths (corresponding to message sequences) that lead to failures. This allows us to assess how many distinct failure-inducing sequences of state changes are captured in the EFSMs. To further compare the four tools, we analyse the sensitivity of each EFSM, calculated using the formula:

$$\text{sensitivity} = \frac{\#\text{accepted}}{\#\text{accepted} + \#\text{rejected}}$$

where `#accepted` and `#rejected` are the number of traces accepted and rejected by the EFSM, respectively. In our context, an EFSM with high sensitivity is desirable as it is less likely to miss possible failure-inducing sequences of control messages. To fairly calculate sensitivity, we elected to create a dataset that maintains a balanced representation of success and failure traces across all tools, thereby reducing potential biases toward a specific tool. To do so, we created a test dataset containing 800 fuzzing results, with an equal split of 400 success traces and 400 failure traces. These fuzzing results were randomly sampled from separate runs of SeqFuzzSDN, FuzzSDN, DELTA, and BEADS, with each tool contributing 200 results, evenly divided into 100 success traces and 100 failure traces. We note that other commonly used evaluation metrics, such as precision and F1-score, are not applicable in our context because an EFSM is built from fuzzed message sequences that are not generated during normal operations of the SDN controller under test. Hence, our datasets lack negative traces—message sequences that should not be produced by the SDN controller—since they are derived from fuzzed message sequences, making it impossible to compute precision and F1-score.

Additionally, we measure the diversity of fuzzed message sequences obtained from the four tools using the Normalised Compression Distance (NCD) for multisets [19]. Recall from Section 3 that the fuzzed message sequences vary in length, message types, and message values, making the application of simple sequence comparison metrics difficult. In our context, a high NCD value indicates that the fuzzed sequences of control messages (i.e., tests) are diverse, reducing the likelihood of redundancy or overly similar tests.

EXP2. To answer RQ2, we compare SeqFuzzSDN to its variant, named SeqFuzzSDN^{NS}. At each learning step, instead of using the sampling technique (see Algorithm 3), SeqFuzzSDN^{NS} uses all the collected event traces to infer an EFSM.

In this experiment, we use the same synthetic systems as those used in EXP1. Our test procedure specifies a pairwise ping test, and our failure detection mechanism identifies unexpected communication breakdowns. This experiment counts the number of iterations of the fuzzing, learning, and planning steps within the time budget and measures the execution time of each step. In addition, we compare SeqFuzzSDN and SeqFuzzSDN^{NS} by measuring the sensitivity of the final EFSMs obtained after the time budget expires. To ensure fair comparisons between SeqFuzzSDN and SeqFuzzSDN^{NS}, we created a test dataset comprising 1000 fuzzing results, evenly split into 500 success traces and 500 failure traces. These results were obtained from separate runs of SeqFuzzSDN, SeqFuzzSDN^{NS}, FuzzSDN^E, DELTA^E, and BEADS^E, with each method contributing 200 results, evenly split into 100 success traces and 100 failure traces. Therefore, this test dataset is not biased toward either SeqFuzzSDN or SeqFuzzSDN^{NS}. Furthermore, we measure the coverage and diversity degrees (defined in Section 3.4.2) of the planned paths (corresponding to message sequences) obtained at the last iteration, allowing us to assess the effectiveness of the EFSMs in generating message sequences that cover diverse states.

EXP3. To answer RQ3, we investigate the correlation between the resource consumption of SeqFuzzSDN and the size of the five synthetic systems described in Section 4.3, each with 1, 2, 4, 8, and 16 switches, controlled by either ONOS or RYU. For this experiment, we use a test procedure that implements the pairwise ping test, similar to EXP1 and EXP2. Compared to EXP1 and EXP2, the sequences of control messages produced by the test procedure in EXP3 differ significantly in terms of their lengths. This is due to the fully connected topology in EXP3, which includes multiple switches. Moreover, when there are more than two switches, the topology introduces switching loops [65], further increasing the number of events in a trace. We measure the time required to configure Mininet and the SDN controller, perform the test procedure, and execute each step of SeqFuzzSDN (i.e., fuzzing, learning, and planning).

4.5 Parameter Setting

As described in Section 3, SeqFuzzSDN takes as input parameters that can be tuned to improve its efficiency and effectiveness. For clarity and reproducibility, this section provides all the parameter values and describes how we set them. We note that, given the extremely long execution time required for applying automated hyperparameter optimisation techniques in our context, we manually set some of the parameters as described below.

In the learning step, the parameters to be tuned are those of the sampling technique (Algorithm 3) and MINT [82]. For the sampling technique, we set the number (n_{ts}) of event traces to 1000, limiting the maximum size of the dataset used by MINT. This configuration allowed SeqFuzzSDN to generate EFSMs in practical time (approximately 100 minutes). For MINT, we configured the parameter values of RIPPER [20] as follows: three folds, a minimal weight of 2.0, and two optimisation runs as specified by the default setting in WEKA [85].

In the planning step, we set the size of a candidate solution (n_s) to 200 in order to match the number of test procedure executions to be performed in each iteration of SeqFuzzSDN. This ensures that a candidate solution contains the 200 traces to be followed during the 200 executions of the test procedure. We set the candidate solution fuzzing probability (μ_f) to 0.5, as we want to strike a balance between exploitation and exploration of the generated EFSM. The crossover probability (μ_c) and the mutation probability (μ_m) in the planning step were set to 0.8 and 0.02, respectively, following published guidelines. The size of the population and archive, n_p , is set to 100 and the

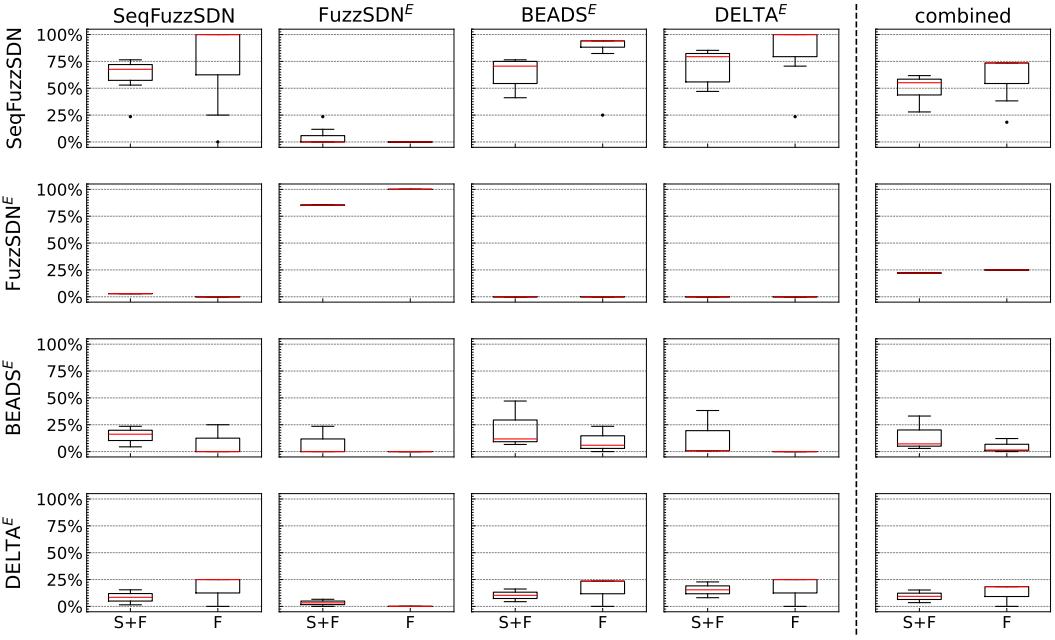


Fig. 7. Comparing the sensitivity of the EFSMs generated by SeqFuzzSDN, FuzzSDN^E , BEADS^E , and DELTA^E , the five plots in each row display the sensitivity of the corresponding tool. The first four columns represent the sensitivity of the EFSMs assessed using the test dataset containing message sequences generated by each tool. Sensitivity is assessed using message sequences that lead to both success and failure, denoted by (S+F), and only failure, denoted by (F). The last column represents the sensitivity assessed using all datasets generated by the four tools. The boxplots (25%-50%-75%) show the distribution of sensitivity over 10 runs of each tool.

search generates 50 populations, allowing the planning step to complete within a reasonable time (on average, 79 minutes for our ONOS study subject, and 44 minutes for our RYU study subject).

The remaining parameters were tuned using hyperparameter optimisation [85], following guidelines from the literature [39, 85]. We evaluated 10 different configurations of SeqFuzzSDN using grid search [85]. As a result of this optimisation process, we set the remaining parameters as follows: the initial fuzzing probability (μ) of a message is 0.3, the minimum merging score (k) of MINT is 1, and the number (n_k) of shortest paths used during the generation of a candidate solution is 15.

The parameters of SeqFuzzSDN used in our experiments could be further refined to improve efficiency and effectiveness. However, the configuration we chose produced results that are satisfactory to support our findings. As a result, we have not included additional experiments aimed at optimising these parameters in this article.

4.6 Experiment Results

To answer the research questions, we assessed the results obtained from both the ONOS and RYU subjects. Since the findings from the ONOS results are consistent with those from the RYU results, this section presents only the ONOS results for brevity. Note that the results for our RYU study subject are presented in Appendix A.

RQ1. Figure 7 compares the sensitivity of the EFSMs measured using the test dataset, which contains message sequences and their test results obtained from the four tools: SeqFuzzSDN,

FuzzSDN^E , BEADS^E , and DELTA^E . The last column of the first row in the figure shows that, when evaluating all the message sequences produced by these tools, on average, SeqFuzzSDN achieves a sensitivity of 49.89% on the message sequences leading to both success and failure (referred to as the combined S+F dataset) and 60.19% on the message sequences leading only to failure (referred to as the combined F dataset). For brevity, we refer to datasets containing message sequences generated by each tool that result in both success and failure as the [tool] S+F dataset and those that result only in failure as the [tool] F dataset. Specifically, as shown in the first row of the figure, SeqFuzzSDN achieves, on average, a sensitivity of 67.1% on the SeqFuzzSDN S+F dataset and 92.3% on the SeqFuzzSDN F dataset, 0.23% on the FuzzSDN^E S+F dataset and 0.00% on the FuzzSDN^E F dataset, 71.27% on the BEADS^E S+F dataset and 86.88% on the BEADS^E F dataset, and 73.30% on the DELTA^E S+F dataset and 89.59% on the DELTA^E F dataset.

For FuzzSDN^E , BEADS^E , and DELTA^E , respectively, the figure (the last column of the 2nd, 3rd, and 4th rows) shows that their EFSMs' sensitivities are, on average, 22.06%, 14.40%, and 9.38% on the combined S+F dataset, and 25.00%, 4.53%, and 12.25% on the combined F dataset. Specifically, as shown in the first column of the figure, starting from the 2nd row, using the SeqFuzzSDN S+F dataset (and the SeqFuzzSDN F dataset), FuzzSDN^E , BEADS^E , and DELTA^E achieve, respectively, on average, sensitivities of 0.84%, 4.78%, and 3.68% (and 0.0%, 0.0%, and 4.55%). Regarding the FuzzSDN^E S+F dataset (and the FuzzSDN^E F dataset), as shown in the 2nd column of the figure, FuzzSDN^E , BEADS^E , and DELTA^E achieve, respectively, on average, sensitivities of 54.62%, 1.35%, and 0.00% (and 66.39%, 0.00%, and 0.00%). For the BEADS^E S+F dataset (and the BEADS^E F dataset), shown in the 3rd column, these three baselines achieve, respectively, on average, sensitivities of 0.00%, 6.37%, and 5.75% (and 0.00%, 0.00%, and 4.28%). Lastly, when using the DELTA^E S+F dataset (and the DELTA^E F dataset), these baselines achieve, respectively, on average, sensitivities of 0.00%, 5.64%, and 8.69% (and 0.00%, 0.00%, and 4.01%).

These results show that SeqFuzzSDN achieves, on average, a higher sensitivity compared to the baselines, and the differences are statistically significant. However, note that the EFSM produced by SeqFuzzSDN rejects most of the failure-inducing message sequences obtained from FuzzSDN^E , as SeqFuzzSDN and FuzzSDN^E use significantly different fuzzing methods. While FuzzSDN^E fuzzes a single message by modifying its fields' values, SeqFuzzSDN fuzzes a sequence of messages using multiple fuzz operators (i.e., delay, modification, duplication, deletion, and insertion). Consequently, the message sequences that lead to failure are significantly different between the two tools, resulting in producing very different EFSMs, which cannot accept the message sequences generated by the other tool. Even when the same failures are triggered, the generated traces differ due to these distinct paths. However, recall that the EFSM produced by FuzzSDN^E rejects most of the message sequences generated by SeqFuzzSDN, BEADS^E , and DELTA^E , indicating that the EFSMs are specific only to FuzzSDN^E .

Figure 8 compares (a) the NCD scores of the message sequences, (b) the number of unique failure-inducing paths in the EFSMs, and (c) the number of message sequences leading to failure, which are obtained from 10 runs of SeqFuzzSDN, FuzzSDN^E , BEADS^E , and DELTA^E . Figure 8a shows that SeqFuzzSDN achieves a higher NCD score, with an average of 0.99, compared to those of the baselines. Figure 8b shows that, on average, SeqFuzzSDN was able to infer an EFSM containing 18 unique loop-free paths that lead to failure, which is significantly higher than the others. From these results, we found that SeqFuzzSDN generates more diverse sequences of control messages that exercise a larger number of state changes compared to the baselines.

However, Figure 8c shows that FuzzSDN^E generates a larger number of message sequences (an average of 265) leading to failure compared to the other tools, while SeqFuzzSDN generates, on average, 140 message sequences leading to failure, thus outperforming BEADS^E and DELTA^E . Even though FuzzSDN^E outperforms SeqFuzzSDN in terms of number of failures, recall from Figure 8a

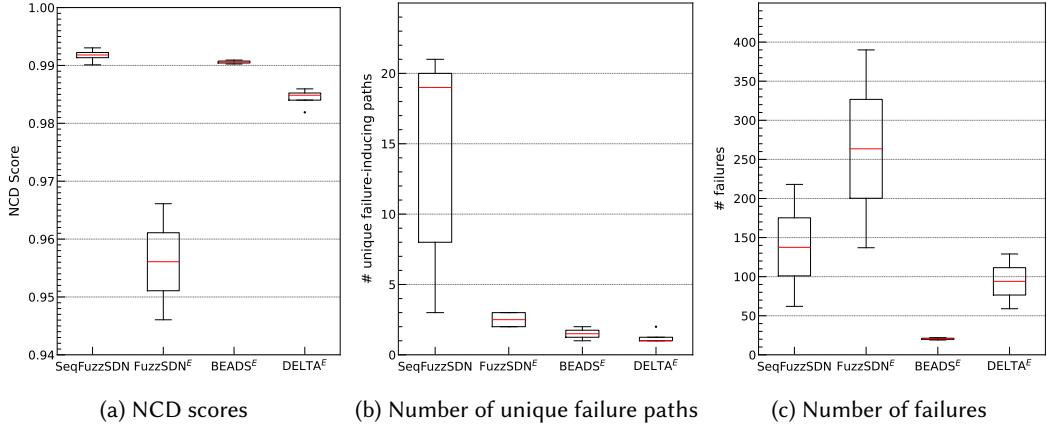


Fig. 8. Comparing (a) the NCD scores of the message sequences, (b) the number of unique failure-inducing paths in the EFSMs, and (c) the number of message sequences leading to failure, all obtained from SeqFuzzSDN, $FuzzSDN^E$, $BEADS^E$, and $DELTA^E$. The boxplots (25%-50%-75%) show the distribution of each metric over 10 runs of each tool.

Table 3. Failure-inducing message sequences discovered by SeqFuzzSDN in EXP1 and whether they were found by existing tools: $FuzzSDN$ and $BEADS$.

ID	Failure-inducing message sequences	$FuzzSDN$	$BEADS$
ID1	Modification of FEATURE_REQUEST or FEATURE_REPLY	Yes	Yes
ID2	Modification of *_STATS_REPLY of *_STATS_REQUEST	No	Yes
ID3	Removal of a non-existing flow	No	No
ID4	Insertion of PACKET_IN after DESC_STATS_REPLY	No	No
ID5	Duplication of ERROR after GET_CONFIG_REQUEST	No	No
ID6	Deletion of GET_CONFIG_REPLY	No	No
ID7	Insertion of PACKET_IN after a flow removed	No	No
ID8	Modification of BARRIER_REQUEST or BARRIER_REPLY	Yes	Yes
ID9	Modification of ECHO_REQUEST or ECHO_REPLY	Yes	No
ID10	Insertion of PACKET_IN after deletion of PORT_STATS_REQUEST	No	No
ID11	Insertion of PACKET_IN after BARRIER_REPLY	No	No
ID12	Duplication of handshake messages	No	Yes
ID13	Delay or deletion of *_STATS_REQUEST of *_STATS_REPLY	No	No
ID14	Duplication of FLOW_MOD	No	No

and Figure 8b that $FuzzSDN^E$ generates message sequences that are less diverse and exercise significantly fewer number of state changes compared to SeqFuzzSDN. Furthermore, as described in Section 3, SeqFuzzSDN aims to generate a balanced number of message sequences that lead to success and failure, rather than focusing solely on the latter.

In addition to the metric-based comparison described above, we reviewed the failure-inducing message sequences produced by SeqFuzzSDN in EXP1 and assessed whether SeqFuzzSDN could discover new failure-inducing sequences compared to existing studies. Specifically, we manually inspected the failure-inducing sequences obtained from five runs of EXP1 (approximately 800 sequences) and categorised them based on their unique characteristics that contribute to failures into 14 cases. Table 3 presents these 14 cases of failure-inducing message sequences generated

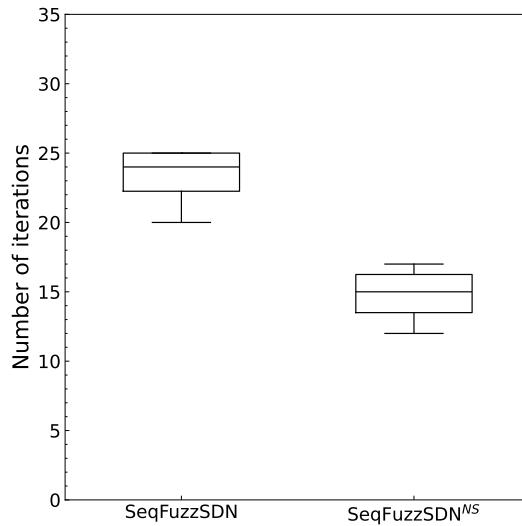


Fig. 9. Comparing the number of iterations completed by SeqFuzzSDN and SeqFuzzSDN^{NS} within a 5-day time budget. The boxplots (25%-50%-75%) show the distribution of iteration counts over 10 runs of each tool.

by SeqFuzzSDN in EXP1, comparing them with those reported by FuzzSDN and BEADS. The first column represents the class ID, the second column describes the characteristics of failure-inducing message sequences, and the third and fourth columns indicate whether or not the corresponding sequence class was identified by FuzzSDN and BEADS, respectively. For example, class ID1 refers to message sequences that modify the FEATURE_REQUEST or FEATURE_REPLY messages. For ID1, both FuzzSDN and BEADS were able to discover the failure-inducing case. As another example, class ID10 refers to message sequences that insert the PACKET_IN message after deleting the PORT_STATS_REQUEST message. In contrast to ID1, neither FuzzSDN nor BEADS were able to discover this failure-inducing case. As shown in the table, ID1 and ID8 were reported in both FuzzSDN and BEADS. ID2 and ID12 were reported in BEADS but not in FuzzSDN. ID9 was reported in FuzzSDN but not in BEADS. IDs 3, 4, 5, 6, 7, 10, 11, 13, and 14 were not reported in either FuzzSDN or BEADS. Hence, the results show that SeqFuzzSDN is effective in identifying new types of failure-inducing message sequences compared to prior work.

The answer to RQ1 is that SeqFuzzSDN significantly outperforms the baselines that extend FuzzSDN, BEADS, and DELTA. In particular, our experiment results indicate that SeqFuzzSDN can generate more diverse sequences of control messages leading to failure than those obtained from the baselines, while also providing EFSMs that accurately capture failure-inducing message sequences.

RQ2. Figure 9 presents a comparison of the number of iterations for the fuzzing, learning, and planning steps completed by SeqFuzzSDN and SeqFuzzSDN^{NS} within a time budget of 5 days. The boxplots show the distributions (25%-50%-75% quantiles) of the number iterations performed by SeqFuzzSDN and SeqFuzzSDN^{NS}, obtained from 10 runs of EXP2. As shown in the figure, SeqFuzzSDN can execute significantly more iterations than SeqFuzzSDN^{NS}. For a time budget of 5 days, SeqFuzzSDN completes, on average, 25 iterations, while SeqFuzzSDN^{NS} completes approximately 15 iterations. This result indicates that the sampling technique, which caps the

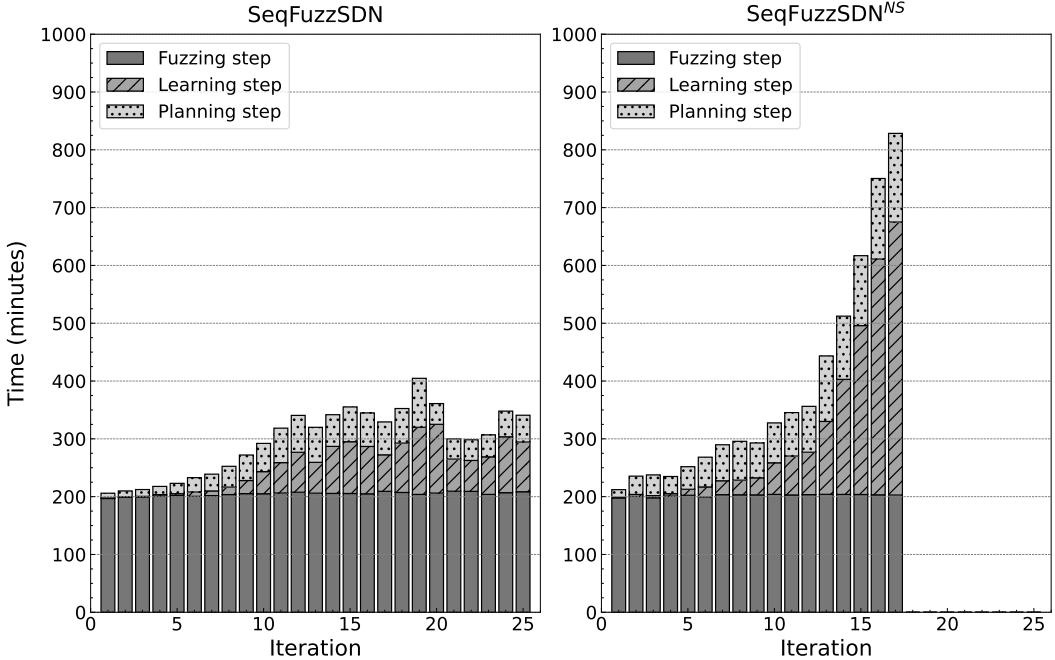


Fig. 10. Comparing the execution time per iteration for the fuzzing, learning, and planning steps of SeqFuzzSDN and SeqFuzzSDN^{NS} within a 5-day time budget. The execution times shown in this figure are the average values observed over 10 runs of EXP2.

maximum size of the dataset for MINT, allows SeqFuzzSDN to complete more iterations within the same time frame. In contrast, SeqFuzzSDN^{NS}, which permits the dataset to grow continuously over iterations, completes fewer iterations. Note that each iteration of SeqFuzzSDN (and SeqFuzzSDN^{NS}) tests the SDN controller 200 times; hence, the sampling technique enables SeqFuzzSDN to test the SDN controller, on average, 1000 times more than SeqFuzzSDN^{NS}.

In addition, Figure 10 compares SeqFuzzSDN and SeqFuzzSDN^{NS} with regard to the execution times per iteration for the fuzzing, learning, and planning steps over a time budget of 5 days. The bar graph shows the average execution times taken by SeqFuzzSDN and SeqFuzzSDN^{NS} for the fuzzing, learning, and planning steps at each iteration, based on 10 runs of EXP2.

The results show that the fuzzing time per iteration remains constant at around 200 minutes for both SeqFuzzSDN and SeqFuzzSDN^{NS}, indicating that the fuzzing step is independent of the tool used. Recall from Section 4.4 that EXP2 uses the pairwise ping test procedure, which is executed at each iteration during fuzzing and does not introduce variance in execution time over iterations. For the planning step, Figure 10 shows that the planning time does not exceed 150 minutes in both SeqFuzzSDN and SeqFuzzSDN^{NS}. Figure 10 also suggests that, for SeqFuzzSDN^{NS}, the time required to learn an EFSM increases exponentially with each iteration due to the growing size of the dataset fed to MINT. Furthermore, we observe that, on the 17th iteration of SeqFuzzSDN^{NS}, the learning time reaches the 12-hour timeout limit, thus preventing SeqFuzzSDN^{NS} from completing any further iterations. This finding aligns with the literature [30, 74, 83], as inferring EFSMs is a complex problem that scales poorly with larger input sizes. In contrast, the results for SeqFuzzSDN indicate that the time required for inferring an EFSM (i.e., the learning step) remains below 115 minutes due to the application of the sampling technique. Thus, based on the results shown in

Table 4. Statistical significance analysis using the Wilcoxon Rank-Sum test for sensitivity, diversity, and coverage results obtained from 10 runs of EXP2.

Metric	Average (SeqFuzzSDN)	Average (SeqFuzzSDN ^{NS})	p-value	Statistical Significance ($\alpha = 0.05$)
Sensitivity	0.542	0.529	0.571	Not Significant
Diversity	0.9925	0.9920	0.297	Not Significant
Coverage	0.5533	0.6599	0.0124	Significant

Figure 10, we can further conclude that applying the sampling technique enables SeqFuzzSDN to overcome the scalability issues associated with the complexity of learning EFSMs.

Furthermore, Table 4 presents the statistical test results for the distributions of sensitivity, diversity, and coverage (described in Section 3) achieved by SeqFuzzSDN and SeqFuzzSDN^{NS} after 10 runs of EXP2, using the Wilcoxon Rank-Sum test [38] with an α value of 0.05. On average, SeqFuzzSDN (resp. SeqFuzzSDN^{NS}) achieves a sensitivity of 54.2% (resp. 52.9%), a diversity of 0.9925 (resp. 0.9920), and a coverage of 0.5533 (resp. 0.6599). We observed that the differences in sensitivity ($p = 0.14$) and diversity ($p = 0.9$) are not significant, while the difference in coverage ($p = 0.01$) is. The results indicate that the use of the sampling technique does not negatively impact the sensitivity of the generated EFSMs nor the diversity of the generated message sequences. However, the coverage achieved by SeqFuzzSDN has significantly improved, suggesting that the states in the EFSM are explored more thoroughly. One possible explanation for the improved coverage is that the increased number of iterations gives SeqFuzzSDN more opportunities to refine EFSMs with respect to the coverage objective targeted at the planning step.

The answer to RQ2 is that the sampling technique introduced in SeqFuzzSDN reduces its computation cost, allowing for more iterations to be performed within a given time budget. This helps overcome scalability issues in inferring EFSMs without compromising the accuracy of the EFSMs and the diversity of the generated message sequences. Additionally, the sampling technique significantly improves SeqFuzzSDN’s coverage, leading to a more thorough exploration of the search space.

RQ3. Figure 11 presents the distributions of execution times (25%-50%-75% boxplots) for the fuzzing, learning, and planning steps of SeqFuzzSDN. These execution times were measured using the five study subjects in EXP3, which consist of 1, 2, 4, 8, and 16 switches controlled by ONOS. As shown in Figure 11, the execution time taken for the fuzzing step is, on average, 203 minutes for the 1-switch configuration, 215 minutes for 2 switches, 235 minutes for 4 switches, 257 minutes for 8 switches, and 274 minutes for 16 switches. The learning step took, on average, 15 minutes for the 1-switch configuration, 14 minutes for 2 switches, 11 minutes for 4 switches, 25 minutes for 8 switches, and 26 minutes for 16 switches. The planning step took, on average, 79 minutes for the 1-switch configuration, 69 minutes for 2 switches, 56 minutes for 4 switches, 56 minutes for 8 switches, and 76 minutes for 16 switches.

The results show that there is no significant difference in the times required for the learning and planning steps across the five study subjects. However, the only time increase occurs during the fuzzing step, where test procedures are executed. This includes the time required to configure and teardown Mininet and the SDN controller. This increasing trend aligns with our expectations, as the execution time for a test procedure increases with its complexity. As described in Section 4.4,

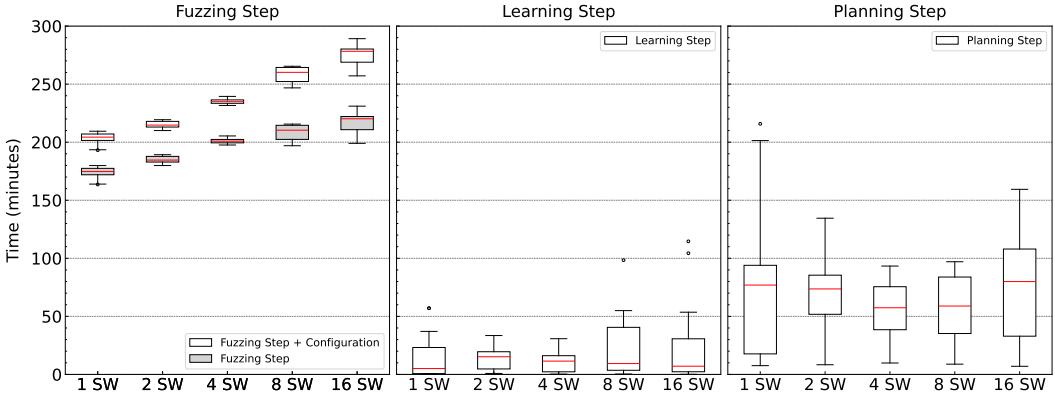


Fig. 11. Boxplots (25%-50%-75%) representing the distributions of time taken in minutes for the fuzzing, learning, and planning steps of SeqFuzzSDN. This figure includes the times observed over 10 runs of SeqFuzzSDN with 1, 2, 4, 8, and 16 switch configurations.

this is primarily due to the increasing number of messages exchanged between the switches and the controller as the number of switches and their connections grows [65]. This increase in time is independent of SeqFuzzSDN, as it solely depends on the complexity of the test procedures executed. Note that when 16 switches are fully connected, the pairwise ping test procedure produces, on average, 30.73 control messages (min: 1, max: 8178). With 8 switches, the average number of control messages decreases to 25.19 (min: 1, max: 4863), while 4 switches produce an average of 16.16 control messages (min: 1, max: 1004). When only 2 switches are connected, the test procedure generates, on average, 12.83 control messages (min: 1, max: 86), and with just 1 switch, the average is further reduced to 10.46 control messages (min: 1, max: 85). The pairwise ping test procedure produces 35 unique types of control messages. However, SeqFuzzSDN halts the execution of the test procedure upon detecting a failure. As a result, the total number of control messages recorded in our experiments could be lower than 35, depending on when failures occur.

The answer to RQ3 is that the primary factor affecting the execution time of SeqFuzzSDN is its fuzzing time, which is influenced by the number of control messages generated by a test procedure. Consequently, SeqFuzzSDN is applicable to complex systems with large networks, provided that the execution time of a test procedure remains within an acceptable time budget.

4.7 Threats to Validity

To address potential threats to internal validity, we compared SeqFuzzSDN against three SOTA tools (DELTA, BEADS, and FuzzSDN), which have been used to generate failure-inducing control messages for testing SDN controllers. However, DELTA, BEADS, and FuzzSDN do not generate failure-inducing models that consider the sequences of messages exchanged between the controller and switches. Consequently, we extended these tools as baselines to produce EFSMs, allowing for a comparative analysis between SeqFuzzSDN and these baselines.

The principal external validity threat to SeqFuzzSDN is the risk that it may not be applicable to different contexts, such as other SDN systems with different switch configurations or controllers. To address this potential threat, we conducted experiments with SeqFuzzSDN against multiple SDNs and two popular SDN controllers found in the literature, namely ONOS and RYU. We varied

our synthetic systems, which comprise five networks with 1, 2, 4, 8, and 16 switches, respectively, managed by either ONOS or RYU.

Additionally, the prototype implementation of SeqFuzzSDN is compatible with OpenFlow, a widely accepted standard protocol for SDNs, which has been used in numerous SDN studies and practices [40, 49–51, 58, 77]. As a consequence, we were able to successfully apply SeqFuzzSDN to real-world SDN controllers (ONOS and RYU) and compare it to existing tools (i.e., DELTA, BEADS, and FuzzSDN), considering their support for OpenFlow. However, to further explore the applicability of our findings, it is essential to conduct additional case studies in various settings. This includes industrial systems that use different SDN protocols and user studies that involve practitioners.

5 Discussion

In this section, we discuss the primary challenges faced by SeqFuzzSDN and considerations for further improvement, including testing time, testing scope, information loss due to sampling, and generalisation to other controllers. In addition, we provide practical guidance on adapting SeqFuzzSDN for other systems

Testing time. SeqFuzzSDN executes an SDN controller for testing, which requires initialisation of the controller, network simulation, and teardown overhead. These required operations are essential for testing the SDN controller in a realistic context. However, further research on optimising these operations, such as reducing initialisation and teardown times and developing lightweight network simulations, is needed to enhance efficiency and scalability.

Testing scope. SeqFuzzSDN takes as input a test procedure and a failure detection mechanism. Hence, the testing scope is limited by these inputs. For example, a pairwise ping test is unlikely to exercise components of the controller responsible for handling backup functionality. Automatically exploring possible use scenarios (i.e., test procedures) and defining corresponding test oracles (i.e., failure detection mechanisms) help engineers reduce their manual efforts in creating them. However, efficiently and effectively exploring the space of test procedures and defining oracles remain hard problems.

Information loss. SeqFuzzSDN employs a sampling technique to use MINT, a model inference tool, in a scalable manner. However, sampling sequences from all recorded message sequences inherently leads to information loss, even if our approach attempts to minimise such loss, as described in Section 3.3.3. To fully address this issue, it is highly desirable to develop a scalable model inference technique capable of handling large volumes of message sequences.

Application to other systems. Although SeqFuzzSDN was evaluated with existing fuzzing tools and SDN controllers that rely on OpenFlow, it may also be applied to other SDN controllers or even other network systems. To facilitate such use cases, we provide practical guidance on applying and adapting SeqFuzzSDN to other systems. First, to utilise SeqFuzzSDN with systems that incorporate other SDN protocols, such as Cisco OpFlex [78] and ForCES [36], it is necessary to modify the sniffing and injection mechanisms of SeqFuzzSDN to decode and encode control messages. However, these modifications do not affect the fuzzing, learning, and planning steps. Therefore, we anticipate that, although such modification necessitates engineering effort to revise the sniffing and injection mechanisms, they are unlikely to impact SeqFuzzSDN’s efficiency and effectiveness. Second, SeqFuzzSDN may also inspire applications for testing network servers. In such use cases, adapting SeqFuzzSDN will require modifying components related to SDNs, such as parsing and modifying control messages and simulating SDN communications. In addition, recall from Section 3 that SeqFuzzSDN uses message fields to define EFSM guards and to fuzz messages. This implies that SeqFuzzSDN is applicable only to network systems where the structure of message fields is known.

6 Related Works

In this section, we discuss related works in the areas of SDN testing, fuzzing, and characterising failure-inducing inputs. Readers familiar with our previous work [62] may notice significant similarities. This is because this work builds upon, and extends our previous research. As such, much of the foundational literature and related work remain relevant and are thus referenced here. We believe this will provide a comprehensive context for both new readers and those familiar with our prior work.

SDN testing. The study of SDN testing in the networking literature focuses on various objectives, such as detecting security vulnerabilities and attacks [4, 10, 17, 40, 50, 59, 89], identifying inconsistencies among the SDN components (i.e., applications, controllers, and switches) [49, 52, 77], and analysing SDN executions [29, 58, 79]. In this discussion, we focus on SDN testing methods that utilise fuzzing, as they are the most relevant to our research. Lee et al. [48, 49] proposed AUDI-SDN, a framework that employs a fuzzing technique to detect policy inconsistencies among SDN components (i.e., controllers and switches). AUDI-SDN relies on the fuzzing of network policies configured by the administrators through the REST APIs of the SDN components. To increase the probability of uncovering inconsistencies, AUDI-SDN restricts valid relationship elements by building rule dependency trees from the specification of the OpenFlow protocol. RE-CHECKER, proposed by Woo et al. [86], is designed to fuzz the RESTful services offered by SDN controllers. It fuzzes an input file in JSON format, which is used by a network administrator to define network policies, such as data forwarding rules. This process generates a large number of malformed REST messages for testing RESTful services in SDN. Dixit et al. [27] introduced AIM-SDN to test the implementation of the Network Management Datastore Architecture (NMDA) in SDN. AIM-SDN uses random fuzzing of REST messages to test the NMDA implementation in SDN, focusing on the availability, integrity, and confidentiality of datastores. Shukla et al. [77] created PAZZ, which is designed to identify faults in SDN switches by fuzzing data packet headers, such as IPv4 and IPv6 headers. Finally, Albab et al. [2] introduced SWITCHV to verify the behaviours of SDN switches. SWITCHV employs fuzzing and symbolic execution to analyse the p4 models that define the behaviours of SDN switches. In contrast to these methods, SeqFuzzSDN fuzzes SDN control messages to test SDN controllers, similar to DELTA, BEADS, and FuzzSDN. Furthermore, SeqFuzzSDN uses learned EFSMs to guide the fuzzing process and characterise the messages sequences that may lead to a system failure.

Fuzzing and Stateful Testing. To efficiently generate effective test data, fuzzing has been widely applied in many application domains [57]. The research strands that most closely relate to ours are stateful fuzzing techniques [5, 33, 60, 66]. Numerous research studies have explored the use of FSMs for testing complex systems. Gascon et al. [33] proposed PULSAR, a stateful black-box fuzzing technique aimed at discovering vulnerabilities in proprietary network protocols. Their proposed approach involves the inference of a Markov model from network traces, which are used to generate test cases using fuzzing primitives (i.e., paths in the Markov model), and finally the selection of the test cases that maximise the coverage of the protocol stack. Pham et al. [66] proposed AFLNET, a grey-box fuzzer for network protocols implementation, based on AFL [87]. Their proposed technique takes a mutational approach and states feedback to guide the fuzzing of network-enabled servers. AFLNET takes as input a corpus of server-client network interactions and subsequently acts as a client. It replays modified versions of the initial message sequence sent to the server, preserving only the alterations that successfully expanded the coverage of the code or state space. From the newly discovered message sequences, AFLNET uses the server's response codes to build an FSM that describes the protocol states. From those inferred FSMs, their approach identifies regions in the state space that have been the least explored and systematically steers the fuzzing

process towards the test of such regions. Natella [60] proposed STATEAFL, a grey-box fuzzing techniques that infers FSMs based on the in-memory states of a server, leveraging compile-time instrumentation and fuzzy hashing techniques; hence, it does not require response codes. During the fuzzing process, STATEAFL guides the generation of new inputs to the server based on the inferred FSMs. It employs both byte-level and message-level fuzz operators, which do not rely on protocol specifications. Kim et al. [42] proposed AMBUSHER, a protocol-state-aware fuzzing technique for testing the “*East-West*” protocol of distributed SDN controllers. AMBUSHER takes as input a test configuration which includes the alphabet of the protocol used as well as the cluster information. In its first phase, AMBUSHER uses a dummy network node to generate queries between the controllers, and a dummy controller to log such queries generated in the network. In its second phase, the logged cluster queries are then used by a FSM learner to infer a FSM of the cluster’s protocol. In its third phase, AMBUSHER explores the inferred FSM to extract message sequences. Those message sequences are then used as seeds for the fuzzing process, in which attack scenarios are generated by randomising the message sequences. In its final phase, AMBUSHER leverages the randomised sequences to test the cluster “*East-West*” interfaces. Among these, AMBUSHER is the most relevant to SeqFuzzSDN, as both take into account the SDN architecture, which differs from the server-client architecture. Compared to AMBUSHER, SeqFuzzSDN fuzzes and infers EFSMs based on sequences of control messages exchanged through the control channel of the SDN (i.e., “*South*” interface). To our knowledge, SeqFuzzSDN is the first SDN testing method that focuses on the “*South*” interface of SDN controllers while accounting for the statefulness of SDNs.

Characterising Failure-Inducing Inputs. Recently, several research efforts have focused on identifying the input conditions that cause a system under test to fail [34, 41]. Gopinath et al. [34] introduced DDTEST, which abstracts inputs that lead to failures. DDTEST is designed to test software programs, such as JavaScript translators and command-line utilities, that accept string inputs. It uses a derivation tree to represent how failure-inducing strings are generated. Kampmann et al. [41] developed ALHAZEN, which identifies the conditions under which software programs fail. ALHAZEN also targets software that processes strings and uses machine learning to learn failure-inducing conditions in the form of decision trees. In the domain of SDN systems, Ollando et al. [62] introduced FuzzSDN, a machine learning-guided Fuzzing method for testing SDN controllers. FuzzSDN learns an interpretable classification model that characterises conditions on a control message’s fields under which the controller fails. We note that these methods do not attempt to create a failure-inducing model for sequential data, which makes those methods not suitable for our objectives. To our knowledge, SeqFuzzSDN is the first approach that applies an EFSM-guided fuzzing approach to infer failure-inducing models, in the form of EFSMs, with a focus on SDNs. Specifically, SeqFuzzSDN tests SDN controllers by accounting for the architecture and protocols unique to SDNs, which differ from other systems (e.g., server-client systems). Further, SeqFuzzSDN tests SDN controllers without requiring any modifications or instrumentation of the controllers or their networks, enabling SDN testing in realistic operational settings.

7 Conclusions

We developed SeqFuzzSDN, a learning-guided fuzzing method for testing stateful SDN controllers. SeqFuzzSDN uses a fuzzing strategy, guided by EFSMs, in order to (1) efficiently explore the space of states of the SDN controller under test and (2) infer EFSMs that characterise the sequence of messages that may make the system fail. SeqFuzzSDN implements an iterative process that fuzzes sequences of control messages, learns an EFSM, and plans how to guide the subsequent fuzzing steps by leveraging the learned EFSM. We evaluated SeqFuzzSDN on several synthetic systems controlled by two different SDN controllers. In addition, we compared SeqFuzzSDN against our extended versions of three SOTA methods for testing SDN controllers, which served as baselines

in our evaluation. Our results show that SeqFuzzSDN significantly outperforms the baselines by generating effective and diverse tests (i.e., sequences of control messages), that cause the system to fail, and by producing accurate EFSMs.

In the future, we will devise a learning technique that will allow SeqFuzzSDN to learn stateful models (i.e., EFSMs) incrementally, thus addressing scalability issues in inferring EFSMs. To our knowledge, no existing solution supports incremental EFSM inference in a form applicable to SeqFuzzSDN. This poses new challenges due to the complexity of continuously updating and maintaining complex dependencies between states and transitions, without losing any of the previously learned information. In addition, we will explore model inference techniques that can capture the asynchronous, distributed, and concurrent nature of an SDN system and apply these techniques to model SDN behaviours, utilising them to guide testing. Since SDN components (i.e., hosts, switches, and controllers) are integrated and operate concurrently, leveraging such techniques will allow us to accurately represent interactions among components and potentially improve test effectiveness. Further, we also aim to confirm the applicability and effectiveness of SeqFuzzSDN by testing it on more SDN systems and performing user studies.

Data Availability

Our evaluation package and the SeqFuzzSDN tool can be accessed online [61] to allow researchers and practitioners to (1) reproduce our experiments and (2) utilize and modify SeqFuzzSDN.

Acknowledgments

This project has received funding from SES and the Luxembourg National Research Fund under the Industrial Partnership Block Grant (IPBG), ref. IPBG19/14016225/INSTRUCT. Lionel Briand was partly funded by the Science Foundation Ireland grant 13/RC/2094-2 and NSERC of Canada under the Discovery and CRC programs. For the purpose of open access, and in fulfilment of the obligations arising from the grant agreement, the author has applied a Creative Commons Attribution 4.0 International (CC BY 4.0) license to any Author Accepted Manuscript version arising from this submission.

References

- [1] Vangalur S. Alagar and Kasingam Periyasamy. 2011. *Specification of Software Systems*. Springer London, Chapter Extended Finite State Machine, 105–128.
- [2] Kinan Dak Albab, Jonathan DiLorenzo, Stefan Heule, Ali Kheradmand, Steffen Smolka, Konstantin Weitz, Muhammad Timarzi, Jiaqi Gao, and Minlan Yu. 2022. SwitchV: automated SDN switch validation with P4 models. In *Proceedings of the ACM SIGCOMM 2022 Conference*. 365–379.
- [3] Talal Alharbi, Dario Durando, Farzaneh Pakzad, and Marius Portmann. 2016. Securing ARP in Software Defined Networks. In *Proceedings of the 41st IEEE Conference on Local Computer Networks (LCN'14)*. 523–526.
- [4] Abdullah M. Alshanqiti, Safi Faizullah, Sarwan Ali, Maria Khalid Alvi, Muhammad Asad Khan, and Imdadullah Khan. 2019. Detecting DDoS Attack on SDN Due to Vulnerabilities in OpenFlow. In *Proceedings of the 2019 International Conference on Advances in the Emerging Computing Technologies*. 1–6.
- [5] Greg Banks, Marco Cova, Viktoria Felmetsgier, Kevin Almeroth, Richard Kemmerer, and Giovanni Vigna. 2006. SNOOZE: toward a Stateful NetwOrk prOtoCol fuzZEr. In *Proceedings of the 9th International Conference on Information Security*. 343–358.
- [6] C.H. Bennett, P. Gacs, Ming Li, P.M.B. Vitanyi, and W.H. Zurek. 1998. Information distance. *IEEE Transactions on Information Theory* 44, 4 (1998), 1407–1423.
- [7] Pankaj Berde, Matteo Gerola, Jonathan Hart, Yuta Higuchi, Masayoshi Kobayashi, Toshio Koide, Bob Lantz, Brian O'Connor, Pavlin Radoslavov, William Snow, and Guru Parulkar. 2014. ONOS: Towards an Open, Distributed SDN OS. In *Proceedings of the 3rd Workshop on Hot topics in Software Defined Networking*. 1–6.
- [8] Seifeddine Bettaieb, Seung Yeob Shin, Mehrdad Sabetzadeh, Lionel C. Briand, Michael Garceau, and Antoine Meyers. 2020. Using machine learning to assist with the selection of security controls during security assessment. *Empirical Software Engineering* 25, 4 (2020), 2550–2582.

- [9] Seifeddine Bettaieb, Seung Yeob Shin, Mehrdad Sabetzadeh, Lionel C. Briand, Gr  gory Nou, and Michael Garceau. 2019. Decision Support for Security-Control Identification Using Machine Learning. In *Proceedings of the 25th International Working Conference on Requirements Engineering: Foundation for Software Quality*, Vol. 11412. 3–20.
- [10] Suman Sankar Bhunia and Mohan Gurusamy. 2017. Dynamic attack detection and mitigation in IoT using SDN. In *Proceedings of the 27th International Telecommunication Networks and Applications Conference*. 1–6.
- [11] Andreas Blenk, Arsany Basta, Martin Reisslein, and Wolfgang Kellerer. 2016. Survey on Network Virtualization Hypervisors for Software Defined Networking. *IEEE Communications Surveys & Tutorials* 18, 1 (2016), 655–685.
- [12] Robert T. Braden. 1989. *Requirements for Internet Hosts - Communication Layers*. Information RFC 1122. Internet Engineering Task Force (IETF).
- [13] J  rgen Branke, Kalyanmoy Deb, Henning Dierolf, and Matthias Osswald. 2004. Finding Knees in Multi-objective Optimization. In *Proceedings of the 8th International Conference on Parallel Problem Solving from Nature (PPSN'04)*. 722–731.
- [14] Caius Brindescu, Iftekhar Ahmed, Rafael Leano, and Anita Sarma. 2020. Planning for untangling: Predicting the difficulty of merge conflicts. In *Proceedings of the 42nd International Conference on Software Engineering*. 801–811.
- [15] Alessandro Cal  , Paolo Arcaini, Shaikat Ali, Florian Hauer, and Fuyuki Ishikawa. 2020. Generating Avoidable Collision Scenarios for Testing Autonomous Driving Systems. In *Proceeding of the 13th IEEE International Conference on Software Testing, Validation and Verification*. 375–386.
- [16] Tao Chen, Ke Li, Rami Bahsoon, and Xin Yao. 2018. FEMOSAA: Feature-Guided and Knee-Driven Multi-Objective Optimization for Self-Adaptive Software. *ACM Transactions on Software Engineering and Methodology* 27, 2 (2018), 1–50.
- [17] Juan Camilo Correa Chica, Jenny Cuatindioy Imbachi, and Juan Felipe Botero. 2020. Security in SDN: A comprehensive survey. *Journal of Network and Computer Applications* 159 (2020), 1–23.
- [18] R. Cilibrasi and P.M.B. Vitanyi. 2005. Clustering by compression. *IEEE Transactions on Information Theory* 51, 4 (2005), 1523–1545.
- [19] Andrew R. Cohen and Paul M.B. Vit  nyi. 2015. Normalized Compression Distance of Multisets with Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 8 (2015), 1602–1614.
- [20] William W. Cohen. 1995. Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning*. 115–123.
- [21] Mauro Conti, Nicola Dragoni, and Viktor Lesyk. 2016. A Survey of Man In The Middle Attacks. *IEEE Communications Surveys & Tutorials* 18, 3 (2016), 2027–2051.
- [22] Leonardo de Moura and Nikolaj B  rner. 2008. Z3: An Efficient SMT Solver. In *Proceeding of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. 337–340.
- [23] Kalyanmoy Deb. 2001. *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons.
- [24] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 2 (2002), 182–197.
- [25] L. Peter Deutsch. 1996. *GZIP file format specification version 4.3*. Information RFC 1952. Aladdin Enterprises.
- [26] Mohan Dhawan, Rishabh Poddar, Kshiteej Mahajan, and Vijay Mann. 2015. SPHINX: Detecting Security Attacks in Software-Defined Networks. In *Proceedings of the 22nd Network and Distributed System Security Symposium*. 1–16.
- [27] Vaibhav Hemant Dixit, Adam Doup  , Yan Shoshitaishvili, Ziming Zhao, and Gail-Joon Ahn. 2018. AIM-SDN: Attacking Information Mismanagement in SDN-datastores. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 664–676.
- [28] Dmitry Drutskoy, Eric Keller, and Jennifer Rexford. 2013. Scalable Network Virtualization in Software-Defined Networks. *IEEE Internet Computing* 17 (2013), 20–27.
- [29] Ramakrishnan Durairajan, Joel Sommers, and Paul Barford. 2014. Controller-agnostic SDN Debugging. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, Aruna Seneviratne, Christophe Diot, Jim Kurose, Augustin Chaintreau, and Luigi Rizzo (Eds.). 227–234.
- [30] S. S. Emam and J. Miller. 2018. Inferring Extended Probabilistic Finite-State Automaton Models from Software Executions. *ACM Transactions on Software Engineering and Methodology* 27, 1 (2018), 1–39.
- [31] David Eppstein. 1998. Finding the k Shortest Paths. *SIAM J. Comput.* 28, 2 (1998), 652–673.
- [32] Ramon Ferr  s, Harilaos Koumaras, Oriol Sallent, George Agapiou, Tinku Rasheed, M-A Kourtis, C Boustie, Patrick G  lard, and Toufik Ahmed. 2016. SDN/NFV-enabled satellite communications networks: Opportunities, scenarios and challenges. *Journal of Physical Communication* 18 (2016), 95–112.
- [33] Hugo Gascon, Christian Wressnegger, Fabian Yamaguchi, Daniel Arp, and Konrad Rieck. 2015. Pulsar: Stateful Black-Box Fuzzing of Proprietary Network Protocols. In *Security and Privacy in Communication Networks*. 330–347.
- [34] Rahul Gopinath, Alexander Kampmann, Nikolas Havrikov, Ezekiel O. Soremekun, and Andreas Zeller. 2020. Abstracting failure-inducing inputs. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ACM, 237–248.

[35] Evangelos Haleplidis, Kostas Pentikousis, Spyros G. Denazis, Jamal Hadi Salim, David Meyer, and Odysseas G. Koufopavlou. 2015. *Software-Defined Networking (SDN): Layers and Architecture Terminology*. Information RFC 7426. Internet Research Task Force (IRTF).

[36] Joel M. Halpern, Robert Haas, Doria Avri, Ligang Dong, Weiming Wang, Hormuzd M. Khosravi, Jamal Hadi Salim, and Ram Gopal. 2010. *Forwarding and Control Element Separation (ForCES) Protocol Specification*. Information RFC 5810.

[37] Fitash Ul Haq, Donghwan Shin, Shiva Nejati, and Lionel C. Briand. 2021. Can Offline Testing of Deep Neural Networks Replace Their Online Testing? *Empirical Software Engineering* 26, 90 (2021), 1–30.

[38] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. 2015. *Nonparametric Statistical Methods*. John Wiley & Sons.

[39] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated Machine Learning: Methods, Systems, Challenges* (1 ed.). Springer.

[40] Samuel Jero, Xiangyu Bu, Cristina Nita-Rotaru, Hamed Okhravi, Richard Skowyra, and Sonia Fahmy. 2017. BEADS: Automated Attack Discovery in OpenFlow-Based SDN Systems. In *Proceedings of the 20th International Symposium on Research in Attacks, Intrusions, and Defenses*. 311–333.

[41] Alexander Kampmann, Nikolas Havrikov, Ezekiel O. Soremekun, and Andreas Zeller. 2020. When does my program do this? learning circumstances of software behavior. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 1228–1239.

[42] Jinwoo Kim, Minjae Seo, Eduard Marin, Seungssoo Lee, Jaehyun Nam, and Seungwon Shin. 2024. Ambusher: Exploring the Security of Distributed SDN Controllers Through Protocol State Fuzzing. *IEEE Transactions on Information Forensics and Security* 19 (2024), 6264–6279.

[43] A. N. Kolmogorov. 1968. Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics* 2, 1–4 (1968), 157–168.

[44] Bob Lantz, Brandon Heller, and Nick McKeown. 2010. A network in a laptop: rapid prototyping for software-defined networks. In *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*. 1–6.

[45] Jaekwon Lee, Seung Yeob Shin, Lionel C. Briand, and Shiva Nejati. 2024. Probabilistic Safe WCET Estimation for Weakly Hard Real-time Systems at Design Stages. *ACM Transactions on Software Engineering and Methodology* 33, 2 (2024), 1–34.

[46] Jaekwon Lee, Seung Yeob Shin, Shiva Nejati, and Lionel C. Briand. 2022. Optimal priority assignment for real-time systems: a coevolution-based approach. *Empirical Software Engineering* 27, 6 (2022), 142:1–142:49.

[47] Jaekwon Lee, Seung Yeob Shin, Shiva Nejati, Lionel C. Briand, and Yago Isasi Parache. 2023. Estimating Probabilistic Safe WCET Ranges of Real-Time Systems at Design Stages. *ACM Transactions on Software Engineering and Methodology* 32, 2 (2023), 37:1–37:33.

[48] Seungssoo Lee, Seungwon Woo, Jinwoo Kim, Jaehyun Nam, Vinod Yegneswaran, Phillip A. Porras, and Seungwon Shin. 2022. A Framework for Policy Inconsistency Detection in Software-Defined Networks. *IEEE/ACM Transactions on Networking* 30, 3 (2022), 1410–1423.

[49] Seungssoo Lee, Seungwon Woo, Jinwoo Kim, Vinod Yegneswaran, Phillip A. Porras, and Seungwon Shin. 2020. AudiSDN: Automated Detection of Network Policy Inconsistencies in Software-Defined Networks. In *Proceedings of the 39th IEEE Conference on Computer Communications*. 1788–1797.

[50] Seungssoo Lee, Changhoon Yoon, Chanhee Lee, Seungwon Shin, Vinod Yegneswaran, and Phillip Porras. 2017. DELTA: A Security Assessment Framework for Software-Defined Networks. In *Proceedings of the 24th Network and Distributed System Security Symposium*. 1–15.

[51] Yahui Li, Zhiliang Wang, Jiangyuan Yao, Xia Yin, Xingang Shi, Jianping Wu, and Han Zhang. 2019. MSAID: Automated detection of interference in multiple SDN applications. *Computer Networks* 153 (2019), 49–62.

[52] Yahui Li, Zhiliang Wang, Jiangyuan Yao, Xia Yin, Xingang Shi, Jianping Wu, and Han Zhang. 2019. MSAID: Automated detection of interference in multiple SDN applications. *Computer Networks* 153 (2019), 49–62.

[53] Yuekang Li, Yinxing Xue, Hongxu Chen, Xiuheng Wu, Cen Zhang, Xiaofei Xie, Haijun Wang, and Yang Liu. 2019. Cerebro: context-aware adaptive fuzzing for effective vulnerability detection. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 533–544.

[54] Jiajia Liu, Yongpeng Shi, Lei Zhao, Yurui Cao, Wen Sun, and Nei Kato. 2018. Joint Placement of Controllers and Gateways in SDN-Enabled 5G-Satellite Integrated Network. *IEEE Journal on Selected Areas in Communications* 36, 2 (2018), 221–232.

[55] Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. 2019. How Complex Is Your Classification Problem? A Survey on Measuring Classification Complexity. *Comput. Surveys* 52, 5 (2019), 107:0–107:34.

[56] Qi Luo, Aswathy Nair, Mark Grechanik, and Denys Poshyvanyk. 2017. FOREPOST: finding performance problems automatically with feedback-directed learning software testing. *Empirical Software Engineering* 22, 1 (2017), 6–56.

[57] Valentin J.M. Manes, HyungSeok Han, Choongwoo Han, Sang Kil Cha, Manuel Egele, Edward J. Schwartz, and Maverick Woo. 2021. The Art, Science, and Engineering of Fuzzing: A Survey. *IEEE Transactions on Software Engineering* 47

(2021), 2312–2331. Issue 11.

[58] Canini Marco, Venzano Daniele, Perešini Peter, Kostić Dejan, and Rexford Jennifer. 2012. A NICE Way to Test OpenFlow Applications. In *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation*. 127–140.

[59] Saurav Nanda, Faheem Zafari, Casimer DeCusatis, Eric Wedaa, and Baijian Yang. 2016. Predicting network attack patterns in SDN using machine learning approach. In *Proceedings of the 2016 IEEE Conference on Network Function Virtualization and Software Defined Networks*. 167–172.

[60] Roberto Natella. 2022. StateAFL: Greybox fuzzing for stateful network servers. *Empirical Software Engineering* 27, 7 (2022), 191.

[61] Raphaël Ollando, Seung Yeob Shin, and Lionel C. Briand. 2024. [Artifact Repository] Learning-Guided Fuzzing for Testing Stateful SDN Controllers. <https://doi.org/10.6084/m9.figshare.27180477>.

[62] Raphaël Ollando, Seung Yeob Shin, and Lionel C. Briand. 2024. Learning Failure-Inducing Models for Testing Software-Defined Networks. *ACM Transaction on Software Engineering and Methodologies* 33, 5 (2024), 113:1–113:25.

[63] Open Networking Foundation. 2015. *OpenFlow Switch Specification, Version 1.5.1*. Specification ONF TS-025. Open Networking Foundation.

[64] Annibale Panichella, Fitsum Meshesha Kifetew, and Paolo Tonella. 2015. Reformulating Branch Coverage as a Many-Objective Optimization Problem. In *Proceeding the 8th IEEE International Conference on Software Testing, Verification and Validation (ICST)*. 1–10.

[65] Larry L Peterson and Bruce S Davie. 2007. *Computer Networks: A Systems Approach*. Morgan Kaufmann.

[66] Van-Thuan Pham, Marcel Böhme, and Abhik Roychoudhury. 2020. AFLNET: A Greybox Fuzzer for Network Protocols. In *2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST)*. 460–465.

[67] David C. Plummer. 1982. *An Ethernet Address Resolution Protocol: Or Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware*. Information. Internet Engineering Task Force (IETF).

[68] Jon Postel. 1981. *Internet Protocol*. Information RFC 791, USC/Information Sciences Institute.

[69] Shisong Qin, Fan Hu, Zheyu Ma, Bodong Zhao, Tingting Yin, and Chao Zhang. 2023. NSFuzz: Towards Efficient and State-Aware Network Service Fuzzing. *ACM Transaction on Software Engineering and Methodologies* 32, 6 (2023), 160:1–160:26.

[70] John Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[71] Wajid Rafique, Lianyong Qi, Ibrar Yaqoob, Muhammad Imran, Raihan Ur Rasool, and Wanchun Dou. 2020. Complementing IoT Services Through Software Defined Networking and Edge Computing: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials* 22, 3 (2020), 1761–1804.

[72] RYU Project Team. 2014. *RYU SDN Framework* (1 ed.). RYU Project Team.

[73] Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.

[74] Donghwan Shin, Domenico Bianculli, and Lionel C. Briand. 2022. PRINS: scalable model inference for component-based system logs. *Empirical Software Engineering* 27, 4 (2022), 1–32.

[75] Seung Yeob Shin, Shiva Nejati, Mehrdad Sabetzadeh, Lionel C. Briand, Chetan Arora, and Frank Zimmer. 2020. Dynamic adaptation of software-defined networks for IoT systems: a search-based approach. In *Proceedings of the 15th IEEE/ACM International Symposium on Software Engineering for Adaptive and Self-Managing Systems*. 137–148.

[76] Seung Yeob Shin, Shiva Nejati, Mehrdad Sabetzadeh, Lionel C. Briand, and Frank Zimmer. 2018. Test case prioritization for acceptance testing of cyber physical systems: a multi-objective search-based approach. In *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*. 49–60.

[77] Apoorv Shukla, Said Jawad Saidi, Stefan Schmid, Marco Canini, Thomas Zinner, and Anja Feldmann. 2020. Toward Consistent SDNs: A Case for Network State Fuzzing. *IEEE Transactions on Network and Service Management* 17, 2 (2020), 668–681.

[78] Michael Smith, Robert Adams Edward, Mike Dvorkin, Youcef Laribi, Vijoy Pandey, Pankaj Garg, and Nik Weidenbacher. 2016. *OpFlex Control Protocol*. Internet Draft draft-smith-opflex-03. Internet Engineering Task Force.

[79] Radu Stoenescu, Dragos Dumitrescu, Matei Popovici, Lorina Negreanu, and Costin Raiciu. 2018. Debugging P4 programs with vera. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*. 518–532.

[80] El-Ghazali Talbi. 2009. *Metaheuristics: From design to implementation* (1 ed.). John Wiley & Sons.

[81] Robert Wagner. 2001. *Address Resolution Protocol Spoofing and Man-In-The-Middle Attacks*. Technical Report. Escal Institute of Advanced Technologies. 1–9 pages.

[82] Neil Walkinshaw, Ramsay Taylor, and John Derrick. 2016. Inferring extended finite state machine models from software executions. *Empirical Software Engineering* 21, 3 (2016), 811–853.

[83] Shaowei Wang, David Lo, Lingxiao Jiang, Shahar Maoz, and Aditya Budi. 2015. Chapter 21 - Scalable Parallelization of Specification Mining Using Distributed Computing. In *The Art and Science of Analyzing Software Data*. Morgan Kaufmann, 623–648.

[84] Tao Wang, Fangming Liu, and Hong Xu. 2017. An Efficient Online Algorithm for Dynamic SDN Controller Assignment in Data Center Networks. *IEEE/ACM Transactions on Networking* 25 (2017), 2788–2801.

[85] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. *Data mining: practical machine learning tools and techniques* (4 ed.). Elsevier.

[86] Seungwon Woo, Seungssoo Lee, Jinwoo Kim, and Seungwon Shin. 2018. RE-CHECKER: Towards Secure RESTful Service in Software-Defined Networking. In *Proceedings of the 2018 IEEE Conference on Network Function Virtualization and Software Defined Networks*. IEEE, Piscataway, NJ, USA, 1–5.

[87] Michał Zalewski. 2016. American Fuzzy Lop – Whitepaper. https://lcamtuf.coredump.cx/afl/technical_details.txt

[88] Andreas Zeller, Rahul Gopinath, Marcel Böhme, Gordon Fraser, and Christian Holler. 2024. *The Fuzzing Book*. CISPA Helmholtz Center for Information Security. <https://www.fuzzingbook.org/> Retrieved 2024-07-01 16:50:18+02:00.

[89] Peng Zhang. 2017. Towards rule enforcement verification for software defined networks. In *Proceedings of the 2017 IEEE Conference on Computer Communications*. IEEE, Piscataway, NJ, USA, 1–9.

A Additional Results for RYU Study Subject

A.1 Results for RQ1



Fig. A.1. Comparing the sensitivity of the EFSMs generated by SeqFuzzSDN, FuzzSDN^E, BEADS^E, and DELTA^E, the five plots in each row display the sensitivity of the corresponding tool. The first four columns represent the sensitivity of the EFSMs assessed using the test dataset containing message sequences generated by each tool. Sensitivity is assessed using message sequences that lead to both success and failure, denoted by (S+F), and only failure, denoted by (F). The last column represents the sensitivity assessed using all datasets generated by the four tools. The boxplots (25%-50%-75%) show the distribution of sensitivity over 10 runs of each tool in EXP1 (RYU).

In EXP1, when ONOS is replaced with RYU, Figure A.1 corresponds to Figure 7. Figure A.1 shows that SeqFuzzSDN achieves a sensitivity of 49.18% on the message sequences leading to both success and failure (referred to as the combined S+F dataset) and 63.37% on the message sequences leading only to failure (referred to as the combined F dataset). Specifically, SeqFuzzSDN achieves, on average, a sensitivity of 60.92% on the SeqFuzzSDN S+F dataset and 90.0% on the SeqFuzzSDN F

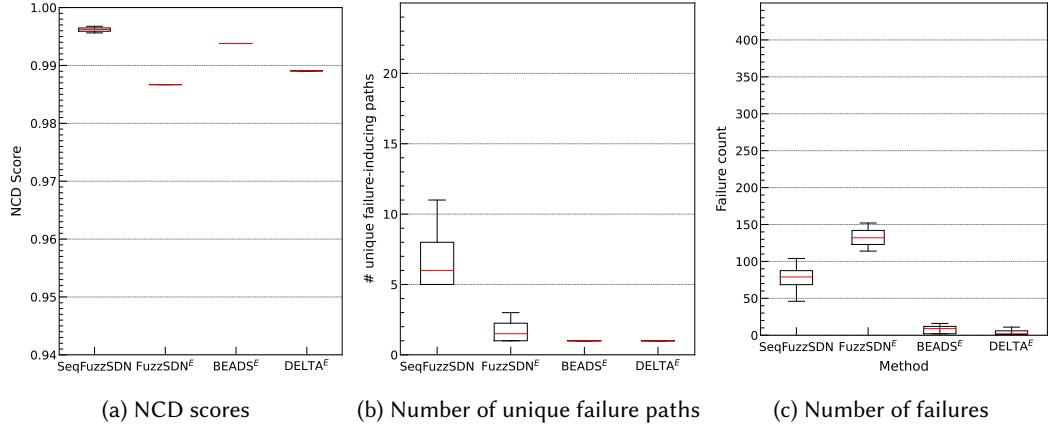


Fig. A.2. Comparing (a) the NCD scores of the message sequences, (b) the number of unique failure-inducing paths in the EFSMs, and (c) the number of message sequences leading to failure, all obtained from SeqFuzzSDN, FuzzSDN^E, BEADS^E, and DELTA^E. The boxplots (25%-50%-75%) show the distribution of each metric over 10 runs of each tool in EXP1 (RYU).

dataset, 5.04% on the FuzzSDN^E S+F dataset and 0.00% on the FuzzSDN^E F dataset, 63.87% on the BEADS^E S+F dataset and 82.56% on the BEADS^E F dataset, and 66.89% on the DELTA^E S+F dataset and 80.91% on the DELTA^E F dataset.

Figure A.1 also shows the average EFSM sensitivity for RYU, for FuzzSDN^E, BEADS^E, and DELTA^E, respectively, of 26.18%, 11.96%, and 5.43% on the combined S+F dataset, and 25.0%, 3.71%, and 1.67% on the combined F dataset. More specifically, using the SeqFuzzSDN S+F dataset (and the SeqFuzzSDN F dataset), these three baselines achieve, respectively, on average, sensitivities of 0%, 20.41%, and 4.87% (and 0%, 0.67%, and 3.20%). Regarding the FuzzSDN^E S+F dataset (and the FuzzSDN^E F dataset), these three baselines achieve, respectively, on average, sensitivities of 92.29%, 1.51%, and 2.26% (and 100%, 0%, and 0%). For the BEADS^E S+F dataset (and the BEADS^E F dataset), these three baselines achieve, respectively, on average, sensitivities of 0%, 18.05%, and 0% (and 0%, 11.50%, and 0%). Lastly, when using the DELTA^E S+F dataset (and the DELTA^E F dataset), these baselines achieve, respectively, on average, sensitivities of 0%, 14.24%, and 19.44% (and 0%, 0%, and 6.67%).

Figure A.2 compares (a) the NCD scores of the message sequences, (b) the number of unique failure-inducing paths in the EFSMs, and (c) the number of message sequences leading to failure, which are obtained from 10 runs of SeqFuzzSDN, FuzzSDN^E, BEADS^E, and DELTA^E for our RYU study subject. Figure A.2a shows that SeqFuzzSDN achieves a higher NCD score, with an average of 0.997, compared to those of the baselines. Figure A.2b shows that, on average, SeqFuzzSDN was able to infer an EFSM containing 6 unique loop-free paths that lead to failure, which is significantly higher than the others. From these results, similarly to our ONOS study subject, we found that SeqFuzzSDN generates more diverse sequences of control messages that exercise a larger number of state changes compared to the baselines.

However, Figure A.2c shows that FuzzSDN^E generates a larger number of message sequences (an average of 141) leading to failure compared to the other tools, while SeqFuzzSDN generates, on average, 76 message sequences leading to failure, thus outperforming BEADS^E and DELTA^E. As with our ONOS study subject, even though FuzzSDN^E outperforms SeqFuzzSDN in terms of number of failures, recall from Figure A.2a and Figure A.2b that FuzzSDN^E generates message sequences that

are less diverse and exercise significantly fewer number of state changes compared to SeqFuzzSDN. Furthermore, as described in Section 3, SeqFuzzSDN aims to generate a balanced number of message sequences that lead to success and failure, rather than focusing solely on the latter.

A.2 Results for RQ2

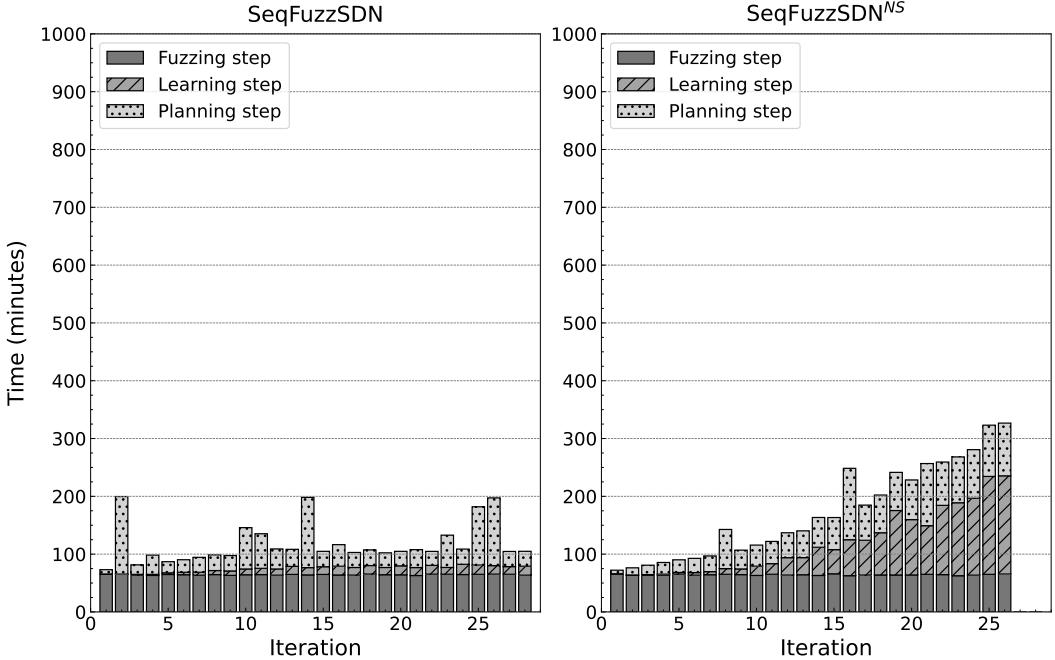


Fig. A.3. Comparing the execution time per iteration for the fuzzing, learning, and planning steps of SeqFuzzSDN and SeqFuzzSDN^{NS} within a 3-day time budget. The execution times shown in this figure are the average values observed over 10 runs of EXP2 (RYU).

Figure A.3 compares SeqFuzzSDN and SeqFuzzSDN^{NS} with regard to the execution times per iteration for the fuzzing, learning, and planning steps over a time budget of 3 days, for our RYU study subject. Similarly to Figure 10, the bar graph shows the average execution times taken by SeqFuzzSDN and SeqFuzzSDN^{NS} for the fuzzing, learning, and planning steps at each iteration, based on 10 runs of EXP2 for RYU.

The results show that the fuzzing time per iteration remains constant at around 70 minutes for both SeqFuzzSDN and SeqFuzzSDN^{NS}, indicating that the fuzzing step is independent of the tool used. For the planning step, Figure A.3 shows that the planning time does not exceed 150 minutes in both SeqFuzzSDN and SeqFuzzSDN^{NS}. Figure A.3 also suggests that, for SeqFuzzSDN^{NS}, the time required to learn an EFSM increases significantly with each iteration due to the growing size of the dataset fed to MINT. Furthermore, we observe that, as opposed to our ONOS study subject, the learning time for SeqFuzzSDN^{NS} does not reach the upper learning limit of 12h, but grows from under 1 minute to above 150 minutes. This finding aligns with the literature [30, 74, 83], as inferring EFSMs is a complex problem that scales poorly with larger input sizes. In contrast, the results for SeqFuzzSDN indicate that the time required for inferring an EFSM (i.e., the learning step) remains below 20 minutes due to the application of the sampling technique. Thus, based on the

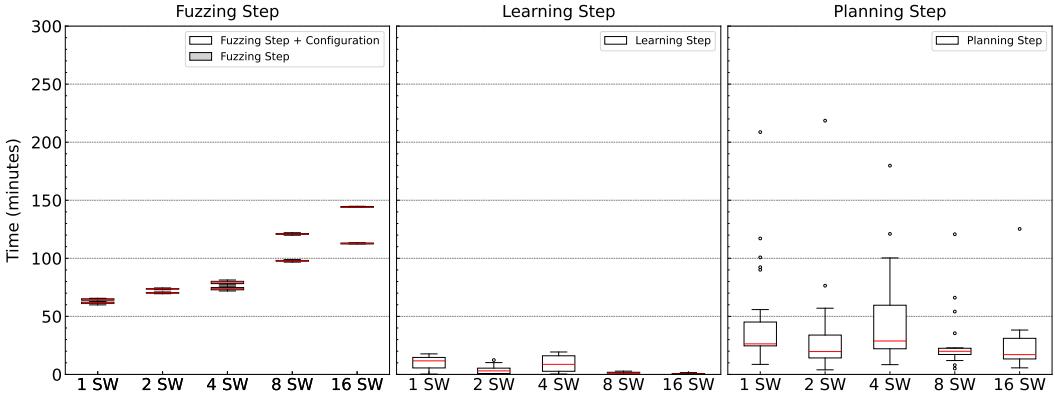


Fig. A.4. Boxplots (25%-50%-75%) representing the distributions of time taken in minutes for the fuzzing, learning, and planning steps of SeqFuzzSDN. This figure includes the times observed over 10 runs of SeqFuzzSDN with 1, 2, 4, 8, and 16 switch configurations controlled by RYU.

results shown in Figure A.3, we can further conclude that applying the sampling technique enables SeqFuzzSDN to overcome the scalability issues associated with the complexity of learning EFSMs.

Table A.1. Statistical significance analysis using the Wilcoxon Rank-Sum test for sensitivity, diversity, and coverage results obtained from 10 runs of EXP2 (RYU).

Metric	Average (SeqFuzzSDN)	Average (SeqFuzzSDN ^{NS})	p-value	Statistical Significance ($\alpha = 0.05$)
Sensitivity	0.553	0.534	0.385	Not Significant
Diversity	0.9976	0.9975	0.987	Not Significant
Coverage	0.5866	0.8248	0.0023	Significant

Furthermore, Table A.1 presents the statistical test results for the distributions of sensitivity, diversity, and coverage (described in Section 3) achieved by SeqFuzzSDN and SeqFuzzSDN^{NS} after 10 runs of EXP2, using the Wilcoxon Rank-Sum test [38] with an α value of 0.05, for our RYU test subject. On average, SeqFuzzSDN (resp. SeqFuzzSDN^{NS}) achieves a sensitivity of 55.3% (resp. 53.4%), a diversity of 0.9976 (resp. 0.9975), and a coverage of 0.5866 (resp. 0.8248). We observed that the differences in sensitivity ($p = 0.18$) and diversity ($p = 0.7$) are not significant, while the difference in coverage ($p = 0.002$) is. The results indicate that the use of the sampling technique does not negatively impact the sensitivity of the generated EFSMs nor the diversity of the generated message sequences, on our RYU test subject. However, the coverage achieved by SeqFuzzSDN has significantly improved, suggesting that, similarly to our ONOS test subject, the states in the EFSM are explored more thoroughly.

A.3 Results for RQ3

Figure A.4 presents the distributions of execution times (25%-50%-75% boxplots) for the fuzzing, learning, and planning steps of SeqFuzzSDN, obtained from EXP3 (RYU). These execution times were measured using the five study subjects in EXP3, which consist of 1, 2, 4, 8, and 16 switches controlled by RYU. As shown in Figure 11, the execution time taken for the fuzzing step is, on

average, 203 minutes for the 1-switch configuration, 60 minutes for 2 switches, 70 minutes for 4 switches, 95 minutes for 8 switches, and 109 minutes for 16 switches. The learning step took, on average, 10 minutes for the 1-switch configuration, 3 minutes for 2 switches, 9 minutes for 4 switches, 2 minutes for 8 switches, and 1 minute for 16 switches. The planning step took, on average, 44 minutes for the 1-switch configuration, 31 minutes for 2 switches, 47 minutes for 4 switches, 28 minutes for 8 switches, and 33 minutes for 16 switches. These results are consistent with our findings from EXP3 (ONOS).