**Gang of Eight: A Defect Taxonomy for Infrastructure as Code Scripts**

Akond Rahman  
Tennessee Tech University  
 Cookeville, Tennessee  
arahman@tntech.edu

Effat Farhana  
NC State University  
Raleigh, North Carolina  
efarhan@ncsu.edu

Chris Parnin  
NC State University  
Raleigh, North Carolina  
cjparnin@ncsu.edu

Laurie Williams  
NC State University  
Raleigh, North Carolina  
lawilli@ncsu.edu

**ABSTRACT**

Defects in infrastructure as code (IaC) scripts can have serious consequences, for example, creating large-scale system outages. A taxonomy of IaC defects can be useful for understanding the nature of defects, and identifying activities needed to fix and prevent defects in IaC scripts. The goal of this paper is to help practitioners improve the quality of infrastructure as code (IaC) scripts by developing a defect taxonomy for IaC scripts through qualitative analysis. We develop a taxonomy of IaC defects by applying qualitative analysis on 1,448 defect-related commits collected from open source software (OSS) repositories of the Openstack organization. We conduct a survey with 66 practitioners to assess if they agree with the identified defect categories included in our taxonomy. We quantify the frequency of identified defect categories by analyzing 80,425 commits collected from 291 OSS repositories spanning across 2005 to 2019.

Our defect taxonomy for IaC consists of eight categories, including a category specific to IaC called idempotency (i.e., defects that lead to incorrect system provisioning when the same IaC script is executed multiple times). We observe the surveyed 66 practitioners to agree most with idempotency. The most frequent defect category is configuration data i.e., providing erroneous configuration data in IaC scripts. Our taxonomy and the quantified frequency of the defect categories can help practitioners improve IaC script quality by prioritizing verification and validation efforts.

**CCS CONCEPTS**

- Software and its engineering → Software defect analysis:

**KEYWORDS**

bug, category, defect, devops, infrastructure as code, puppet, quality

**ACM Reference format:**

https://doi.org/10.1145/nnnnnnn.mnnnnnn

1 INTRODUCTION

Infrastructure as code (IaC) is the practice of automatically maintaining system configurations and provisioning deployment environments using source code [26]. IaC scripts are also known as configuration as code scripts [57], or configuration scripts [64]. Information technology (IT) organizations use tools, such as Chef 1 and Puppet 2 to implement the practice of IaC. The use of IaC has yielded benefits for IT organizations, such as General Motors (GM) and the European Organization for Nuclear Research (CERN). Using Chef, GM increased software deployment frequency by a factor of 21 [21]. CERN uses Puppet to manage 15,000 servers and to process 2,000 terabytes of data everyday [9]. Puppet has helped CERN to minimize service disruptions and reduce deployment time [2].

Despite the above-mentioned benefits, defects with serious consequences can appear in IaC scripts. For example, a defect in an IaC script created an outage resulting in business losses worth of 150 million USD for Amazon Web Services [24]. Other example consequences of defects in IaC scripts include outage for GitHub [20] and deletion of user directories for ~270 users in cloud instances maintained by Wikimedia Commons [14].

A defect taxonomy for IaC scripts can help practitioners understand the nature of defects, and identify possible development activities for defect mitigation. Figure 1 presents an example of a security defect, which exposes users’ passwords in logs 3. The defect is mitigated by adding ‘secret=true’, which prevents the password getting exposed in logs 3. Identification of certain defect categories, such as security defects similar to Figure 1, can help practitioners to make informed decisions on what development activities could be adopted to improve the quality of IaC scripts.

The importance of defect categorization has been recognized by the research community [8, 39, 79, 82]. For example, Linraes-Vasquez et al. [39] stated categorizing vulnerabilities can help Android developers “in focusing their verification & validation (V&V) activities”. According to Catolino et al. [8], “understanding the bug type represents the first and most time-consuming step to perform in the process of bug triage, since it requires an in-depth analysis”. A defect taxonomy for IaC, an area that remains unexplored, can help practitioners in understanding the nature of defects, and help in improving IaC script quality through activities such as triaging defects, prioritizing V&V efforts, and measuring IaC script quality.

We observe practitioners asking about defect categories for IaC scripts in online forums such as Reddit: “I want to adopt Puppet in my organization. Before adoption I want to be aware of the quality issues that may arise in Puppet scripts. Can someone give me some pointers in 1

---

1https://www.chef.io/  
2https://puppet.com/  
3https://bugs.launchpad.net/puppet-ceilometer/+bug/1328448
on what type of bugs/defects appear for Puppet scripts? " [17]. While the forum members offered suggestions on possible categories of defects, e.g., syntax-related defects, these suggestions lack substantiation. We hypothesize that through systematic empirical analysis we can develop a taxonomy, and derive defect categories for IaC using open source software (OSS) repositories.

The goal of this paper is to help practitioners improve the quality of infrastructure as code (IaC) scripts by developing a defect taxonomy for IaC scripts through qualitative analysis.

We answer the following research questions:

- [Categorization] RQ1: What categories of defects appear in infrastructure as code scripts?
- [Perception] RQ2: How do practitioners perceive the identified defect categories for infrastructure as code scripts?
- [Frequency] RQ3: How frequently do the identified defect categories appear in infrastructure as code scripts?

We derive a defect taxonomy for IaC by applying descriptive coding [62] on 1,448 defect-related commit messages collected from OSS repositories maintained by the Openstack organization [44]. We survey 66 practitioners to assess how practitioners perceive the identified defect categories. To automatically identify IaC defect categories, we develop Automated Categorizer of Infrastructure as code Defects (ACID), and apply ACID on 80,415 commits collected from 291 OSS repositories hosted by GitHub, Mozilla [42], Openstack [45], and Wikimedia Commons [15].

We list our contributions as following:

- A taxonomy that includes eight defect categories for IaC scripts (Section 3.2);
- An evaluation of how practitioners perceive the identified defect categories (Section 3.3.2);
- An analysis of how frequently the identified defect categories occur (Section 5.2.2); and
- A tool called that automatically identifies instances of defect categories from repositories with IaC scripts (Section 4).

We organize rest of the paper as following: we discuss background and related work in Section 2. We describe the defect categories and survey results in Section 3. We describe the construction and evaluation of ACID in Section 4. We describe our empirical study in Section 5. We discuss our findings and limitations respectively, in Sections 6 and 7. We conclude the paper in Section 8.

2 BACKGROUND AND RELATED WORK

Here, we provide necessary background and discuss related work.

---

Figure 1: Example of a security-related defect where a password is exposed in logs. The defect is mitigated by adding `secret=>true`.

Figure 2: Annotation of an example Puppet script.
identified 21,201 occurrences of security smells i.e., coding patterns indicative of security weaknesses that also included 1,326 occurrences of hard-coded passwords. The above-mentioned discussion highlights the lack of studies that investigate defect categories for IaC and motivate us further to investigate defect categories of IaC.

3 DEFECT TAXONOMY FOR INFRASTRUCTURE AS CODE SCRIPTS

In this section, we answer RQ1: *What categories of defects appear in infrastructure as code scripts?*. First, we describe our methodology, then we present the identified defect categories.

3.1 Methodology to Develop Defect Taxonomy

Qualitative Analysis: We use descriptive coding [62] on 1,448 defect-related commits. Defect-related commits have commit messages that indicate an action was taken related to a defect [58]. We derive defect-related commits by manually inspecting 7,808 commits that map to 1,386 Puppet scripts. We collect commits from 61 OSS repositories maintained by the Openstack organization [45], as Openstack provides cloud-based infrastructure services, and our assumption is that an analysis of commits collected from Openstack could give us insights on defect categories. We use commits because commits summarize changes that are made to a script and could identify the types of changes that are being performed on a script. Descriptive coding is a qualitative analysis technique that summarizes the underlying theme from unstructured text [62]. We select descriptive coding because we can obtain (i) a summarized overview of the defect categories that occur for IaC scripts; and (ii) context on how the identified defect categories can be automatically identified. We use Puppet scripts to construct our dataset because Puppet is considered as one of the most popular tools for configuration management [31] [63], and has been used by companies since 2005 [40].

We extract commit message text from the 1,448 defect-related commits, as well as any existing identifier to a bug report in the commit message. We combine the commit message with any existing bug report description and refer to the combination as enhanced commit message (ECM). If no bug identifier is present in the commit message, then the commit message becomes the ECM. We use ECMs to derive defect categories as following: first we identify text patterns that describe a reason and/or a symptom of a defect, where defect is defined as—an imperfection that needs to be replaced or repaired [28]. Figure 3 provides an example of our qualitative analysis process. We first analyze the ECM for each commit where an IaC script is modified, and extract snippets that correspond to a reason or symptom of a defect. From the snippet provided in the bottom left corner, we extract raw text: ‘fix config options’. Next, we generate the initial category e.g., we generate ‘fixing config options’ from this raw text. Finally, we determine the defect category ‘Configuration Data’ by combining initial categories. We combine these two initial categories, as both correspond to a common pattern of fixing configuration data defects. Multiple defect categories can be identified in an ECM.

The first and second author, individually, conduct a descriptive coding process on 1,448 defect-related ECMs. Upon completion of this process, we record the agreements and disagreements for the identified defect categories. The first and second author, respectively, identified eight and ten categories. The Cohen’s Kappa [13] is 0.8, which according to Landis and Koch [37] is ‘substantial agreement’. The second author identified two additional categories: hard-coded values and network setting. Upon discussion, the first and second authors agreed that both hard-coded values and network setting can be merged with the category ‘configuration data’, as defects related to configuration data include both.

3.2 Answer to RQ1: Defect Taxonomy for IaC

Our developed taxonomy includes eight defect categories. We observe one category, idempotency, to be unique to IaC, whereas, the other defect categories have been observed for other software systems as reported in prior literature [8, 30, 76, 80]. We report each defect category in an alphabetical order, with examples below:

**Conditional**: This category represents defects that occur due to erroneous logic and/or conditional values used to execute one or multiple branches of an IaC script. Conditional logic defects also appear for other types of software systems such as, machine learning [72], relational databases [16], [51], text editor software [16], and deep learning software [30].

**Example**: Conditional logic defects can lead to erroneous status output. For example, for a Wikimedia Commons project ‘cdh’ [78], when checking the status of mysqldadmin, a conditional logic defect caused the output to be 0 (zero), both in the case of success and failure of mysqldadmin’s ‘ping’ command.

**Configuration Data**: This category represents defects that happen due to erroneous configuration data that reside in IaC scripts. An example configuration data defect downloaded from an OSS repository [46] is ‘$config_dir="/etc/service_name"’, where wrong configuration data is provided. The fix is ‘$config_dir="/etc/hecad"’. Practitioners have reported configuration data defects in IaC scripts to cause deployment problems for the Google App Engine [52], and inavailability of StackExchange [68], an online question and answer platform [67]. ‘Configuration data’ includes five sub-categories: (i) data for storage system such as MySQL, MongoDB, and SQLite; (ii) data for file system such as specifying file permissions and file names; (iii) data for network setup such as TCP/DHCP ports and addresses, MAC addresses, and IP table rules; (iv) data for user credentials such as usernames; and (v) data for caching systems such as Memcached.

Prior research has observed configuration data defects to appear for machine learning software [72], build systems [79], blockchain software [76], Eclipse software projects [8], and Mozilla projects [8].

**Example**: Configuration data-related defects cause deployment failures: in the case of Openstack Bug#1592842 ⁴, value of an attribute ‘host_ip’ was not set to the correct IP address, which lead to a deployment task to fail. Other example consequences of configuration data defects include provisioning errors for object storage systems such as Openstack Swift ⁵.

**Dependency**: This category represents defects that occur when execution of an IaC script is dependent upon an artifact, which is either missing or incorrectly specified. The artifact can be a file,
class, package, Puppet manifest, or a module. Previously, practitioners have reported that managing dependencies in IaC scripts leads to complexity, stating dependency management as a ‘nightmare’ [23], because incorrectly specified dependencies could lead to execution of IaC scripts in the wrong order. Dependency-related defects have been reported for machine learning [72], and audio processing software [16].

Example: Dependency defects cause deployment errors. According to the bug report6, a user faces an error when installing an Openstack project called ‘Neutron’ [49], which provides networking as a service in Debian and Ubuntu. The error occurred because of specifying an incorrect package as a dependency. Dependency defects can also cause installation errors [4].

Documentation: This category represents defects that occur when incorrect information about IaC scripts are specified in source code comments, in maintenance notes, and in documentation files such as README files. Documentation-related defects have been reported in prior studies: Chillarege et al. [11] identified documentation as one of the eight defect categories for ODC. Storey et al. [69] and Tan et al. [71] have reported that erroneous information expressed in source comments can lead to defects in source code.

Example: According to a bug report7, the format of specifying policies for the Openstack Nova project is incorrectly specified in the comments of an IaC script, which can lead to Puppet runtime errors. Other examples of documentation-related defects are providing incompatible license information in comments, and missing license headers [4], which can have negative implications on the distribution of software [75].

Idempotency: This category represents defects that violate the idempotency property for IaC scripts. For IaC, idempotency is the property which ensures that even after n executions, where n > 1, the provisioned system’s environment is exactly the same as it was after the first execution of the relevant IaC scripts. We use a Reddit post8 to explain further. In the post, the user mentions that a new string is appended every time the IaC script is executed. Such execution becomes problematic when the script is run multiple times, because the desired behavior is that the new string will only be added once at the first execution of the script. Idempotency-related defects have not been reported in prior research related to defect categorization for non-IaC software. However, in IaC-related publications, researchers have reported the existence of idempotency defects for Chef scripts [27] and for CFEngine scripts [6].

Example: An idempotency defect caused unwanted changes to artifacts that should not be modified. Idempotency-related defects can cause file backup problems, package update problems, database setup problems, logging problems, and network setup problems [4].

Security: This category represents defects that violate confidentiality, integrity or availability for the provisioned system. Example security-related defects include exposing secrets in logs; specifying SSL certificates that lead to authentication problems; and hard-coding secrets, such as passwords [4]. Prior research on defect categorization has also observed security-related defects to appear in other types of software systems e.g., in blockchain projects [76], video game software [50], cloud management software [80], and OSS Apache, Linux, and Mozilla projects [8] [70].

Example: Setting up user accounts is common in IaC scripts [58], but in the process user confidentiality may be breached. For example, the security defect from Section 1 is listed in a bug report [81]. The defect leaks passwords using variables such as rabbit_password in Puppet logs. The defect impacts seven Openstack projects and is fixed by adding the ‘secret’ attribute10. Another example is keeping a token alive indefinitely, which provides access to anyone without authentication and lasts indefinitely if not disabled11.

Service: This category represents defects related to improper provisioning and inadequate availability of computing services, such as load balancing services and monitoring services. Service-related defects can be caused by improper specification of attributes while using the ‘service’ resource for Puppet [36], or Chef [10], or the ‘service’ module for Ansible [3]. Service-related defects have previously reported for cloud management software [80]. Rahman and Williams [58] reported provisioning of web service and database services to appear in defective IaC scripts.

Example: Service-related-defects can cause provisioned services to stop unexpectedly, as it happened for a provisioned MySQL

---

6https://bugs.launchpad.net/puppet-neutron/+bug/1288741
7https://bugs.launchpad.net/puppet-nova/+bug/1490897
8https://www.reddit.com/r/Puppet/comments/478deb/
9https://bugs.launchpad.net/puppet-nova/+bug/1490897
10https://bugs.launchpad.net/puppet-nova/+bug/1490897
11https://bugs.launchpad.net/puppet-nova/+bug/1490897
service. The defect occurred as the provisioned service was not working with a title provided as input. Other examples of service-related defects include not being able to start monitoring services, such as Nagios, and load balancing services, such as HAProxy.

**Syntax**: This category represents defects related to syntax in IaC scripts. Syntax-defects have also been reported in prior research studies: Ray et al. [60] and Pascarella et al. [50] in separate studies identified generic programming defects for OSS GitHub projects, which included programming syntax errors. Islam et al. [30] also identified syntax-related defects for deep learning projects.

**Example**: The Openstack Fuel plugin development workflow [47] uses a continuous integration pipeline, which generates errors if IaC scripts fail style checks. Other examples of syntax-related defects include specifying wrong types for variables [4].

### 3.3 RQ2: Practitioner Perception

We present the methodology and findings for RQ2: **How do practitioners perceive the identified defect categories for infrastructure as code scripts?**

**Methodology for RQ2**: Practitioner agreement with the identified defect categories in Section 3.2 can indicate the relevance of the eight categories. We answer RQ2 by deploying an online survey to practitioners. In the survey, we ask practitioners how many years they worked with IaC scripts. Then, we provide definitions and examples for each category. Next, we asked “We have identified eight defect categories for IaC (Puppet) scripts. Each of these categories are listed below. To which extent do you agree with our list?”

We construct the survey following Kitchenham and Pfleeger’s guidelines [34]: (i) use Likert scale to measure agreement levels: strongly disagree, disagree, neutral, agree, and strongly agree; (ii) add explanations related to the purpose of the study, how to conduct the survey, preservation of confidentiality, relevance to participants, and an estimate of completion time; and (iii) conduct a pilot survey to get initial feedback. From the feedback of the pilot survey, we added an open-ended text box so that surveyed practitioners can further respond. The survey questionnaire is available and included in our verifiability package [5].

We offered a drawing of two 50 USD Amazon gift cards as an incentive for participation following Smith et al. [65]’s recommendations on incentivizing surveys. We deploy the survey to 750 practitioners from April 19, 2019 to August 15, 2019. We distributed the survey to practitioners via e-mail following the Internal Review Board (IRB) protocol#16813. We collect practitioner e-mail addresses from the OSS repositories used in our empirical analysis.

**Answer to RQ2**: Of the 750 practitioners, 66 responded. We observe the eight categories to have relevance amongst practitioners: 50%~79% of the 66 survey respondents agree with one of the eight categories. Respondents agree most with idempotency, with an agreement rate of 79%. The least agreed upon category is syntax-related defects. Details of our survey results is listed in Figure 4, where the defect categories are sorted from top to bottom based on agreement rate. The percentage of respondents who agreed or strongly agreed with each category is listed on the right.

Practitioners provide reasoning on why they agree, disagree, or remain neutral. For example, one practitioner considered syntax-related defects to be less important, as syntax-related defects should be found “during initial testing”. The practitioner identified idempotency and conditional defects as “two key ones”. Another practitioner highlighted the importance of idempotency, stating idempotency defects “may not produce functional issues, but can cause product indirect issues”. We notice contrary views related to idempotency as well. One practitioner stated “I have never run into defects in Puppet’s idempotency mechanisms”, instead the practitioner claimed dependency-related defects to be prevalent stating “Debugging dependency loops is tricky and happens often”. Another practitioner indicated limitations of IaC tools as a possible explanation on why dependency defects could be frequent “the problem [is] with these [IaC] tools it is not clear how to do it [dependency management]”. For practitioners who provided neutral responses, the presence of defects is more relevant than the category “I am more concerned about the defect itself rather than its category. If I have that info, it will be good, but it is not mandatory”.

In summary, we observe our identified defect categories to have relevance amongst practitioners. Surveyed practitioners have varying opinions on what defect categories are more frequent in IaC development, which are possibly formed by their own experiences.

### 4 AUTOMATED CATEGORIZER FOR INFRASTRUCTURE AS CODE DEFECTS

We construct an automated tool called Automated Categorizer for Infrastructure as Code Defects (ACID) to automatically identify IaC defect categories at scale. ACID analyzes each ECM and determines whether or not any of the eight identified defect categories can be identified from the ECM. ACID takes one or multiple repositories as input, and as output reports the defect category for each ECM. If none of the eight categories are identified, ACID reports ‘NO DEFECT’, indicating no defect is identified. ACID can report an ECM to belong to multiple defect categories.

We first describe how we construct the rules used by ACID to detect defect categories. Next, we use a running example to illustrate the components of ACID, and how ACID determines a defect category. Finally, we evaluate the accuracy of ACID by using an oracle dataset.

**Rules used by ACID**: We construct rules needed for ACID by abstracting patterns that appear in the 1,448 ECMS and their corresponding diffs, obtained from the 61 Openstack OSS repositories. Our hypothesis is that we can abstract patterns from the 1,448
ECMs and their corresponding diffs to automatically detect defect categories for other datasets.

We use Table 1 to demonstrate our approach. The ‘ECM’ column presents a set of ECMs that include ‘fix’, a string pattern that represents a defect-related action. The ‘Dependent’ column shows the string patterns upon which the defect-related action is applied. For example, for ‘fix catalog compilation when not configuring endpoint’, ‘fix’ is applied upon ‘compilation’. We can abstract patterns using ‘fix’ and tokens in the ‘Dependent’ column to construct a rule to identify a defect category. For example, in Table 1, the keywords ‘compilation’, ‘warning’, ‘lint’ and, ‘typo’ is linked with ‘fix’. The first author looked into the set of 1,448 ECMs to determine what string patterns need to be captured as dependents to construct rules.

Some ECMs such as ‘Various small fixes’, which is downloaded from an OSS repository called ‘fuel-plugin-contrail’\textsuperscript{13}, indicate that a defect-related action occurred but do not clearly express what defect category could be identified. We address this challenge by inspecting diffs of ECMs, in addition to dependency parsing. From the diffs of ECMs we can identify code elements and use them in rules for any of the eight categories. As an example, in Figure 5 we present the changeset for the ECM ‘Various small fixes’. We observe that configuration data are changed using three variables ‘network\_scheme’, ‘cidr\_short’, and ‘netmask’.

An ECM can include a sentence that expresses defect-related actions for multiple defect categories. For example, for ‘Fix dependencies and various typos’, which is downloaded from an OSS repository \cite{48}, we observe actions taken to resolve two defect categories: dependency defect and syntax defect. For ECMs with one or multiple sentences, ACID maps an ECM to multiple defect categories if multiple rules are satisfied. The rules ACID uses to detect each of the eight defect categories is listed in Table 2. In Table 2, the ‘Category’ and ‘Rule’ column respectively presents the defect categories, and corresponding rules to detect the category. The presented rules use functions. The functions use string pattern matching to check if sentences and/or diffs of an ECM satisfy a certain condition. A mapping of the functions and stemmed string patterns are provided in Table 3. For example, the has\_Defect\(x\).\(sen\) function returns true if any of the provided string patterns listed in Table 3, appear for a sentence in an ECM x.sen. In Table 2, functions changed\_Include\(x\).\(diff\), changed\_Comment\(x\).\(diff\), and changed\_Service\(x\).\(diff\), identify if the diff of ECM x consists of changes in include statements, comments, and service resources, respectively. changed\_Include\(x\).\(diff\) identifies if the diff includes a change in configuration data. Except for function has\_Defect\(x\), string patterns for all other functions are derived from our qualitative analysis process. We derive string patterns for has\_Defect\(x\) from prior work \cite{60,59} and ACID does not use diff content because our qualitative analysis did not identify any Puppet-specific code element in the diffs that express the defect category of interest.

**Execution of ACID:** We construct ECMs using Git, Mercurial, and bug report APIs, and feed the ECMs as input to ACID. ACID uses four steps to identify a defect category: sentence tokenization, text pre-processing, dependency parsing, and rule matching. We use Table 4 to describe how ACID processes an example ECM.

---

**Table 1: An Example of Identifying Dependents in ECMs for Rule Construction**

<table>
<thead>
<tr>
<th>ECM</th>
<th>Dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>fix catalog compilation when not configuring endpoint</td>
<td>compilation</td>
</tr>
<tr>
<td>fix puppet lint 140 characters</td>
<td>lint</td>
</tr>
<tr>
<td>typo fix</td>
<td>typo</td>
</tr>
</tbody>
</table>

- − $netmask = pick(get_network_role_property('neutron/mesh', 'netmask'))
- + $netmask_short = netmask_to_cidr($netmask)

**Figure 5:** Diff of ECM 'Various small fixes'. Configuration data are changed in the diff indicating the ECM to be related with category 'configuration data'.

**Table 2: Rules to Detect Defect Categories**

<table>
<thead>
<tr>
<th>Category</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional</td>
<td>hasDefect(x).(sen) ∧ hasCond(x).(sen).(dep)</td>
</tr>
<tr>
<td>Data</td>
<td>hasDefect(x).(sen) ∧ (hasStor(x).(sen).(dep) ∨ hasFile(x).(sen).(dep))</td>
</tr>
<tr>
<td>Data</td>
<td>hasNet(x).(sen).(dep) ∨ hasUser(x).(sen).(dep) ∨ hasCach(x).(sen).(dep)</td>
</tr>
<tr>
<td>Dependency</td>
<td>hasDefect(x).(sen) ∧ (hasDep(x).(sen).(dep) ∨ changed_Include(x).(diff))</td>
</tr>
<tr>
<td>Documentation</td>
<td>hasDefect(x).(sen) ∧ (hasDoc(x).(sen).(dep) ∨ changed_Comment(x).(diff))</td>
</tr>
<tr>
<td>Service</td>
<td>hasDefect(x).(sen) ∧ (hasServ(x).(sen).(dep) ∨ changed_Service(x).(diff))</td>
</tr>
<tr>
<td>Syntax</td>
<td>hasDefect(x).(sen) ∧ hasSynt(x).(sen).(dep)</td>
</tr>
</tbody>
</table>

**Table 3: String Patterns Used for Functions in Rules**

<table>
<thead>
<tr>
<th>Function</th>
<th>String Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasDefect(x)</td>
<td>'error', 'bug', 'fix', 'issue', 'mistake', 'incorrect', 'fault', 'defect', 'flaw'</td>
</tr>
<tr>
<td>hasCond(x)</td>
<td>'logic', 'conditi', 'boolean'</td>
</tr>
<tr>
<td>hasStor(x)</td>
<td>'sql', 'db', ' databases'</td>
</tr>
<tr>
<td>hasFile(x)</td>
<td>'file', 'permiss'</td>
</tr>
<tr>
<td>hasNet(x)</td>
<td>'network', 'ip', 'address', 'port', 'tcp', 'dhep'</td>
</tr>
<tr>
<td>hasUser(x)</td>
<td>'user', 'username', 'password'</td>
</tr>
<tr>
<td>hasCach(x)</td>
<td>'cach'</td>
</tr>
<tr>
<td>hasDep(x)</td>
<td>'require', 'depend', 'relat', 'order', 'sync', 'compat', 'ensur', 'in-herit'</td>
</tr>
<tr>
<td>hasDoc(x)</td>
<td>'doc', 'comment', 'spec', 'licens', 'copyright', 'notic', 'header', 'readin'</td>
</tr>
<tr>
<td>hasIdem(x)</td>
<td>'idempot'</td>
</tr>
<tr>
<td>hasSecu(x)</td>
<td>'vulner', 'ssl', 'secr', 'authent', 'password', 'secur', 'cve'</td>
</tr>
<tr>
<td>hasServ(x)</td>
<td>'servic', 'server'</td>
</tr>
<tr>
<td>hasSynt(x)</td>
<td>'compil', 'lint', 'warn', 'typo', 'spell', 'indent', 'regex', 'variabl', 'whitespac'</td>
</tr>
</tbody>
</table>
**Table 5: ACID’s Accuracy for Oracle Dataset**

<table>
<thead>
<tr>
<th>Category</th>
<th>Occur.</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional</td>
<td>6</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Configuration Data</td>
<td>29</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>Dependency</td>
<td>3</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Documentation</td>
<td>6</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Idempotency</td>
<td>3</td>
<td>0.75</td>
<td>1.00</td>
</tr>
<tr>
<td>Security</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Service</td>
<td>13</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>Syntax</td>
<td>6</td>
<td>0.86</td>
<td>1.00</td>
</tr>
<tr>
<td>No defect</td>
<td>65</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.84</td>
<td>0.96</td>
</tr>
</tbody>
</table>

**Evaluation of ACID:** We use raters who are not authors of the paper to construct an oracle dataset to evaluate ACID. To construct the oracle dataset, raters perform closed coding \[62\], where they map each assigned ECM to any of the categories identified in Section 3.2. The mapping task was made available using a website. Each rater was provided with the IEEE Standard Classification for Software Anomalies \[28\] and a handbook on Puppet \[36\] for reference.

We recruit raters from a graduate-level class of 60 students, where 22 students volunteered to participate in constructing the oracle dataset. The graduate class is focused on DevOps principles. We used balanced incomplete block design \[19\] to select and distribute the 132 ECMs amongst the 22 raters, so that each ECM is reviewed by at least two raters. Upon completion of the mapping task, we observe the agreement rate to be 71.9%, and Cohen’s Kappa to be 0.61, which is ‘substantial agreement’ according to Landis and Koch \[37\]. The first author resolved disagreements.

Upon completion of constructing the oracle dataset, we evaluate ACID by computing the precision and recall of ACID for the oracle dataset. Precision refers to the fraction of correctly-identified categories among the total identified defect categories, as determined by ACID. Recall refers to the fraction of correctly-identified defect categories that have been retrieved by ACID over the total amount of defect categories. We report the precision and recall values for each defect category in Table 5. We observe the average precision and recall to be respectively 0.84 and 0.96 across all categories.

We notice that false negatives occur when the dependency parsing technique incorrectly identify tokens in the ECM that have no relationship to a defect category or identifies tokens that are related to an incorrect category. For example, for the ECM ‘fix following
warnings’, instead of ‘warnings’ dependency parsing incorrectly identifies ‘following’ as the dependent. We observe keyword matching to contribute to false positives, for example ACID incorrectly identifies the ECM ‘Bug 1085520 - Support instance-store backed AMIs for builders. Add new secrets.’ as a security-related defect, even though the ECM corresponds to adding a new feature.

Verifiability: All constructed datasets and ACID’s source code are available online [5].

5 EMPirical analysis

In this section, we provide the methodology and findings for the research question: **RQ3**: How frequently do the identified defect categories appear for infrastructure as code scripts?

### 5.1 Datasets

We conduct our empirical analysis with four datasets of Puppet scripts. We construct three datasets from the OSS repositories maintained by three organizations: Mozilla [42], Openstack [45], and Wikimedia Commons [15]. We select repositories from these three organizations because these organizations create or use cloud-based services. We construct the other dataset from OSS repositories hosted on GitHub, as companies tend to host their popular OSS projects on GitHub [35] [1].

Munaiah et al. [43] advocated for curation of OSS repositories before conducting empirical analysis. We apply the following criteria to curate the collected repositories: **Criteria-1**: At least 11% of the files belonging to the repository must be IaC scripts. Jiang and Adams [31] reported that in OSS repositories IaC scripts co-exist with other types of files, such as source code files. They observed a median of 11% of the files to be IaC scripts. By using a cutoff of 11% we assume to collect repositories that contain sufficient amount of IaC scripts. **Criteria-2**: The repository is not a copy of another repository. **Criteria-3**: The repository must have at least two commits per month. Muniaiah et al. [43] used the threshold of at least two commits per month to determine which repositories have enough software development activity. We use this threshold to filter repositories with limited activity. **Criteria-4**: The repository has at least 10 contributors. Our assumption is that the criteria of at least 10 contributors may help us to filter out irrelevant repositories. Previously, researchers have used the cutoff of at least nine contributors [55] [1] as a curation criterion. As shown in Table 6, 291 repositories met our filtering criteria. Attributes of the collected repositories are available in Table 7.

### 5.2 Sanity Check for ACID

Before applying ACID for all datasets, we apply a sanity check to assess ACID’s detection performance for ECMs not included in the oracle dataset. The first author randomly selected 500 ECMs from each of the four datasets, and applied ACID on 2,000 ECMs. We report the precision and recall in Table 8. We observe the recall for all defect categories is $\geq 0.91$, which indicates ACID’s ability to detect most existing defect categories, but generate false positives.

### 5.3 RQ3: Frequency of Defect Categories

In this section, we answer **How frequently do the identified defect categories appear in infrastructure as code scripts?**

#### 5.3.1 Methodology for Defect Category Frequency

We answer RQs by reporting ‘defect proportion’ and ‘script proportion’ values, as well as, temporal frequency for each defect category. ‘Defect Proportion’ refers to the percentage of commits in the dataset that include at least one category. Practitioners can find ‘script proportion’ values useful, as the metric quantifies how many scripts each category of defects evolve with time, and summarizes temporal trends. We report the temporal frequency for each year using Equation 1, where we calculate the metric ‘Defect/Year’ for a certain category $x$ that occurs in year $y$.

\[
\text{Defect/Year}(x, y) = \frac{\text{number of ECMs in year } y \text{ and marked as defect category } x}{\text{total ECMs in year } y}
\]

#### 5.3.2 Answer to RQ3: How frequently do the identified defect categories appear in infrastructure as code scripts?

We report defect proportion and script proportion values for the eight categories in Table 9. Configuration data defects is the most frequently occurring category based on the defect proportion metric. Network is the most frequent subcategory: 75.3%–88.2% of the identified
configuration data defects are network defects. Frequency of subcategories for configuration data defects is available online [5]. We observe 23.5%–33.9% scripts to be modified in commits related to configuration data, which is the highest across all defect categories. As reported in the ‘Total’ row in Table 9, we observe 16.6%, 10.9%, 18.7%, and 15.2% of all commits to include at least one of the eight defect categories, respectively, for GitHub, Mozilla, Openstack, and Wikimedia. Also, we observe 43.6%, 47.4%, 48.6%, and 49.2% of all scripts to include at least one defect category respectively, for GitHub, Mozilla, Openstack, and Wikimedia. Similar to defect proportion, based on script proportion values, we observe configuration data-related defects to be the dominant category across all four datasets.

In Figure 7, the x-axis presents years, and the y-axis presents defect/year values for each year. From Figure 7, over time we observe defects to be reported across all datasets as indicated by ‘TOTAL’. For configuration data-related defects, the defect/year value does not reduce to zero, indicating configuration data defects to be reported throughout the entire lifetime period for all four datasets.

![Figure 7: Evolution of defect proportion for eight categories; GitHub: 7a, Mozilla: 7b, Openstack: 7c, and Wikimedia: 7d. For each dataset, ‘Total’ presents the proportion of commits, which includes at least one category of defect.](Pre-Print)
Table 9: Defect and Script Proportion for Defect Categories

<table>
<thead>
<tr>
<th>Categ.</th>
<th>GHB</th>
<th>MOZ</th>
<th>OST</th>
<th>WIK</th>
<th>GHB</th>
<th>MOZ</th>
<th>OST</th>
<th>WIK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond.</td>
<td>0.3</td>
<td>0.04</td>
<td>0.3</td>
<td>0.03</td>
<td>1.9</td>
<td>0.3</td>
<td>1.8</td>
<td>1.4</td>
</tr>
<tr>
<td>Conf.Data</td>
<td>9.5</td>
<td>3.8</td>
<td>11.5</td>
<td>7.2</td>
<td>29.2</td>
<td>23.5</td>
<td>33.9</td>
<td>29.6</td>
</tr>
<tr>
<td>Depe.</td>
<td>1.8</td>
<td>1.3</td>
<td>2.4</td>
<td>1.7</td>
<td>10.6</td>
<td>12.4</td>
<td>16.3</td>
<td>17.1</td>
</tr>
<tr>
<td>Docu.</td>
<td>1.7</td>
<td>0.8</td>
<td>1.6</td>
<td>1.5</td>
<td>13.7</td>
<td>7.2</td>
<td>14.0</td>
<td>13.6</td>
</tr>
<tr>
<td>Idem.</td>
<td>0.1</td>
<td>0.01</td>
<td>0.3</td>
<td>0.03</td>
<td>1.0</td>
<td>1.6</td>
<td>3.5</td>
<td>0.1</td>
</tr>
<tr>
<td>Secu.</td>
<td>0.1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1</td>
<td>0.9</td>
<td>5.5</td>
<td>2.8</td>
<td>1.1</td>
</tr>
<tr>
<td>Serv.</td>
<td>1.4</td>
<td>2.6</td>
<td>1.8</td>
<td>3.0</td>
<td>4.9</td>
<td>23.1</td>
<td>9.3</td>
<td>12.4</td>
</tr>
<tr>
<td>Synt.</td>
<td>2.8</td>
<td>1.7</td>
<td>2.3</td>
<td>2.4</td>
<td>16.3</td>
<td>9.2</td>
<td>19.2</td>
<td>17.4</td>
</tr>
<tr>
<td>Total</td>
<td>16.6</td>
<td>10.9</td>
<td>18.7</td>
<td>15.2</td>
<td>43.6</td>
<td>47.4</td>
<td>48.6</td>
<td>49.2</td>
</tr>
</tbody>
</table>

6 DISCUSSION

We discuss our findings in this section.

Practitioner Perception and Observed Evidence: According to Srikant and Menzies [7], "documenting developer beliefs should be the start, not the end, of software engineering research. Once prevalent beliefs are found, they should be checked against real-world data", suggesting researchers to complement survey data with software repository data. We too have complemented survey data with analysis of OSS data. We have reported varying practitioner perceptions for the identified categories in Section 3.3.2. We notice congruence for two categories: configuration data and dependency are the second most agreed upon category and also frequently occurs in OSS datasets. On the other hand, service defects are least agreed upon, but they are more frequent than idempotency – the defect category surveyed practitioners most agreed upon.

In the Reddit post [17] mentioned in Section 1, practitioners reported only one defect category–syntax. Along with syntax, our taxonomy includes seven other defect categories. Suggestions from online forums could be inconclusive, and practitioners can find our taxonomy helpful.

Mitigation: Companies can mitigate the occurrence of defects by incorporating tools that target one or more of the identified categories during IaC development. For example, companies can adopt `spec-puppet’ [61] to reduce conditional and service defects. Tools such as Tortoise [77], ‘librarian-puppet’ [38], and ‘puppet-strings’ [29] might be helpful respectively, in mitigating configuration data, dependency, and documentation defects. Static analysis tools such as SLIC [56] and ‘puppet-lint’ [54] might respectively, be helpful in mitigating security and syntax defects. Mitigation of idempotency defects might be possible through early detection of idempotency with Hummer et al. [27]’s approach that uses model-based testing [73] to detect idempotency.

Implications: Our paper can be helpful for practitioners in the following manner: (i) better understanding of defects: use definitions and examples in Section 3.2 to understand the consequences of IaC defect categories, and activities needed to mitigate each defect category; (ii) prioritizing V&V efforts: if frequency is considered as a prioritization metric, then practitioners can prioritize V&V efforts by fixing configuration data defects that occur in 23.5%−33.9% of IaC scripts; (iii) triaging defects: practitioners can find what defect categories are being reported for a long period, helping them make informed decisions on defect triaging; (iv) measuring IaC script quality: use our reported frequency in Table 9 as a reference; and (v) saving manual effort: Zou et al. [82] identified three reasons why practitioners find automated defect categorization important: better resource allocation, saving manual work in classifying defects, and facilitating postmortem analysis. For automated categorization of IaC defects ACID can be helpful for practitioners.

Future Work: Researchers can investigate if above-mentioned recommendations can actually reduce defects in IaC scripts. The coding patterns that ACID use, could be further leveraged in investigating if defect categories for IaC, such as configuration data, can be detected at compile time.

7 THREATS TO VALIDITY

We briefly describe the limitations of our paper in this section.

Conclusion Validity: The derived defect categories and the oracle dataset are susceptible to rater bias. In both cases, we mitigate rater bias by allocating multiple raters. Also, we use the content of commit messages to determine the defect categories for IaC scripts, which is limiting as commit messages do not always express a defect-related action. Practitioners may use other keywords that we did not include. We mitigate this limitation by using a set of defect-related keywords, derived from prior research, and shown to be adequate in detecting defect-related commits [60] [59] [41]. Also, ACID uses dependency parsing relies on annotated datasets mined from news articles [32], which can be limiting to capture dependencies.

External Validity: We have not analyzed proprietary repositories, and our findings are limited to OSS datasets with Puppet scripts. We mitigate this limitation by mining OSS repositories from GitHub and three organizations.

Internal Validity: The defect category list is not comprehensive, as the derivation process is dependent on the collected commits and rater judgment. We mitigate this limitation with 1,448 ECMs and two raters to derive defect categories. We acknowledge the limitations of the rules presented in Table 2, as the construction is dependent upon the ECMs and diffs of the collected commits, along with the first author’s judgment. Practitioners can use certain string patterns to describe a category that we did not list. We mitigate this limitation by inspecting 1,448 ECMs and diffs to derive the used string patterns.

8 CONCLUSION

Defects in IaC scripts can have serious consequences. Defect categorization can help practitioners to make informed decisions on how to mitigate IaC defects. We applied qualitative analysis on 1,448 defect-related commits to identify defect categories for IaC scripts. We surveyed 66 practitioners to assess if practitioners agree with the identified defect categories. Next, we constructed a tool called ACID and apply ACID on 80,415 commits from 291 repositories to automatically identify defect categories in IaC scripts.

Our derived taxonomy consists of eight defect categories that included idempotency, a category specific to IaC. Amongst 66 survey respondents, 79% of the practitioners agreed with idempotency-related defects, and 49%−79% of the practitioners agreed with one of the identified defect categories. We observe configuration data-related defects to be the most frequent defect category, whereas idempotency is the least frequently occurring defect category. Using our reported defect category frequency results, practitioners
can prioritize V&V efforts by fixing configuration data defects that occur in 23.5%–33.9% of IaC scripts. Our research can be helpful for practitioners to improve IaC script quality, as they can use ACID to identify defect categories automatically, and use our definitions and examples of each identified defect category to assess the importance of the identified defect categories. Our taxonomy of IaC defects can help practitioners in understanding the nature of defects and guide them in triaging defects, prioritizing V&V efforts, and measuring IaC script quality. We hope our findings will facilitate further research in the domain IaC script quality.

REFERENCES

[12] Marcelo Cinque, Dominico Cotroneo, Raffaele D. Corte, and Antonio Pecchia. 2017. Asserting IaC script quality. We hope our findings will facilitate further research in the domain IaC script quality.

**REFERENCES**
