Checking Security Properties of Cloud Service REST APIs

Vaggelis Atlidakis\textsuperscript{*}  
Columbia University  

Patrice Godefroid  
Microsoft Research  

Marina Polishchuk  
Microsoft Research

\textbf{Abstract}—Most modern cloud and web services are programmatically accessed through REST APIs. This paper discusses how an attacker might compromise a service by exploiting vulnerabilities in its REST API. We introduce four security rules that capture desirable properties of REST APIs and services. We then show how a stateful REST API fuzzer can be extended with active property checkers that automatically test and detect violations of these rules. We discuss how to implement such checkers in a modular and efficient way. Using these checkers, we found new bugs in several deployed production Azure and Office-365 cloud services, and we discuss their security implications. All these bugs have been fixed.

\textbf{Keywords}-Test generation; Security; Cloud and Web services; REST APIs

I. \textbf{Introduction}

Cloud computing is exploding. Over the last few years, thousands of new cloud services have been deployed by cloud platform providers, like Amazon Web Services [2] and Microsoft Azure [13], and by their customers who are “digitally transforming” their businesses by modernizing their processes while collecting and analyzing all kinds of new data.

Today, most cloud services are programmatically accessed through REST APIs [9]. REST APIs are implemented on top of the ubiquitous HTTP/S protocol, and offer a uniform way to create (PUT/POST), monitor (GET), manage (PUT/POST/PATCH) and delete (DELETE) cloud resources. Cloud service developers can document their REST APIs and generate sample client code by describing their APIs using an interface-description language such as Swagger (recently renamed OpenAPI) [25]. A Swagger specification describes how to access a cloud service through its REST API, including what requests the service can handle, what responses may be received, and the response format.

How secure are all those APIs? Today, this question is still largely open. Tools for automatically testing cloud services via their REST APIs and checking whether these services are reliable and secure are still in their infancy. Some tools available for testing REST APIs capture live API traffic, and then parse, fuzz, and replay the traffic with the hope of finding bugs [4], [21], [6], [26], [3]. Recently, \textit{stateful REST API fuzzing} [5] was proposed to specifically test more deeply services deployed behind REST APIs. Given a Swagger specification of a REST API, this approach automatically generates \textit{sequences of requests}, instead of single requests, in order to thoroughly exercise the cloud service deployed behind that API, with the goal of finding unhandled exceptions (service crashes) that can be detected by a test client as “500 Internal Server Errors”. While that work looks promising and reports many new bugs found, its scope is restricted to the detection of unhandled exceptions.

In this paper, we introduce four security rules that capture desirable properties of REST APIs and services.

- \textbf{Use-after-free rule.} A resource that has been deleted must no longer be accessible.
- \textbf{Resource-leak rule.} A resource that was not created successfully must not be accessible and must not “leak” any side-effect in the backend service state.
- \textbf{Resource-hierarchy rule.} A child resource of a parent resource must not be accessible from another parent resource.
- \textbf{User-namespace rule.} A resource created in a user namespace must not be accessible from another user namespace.

Violations of such rules might allow an attacker to hijack cloud resources or bypass quotas (\textit{Elevation-of-Privilege} attack), or to steal information from other users (\textit{Information-Disclosure} attack), or to corrupt the backend service state so that it no longer operates properly (\textit{Denial-of-Service} attack), as will be discussed later.

We show how a stateful REST API fuzzer can be extended to test and detect violations of such rules. For each rule, we define an active property checker which (1) generates new API requests to test specific rule violations and (2) detects any such rule violation. In other words, each checker actively tries to break its rule in addition to monitoring for any rule violation. We discuss how to implement such checkers in a modular way, so that checkers do not interfere with each other. Since each checker generates new tests, in addition to an already-large state space exploration, we also discuss how to implement each individual checker efficiently, by eliminating likely-redundant tests whenever possible.

By construction, these checkers can find security rule violations beyond the “500 Internal Server Errors” that can be detected by baseline \textit{stateful REST API fuzzing}. Using these checkers, we found new bugs in several production Azure and Office-365 cloud services. The use of security checkers increases the value of REST API fuzzing by detecting more types of bugs at a modest incremental testing cost.

This paper makes the following contributions:

\textsuperscript{*}The work of this author was mostly done while visiting Microsoft Research.
We introduce rules that describe security properties of REST APIs.
- We design and implement active checkers to test and detect violations of these rules.
- We present detailed experimental results evaluating the performance and effectiveness of these active checkers on three production cloud services.
- With these checkers, we found new bugs in several production Azure and Office-365 cloud services, and we discuss their security implications.

The rest of the paper is organized as follows. In Section II, we recall background information on stateful REST API fuzzing. In Section III, we introduce rules that capture desirable properties of secure REST APIs and present active checkers to test and detect violations of these rules. In Section IV, we present experimental results with active checkers on production cloud services. In Section V, we discuss new bugs found by these checkers and their security implications. In Section VI, we discuss related work, and we conclude the paper in Section VII.

II. STATEFUL REST API FUZZING

In this section, we recall the definition of stateful REST API fuzzing [5], before introducing in Section III security property checkers that can be implemented as extensions of this basic scheme.

We consider cloud services accessible through REST APIs. A client program sends messages, called requests, to a service and receives messages back, called responses. Such messages are sent over the HTTP/S protocol. Each response is associated with a single HTTP status code which is either in the 2xx, 3xx, 4xx or 5xx ranges.

Swagger [25], also known as OpenAPI, is an example of specification language to define REST APIs. A Swagger specification describes how to access a service through its REST API, including what requests the service can handle, what responses may be received, and the respective response format.

We define a REST API as a finite set of requests. Each request \( r \) is a tuple of the form \((a, t, p, b)\) where
- \( a \) is an authentication token,
- \( t \) is the request type,
- \( p \) is a resource path, and
- \( b \) is the request body.

A request type \( t \) is any of the following five REST-allowed values: PUT (create or update), POST (create or update), GET (read, list or query), DELETE (delete), PATCH (update). The resource path \( p \) is a string identifying a cloud resource and its parent hierarchy. Typically, \( p \) is a (non-empty) sequence matching the regular expression

\[
/\langle\text{resourceType}\rangle/\langle\text{resourceName}\rangle/+/\]

where resourceType denotes the type of a cloud resource and resourceName is the specific name of the resource of that type. The last resource named in the path is typically the specific resource that the request tries to create, access, or delete. The request body \( b \) may include additional parameters and their values that may be required or optional for the request to be executed successfully.

For instance, here is a request to get the properties of a specific Azure DNS zone [14] (shown on multiple lines):

\[
\{ \text{User-auth-token} \} \ GET
\]

https://management.azure.com/

subscriptions/{subscriptionId}/

resourceGroups/{resourceGroupName}/

providers/Microsoft.Network/

dnsZones/{zoneName}

?api-version=2018-03-01 \{}
\]

This request is of type GET, its path requires three resource names, namely a subscriptionID, a resourceGroupName, and a zoneName, and its body (at the end) denoted by \{ \} is empty.

REST API requests of type PUT or POST typically create new resources, while DELETE requests destroy existing resources. A request whose execution creates a new resource of type \( T \) is called a producer for the resource type \( T \). A newly created resource is represented by its identifier, or id for short. Because resources are dynamically created, we will sometimes call them dynamic objects. A request which requires a resource name of type \( T \) in its path or in its body is called a consumer for the resource type \( T \). We will sometimes refer to the resource name of type \( T \) as the dynamic object type. In the Azure DNS zone example above, the GET request shown consumes three resources of type subscriptions, resourceGroups, and dnsZones respectively, but does not produce any new resource.

Inside resource paths or request bodies of individual requests, the user is allowed to specify that some specific values, called fuzzable values, are to be chosen randomly among a (small finite) set of specific values. For instance, a user might specify that a given integer value in the body of a request may be, say, either 0, 10, 1000000, or -10. Such a set of values is called a fuzzing dictionary. Given a request with fuzzable values, a rendering of that request denotes a mapping of each fuzzable value to a single concrete value selected in its fuzzing dictionary. Thus, a request with \( n \) fuzzable values which can each take \( k \) possible values results in \( n^k \) possible renderings. A rendering is called valid if the execution of the corresponding request returns a valid response (defined in the next paragraph). Users are responsible for identifying values they want to fuzz and their associated fuzzing dictionaries.

We define the state space of a service as a directed graph where nodes represent service states and edges are transitions between these. Given a state \( s \) of the service, executing a single request \( r \) leads to a successor state \( s' \); this execution is denoted by \( s \xrightarrow{r} s' \). The execution of a request \( r \) in a state \( s \) is either valid if it triggers a 2xx response, invalid if it triggers a 3xx or 4xx response, or a bug if it triggers a 5xx response.

Given an initial state where no resources exist, the state space of the service reachable from that initial state can be explored by executing sequences of requests. Such an exploration is stateful when it attempts to explore service states
that are reachable only using sequences of multiple requests: earlier requests in a sequence may produce resources that are consumed in subsequent requests in that sequence in order to exercise more requests and reach deeper service states.

State-space exploration can be performed using various search strategies, e.g., a systematic breadth-first search or a random search [5]. State spaces can be large, even infinite, because the length of request sequences is not bounded, because the sets of possible renderings can be very large, and because the service under test is viewed as a blackbox. Fortunately, the sets of possible renderings can be very large, and because a partial state-space exploration may be sufficient to reveal interesting bugs. In our context, a bug is defined as a partial state-space exploration algorithm of this section the main driver of stateful REST API fuzzing.

III. SECURITY CHECKERS FOR REST APIs

In this section, we define and describe active checkers for security rules of REST APIs. First, in Section III-A, we introduce four REST API security rules. In Section III-B, we describe how to implement active checkers for testing and detecting security rule violations. Each active checker focuses on a single type of security rule violation. In Section III-C, we discuss how each checker can be combined in a modular way with the other checkers and with the main driver of stateful REST API fuzzing. In Section III-D, we propose a new search strategy for scalable test generation with property checkers. In Section III-E, we describe how to group together checker violations in order to avoid reporting the same bug multiple times to the user.

A. Security Rules

We introduce four security rules that capture desirable properties of REST APIs and services. We illustrate each rule with an example and discuss its security implications. All four rules are inspired by past real bugs in deployed cloud services, which were found either by manual penetration testing or by root cause analysis of customer-visible incidents. Examples of new, previously-unknown bugs we found as rule violations in deployed production Azure and Office-365 services are presented later in Section V.

Use-after-free rule. A resource that has been deleted must no longer be accessible. In other words, after a successful DELETE operation on any resource, any subsequent operation – like read, update, or delete – on that resource must fail.

For example, after issuing a DELETE request to URI /users/user-id1 in order to delete the account with identifier user-id1, all subsequent attempts to use user-id1 must fail and thus return a “404 Not Found” HTTP status code in their response.

A use-after-free violation occurs when a resource that has been deleted still remains accessible through the API. This must never happen. It is a clear bug that may lead to bypassing resource quotas and corrupting the service backend state.

Resource-leak rule. A resource that was not successfully created must not be accessible, and must not “leak” any associated resources in the backend service state. In other words, if the execution of a PUT or POST request to create a new resource fails (for any reason), any subsequent operation on that resource must also fail with a 4xx response. Furthermore, no side-effects associated with successful creation of that resource type must occur in the backend service state and be visible to the user. For instance, a failed-to-be-created resource must not be counted in the user’s resource counter towards service quotas, and the name of the failed-to-be-created resource must be reusable by the user.

As an example, after issuing a malformed PUT request to create URI /users/user-id1, a 4xx response must be received. Any subsequent request to access (read, update, or delete) this URI must also fail.

A resource-leak violation occurs when a resource that was not successfully created nevertheless “leaks” some side-effect in the backend service state. For instance, the resource may be listed by a subsequent GET request, yet it cannot be deleted with a DELETE request, or subsequent attempts to re-create this resource return “409 Conflict” responses. Such violations must never happen, as they may have unintended consequences on the capacity for that resource type (e.g., if resource quota limits are reached and no new resources can be created) and on the performance of the service (e.g., due to unnecessarily large database tables).

Resource-hierarchy rule. A child resource of a parent resource must not be accessible from another parent resource. In other words, if a resource child is successfully created from a resource parent and identified as such in service resource paths of the form {parentType}/parent/{childType}/child/, the child resource must not be accessible (i.e., must not be successfully read, updated or deleted) when substituting the parent resource by any other parent resource.

For example, after issuing POST requests to URIs /users/user-id1, /users/user-id2, and /users/user-id1/reports/report-id1 to create users user-id1, user-id2, and then add report report-id1 to user user-id1, subsequent requests to URI /users/user-id2/reports/report-id1 must fail since, according to the resource-hierarchy rule, report report-id1 belongs to user user-id1 but not to user user-id2.

A resource-hierarchy violation occurs when a sub-resource originally created from a parent resource is accessible from a different parent resource with no parent-child relationship. When such violations are possible, an attacker might be able to provide an unauthorized parent object identifier.
Fig. 1: Use-after-free checker.

We enforce the first principle by running all the checkers whenever the main driver has finished executing a new test case. We enforce the second principle by prioritizing the order of applying checkers based on their semantics, so that they operate on different test cases and do not interfere with each other (more on this later in this section). In what follows, we present implementation details of each checker as well as optimizations to limit state-space explosion.

Use-after-free checker. The implementation of the use-after-free rule checker is described in Figure 1 in python-like notation. The algorithm is called after the main driver executes a DELETE request (see Figure 4) and takes three inputs: a sequence seq of requests, which is the latest test case executed by the main driver; the global cache of dynamic objects, denoted global_cache, which contains the most recent object types and ids for the dynamic objects created so far; and the request collection, denoted reqCollection, which is the set of all available API requests.

First, the types of the dynamic objects consumed by the last request are retrieved (line 5) and the id of the last object type, denoted target_obj_type, is stored in a temporary variable, denoted target_obj_id. Although the last request may be consuming more than one object type, we consider the last type in seq_obj_types as the actual type of the deleted object. (For example, a DELETE request on the URI /users/userId1/reports/reportId1 consumes two object types (users and reports) but only deletes report objects.) After this initial setup, the for-loop (line 12) iterates over all requests available in reqCollection and skips those that do not consume the target object type (line 14). Once a request, req, that consumes the target object id is found, the target object id is restored in the global cache of dynamic objects (line 17) and is therefore used by the function EXECUTE (line 19) which executes request req. Note that the target object id is repeatedly restored in the global cache because the function EXECUTE uses object ids available in global_cache when executing a request. If any of these requests succeeds, line 20 will trigger a use-after-free violation (see Section III-A).
Finally, in order to limit the number of additional tests generated for each request sequence, the inner loop (optionally) terminates when one request for each target object type is found (line 21). This option is used if the variable mode is not set to value exhaustive. We present detailed experimental results regarding the impact of this optimization in Section IV.

Resource-leak checker. The resource-leak rule checker is described in Figure 2. The algorithm takes the same three inputs as the use-after-free checker. This checker operates on request sequences executed by the main driver whose last request led to an invalid HTTP status code in the response (see Figure 4). Initially, the algorithm identifies the dynamic object types produced by the whole sequence, denoted seq_obj_types, and produced by the last request, denoted target_obj_types (lines 4 and 5). The main logic of the algorithm is implemented in three nested for loops. The first loop (line 6) iterates over all object types produced by the last request. The second loop (line 7) iterates over object ids “guessed” for the current object type for which an invalid HTTP status code was received. The function GUESS takes as argument an object type and returns a set of possible object ids matching this type and which were not created successfully. For instance, if the creation of a dynamic object with object type “x” and object id “obj1” fails through the API (according to the response received), the checker will attempt to execute any request that consumes the object type “x” and assert it fails when using the object id “obj1”. Note that the total number of guessed values per object id is limited to a user-provided parameter value in order to avoid an explosion in the number of additional tests.

In line 8, a guessed object-id value is temporarily added to the global cache of properly-created dynamic objects. Then the inner loop (line 9) iterates over all requests in reqCollection to find requests that are executable (given the object types produced by the current sequence) and that consume the given target object type. These requests are executed (line 17) using the “guessed” object ids previously registered in the global cache. This way, the algorithm tries to trigger a resource-leak violation (see Section III-A) or asserts that no such violation occurs for the given request sequence (line 18).

Finally, in order to limit the number of additional tests generated for each input sequence, the inner loop (optionally) terminates when one request for each guessed object is found (line 19). We evaluate this optimization in Section IV.

Resource-hierarchy checker. The implementation of the resource-hierarchy rule checker is described in Figure 3. The algorithm takes two inputs: a sequence of requests, denoted seq, which is the latest test case executed by the main driver and the current global cache of dynamic objects, denoted global_cache. First, the algorithm records the object types consumed by the last request of the current sequence, denoted target_obj_types (line 6), and the object types consumed by all other requests of the sequence before the last request, denoted predecessor_obj_types (line 7). Afterwards, the ids of the objects consumed only by the last request are stored locally (lines 12 and 13). These are the child objects whose hierarchy the checker will try to violate by executing requests that try to access them using invalid parent objects. To this end, in line 15, the current sequence is executed up to (and not including) the last request. Finally, the old child object ids are restored (lines 18 and 19) and the last request is executed using the old child object ids on top of new parent object ids (line 20). These parent object ids are not proper parent objects of the restored child object ids. This way, the algorithm tries to trigger a resource-hierarchy violation (see Section III-A) or asserts that no such violation occurs for the given request sequence (line 21).

User-namespace checker. Due to space constraints, we omit a detailed presentation of this checker. In a nutshell, this checker attempts to re-execute the valid last request of any test case executed by the main driver using a different authentication token. If this succeeds, an attacker with a different authentication token could hijack the objects used in the last request, and a user namespace violation (see Section III-A) is reported.

C. Combining All Checkers

The four checkers defined in the previous section are executed as follows. Whenever the stateful REST API fuzzer reaches a new state (as defined in Section II), its main driver calls the code shown in Figure 4. Depending on the last request executed, this code activates the checkers that are applicable to the current state. We now discuss important properties of these checkers and of their combination.
Contribution beyond stateful REST API fuzzing. The checkers extend the main driver of baseline stateful REST API fuzzing in two ways: (1) they extend the state space by executing additional tests and (2) they check for responses other than 5xx and can flag unexpected 2xx responses as rule-violation bugs. Thus, they clearly increase the bug-finding capabilities of the main driver: they can find bugs that the main driver alone would not find.

Active property checking versus passive monitoring. As discussed earlier, the checkers we define extend the search space explored by the main driver with additional test cases aimed at triggering and detecting specific rule violations. In contrast, passive runtime monitoring of these rules in conjunction with the main driver, i.e., without executing those new tests, would likely be unable to detect rule violations. Specifically, use-after-free and resource-leak rule violations would likely not be detected with passive monitoring alone because the default state space exploration, performed by the main driver, would likely not attempt to re-use deleted resources or resources after a failure, respectively. Similarly, resource-hierarchy and user-namespace rule violations would not be detected by passive monitoring either because the baseline main driver does not attempt to substitute object identifiers or authentication tokens, respectively. In other words, the additional test cases generated by the checkers are necessary to find rule violations and are not redundant with respect to non-checker tests.

Complementarity among the checkers. The four checkers we define complement each other: no two checkers will ever generate the same new tests, by construction, because their preconditions are all mutually exclusive. First, the use-after-free checker is the only checker activated by request sequences that end in a DELETE request. Second, the resource-leak checker is the only checker activated when the last request executed returns an invalid HTTP status code. Third, the resource-ownership checker is the only checker activated on request sequences with valid renderings that do not end in a DELETE request. Fourth and last, the user-namespace checker executed tests using an attacker token different from the authentication token used by the main driver and all other checkers, so it clearly extends the state space in another, orthogonal dimension.

D. Search Strategies for Checkers

The main search strategy used for test generation in stateful REST API fuzzing [5] is a breadth-first search (BFS) in the search space defined by all possible request sequences. This search strategy provides full grammar coverage both with respect to all possible renderings of each individual request and with respect to all possible request sequences up to a given sequence length. However, since the search space explored by BFS is typically enormous, the search does not scale well as the sequence length increases. Therefore, an optimization called BFS-Fast was introduced. With BFS-Fast, whenever the search depth increases to a new value $n + 1$, each request is appended to at most one request sequence of length $n$, instead of to all of them as in BFS [5]. BFS-Fast provides full grammar coverage only with respect to all possible renderings of individual requests but does not explore all request sequences of a given sequence length.

Although BFS-Fast scales better compared to BFS, it does so by exploring only a subset of all possible request sequences. Unfortunately, this limits the number of violations the security checkers can actively check. To alleviate this limitation, we introduce a new search strategy, called BFS-Cheap.

BFS-Cheap follows the inverse trade-off of BFS-Fast: it sacrifices full coverage of all possible request renderings at every state but explores all possible request sequences for a given sequence length, albeit not with all possible renderings. Specifically, given a set of sequences of length $n$, called $\text{seqSet}$, and a set of requests, called $\text{reqCollection}$, BFS-Cheap operates as follows:

1. For each sequence $\text{seq} \in \text{seqSet}$, append each $\text{req} \in \text{reqCollection}$ to the end of $\text{seq}$, execute the new sequence while considering the possible renderings of $\text{req}$, and add to $\text{seqSet}$ at most one valid (if any) and one invalid (if any) sequence rendering.

Valid renderings are used by the use-after-free, resource-hierarchy, and user-namespace checkers, while invalid renderings are used by the resource-leak checker.

BFS-Cheap thus provides a middle-ground between BFS and BFS-Fast (see Section IV-B for an experimental evaluation). It explores all possible request sequences up to a given sequence length (like BFS) and adds at most two new renderings for each sequence in order to avoid an enormous $\text{seqSet}$ (like BFS-Fast). Two new renderings per sequence explored allow for active checking of all the security rules defined in Section III-A while maintaining a tractable number of sequences in $\text{seqSet}$ as the sequence length increases.

Note that the suffix “cheap” comes from the fact BFS-Cheap is a cheaper version of BFS where at most one valid rendering is added to the BFS “frontier” $\text{setSeq}$ for each new sequence. This leads to the creation of fewer resources than those created when all valid renderings of each request sequence are explored, as in BFS. For instance, imagine a request definition with an enum type describing ten different flavours of the same resource type. BFS-Cheap will stop creating resources once one resource of one flavour is successfully created. In contrast, BFS and BFS-Fast, will create ten resources of the same type with ten different flavours.

E. Bug Bucketization

Before discussing examples of real violations found with active checkers, we define the bucketization scheme used to group together similar violations. In the context of active checkers, we define “bugs” as rule violations. Each bug is associated with the request sequence that was executed to trigger it. Given this property, we use the following procedure to create per-checker bug buckets:

Whenever a new bug is found, compute all non-empty suffixes of the request sequence that triggers...
TABLE I: Comparison of BFS, BFS-Fast and BFS-Cheap. Shows the maximum sequence length (Max Len.), the number of requests sent (Tests), the percentage of tests generated by the main driver (Main) and by all four checkers combined (Checkers) and individually, with each search strategy after 1 hour of search. The second column shows the total number of requests in each API.

<table>
<thead>
<tr>
<th>API</th>
<th>Total Req.</th>
<th>Search Strategy</th>
<th>Max Len.</th>
<th>Tests</th>
<th>Main</th>
<th>Checkers</th>
<th>Checker Stats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azure A</td>
<td>13</td>
<td>BFS</td>
<td>3</td>
<td>3255</td>
<td>48.1%</td>
<td>51.9%</td>
<td>11.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BFS-Cheap</td>
<td>4</td>
<td>4050</td>
<td>55.0%</td>
<td>45.0%</td>
<td>10.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BFS-Fast</td>
<td>9</td>
<td>4347</td>
<td>59.2%</td>
<td>40.8%</td>
<td>15.5%</td>
</tr>
<tr>
<td>Azure B</td>
<td>19</td>
<td>BFS</td>
<td>5</td>
<td>7721</td>
<td>46.4%</td>
<td>53.6%</td>
<td>3.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BFS-Cheap</td>
<td>5</td>
<td>7979</td>
<td>46.2%</td>
<td>53.8%</td>
<td>3.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BFS-Fast</td>
<td>40</td>
<td>17416</td>
<td>65.3%</td>
<td>34.7%</td>
<td>0.3%</td>
</tr>
<tr>
<td>O-365 C</td>
<td>18</td>
<td>BFS</td>
<td>3</td>
<td>11693</td>
<td>89.4%</td>
<td>10.6%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BFS-Cheap</td>
<td>4</td>
<td>10982</td>
<td>95.9%</td>
<td>4.1%</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BFS-Fast</td>
<td>33</td>
<td>18120</td>
<td>66.9%</td>
<td>33.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

IV. EXPERIMENTAL EVALUATION

In this section, we report results of experiments with three production cloud services. These services and our experimental setup are described in Section IV-A. Then, we compare in Section IV-B the three search strategies described in Section III-D. Next, we present results showing the number of rule violations reported by each checker on the three cloud services as well as the impact of various optimizations (Section IV-C).

A. Experimental Setup

We report results of experiments performed with three cloud services, whose names are anonymized (to avoid targeting them): Azure A and Azure B are two Azure [13] management services, and O-365 C is an Office365 [16] messaging service. The number of requests in the REST API of each of these three services ranges from 13 to 19 requests. We selected those three services because their size and complexity are representative among the cloud services we analyzed. So far, we have performed similar experiments with about a dozen production services, and our general experience with these other services is summarized in Section V.

Every service we consider has a publicly-available Swagger specification [15]. For each service, we compile its specification to produce a test-generation grammar, similarly to prior work [5]. Each grammar is encoded as executable python code. For a given service and API, the same grammar and fuzzing dictionaries were used across all the experiments reported in this section. There is no randomness in the renderings generated. We ran our fuzzing experiments using a single-threaded fuzzer running on a PC connected to the internet and a valid service subscription that allows access to each service API. No other special test setup or service knowledge was required. As in [5], our fuzzer includes a garbage-collector that deletes no-longer-used resources (dynamic objects) in order to avoid exceeding service quota limits.

We fuzz production services already deployed and accessible to anyone with a valid subscription, but we have no visibility as to what happens inside the backend of the services we test. Our fuzzer only observes the HTTP status codes of the responses it receives. All client-side requests are sent over the internet to the target services, and responses are parsed when they are received. Because we do not control the deployment of these services, the experiments reported in this section are not fully controlled. However, we performed these experiments several times and the results did not vary significantly.

B. Comparing Search Strategies

We now compare our new search strategy, BFS-Cheap, with BFS and BFS-Fast when using security checkers to fuzz real services. We present results of experiments with two Azure and one Office-365 services, denoted by Azure A, Azure B, and O-365 C respectively.

Table I shows individual experiments with the three search strategies on each service, over a fixed time budget of one hour per experiment. For each experiment, we report the total number of requests in the API (Total Req.), the maximum sequence length generated (Max Len.), the number of requests sent (Tests), the percentage of the requests sent by the main driver (Main) and by all four checkers combined (Checkers) and individually, with each search strategy after 1 hour of search. The second column shows the total number of requests in each API.

...
to be due to a significantly lower number of failed requests generated by BFS-FAST for these two services compared to BFS and BFS-Cheap. Such failed requests are sent back to the client (our fuzzer) with larger time delays. Delaying responses to failed requests is a well-known mechanism used by services to throttle future requests, i.e., to try to slow them down. For Azure B, BFS-Fast executes more tests because its request sequences are deeper but include many DELETE requests which are faster to execute (their responses are received almost instantly); BFS-Fast executes about 9 times more DELETE requests than BFS or BFS-Cheap.

The total percentage of checker tests (Checkers) is the highest for BFS and the lowest for BFS-FAST, while BFS-Cheap is again in between. Indeed, while BFS-Fast generates the largest number of tests, its search space is pruned and activates checkers less often, as discussed in Section III-D – this is the precise motivation for introducing BFS-Cheap in that section. An exception is the 33% spike in checker-generated tests by BFS-FAST for O-365 C. This spike seems to be due to a larger number of successful requests (see the previous paragraph), which in turn led to more checker tests.

From the individual checker statistics in Table I, we observe that the number of tests they each generate varies from service to service. This number depends on the number of DELETE requests executed for the use-after-free checker, the number of failed resource-creation requests for the resource-leak checker, and the depth of the object hierarchy for the resource-hierarchy checker. In contrast, the user-namespace checker is triggered more consistently more often and contributes the largest percentage of checker-generated tests.

For all three services, the number of bugs found is nearly the same for all three search strategies and is discussed next.

C. Comparing Checker Optimizations

We now compare the performance of the two modes optimized and exhaustive discussed in Section III.

Table II shows how many requests were sent in one hour of fuzzing with BFS-Cheap in the Tests column, and what percentage of those requests were generated by either the main driver of Section II or by any of the four checkers. The table also shows how many unique bugs (bug buckets) were found in one hour of search by the main driver and by each of the checkers. Results are presented for both the optimized and the exhaustive modes previously discussed.

We observe that the number of tests varies for different services and checker modes. However, the percentage of tests generated by the checkers is always higher with the exhaustive mode, as expected. Since in the optimized mode the checkers produce fewer tests per visited state, the main driver is allowed to explore more states faster. Yet, despite the lower number of checker tests per visited state, for all three services considered, the optimized mode finds all the unique bugs (bug buckets) found by the exhaustive mode. Also, for the O-365 C service, the main driver finds one more bug with the optimized mode compared to the exhaustive mode within one hour of search.

Table II reveals an interesting inversion that further demonstrates the value of the optimized checkers mode. In Azure A, we observe that the optimized mode produces almost twice as many tests than than the exhaustive mode (4050 versus 2174). At first sight, this is counter-intuitive. After a deeper investigation, we discovered that some of the tests produced by the exhaustive mode of the user-namespace checker have significantly larger response times for service Azure A. Indeed, this specific checker in exhaustive mode executes additional tests compared to the optimized mode, but containing expensive operations (i.e., high latency) that slow down the overall test throughput.

During the course of all experiments with these three services, we found and reported a total of 7 unique bugs to the developers of those services, including 4 500 bugs found by the main driver and 3 bugs found by each of the checkers except the user-namespace checker. In the next section, we discuss several interesting bugs found thanks to the checkers introduced in this paper.

V. EXAMPLES OF REST API SECURITY VULNERABILITIES

At the time of this writing, we have fuzzed nearly a dozen production Azure and Office-365 cloud services of size and complexity similar to the three services used in the previous section. In almost all cases, our fuzzing was able to find about a handful of new bugs in each of these services. About two thirds of those bugs are “500 Internal Server Errors”, and about one third are rule violations reported by our new security checkers. We reported these bugs to the service owners, and all have been fixed.

We emphasize that, even when the security checkers do not find any bugs, they increase confidence that the rules
they check cannot be violated and therefore they increase confidence in the overall service reliability and security.

This section presents examples of real bugs found in deployed Azure and Office-365 services and discuss their security relevance. We anonymize the name of those services and key details not to target any specific service.

**Use-after-free violation in Azure.** In an Azure service, we found the following use-after-free violation.

1) Create a new resource R (with a PUT request).
2) Delete resource R (with a DELETE request).
3) Create a new child resource of the deleted resource R and of a specific type (with another PUT request).

This sequence of requests results in a “500 Internal Server Error”. The Use-after-free checker catches this as (1) it attempts to re-use in Step 3 the deleted resource in Step 2 and (2) the response of Step 3 is different from the expected “404 Not Found” response.

**Resource-hierarchy violation in Office365.** In an Office365 messaging service where users can post messages and then reply and edit these, the resource-hierarchy checker detected the following bug.

1) Create a first message msg-1 (with a request POST /api/posts/msg-1).
2) Create a second message msg-2 (with a request POST /api/posts/msg-2).
3) Create a reply reply-1 to the first message (with a request POST /api/posts/msg-1/replies/reply-1).
4) Edit the reply reply-1 with a PUT request using msg-2 as message identifier (with a request PUT /api/posts/msg-2/replies/reply-1).

Surprisingly, the last request in Step 4 returns a “200 Allowed” response while it must have returned a “404 Not Found” response. This rule violation reveals that the implementation of the API that posts a reply does not analyze the full hierarchy when checking permissions for a reply. Missing hierarchy validation checks are potential security vulnerabilities: an attacker might be able to exploit them to access child objects by bypassing the parent hierarchy.

**Resource-leak violation in Azure.** In another Azure service, the resource-leak checker triggered the following bug.

1) Create a new resource of type CM and of name X with a specific malformed body (with a PUT request). This returns a “500 Internal Server Error”, which is already a bug.
2) Get a list of all resources of type CM: the returned list is empty.
3) Create a new resource of type CM with the same name X as in Step 1 with a well-formed body but in a different region (e.g., US-West versus US-Central) with a PUT request.

Unexpectedly, the last request in Step 3 returns a response “409 Conflict” instead of an expected “200 Created”. This behavior means that the service has reached an inconsistent state: the failed request in Step 1 has left unintended side-effects on the service state. Indeed, the GET request in Step 2 shows that the user view is correct: the CM resource named X attempted to be created in Step 1 has not been created. However, the second PUT request in Step 3 proves that the service still remembers the failed creation of the CM resource named X attempted in the first PUT request of Step 1. This bug is potentially dangerous: an attacker could create an unbounded number of such “zombie” resources by repeating Step 1 using many different names, and exceed his/her official quota since such failed resource creations are (correctly) not counted towards the user’ resource quota. Yet, they are clearly remembered (incorrectly) somewhere in the backend service.

Our fuzzing tool uses a garbage collector not to exceed quotas for the cloud resources created during fuzzing. For instance, if a default quota for a resource type Y is 100, at most 100 resources of that type can be created at any time, and our garbage collector makes sure that the number of live resources never exceeds quotas by deleting (using a DELETE request) resources that are no longer used. Without garbage collection, our fuzzing tool would typically reach quota limits in minutes and would not be able to continue state-space exploration.

In this specific Azure service, any PUT request to create a resource of a specific type, let us call it IM, returns a response quickly but actually triggers other tasks that take minutes to complete in the service backend. Similarly, a DELETE request for an IM resource also returns quickly but also triggers delete tasks that also take minutes to complete. However, such PUT and DELETE requests for IM resources update counters towards quotas eagerly, too quickly, without waiting for the several minutes actually needed to fully complete these tasks. As a result, an attacker could create-then-delete quickly many IM resources without exceeding his/her quota, while triggering a huge number of backend tasks, hence literally flooding the backend service. Such a Denial-of-Service attack was accidentally triggered by our fuzzing tool.

A fix to this vulnerability is to update usage counters towards quotas for DELETE requests only when all delete backend operations have been completed, i.e., minutes later in the case of IM resources. This way, the amount of backend tasks is still linearly bounded by the official quota, since subsequent IM resource-creation PUT requests will be blocked until preceding DELETE requests have been fully completed.

**VI. RELATED WORK**

Our work extends stateful REST API fuzzing [5]. Given a Swagger specification of a REST API, this specification is compiled into a fuzzing grammar, which is then used to automatically generate sequences of requests that satisfy the specification. Stateful REST API fuzzing automates the generation of a fuzzing grammar compared to traditional grammar-based fuzzing [20], [22], [24] where the user manually writes a grammar. The BFS and BFS-Fast search strategies are inspired by test generation algorithms used in model-based testing [27].
Given the widespread use of REST APIs, there is surprising little guidance on secure REST API usage. Most of the security guidance is provided by organizations like OWASP [19] (Open Web Application Security Project) or from books on REST APIs [1] or micro-services [17] about managing authentication tokens and API keys. However, these tools do not perform any global analysis of Swagger specifications, including requests, responses, and their contents (like embedded JSON data), and then fuzz those using either pre-defined heuristics [4], [21] or user-defined rules [23], [6]. Some tools to capture, parse, fuzz, and replay HTTP traffic have recently been extended to leverage Swagger specifications in order to construct HTTP requests and guide their fuzzing [4], [21], [26], [3]. However, these tools do not perform any analysis of Swagger specifications and therefore cannot generate new sequences of requests: their fuzzing is stateless, i.e., restricted to fuzzing parameter values of individual requests. Therefore, adding active checkers to stateless fuzzers is problematic. In contrast, our work extends stateful REST API fuzzing with active checkers targeting specific REST API rule violations.

The widespread use of REST APIs, there is surprising little guidance on secure REST API usage. Most of the security guidance is provided by organizations like OWASP [19] (Open Web Application Security Project) or from books on REST APIs [1] or micro-services [17] about managing authentication tokens and API keys. However, these tools do not perform any global analysis of Swagger specifications, including requests, responses, and their contents (like embedded JSON data), and then fuzz those using either pre-defined heuristics [4], [21] or user-defined rules [23], [6]. Some tools to capture, parse, fuzz, and replay HTTP traffic have recently been extended to leverage Swagger specifications in order to construct HTTP requests and guide their fuzzing [4], [21], [26], [3]. However, these tools do not perform any analysis of Swagger specifications and therefore cannot generate new sequences of requests: their fuzzing is stateless, i.e., restricted to fuzzing parameter values of individual requests. Therefore, adding active checkers to stateless fuzzers is problematic. In contrast, our work extends stateful REST API fuzzing with active checkers targeting specific REST API rule violations.

We introduced four security rules that capture desirable properties of REST APIs and services. We then showed how a stateful REST API fuzzing approach can be extended with active property checkers that automatically test and detect violations of these rules. So far, we have fuzzed nearly a dozen production Azure and Office-365 cloud services using the fuzzing and checkers described in this paper. In almost all cases, our fuzzing was able to find a handful of new bugs in each of these services. About two thirds of those bugs are “500 Internal Server Errors”, and about one third are rule violations reported by our new security checkers. We reported all these bugs to the service owners, and all have been fixed.

Indeed, violations of the four security rules introduced in this paper are clearly potential security vulnerabilities. The bugs we found have all been taken seriously by the respective service owners: our current bug “fixed/found” ratio is nearly 100%. Moreover, it is safer to fix these bugs rather than risk a live incident – provoked intentionally by an attacker or triggered by accident – with unknown consequences. Finally, it helps that these bugs are easily reproducible and that our fuzzing approach reports no false alarms.

How general are these results? To find out, we need to fuzz more services through their REST APIs and check more properties to detect different kinds of bugs and security vulnerabilities. Given the recent explosion of REST APIs for cloud and web services, there is surprisingly little guidance about REST API usage from a security perspective. Our paper makes a step in that direction by contributing four rules whose violations are security-relevant and which are non-trivial to check and satisfy.
REFERENCES