

ConfInLog: Leveraging Software Logs to Infer Configuration Constraints

Shulin Zhou, Xiaodong Liu, Shanshan Li, Zhouyang Jia,
Yuanliang Zhang, Teng Wang, Wang Li, Xiangke Liao
National University of Defense Technology, Changsha, China
{zhoushulin, liuxiaodong, shanshanli, jiazhouyang}@nudt.edu.cn
{zhangyuanliang13, wangteng13, liwang2015, xkliao}@nudt.edu.cn

Abstract—Misconfigurations have become the **dominant** causes of software failures in recent years, drawing tremendous attention for their increasing prevalence and severity. Configuration constraints can **preemptively** avoid misconfiguration by defining the conditions that configuration options should satisfy. Documentation is the main source of configuration constraints, but it might be incomplete or inconsistent with the source code. In this regard, prior researches have focused on obtaining configuration constraints from software source code through static analysis. However, the difficulty in pointer analysis and context comprehension prevents them from collecting accurate and comprehensive constraints. In this paper, we observed that software logs often contain configuration constraints. We conducted an empirical study and summarized patterns of configuration-related log messages. Guided by the study, we designed and implemented *ConfInLog*, a static tool to infer configuration constraints from log messages. *ConfInLog* first selects configuration-related log messages from source code by using the summarized patterns, then infers constraints from log messages based on the summarized natural language patterns. To evaluate the effectiveness of *ConfInLog*, we applied our tool on seven popular open-source software systems. *ConfInLog* successfully inferred 22~163 constraints, in which 59.5%~61.6% could not be inferred by the state-of-the-art work. Finally, we submitted 67 documentation patches regarding the constraints inferred by *ConfInLog*. The constraints in 29 patches have been confirmed by the developers, among which 10 patches have been accepted.

Index Terms—Misconfiguration, Configuration Constraints, Log

反馈用户手册不完善

I. INTRODUCTION

In recent years, misconfigurations have become one of the dominant causes of software failures, drawing tremendous attention for their increasing prevalence and severity [1]–[11]. One major reason for the prevalent misconfiguration issues is the ever-increasing complexity of configuration [12]–[16]. It is challenging for novices or even proficient to correctly understand the constraints of configuration options without expert domain knowledge [12], [17]. To preemptively avoid misconfiguration, documentation is the main source for users to comprehend configuration constraints. However, as mentioned by [13], [15], [18], documentation is often incomplete or outdated, making it difficult for users to learn configuration constraints. For instance, the value range (0~32767) of configuration option "LimitRequestFields" in Apache Httpd [19] is a legacy mistake in documentation [20].

手册错误案例

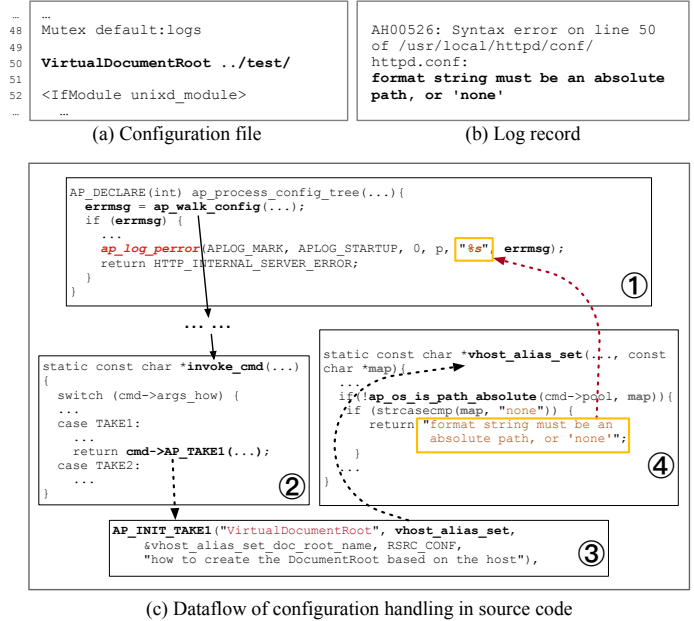


Fig. 1. A motivated example of configuration constraint in log.

In order to help users understand the configuration constraints, there are prior methods [13], [18], [21] focusing on inferring configuration constraints from software source code. The existing methods typically have two main steps. First, they map configuration options to program variables, and then track usages of the variables by using static program analysis. Second, they infer constraints when the variables occur in some predefined code patterns. For example, a typical pattern is to extract constraints from conditional expressions of branch statements, which contain the usages of the variables.

This process, however, may be limited in practice. On one hand, the pointer mechanism in programming languages like C/C++ may bring great barriers in static program analysis, especially for large-scale software systems, limiting the analysis scope and accuracy [22]. On the other hand, the quality of these predefined patterns is highly depended on personal experience. As a result, they may miss the constraints that are hard to be tracked by static analysis, or miss the constraints that do not belong to any pattern.

In this paper, we observe that the *log records* (referring to the text outputs at runtime) of software systems usually contain

程序分析的难点

constraints of configuration options, as exemplified in Fig.1. In Fig.1(a), the configuration option "VirtualDocumentRoot" is set to "../test/" in the configuration file of Apache Httpd; Fig.1(b) presents the log record after running the **official test suite**, indicating that "VirtualDocumentRoot" should be an absolute path or "none". This constraint is clearly defined in the Return statement in code snippet ④ of Fig.1(c). This Return statement is controlled by two branch statements. However, it is difficult to infer constraint from the conditions of the branch statements by using the traditional static analysis method, since it is hard to understand the semantics of function `ap_os_is_path_absolute()`, and the comparison between `map` and "none" does not match any predefined code patterns.

To figure out the prevalence of configuration constraints in logs, we carried out an empirical study on four popular open-source software systems. The result shows that 57.0% (85/149) of configuration options have constraint descriptions in log records. By tracing log records back to source code, we found that *log messages* (referring to the string constants in source code that will be printed by log statements) always indicate the key information of constraints in log records. Yet it is non-trivial to identify whether a log message is related to a configuration option, since **log messages of 77.6% (66/85) configuration options do not contain the option names**. With a further analysis, we summarized patterns which could be used to effectively identify configuration-related log messages.

Guided by our findings, we proposed *ConfInLog*, a practicable tool to **statically** infer configuration constraints from log messages in source code. *ConfInLog* first selects configuration-related log messages from source code through matching the patterns of how log messages relate to configuration options. Then, *ConfInLog* infers the constraints from configuration-related log messages based on natural language patterns summarized from existing constraints descriptions. An experiment on seven popular open-source software systems proved the effectiveness of *ConfInLog*. *ConfInLog* could infer 22~163 constraints from the software systems, and 59.5%~61.6% of them could not be inferred by the state-of-the-art work. Also, after manually checking the documentation of those software systems, we submitted 67 documentation patches regarding the constraints inferred by *ConfInLog*. The constraints in 29 patches have been confirmed by the developers, among which 10 patches have been accepted. The others are still pending for responses at the time of writing.

In summary, this paper makes the following contributions:

- We observed that logs contain configuration constraints based on an empirical study on four popular open-source software systems, and summarized patterns to select the configuration-related log messages.
- We designed and implemented *ConfInLog*, an automated tool to infer configuration constraints from log messages in source code, which could be an effective complement to previous work with static program analysis.
- We evaluated *ConfInLog* on seven open-source software systems. *ConfInLog* could infer 22~163 configuration constraints, in which 59.5%~61.6% could not be inferred

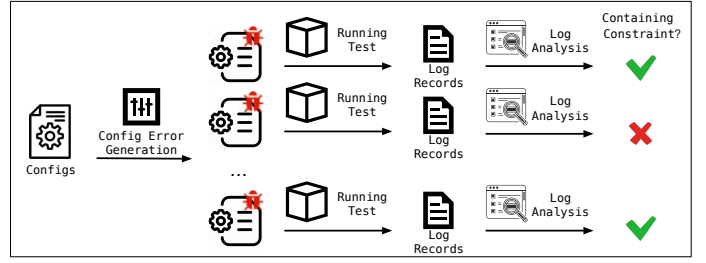


Fig. 2. Workflow of Configuration Error Injection.

by the state-of-the-art work. We submitted 67 patches based on our inferred constraints. The constraints in 29 patches have been confirmed by the developers, among which 10 patches have been accepted.

The remainder of this paper is organized as follows. In Section II, we present the empirical study. In Section III, we first provide an overview of *ConfInLog*, and then describe the details of each step. In Section IV, we elaborate the experimental setup and report the experimental results. In Section V, we discuss **the threats to validity**. In Section VI, we briefly review the related work. Finally, we conclude our study and future work in VII.

II. EMPIRICAL STUDY

In this section, we take an in-depth look into the configuration constraints in logs through an empirical study. We **manually** analyzed the prevalence and characteristics of constraint-related logs. In our observation, we found that configuration constraints are widely existed in the log records of software systems, and there are characteristics to select log messages related to configuration options.

A. Prevalence of Configuration Constraints in Logs

To acquire log records from software systems, we chose to **inject configuration errors**, as the workflow shown in Fig.2. There are mainly two benefits. It is not necessary to comprehend the details of code logics, and it is more efficient to select the log records related to configuration options.

i) Data Set: As shown in Table I, we choose four well-known software systems for our study. All of these software systems are open-source server software and widely deployed in the field, and provide official test suite to run different workloads. Considering the large number of configuration options provided by these software systems, we choose the core modules or functions of each system to conduct configuration error injection. For Httpd and Nginx, we focus on the configuration options in Core modules; for MySQL, we focus on the configuration options related to MyISAM storage module; for PostgreSQL, we test the configuration options related to file locations, connections, authentication, and resource consumption.

ii) Configuration Error Injection: Prior studies [18], [23]–[26] on configuration error generation, mainly fall into two categories, **namely mutation based [24]–[26] and constraint (specification) violation based [18], [23]**. Considering that the constraint violation based methods require the configuration

两种错误生成方式：基于变异、基于约束，这里采用基于变异的

TABLE I
PROPORTION OF CONFIGURATION OPTIONS WITH DIFFERENT REACTION IN LOG RECORDS.

Software	# of Config Options Under Test	# of Config Options with Related Logs		
		Total	with Constraints	without Constraints
Httpd	46	43	34	9
Nginx	28	25	9	16
MySQL	23	23	11	12
PostgreSQL	58	58	31	27
Total	155	149	85	64

constraints beforehand, we choose to implement the mutation rules in the state-of-the-art mutation-based method [24] to generate configuration errors in our study. Then we inject one configuration error each time, and run the official test suite to collect relative log records.

iii) **Log Analysis:** After running the test suites, there are plenty of log records generated by the software system under test. To figure out whether there is configuration constraint in log record, we apply a semi-automatically approach to analyze the log records. First, for a configuration option under test, we preprocess the configuration option for stemming and tokenization to a word set ψ . If there is any match of a) word in ψ , b) the configuration file name, or c) keywords "config", "parameter", "directive", with the log records, we manually inspect the log records to check whether there is any constraint description of the configuration option under test. We regard a log record as related to the option under test if it satisfies one of the following conditions: a) the log record contains the name of the option under test; b) it contains all the words in the option connected with common delimiters; c) it contains the location information (line number) to find the option in configuration file. Moreover, if the log record is related to the option under test, and provides guidance for fixing the configuration error, we regard it as containing configuration constraints.

Finding 1: Up to 57.0% of the configuration options have log records containing constraints description, which provides opportunity to infer configuration constraints.

As shown in Table I, we tested a total of 155 configuration options, among which 96.1% (149/155) have related log records. More importantly, 57.0% (85/149) have explicit constraints description in the log records. For those options without constraints in the log records (64/149), 7.8% (5/64) have descriptions about the purpose of configuration options which, however, are not enough to guide the fixing. For example, the log records of configuration option "UseCanonicalName" indicate that it defines "how to work out the ServerName : Port when constructing URLs", but there is no description about what value should be set. 92.2% (59/64) have log records describing the error status without further guidance of error fixing. For example, configuration option "worker_aio_requests" in Nginx only outputted "invalid value" in the related log records.

if (*errp conf->limit_req_body < 0) { return "LimitRequestBody requires a non-negative integer."; }	Log message in Return stmt
if (len > sizeof(saun->sun_path)) { u->err = "too long path in the unix domain socket"; return NGX_ERROR; }	Log message in Assign stmt
AP_INIT_TABLE2("LimitInternalRecursion", set_recursion_limit, NULL, RSRC_CONF, "maximum recursion depth of internal redirects and subrequests"),	Log message in complex structure

Fig. 3. Different existing styles of log messages in source code.

B. Analysis of Constraints-Related Logs

In light of the wide existence of configuration constraints in log records, we hold the hypothesis that we could infer configuration constraints from logs. Nevertheless, it is challenging to trigger all the log records, considering the various workloads and the limited capability of current methods for generating configuration errors (we will discuss the circumstances in Section IV-D2). Consequently, alternative static analysis strategies are taken into consideration. We took a further look at the logs in source code.

1) **The Existing Styles of Log Messages:** To infer configuration constraints from logs in the source code, a straightforward method is to locate the *log statements* (referring to the program statements in the source code that perform the log behaviors), and check whether there are possible configuration constraints. However, this approach is not feasible in practice. First, as mentioned by [27], the huge number of log functions in large-scale software and the collaborative development make it a non-trivial task to recognize all the log statements in source code. Second, even if we could locate all the log statements, it is still difficult to obtain the corresponding log message. Take Fig.1(c) as example, the log message is not explicitly presented in the log statement `ap_log_perror()` in code snippet ①, but determined by the return value `errmsg` of function `ap_walk_config()`, which is indeterministic due to the invocations of different function pointers in function `invoke_cmd()` in code snippet ③, and the outputted log message is existed in the Return Statement in code snippet ④.

Finding 2: 12.5% of the log messages are not explicitly presented in the log statements, but exist in various styles with implicit dataflow to log statements..

1) 确定
日志是否
和配置相
关
2) 确定
是否含有
约束

如何触
发全部
的日志
记录

TABLE II
PROPORTION OF DIFFERENT EXISTING STYLES OF LOG MESSAGES IN SOURCE CODE.

Software	# of Sampled Log Stmt	# of Log Msg in Log Stmt	# of Log Msg in Return Stmt	# of Log Msg in Assign Stmt	# of Log Msg in Structure
Httpd	100	78	16	3	3
Nginx	100	82	8	10	0
MySQL	100	93	3	4	0
PostgreSQL	100	97	0	3	0
Total	400	350	27	20	3

In order to explore the characteristics of log messages in the source code, we randomly sampled 100 log statements in source code for each studied software system, and manually inspected the context and dataflow to find the corresponding log messages. As shown in Table II, except for the 87.5% (350/400) of log statements that contain log messages directly inside the statement, the remaining 12.5% (50/400) are implicitly existed in three different styles, as the code examples shown in Fig.3. Besides, due to the usage of **function pointers**, a log statement may be reused to output different log messages by calling different function pointers, as the motivating example in Fig.1 shown. These circumstances all prevent us from collecting all the log messages directly from log statements.

2) Identification of Configuration-related Log Messages:

Finding 3: Up to 77.6% of log messages do not directly specify the the configuration option involved, which limits the identification of configuration-related log messages.

To infer configuration constraints from log messages, we need to identify whether the log message is related to specific configuration option at first. Nevertheless, a simple keyword matching for options' name is far from enough. Among all the 85 configuration options with constraints in log records found in Section II-A, up to 77.6% of the configuration-related log messages cannot be identified by directly matching the options' name.

Finding 4: 51.8% of configuration-related log messages could be identified by directly matching option names in log messages, or finding relevant functions or variables, which implies the feasibility of identifying configuration-related log messages.

With a further manual analysis in source code, we found that, except for the 22.4%(19/85) of log messages that directly contain name information inside, 29.4%(25/85) of log messages fall into two code patterns in Fig.4, which are helpful in identifying the configuration-related log messages without name information.

In the first pattern, the log message is influenced by the program variable related to specific configuration option. Take Fig.4(a) as example, the log message is under control of program variable `clcf->disable_symlinks` in the branch condition, which stores the value of the configuration option

```
1 if (clcf-> disable_symlinks == NGX_CONF_UNSET_UINT) {
2     ngx_conf_log_error(NGX_LOG_EMERG, cf, 0,
3         "\"%V\" must have \"off\", \"on\" "
4         "or \"if_not_owner\" parameter",
5         &cmd->name);
6     return NGX_CONF_ERROR;
7 }
```

(a) The log message is influenced by variable related to configuration option.

```
1 AP_INIT_TAKE1(" AllowEncodedSlashes ", set_allow2f, NULL,
2     RSRC_CONF,
3     "Allow URLs containing '/' encoded as '%2F'"),
4 static const char * set_allow2f (cmd_parms *cmd, void *d_,
5     const char *arg)
6 {
7     if (0 == strcasecmp(arg, "on")) {
8         d->allow_encoded_slashes = 1;
9         d->decode_encoded_slashes = 1;
10    }
11    ...
12    else {
13        return apr_pstrcat(cmd->pool,
14            cmd->cmd->name, " must be On, Off, or
15            NoDecode",
16            NULL);
17    }
18 }
```

(b) The log message is influenced by function related to configuration option.

Fig. 4. Different patterns of log messages related to configuration options.

"`disable_symlinks`". In the second pattern, the log message is influenced by the function related to specific configuration option. Fig4(b) shows an example, the log message is controlled by the argument `arg` of function `set_allow2f()` in If/Else branch statement, while this function is related to configuration option "`AllowEncodedSlashes`" with definition in the structure of `AP_INIT_TAKE1` macro.

For the remaining 48.2% (41/85) of configuration options that could not identify the related log messages based on finding 4, all of them are numeric or boolean types, and handled by general logic with predefined information in specific structures. For instance, the configuration options with integer value in PostgreSQL are checked by the `min/max` bound defined in struct `config_int`. These constraints could be easily learned in prior work [13], without the need to infer from log messages.

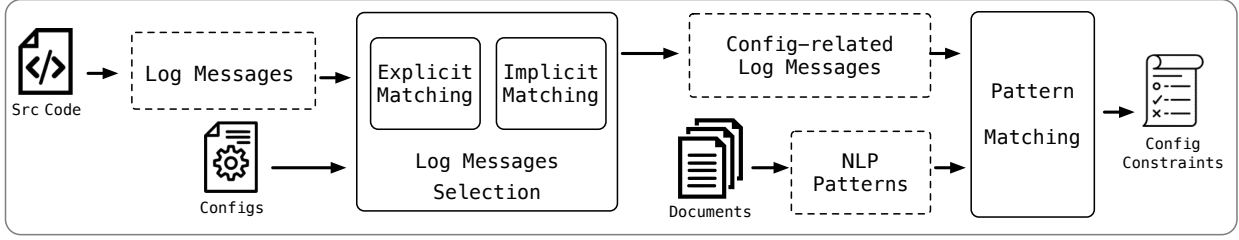


Fig. 5. Workflow of *ConfInLog*.

III. DESIGN AND IMPLEMENTATION

In this section, we present the design and implementation of *ConfInLog*. Fig.5 illustrates the workflow. *ConfInLog* takes source code and configuration options as input, and outputs a list of configuration constraints corresponding to different options. *ConfInLog* consists of two major steps: Log messages selection and constraints inferring. First, *ConfInLog* collects the log messages in target software and selects the configuration-related ones. Then, based on summarized NLP patterns, *ConfInLog* infers the configuration constraints from selected log messages.

A. Log Messages Selection

As mentioned in Section II-B1, not all the log messages are explicitly present in the log statements. In order to get as much log messages as possible, *ConfInLog* considers all the string constants in the code as candidate log messages. The false positive messages will be filtered out in later steps. Since a log message may be spliced by several string constants, variables and macros, *ConfInLog* first converts the variable to a special label "VARIABLE" and replaces the constant and macros, then joins them together. As shown in the following code snippet, the log message that *ConfInLog* collects is "*LimitRequestFields VARIABLE must be a non-negative integer (0 = no limit)*".

```

1 apr_pstrcat(cmd->temp_pool, "LimitRequestFields \"", arg,
2 "\" must be a non-negative integer (0 = no limit)", NULL
);

```

After collecting candidate log messages, *ConfInLog* takes a two-step approach to select configuration-related log messages.

1) **Finding Configuration Options in Log Messages:** The direct method to identify configuration-related log messages is to check whether the log messages contain option name information inside. First, *ConfInLog* splits each configuration option into words, and generates a regular expression with possible delimiters between the words. Take "*data_directory*" as an example, *ConfInLog* first splits it into {"*data*","*directory*"} and generates a regex expression as "*data[.-]directory*". Then, *ConfInLog* tries to match the regex expression with the log message. If there is any match, *ConfInLog* regards the log message as related to this configuration option.

However, If the name of a configuration option is too simple, such as configuration option "*use*" in Nginx, there will be too many false positives by directly matching. For a configuration option with a single word name *w*, *ConfInLog*

identifies whether *w* is a option name when matched in a log message by the following criteria:

- Is *w* quoted in the log message? (For example, the configuration option "*alias*" in log message "*'alias' cannot be used in location where URI was rewritten*")
- Is there any specific modifier for *w* in the message that could indicate the identity. (namely "configuration", "config", "option", "directive", "parameter")

2) **Finding Configuration Options implicitly related to Log Messages:** As mentioned in Section II-B2, many log messages do not directly contain the information of the corresponding configuration option, but are influenced by variables and functions related to the option. Consequently, if a log message doesn't match any option name explicitly, *ConfInLog* tries to find the options that are implicitly related to this log message based on the summarized code patterns in Section II-B2.

i) Collecting Related Code Elements of Configuration Options

To identify whether a message is related to configuration options, the first thing is to figure out the code elements related to configuration options in source code. Considering the summarized patterns of configuration-related log messages in Fig.4, *ConfInLog* mainly focuses on the functions and variables related to configuration options. For ease of description, we refer to the variables related to configuration options as configuration variables, and functions related to configuration options as configuration functions.

Prior researchers [18] observed that developers often use clean interface to manage the configuration-to-variable mapping information, namely structure, container, and comparison interface. *ConfInLog* inherits and improves the method in [28] to automatically collect program variables and functions related to configuration options.

In detail, if the name string of configuration options are used in the structure interface defined for the configuration-variable mapping, *ConfInLog* uses a minimal-common-ancestor algorithm to collect the configuration variables and configuration functions. The main idea is to find the minimal common subtree of the configuration option and its related code elements in AST (Abstract Syntax Tree), without interfering with other configuration options. Take Fig.6 as example, *ConfInLog* finds that Node-① is the common ancestor of configuration option "*default_tablespace*" and its related code elements, then collects the related variables and functions as follows:

```

Variable: "default_tablespace"
Function: "check_default_tablespace".

```

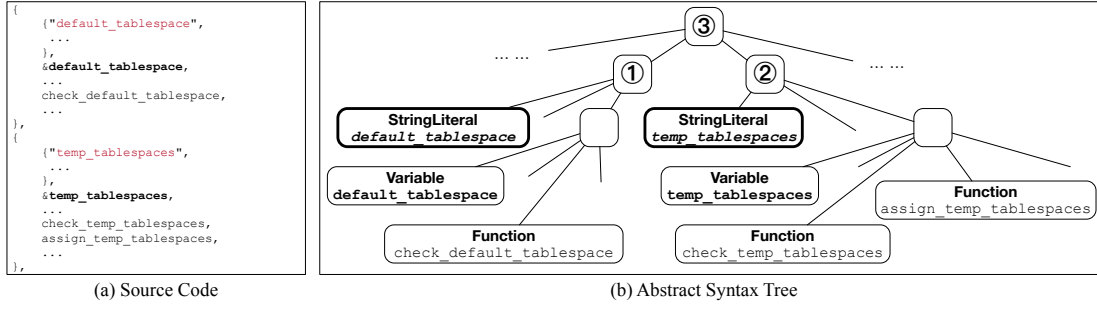


Fig. 6. Example of collecting variables and functions based on minimal common ancestor algorithm.

If the configuration options are mapped to variables with container interface (reading configuration value based on a getter api), *ConfInLog* collects the return variable as configuration variable. Moreover, for comparison interface, if an configuration option is used in strcmp-like conditions, and the succeeding block is less than 3 lines, *ConfInLog* collects the left variable of assign statement in the succeeding block as configuration variable. That makes sense in most cases, since longer handling lines in succeeding block are more likely to contain other irrelevant objects.

There are situations that *ConfInLog* could not collect any configuration variable during this procedure, for example, Nginx only declares the configuration functions in the structure interface, and assigns configuration variable with offset of structure pointers. *ConfInLog* handles these situations in the iii) step.

ii) *Backward Slicing on Log Messages*

Once the configuration variables and functions are collected, *ConfInLog* applies a backward slicing on log messages to identify whether there is any configuration option related to this log message.

ConfInLog begins the slicing from the statement that contains a log message. Once a variable or function is got in the slice, *ConfInLog* checks whether it is a configuration variable or function, and regards the log message as related to this configuration option if it is. Take the code snippet in Fig.1 as example, *ConfInLog* first locates the log message in the Return statement, then orderly checks whether variable `map`, function `ap_os_is_path_absolute()` and `cmd->pool` are configuration variable or function. Finally, when checking the function `vhost_alias_set()` in the slice, *ConfInLog* finds that it is the configuration function of option "VirtualDocumentRoot". So *ConfInLog* regards the log message "format string must be an absolute path, or 'none'" as related to "VirtualDocumentRoot".

However, If the log message is in the structure outside the function declaration as shown below:

```

1 AP_INIT_TAKE1("AddDefaultCharset",
2   set_add_default_charset, NULL, OR_FILEINFO,
3   "The name of the default charset to add to any
4   Content-Type without one or 'Off' to disable"),
5 AP_INIT_TAKE1("AcceptPathInfo",
6   set_accept_path_info, NULL, OR_FILEINFO,
7   "Set to on or off for PATH_INFO to be accepted by
8   handlers, or default for the per-handler preference"),

```

ConfInLog traverses the parent subtree of this log message in AST until it meets the name string of any configuration option or reaches the top of current structure's subtree. In this example, *ConfInLog* first locates the StringLiteral AST node containing log message "The name of the default charset to ..." in line 3~4, then traverses the parent subtree of current node, until reaches the subtree of structure `AP_INIT_TAKE1`, i.e. line 1~4. During this procedure, *ConfInLog* finds the option "AddDefaultCharset", so *ConfInLog* regards the log messages as related to "AddDefaultCharset".

iii) *Complementary Similarity Comparison*

As mentioned in step i), there are situations that no configuration variable is collected by *ConfInLog*. If there is either no matched configuration function in step ii), *ConfInLog* applies an approximate approach to identify configuration-related log messages based on the similarity between configuration options and the variables in the slice.

Developers usually tend to name the configuration variable similar to the option [29]–[34]. Thus, *ConfInLog* applies the Jaccard Index [35] to calculate the similarity between variable and configuration option and infer whether they are relevant. If a variable in the slice of a log message has a high similarity with one configuration option, *ConfInLog* regards this log message as related to the option. In detail, after doing segmentation and lemmatization for configuration option and variable, *ConfInLog* applies the *idf* [36] (inverse document frequency) to calculate the weight of every word, by taking each configuration option as a document, and all the configuration options as the corpus. We selected the *idf* measure since it has been theoretically proved its effectiveness in weighting [36] and practically applied in prior work [24]. Then *ConfInLog* calculates the similarity as follows:

$$\begin{aligned}
Sim(Conf_Name, Var_Name) = & \\
& \frac{\sum idf(word_i) \text{ foreach } word_i \in Cwords \cap Vwords}{\sum idf(word_j) \text{ foreach } word_j \in Cwords \cup Vwords}
\end{aligned}$$

The `Cwords` and `Vwords` in the equation refer to the word set of configuration option and variable after segmentation and lemmatization correspondingly. To evaluate whether a variable is related to the configuration option, a threshold μ for the similarity is needed. We will evaluate the proper value in Section IV-C.

TABLE III
POS PATTERNS OF CONSTRAINTS' DESCRIPTION.

Pattern	Percentage.	Example
NN MD(must/should/...)† [not]‡ VB(be/have/...)	56.3%	This value must be greater than 0.
NN VB(accept/require/...) NN	19.3%	operationProfiling.slowOpSampleRate accepts values between 0 and 1.
JJ(acceptable/valid/...) NN VB(be)	14.7%	Valid options are 'ALWAYS', 'NEVER', and 'ONCE'.
NN VB(be) [not] JJ(acceptable/valid/...)	7.0%	Negative value or 0 are invalid values and will fail NN startup.
NN(ERROR_STATUS), NN(CONFIG) VB(be/have/...) JJ(less/greater/...)	2.8%	Error: olcIdleTimeout is less than 0.

† () means one of them. ‡ [] means optional

B. Constraints Inferring

After the selection step, *ConfInLog* obtains the configuration-related log messages. Then, *ConfInLog* mainly infers the constraints description from log messages. In general, the log messages written in natural language are often unstructured and with various purposes, such as monitoring task progress, pinpointing error status, and guiding users' fixing. To infer the constraints in log messages, we applied a NLP (Natural Language Processing) pattern-based method to represent the description of configuration constraints. We studied and summarized the patterns of constraint description from 11 open-source software systems¹. They are different from those we studied in Section II with different functionalities.

Two authors conducted the job to collect constraints description from documents and log messages. For documents, they manually read the documents and located the description about the usage of each configuration option, then collected the text part that specify the constraints. For log messages in source code, they located the log messages containing configuration options' name at first, then manually comprehended the context and meaning of the log message to judge whether it is a description of configuration constraint. After the collection, the two authors cross checked the constraints' description. If there was a disagreement, a third author was consulted for additional discussion until consensus was reached. Then, we combined their results and replaced the configuration names with keyword "CONFIG" to avoid the interference of option names. Finally, we collected 338 constraints' description in total.

In order to summarize the patterns from those descriptions, we used POS (Part-Of-Speech) tag sequences to represent the descriptions and mine the common sequences as our pattern. Firstly we collected the words related to error from WordNet [37] to build a list λ . For each description of configuration constraints, we generated the POS tag sequences with spaCy [38]. For example, the constraints' description in the motivating example of Fig.1 is transformed as "NN(format) NN(string) MD(must) VB(be) DT(an) JJ(absolute) NN(path) , CC(or) " NN(none) " ". We removed the DT, SYM tags in tag sequences, and merged the successive same tags for NN, JJ as one. If there was any word of a NN matched in λ , we kept the tag NN and replaced the word with

"ERROR_STATUS". For instance, the above tag sequence is transformed as "NN(format string) MD(must) VB(be) JJ(absolute) NN(path) CC(or) NN(none)". Finally, we mined the common sequence among all the tag sequences, and found that 96.7% of the descriptions fall into the patterns illustrated in Table III.

When applying the patterns in recognizing log messages that describe configuration constraints, *ConfInLog* first replaces the configuration name with "CONFIG" if it exists in a log message, and replaces the word matched in λ with "ERROR_STATUS". Then *ConfInLog* generates the corresponding POS tag sequence of the log message. Finally, if there is any match of the patterns illustrated in Table III, *ConfInLog* regards it as a configuration constraints.

IV. EVALUATION

To evaluate *ConfInLog*, we try to answer three Research Questions as follows.

RQ1: How effective is *ConfInLog* on inferring configuration constraints?

RQ2: How accurate is *ConfInLog* on selecting configuration-related log messages?

RQ3: How does *ConfInLog* compared with the state-of-the-art tool for inferring configuration constraints?

For RQ1, we evaluate the effectiveness of *ConfInLog* in inferring configuration constraints. For RQ2, we discuss how to set the parameter μ in *ConfInLog*, and the final precision of *ConfInLog* to select configuration-related log messages. For RQ3, we compared *ConfInLog* with the state-of-the-art tool SPEX [18], and the dynamic method of configuration error injection [24].

A. Experiment Setup

We conducted experiments on seven software systems listed in Table IV to evaluate *ConfInLog* from different aspects. In case of the bias from our empirical study, we added three software systems to prove the generalization. All these seven software systems are widely-deployed, and representative across a number of system software. We set the threshold μ during the selection of configuration-relation log messages as 0.63, and introduce how to set this threshold in Section IV-C.

B. RQ1: Effectiveness of Inferring Configuration Constraints

To answer RQ1, we manually checked the results of *ConfInLog* in inferring configuration constraints. The statistics is shown in Table IV. Generally, *ConfInLog* could infer 22~164

¹The 11 open-source software systems are Hadoop, HDFS, Yarn, Alluxio, Cassandra, Spark, Hypertable, MongoDB, AOLServer, Subversion, OpenLDAP.

TABLE IV
THE EFFECTIVENESS OF INFERRING CONFIGURATION CONSTRAINTS

Software	# of Config Constraints		
	True Positive	False Positive	False Negative
Httpd	164	41	23
Nginx	80	23	8
MySQL	33	9	10
PostgreSQL	35	6	7
Lighttpd	63	19	12
Squid	30	13	8
Postfix	22	8	12
Total	427	119	80

TABLE V
SUBMITTED PATCHES TO ENHANCE DOCUMENTATION.

Software	# of Submitted Patches [†]	# of Confirmed Patches	# of Accepted Patches
Httpd	25	4	3
Nginx	14	13	0
MySQL	7	7	3
PostgreSQL	2	0	0
Lighttpd	4	4	4
Squid	8	1	0
Postfix	7	0	0
Total	67	29	10²

configuration constraints from log messages, in which the average precision is 72.1% (427/546), and the average recall is 18.7%.

The false positives are mainly introduced by the incorrect selection of configuration-related log messages, and the mismatches of POS patterns in log messages not describing constraints. For instance, the log message "*Symbolic link is not allowed or link target not accessible*" is mis-matched with configuration option "*AcceptPathInfo*" due to the wrong configuration variable `r->path_info`, and this log message is also mistakenly inferred as an constraint of "*AcceptPathInfo*". The main reason for the false negatives is that the descriptions of these constraints are mainly describing the error status of configuration, which could not match our POS patterns, but can be inferred by manual comprehension. For instance, the log message "*auth.backend.ldap.filter is missing a replacement operator '?'*" of option "*auth.backend.ldap.filter*" in Lighttpd indicates that there should be a "?", which could not be matched with any of the POS patterns.

To prove the effectiveness of inferring configuration constraints from source code by *ConfInLog*, we manually checked whether the constraints are recorded in the documentation. As shown in Table V, we finally submitted 67 patches to enhance the documentation about description of configuration constraints. The constraints in 29 patches have been confirmed by the developers, among which 10 patches have been accepted. The other 38 patches are still pending for responses at the time of writing. For the 19 confirmed patches that have not

[†]Accepted bug IDs: Httpd-{64893, 64904, 64909}, MySQL-{101513, 101514, 101520}, Lighttpd-{3035, 3036, 3038, 3040}.

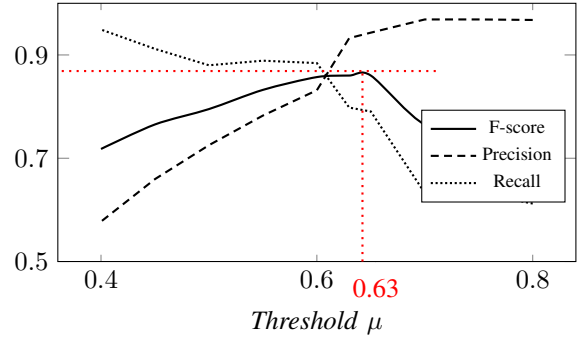


Fig. 7. F1-score of different threshold μ in calculating similarity between configuration options and program variables.

TABLE VI
PRECISION OF *ConfInLog* IN SELECTING CONFIGURATION-RELATED LOG MESSAGES.

Software	# of Config-related Log Messages	Precision (in 100 samples)
Httpd	1545	72.4%
Nginx	773	78.1%
MySQL	1097	73.5%
PostgreSQL	752	74.7%
Lighttpd	396	65.9%
Squid	333	63.1%
Postfix	470	68.5%
Total	5366	70.7%

been accepted, 4 patches are waiting for developers' update in documentation, and the other 15 patches are not accepted mainly due to developers' concern about complicating the documentation. Nevertheless, these constraints could be applied to guide the checking of misconfiguration in production.

C. RQ2: Precision of Selecting Configuration-Related Log Messages

To answer RQ2, we evaluated the precision of *ConfInLog* in selecting Configuration-Related Log Messages. During the procedure of *ConfInLog*, the threshold μ of similarity between configuration option and variable related to log message requires to be predefined. We first tune the threshold μ , then evaluate the precision.

To determine the proper value of threshold μ , we randomly sampled 100 configuration options from Httpd and Nginx, each for 50, and manually inspected the corresponding program variables in source code as oracle. If the option doesn't have a corresponding variable, like the structural option "<Directory>" in Httpd, we sample another option. Then, we manually set the threshold μ to different value, and run *ConfInLog* to find the corresponding variables for the 100 configuration options. Finally, we verified the variables found by *ConfInLog* with the oracle, and calculated the F1-score by $\frac{2 * Precision * Recall}{Precision + Recall}$. Fig.7 shows the F1-score with different value of threshold μ , and when $\mu = 0.63$, F1-score reaches the highest value. Based on the result, we set threshold $\mu = 0.63$ in *ConfInLog*.

TABLE VII
COMPARISON WITH THE STATE-OF-THE-ART.

Software	Constraints that SPEX Could Infer	Constraints Inferred by <i>ConfInLog</i> While SPEX Could NOT Infer	Constraints Inferred by <i>ConfInLog</i> While Dynamic Method Could NOT Infer	Constraints that Could Only be Inferred by <i>ConfInLog</i>
Httpd	78	88 (53.7%)	46 (28.0%)	41 (25.0%)
Nginx	182	42 (51.3%)	73 (91.3%)	39 (48.8%)
MySQL	351	17 (51.5%)	30 (90.9%)	15 (45.5%)
PostgreSQL	263	27 (85.7%)	26 (74.3%)	19 (54.3%)
Lighttpd	95	48 (76.2%)	48 (76.2%)	37 (58.7%)
Squid	189	22 (73.3%)	-†	-
Postfix	137	19 (86.4%)	-†	-
Total	1295	263 (61.6%)	223 (59.5%)	151 (40.3%)

† Don't have official test suite.

We ran *ConfInLog* on seven software systems to select the log messages that related to configuration options. To evaluate the precision, we randomly sampled 100 log messages for each software system to inspect whether the log message is really related to the specific configuration options. The statistics shown in Table VI illustrate the precision of configuration-related log messages selection.

D. RQ3: Comparison with the State-of-the-art

1) *Comparison with SPEX*: We compared *ConfInLog* with SPEX, which is one of the most effective constraint inferring tools. Follow the design of SPEX, we tried our best to annotate the mapping interfaces in source code, and implemented the constraints extraction. The results are shown in the third column of Table VII. SPEX mainly infers configuration constraints based on searching code patterns for different predefined constraints types. Though the total amount of configuration constraints that SPEX could infer is relative large as shown in the second column, *ConfInLog* could still infer up to 61.6% (263/427) of the constraints that SPEX could not.

Among the constraints that could not be inferred by SPEX, 55.1% (145/263) are descriptions about constraints related to the semantic information of complicated context. For example, the constraint "*binlog_checksum should be NONE for Group Replication*" in MySQL, and "*the 'alias' directive cannot be used inside the named location*" in Nginx, all include the summary of context information. 35.7% (94/263) are descriptions about constraints related to the syntax format of configuration values. For example, the configuration option "*server.modules*" in Lighttpd requires "*list of 'mod_XXXXX' strings*". The remaining 9.1% (24/263) are various causes. For example, the value range of configuration option "*MaxRangeOverlaps*" in Httpd is a combination of enumerated and numeric values, requiring "*none*", "*default*", "*unlimited*" or a positive integer", which is not covered by SPEX's predefined constraint types; for configuration option "*old_passwords*" in MySQL, the valid values are "2" and "0" without "1", but SPEX can only infer that it could not be "1" and didn't know what is the valid values.

2) *Comparison with Dynamic Method*: As mentioned in Section II-A, we applied a mutation based configuration error

injection method, which we refer to as dynamic method here, on studied software systems, and obtained a few configuration constraints from log records. We compared *ConfInLog* with dynamic method as follows. Considering that Squid and Postfix don't provide official test suite, we only conduct the dynamic method on other five software systems. We applied the mutation based method to generate configuration errors for the configuration options that *ConfInLog* could infer constraints, then ran the test suite to obtain log records and check whether there is any constraints description about the configuration options. The detailed statistics are listed in the fourth column of Table VII. On average, 59.5% (223/375) of configuration constraints inferred by *ConfInLog* could not be triggered by dynamic method.

We further analyzed the causes, and found the reasons are mainly threefold. First, the configuration values often have syntax requirements. For example, 31 options in Lighttpd require a list of values, expressed as "(..., ...)". When generating a random string for these configuration options, the program output a log record as "*should have been a array of strings like ... = ("...")*", and only when the configuration value is in the required format with a wrong value, the program could indicate the constraints like "*expected list of 'path' strings*" in the log record. Second, the constraints that describe the dependency between multiple configuration options are difficult to trigger when generating configuration error for single option. For instance, the configuration option "*proxy_busy_buffers_size*" in Nginx must be less than the size of all "*proxy_buffers*" minus one buffer, this constraint was obtained by dynamic method in our experiment, while a similar option pair "*fastcgi_busy_buffers_size*" and "*fastcgi_buffers*" didn't trigger the relevant log statements, because randomly changing value of "*fastcgi_busy_buffers_size*" cannot ensure to break the relationship. Last but not least, the various workload and limited test cases prevent the triggering of log records. For instance, in PostgreSQL, only when trying to make an online backup, the program checks the configuration option "*wal_level*" and indicates that the WAL level is not sufficient, which must be set to `replica` or `logical` at server start. As for the test cases, we ran the official test suite of the former five software system in Table VII and measured the coverage

with `gcov`. The statistics show that the code coverage could only reaches 19.2%~65.1%.

3) *Comparison with the Combination*: Finally, we compared the capability of *ConfInLog* with the combined results of SPEX and dynamic method. As shown in the last column of Table VII, 40.3% of the constraints that inferred by *ConfInLog* could neither be inferred by SPEX nor dynamic method. This emphasizes that *ConfInLog* could be an effective complementary to existing methods.

V. THREATS TO VALIDITY

Quality of log messages: *ConfInLog* infers configuration constraints from log messages in source code. The process works under the premise that the log messages contain descriptions of constraints. We believe that logs written by experienced developers in large-scale software systems usually contain description about system status or guidance for the failure diagnosis. In such a setting, *ConfInLog* could infer configuration constraints from log messages.

Generalization of summarized patterns: *ConfInLog* selects the configuration-related log messages and infers constraints from log messages based on summarized patterns. The generalization of those patterns influence the effectiveness of *ConfInLog*. To cover different situations as many as possible, we chose different kinds of software systems to summarize the patterns of configuration-related log messages. As for the nlp patterns of constraints' description, to avoid over-fitting of our studied software systems, we chose 11 software systems different from the evaluated ones to summarize the patterns.

VI. RELATED WORK

A. Misconfiguration Detection

The prevalence and severity of software misconfigurations have promoted the development of many diagnosis and detection technologies [2], [16], [18], [24], [39]–[52]. Misconfiguration diagnosis refers to locate root causes of misconfigurations that caused failures, performance issues, and incorrect behaviors, while detection aims at detecting misconfigurations before deployment. Most of the detection work discovers configuration errors by either checking against (a) rules mined from numbers of configuration files and system behaviors, or (b) constraints inferred from source code and documentation. Our work targets at inferring configuration constraints from the log messages in source code, in order to prevent and detect misconfigurations.

EnCore [44] uses machine-learning method to infer the configuration rules between configuration settings and the executing environment for misconfiguration detection. ConfigV [16] learns specification from configuration files based on the association rule learning algorithm to detect ordering errors, integer correlation errors, type errors, and missing entry errors. CODE [47] exploits the registry's access rules in Windows to detect configuration errors by finding violations of the rules. These approaches rely on the correctness and completeness of the learned samples.

Prior work [13], [15], [18], [21] proposes static analysis approaches to infer configuration constraints from source code. SPEX [18] searches code patterns for different predefined constraints types, such as data range and value relationship. CDep [21] focuses on five types of configuration dependency between multiple configuration parameters. However, these researches are limited by the predefined code patterns and accuracy of program analysis. ConfTest [45] summarizes a fine-grained classification of option types and infers syntactic and semantic constraints of each type based on the classification. PracExtractor [46] employs Natural Language Processing techniques to automatically infer configuration specifications from software manuals. Our work mainly focus constraints in configuration-related log messages in source code, which is complementary to the prior work.

B. Analyzing Logs

Logs contain rich information which helps people understand the running state of the program. Prior work on analyzing logs mainly focus on debugging failures and performance problems.

The literature [53]–[60] uses different methods to diagnose functional system failures and performance faults using system log information. For instance, SherLog [53] analyzes the logs for deterministic replay in conjunction with source code to locate the execution paths and contexts when software errors occur. Lprof [59] associates dispersed log entries to specific individual requests and automatically reconstructs the execution flow of each request to detect performance anomalies. Similarly, Panorama [60] leverages logs to observe interactions between a system's components to find out the gray failures.

Our study mainly focuses on analyzing logs to infer configuration constraints inside the log messages. To the best of our knowledge, few research has been conducted before.

VII. CONCLUSION

Given that software logs often contain configuration constraints, this paper conducted an empirical study and summarized several patterns of configuration-related log messages. Guided by the study, we designed *ConfInLog*, a static tool to infer configuration constraints from log messages. An experiment on seven popular open-source software systems proves its effectiveness. *ConfInLog* could infer 22~163 constraints for the software systems, in which 59.5%~61.6% could not be inferred by the state-of-the-art work. Finally, we submitted 67 documentation patches regarding the constraints inferred by *ConfInLog*. The constraints in 29 patches have been confirmed by the developers, among which 10 patches have been accepted.

ACKNOWLEDGEMENT

This research was substantially supported by National Key R&D Program of China (Project No. 2017YFB1001802); National Natural Science Foundation of China (Project No. 61872373 and No. 61872375).

REFERENCES

- [1] L. A. Barroso, U. Hölzle, and P. Ranganathan, *The datacenter as a computer: Designing warehouse-scale machines*. Morgan & Claypool Publishers, 2018.
- [2] C. Tang, T. Kooburat, P. Venkatachalam, A. Chander, Z. Wen, A. Narayanan, P. Dowell, and R. Karl, “Holistic configuration management at facebook,” in *Proceedings of the 25th Symposium on Operating Systems Principles (SOSP)*, 2015, pp. 328–343.
- [3] B. Maurer, “Fail at scale: Reliability in the face of rapid change,” *Queue*, vol. 13, no. 8, pp. 30–46, 2015.
- [4] J. Shieber, “Facebook blames a server configuration change for yesterday’s outage,” <https://techcrunch.com/2019/03/14/facebook-blames-a-misconfigured-server-for-yesterdays-outage/>, 2019.
- [5] B. T. Sloss, “An update on sunday’s service disruption,” <https://cloud.google.com/blog/topics/inside-google-cloud/an-update-on-sundays-service-disruption>, 2019.
- [6] J. Foust, “SpaceX starship prototype destroyed after static-fire test,” <https://spacenews.com/spacex-starship-prototype-destroyed-after-static-fire-test/>, 2020.
- [7] G. Amvrosiadis and M. Bhadkamkar, “Getting back up: Understanding how enterprise data backups fail,” in *Proceedings of the 2016 USENIX Annual Technical Conference (ATC)*, 2016, pp. 479–492.
- [8] H. S. Gunawi, M. Hao, R. O. Suminto, A. Laksono, A. D. Satria, I. Adityatama, and K. J. Eliazar, “Why does the cloud stop computing? lessons from hundreds of service outages,” in *Proceedings of the Seventh ACM Symposium on Cloud Computing (SoCC)*, 2016, pp. 1–16.
- [9] W. Jiang, C. Hu, Y. Zhou, and A. Kanevsky, “Are disks the dominant contributor for storage failures? a comprehensive study of storage subsystem failure characteristics,” *ACM Transactions on Storage (TOS)*, vol. 4, no. 3, pp. 1–25, 2008.
- [10] S. Kendrick, “What takes us down,” *USENIX; login*, vol. 37, no. 5, pp. 37–45, 2012.
- [11] A. Rabkin and R. H. Katz, “How hadoop clusters break,” *IEEE Software*, vol. 30, no. 4, pp. 88–94, 2013.
- [12] T. Xu, L. Jin, X. Fan, Y. Zhou, S. Pasupathy, and R. Talwadker, “Hey, you have given me too many knobs!: understanding and dealing with over-designed configuration in system software,” in *Proceedings of the 2015 ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2015, pp. 307–319.
- [13] X. Liao, S. Zhou, S. Li, Z. Jia, X. Liu, and H. He, “Do you really know how to configure your software? configuration constraints in source code may help,” *IEEE Transactions on Reliability (TR)*, vol. 67, no. 3, pp. 832–846, 2018.
- [14] S. Nadi, T. Berger, C. Kästner, and K. Czarnecki, “Mining configuration constraints: Static analyses and empirical results,” in *Proceedings of the 36th IEEE/ACM International Conference on Software Engineering (ICSE)*, 2014, pp. 140–151.
- [15] A. Rabkin and R. Katz, “Static extraction of program configuration options,” in *Proceedings of the 33rd IEEE/ACM International Conference on Software Engineering (ICSE)*, 2011, pp. 131–140.
- [16] M. Santolucito, E. Zhai, R. Dhodapkar, A. Shim, and R. Piskac, “Synthesizing configuration file specifications with association rule learning,” *Proc. ACM Program. Lang.*, vol. 1, no. OOPSLA, 2017.
- [17] Z. Yin, X. Ma, J. Zheng, Y. Zhou, L. N. Bairavasundaram, and S. Pasupathy, “An empirical study on configuration errors in commercial and open source systems,” in *Proceedings of the 23rd ACM Symposium on Operating Systems Principles (SOSP)*, 2011, pp. 159–172.
- [18] T. Xu, J. Zhang, P. Huang, J. Zheng, T. Sheng, D. Yuan, Y. Zhou, and S. Pasupathy, “Do not blame users for misconfigurations,” in *Proceedings of the 24th ACM Symposium on Operating Systems Principles (SOSP)*, 2013, pp. 244–259.
- [19] Httpd, “Apache http server project,” <https://httpd.apache.org/>.
- [20] Bug-64893, https://bz.apache.org/bugzilla/show_bug.cgi?id=64893, 2020.
- [21] Q. Chen, T. Wang, O. Legunsen, S. Li, and T. Xu, “Understanding and discovering software configuration dependencies in cloud and datacenter systems,” in *Proceedings of the 2020 ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, 2020, pp. 362–374.
- [22] W. Landi and B. G. Ryder, “Pointer-induced aliasing: A problem classification,” in *Proceedings of the 18th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, 1991, p. 93–103.
- [23] S. Li, W. Li, X. Liao, S. Peng, S. Zhou, Z. Jia, and T. Wang, “Con-fvd: System reactions analysis and evaluation through misconfiguration injection,” *IEEE Transactions on Reliability (TR)*, vol. 67, no. 4, pp. 1393–1405, 2018.
- [24] S. Zhang and M. D. Ernst, “Proactive detection of inadequate diagnostic messages for software configuration errors,” in *Proceedings of the 2015 International Symposium on Software Testing and Analysis (ISSTA)*, 2015, pp. 12–23.
- [25] F. A. Arshad, R. J. Krause, and S. Bagchi, “Characterizing configuration problems in java ee application servers: An empirical study with glassfish and jboss,” in *Proceedings of the 24th IEEE International Symposium on Software Reliability Engineering (ISSRE)*, 2013, pp. 198–207.
- [26] L. Keller, P. Upadhyaya, and G. Candea, “Conferr: A tool for assessing resilience to human configuration errors,” in *Proceedings of the 2008 IEEE International Conference on Dependable Systems and Networks With FTCS and DCC (DSN)*, 2008, pp. 157–166.
- [27] Z. Jia, S. Li, X. Liu, X. Liao, and Y. Liu, “Smartlog: Place error log statement by deep understanding of log intention,” in *Proceedings of the 25th IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2018, pp. 61–71.
- [28] S. Zhou, X. Liu, S. Li, W. Dong, X. Liao, and Y. Xiong, “Confmapper: Automated variable finding for configuration items in source code,” in *Proceedings of the 2016 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, 2016, pp. 228–235.
- [29] D. Lawrie, C. Morrell, H. Feild, and D. Binkley, “What’s in a name? a study of identifiers,” in *Proceedings of 14th IEEE International Conference on Program Comprehension (ICPC)*, 2006, pp. 3–12.
- [30] H. Liu, Q. Liu, C.-A. Staicu, M. Pradel, and Y. Luo, “Nomen est omen: Exploring and exploiting similarities between argument and parameter names,” in *Proceedings of the 38th IEEE/ACM International Conference on Software Engineering (ICSE)*, 2016, pp. 1063–1073.
- [31] A. Rice, E. Aftandilian, C. Jaspán, E. Johnston, M. Pradel, and Y. Arroyo-Paredes, “Detecting argument selection defects,” *Proc. ACM Program. Lang.*, vol. 1, no. OOPSLA, 2017.
- [32] R. Pandita, X. Xiao, H. Zhong, T. Xie, S. Oney, and A. Paradkar, “Inferring method specifications from natural language api descriptions,” in *Proceedings of the 34th IEEE/ACM International Conference on Software Engineering (ICSE)*, 2012, pp. 815–825.
- [33] V. Raychev, M. Vechev, and E. Yahav, “Code completion with statistical language models,” in *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, 2014, pp. 419–428.
- [34] V. Raychev, M. Vechev, and A. Krause, “Predicting program properties from “big code”,” *ACM SIGPLAN Notices*, vol. 50, no. 1, pp. 111–124, 2015.
- [35] P.-N. Tan, S. Michael, and K. Vipin, *Introduction to data mining*. Pearson Education India, 2016.
- [36] K. S. Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of documentation*, 1972.
- [37] WordNet, “A lexical database for english,” <https://wordnet.princeton.edu/>.
- [38] spaCy, “Industrial-strength natural language processing in python,” <https://spacy.io/>.
- [39] M. Attariyan and J. Flinn, “Automating configuration troubleshooting with dynamic information flow analysis,” in *Proceedings of the 2010 USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2010, pp. 237–250.
- [40] M. Attariyan, M. Chow, and J. Flinn, “X-ray: Automating root-cause diagnosis of performance anomalies in production software,” in *Proceedings of the 2012 USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2012, pp. 307–320.
- [41] A. Rabkin and R. Katz, “Precomputing possible configuration error diagnoses,” in *Proceedings of the 26th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2011, pp. 193–202.
- [42] S. Zhang and M. D. Ernst, “Automated diagnosis of software configuration errors,” in *Proceedings of the 35th IEEE/ACM International Conference on Software Engineering (ICSE)*, 2013, pp. 312–321.
- [43] L. Bauer, S. Garriss, and M. K. Reiter, “Detecting and resolving policy misconfigurations in access-control systems,” *ACM Transactions on*

- [44] J. Zhang, L. Renganarayana, X. Zhang, N. Ge, V. Bala, T. Xu, and Y. Zhou, “Encore: Exploiting system environment and correlation information for misconfiguration detection,” in *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2014, pp. 687–700.
- [45] W. Li, S. Li, X. Liao, X. Xu, S. Zhou, and Z. Jia, “Conftest: Generating comprehensive misconfiguration for system reaction ability evaluation,” in *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 2017, pp. 88–97.
- [46] C. Xiang, H. Huang, A. Yoo, Y. Zhou, and S. Pasupathy, “Pracextractor: Extracting configuration good practices from manuals to detect server misconfigurations,” in *Proceedings of the 2020 USENIX Annual Technical Conference (ATC)*, 2020, pp. 265–280.
- [47] D. Yuan, Y. Xie, R. Panigrahy, J. Yang, C. Verbowski, and A. Kumar, “Context-based online configuration-error detection,” in *Proceedings of the 2011 USENIX Annual Technical Conference (ATC)*, 2011, pp. 313–326.
- [48] M. Attariyan and J. Flinn, “Using causality to diagnose configuration bugs,” in *Proceedings of the 2008 USENIX Annual Technical Conference (ATC)*, 2008, pp. 281–286.
- [49] D. Zhen, A. Artur, and S. Kun, “Practical and accurate pinpointing of configuration errors using static analysis,” in *Proceedings of the 2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2015, pp. 171–180.
- [50] P. Huang, W. J. Bolosky, A. Singh, and Y. Zhou, “Confvalley: A systematic configuration validation framework for cloud services,” in *Proceedings of the 10th European Conference on Computer Systems (EuroSys)*, 2015, pp. 1–16.
- [51] T. Xu, X. Jin, P. Huang, Y. Zhou, S. Lu, L. Jin, and S. Pasupathy, “Early detection of configuration errors to reduce failure damage,” in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 619–634.
- [52] C. Xiang, Y. Wu, B. Shen, M. Shen, H. Huang, T. Xu, Y. Zhou, C. Moore, X. Jin, and T. Sheng, “Towards continuous access control validation and forensics,” in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019, pp. 113–129.
- [53] D. Yuan, H. Mai, W. Xiong, L. Tan, Y. Zhou, and S. Pasupathy, “Sherlog: Error diagnosis by connecting clues from run-time logs,” in *Proceedings of the 15th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2010, pp. 143–154.
- [54] I. Beschastnikh, Y. Brun, M. D. Ernst, and A. Krishnamurthy, “Inferring models of concurrent systems from logs of their behavior with csight,” in *Proceedings of the 36th IEEE/ACM International Conference on Software Engineering (ICSE)*, 2014, pp. 468–479.
- [55] W. Xu, L. Huang, A. Fox, D. Patterson, and M. I. Jordan, “Detecting large-scale system problems by mining console logs,” in *Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles (SOSP)*, 2009, pp. 117–132.
- [56] L. Mariani and F. Pastore, “Automated identification of failure causes in system logs,” in *Proceedings of the 19th IEEE International Symposium on Software Reliability Engineering (ISSRE)*, 2008, pp. 117–126.
- [57] J.-G. Lou, Q. Fu, S. Yang, Y. Xu, and J. Li, “Mining invariants from console logs for system problem detection,” in *Proceedings of the 2010 USENIX Annual Technical Conference (ATC)*, 2010, pp. 1–14.
- [58] X. Zhao, K. Rodrigues, Y. Luo, M. Stumm, D. Yuan, and Y. Zhou, “Log20: Fully automated optimal placement of log printing statements under specified overhead threshold,” in *Proceedings of the 26th Symposium on Operating Systems Principles (SOSP)*, 2017, pp. 565–581.
- [59] X. Zhao, Y. Zhang, D. Lion, M. F. Ullah, Y. Luo, D. Yuan, and M. Stumm, “lprof: A non-intrusive request flow profiler for distributed systems,” in *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2014, pp. 629–644.
- [60] P. Huang, C. Guo, J. R. Lorch, L. Zhou, and Y. Dang, “Capturing and enhancing in situ system observability for failure detection,” in *Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2018, pp. 1–16.