A Comprehensive Survey of Logging in Software: From Logging Statements Automation to Log Mining and Analysis

Sina Gholamian, Student Member, IEEE, Paul A. S. Ward, Member, IEEE

Abstract—Logs are widely used to record runtime information of software systems, such as the timestamp and the importance of an event, the unique ID of the source of the log, and a part of the state of a task’s execution. The rich information of logs enables system developers (and operators) to monitor the runtime behaviors of their systems and further track down system problems and perform analysis on log data in production settings. However, the prior research on utilizing logs is scattered and that limits the ability of new researchers in this field to quickly get to the speed and hampers currently active researchers to advance this field further. Therefore, this paper surveys and provides a systematic literature review and mapping of the contemporary logging practices and log statements’ mining and monitoring techniques and their applications such as in system failure detection and diagnosis. We study a large number of conference and journal papers that appeared on top-level peer-reviewed venues. Additionally, we draw high-level trends of ongoing research and categorize publications into subdivisions. In the end, and based on our holistic observations during this survey, we provide a set of challenges and opportunities that will lead the researchers in academia and industry in moving the field forward.

Index Terms—survey, systematic literature review (SLR), systematic mapping (SM), software systems, logging, log statement, log file, log automation, log analysis, log mining, logging cost, anomaly detection, failure detection and diagnosis.

1 INTRODUCTION

SOFTWARE systems are pervasive and play important and often critical roles in the society and economy such as in airplanes or surgery room patient monitoring systems. Gathering feedback about software systems’ states is a non-trivial task and plays a crucial role in system diagnosis in the case of a failure. In the interest of higher availability and reliability, software systems regularly generate log files of their status and runtime information. Developers insert logging statements into the source code which are then printed in the log files, also known as execution logs and event logs [1]. Then, at a later time, while the system is running or postmortem, developers or operators would analyze the log files for various tasks. For example, the content of log files has been studied to achieve a variety of goals such as anomaly and fault detection [2, 3, 4, 5], online or postmortem performance and failure diagnosis [6, 7, 8, 9, 10], pattern detection [11, 12, 13], profile building [14], business decision making [14], and user’s behavior observation [15].

Conventionally, software developers and practitioners apply testing and monitoring techniques to analyze the software systems. System testing happens during the development phase by developers, while practitioners utilize system monitoring techniques to understand the behavior of the system in the deployed environment [15]. As such, it is a common practice to have running programs report on their internal state and variables, through log files that system administrators and operators can analyze [17] for different purposes. This continuous cycle of development and deployment of the software and looking at the system logs has also initiated and thrived adjacent fields of research such as DevOps [18, 19, 20]. That is, the importance of log analysis and its computational intensity has also brought in other tools to scale up the effort. For example, considering the advancements of machine learning (ML) and artificial intelligence (AI) and the vast size of the log files, research has proposed the use of ML for automated operation analysis (AIOps) [21] of execution logs. From the commercialization perspective, the widespread need for log analysis has also contributed to the emergence of commercial products such as Splunk [22] and Elastic Stack [23]. Splunk makes the large-scale logs accessible by extracting patterns and correlating system metrics to diagnose problems and provide insight for business decisions. Elastic Stack [23], a.k.a. ELK, consists of three different subsystems of Elasticsearch [24], Logstash [25], and Kibana [26]. Works to ingest and process logs from different sources by Logstash, in a searchable format accomplished by Elasticsearch, and Kibana lets users visualize data with charts and graphs.

From the system’s diagnosis perspective, the information provided through the logging statements combined with other system metrics, such as CPU, memory, and I/O utilization, serves an important role in anomaly detection and understanding and diagnosing the system’s runtime behavior in the case of a failure. Despite the tremendous potential value hidden in execution logs, the inherent characteristics of logs, such as their heterogeneity and voluminosity [16], make the analysis of them difficult on a large scale and poses several challenges. Some of the associated challenges with logging statements and their analysis in software systems

- S. Gholamian and P. Ward are with University of Waterloo, Waterloo, Canada, N2L 3G1.

E-mail: {sgholamian, pasward}@uwaterloo.ca
Providing proper logging statements inside the source code remains a manual, ad-hoc, and non-trivial task in many cases [27] due to the free-form text format of log statements and lack of a general guideline for logging.

As the size of computer systems increases and software becomes more complex and distributed, manual inspection of log files becomes cumbersome and impractical, and it calls for automated analysis of logs.

Log data can be heterogeneous and voluminous, as within a large software system multiple subsystems may potentially generate a plethora of logs in different formats. Additionally, logging inefficiently introduces overhead on multiple subsystems such as I/O, network, and storage.

Developers and automatic logging analysis tools that aim to automate the logging process always face challenging questions of “what, where, how, and whether to log?”.

Considering the aforementioned challenges, prior and ongoing research has made numerous efforts to mine and understand log statements in the source code and execution of logs to either gain more insights about logging practices, troubleshoot the software, or automate the logging process [28, 29, 30]. Figure 1 depicts a framework in which the creation process and analysis of log files are illustrated. After the system’s architecture is decided and programmers implement the source code with logging statements, the operators run the system by selecting proper runtime configuration parameters (e.g., logging verbosity level in Log4j [31]). While the software is running, events that are logged in the source code generate records within the log files. Next, administrators, practitioners, and automated log analyzer tools may review the files and feedback the outcome to the designer, programmers, and operators to make adjustments to the architecture, source code, and system configuration if needed, respectively.

Fig. 1: Framework for creation and analysis of log files [32].

As the size of computer systems increases, the manual process of developers placing logging statements into the source code and administrators reviewing the log files and detecting problems negatively affecting the system becomes less practical and less effective. Consequently, being able to automatically detect logging points and insert appropriate logging statements in the source code, as well as systematically analyze log files for detecting system issues are very beneficial and high-in-demand research topics [30]. Thus, researchers have dedicated a significant amount of studies in the area of software logging and log analysis techniques throughout the last decade [16] to propose various approaches, including automated analysis and application of machine learning to process large-scale log files. However, after reviewing the prior work, we noticed what has not been addressed is a clear and comprehensive review of the current progress in software systems’ logging and log analysis research. It is, therefore, hard for researchers to recognize how their current and future work will fit in the big picture of present-day logging research. Understanding where we are at the moment and creating a snapshot of the current research is a fundamental step towards understanding where we should go from here and what the necessary next steps of the research would be. Learning from our own experiences and the obstacles that we have to go through to holistically picture the current research in the field, this survey aims to pave the road obstacles and provide a methodological review of logging, its practices, and its automation techniques and tools for software systems. Additionally, in this survey, we review the current state of logging in software to discover solutions for the aforementioned challenges and highlight the next steps for future research efforts. We review and study a vast number of peer-reviewed conference and journal papers from related research areas including software and distributed systems, dependability and reliability, and machine (and deep) learning. Moreover, we aim to build knowledge [34] and trends by connecting and combining findings from multiple research. Thus, we examine and categorize the prior research for logging costs, logging practices, automation of log analysis, and efforts to automate the insertion and improvement of logging statements inside the source code. Finally, we provide trends and opportunities for future work based on the insights we gain during this survey. Before we dive deeper into the survey, we review the necessary vocabulary in the following.

1.1 Terminology

Log printing statement (LPS). LPSs are the log statements in the source code added by developers. We use "log printing statement", "log statement", and "logging statement" interchangeably, as the prior work has used all of the variations [4], [33], [35], [36], [37].

Log message. A log message, typically a single line, is the output of the LPS in the log file. Prior work also makes a subtle distinction between a log message and a log entry, and defines a "log entry" or a "log record" as a line in the log file composed of a log header and a log message [38]. Log header contains timestamp, verbosity level, and source component, and its format is usually defined by the logging framework, e.g., log appenders [39], whereas a log message is written by the developer and consists of the amalgamation of the static part of the log message from the source code and the dynamic value of the variables during the runtime. For our purpose in this survey, log message, log entry, log record, and log event are used interchangeably [37].

Log file. Log file(s) is a collection of log messages stored on a storage medium, also called "event logs", and "execution logs", or simply just "logs" [4], [41], [42]. In most cases, these terms can be used interchangeably, and we commonly use log file(s) as an umbrella term to cover the different naming variations throughout the survey. As a minor point, in special cases that we mean to refer to a set of log lines in general (i.e., without binding them to specific files), it is more appropriate to refer to them as execution logs or log records (e.g., execution logs are used for anomaly detection).

Additionally, we might refer to computer (computing) systems and software systems interchangeably on some
1.2 Survey Organization

The rest of this survey is organized as follows. Section 2 provides the background for log statements, messages, and files. Section 3 provides the details of our study design and research questions (RQs) based on the SLR and SM conventions in performing evidence-based surveys in software engineering [43]. In Section 4 we provide our findings for RQ1 and categorize the prior logging research. In Section 5 we present our findings for RQ2 and present the publication trends for different topics, years, and venues. Then, Section 6 reviews the prior research in each category of logging in detail and provides our findings for RQ3. In Section 7, we provide our findings for RQ4 and describe open problems and opportunities for future work, and Section 8 concludes the survey.

2 LOG STATEMENTS AND LOG FILES

Logging is the process of recording and keeping track of the events of interest, e.g., to developers, practitioners, system admins, and end users, while the software is running. As such, log messages aim to achieve this goal and record the events of interest that happen during the software system’s execution and store them in the log files. Generally, the logging process starts with software developers (i.e., programmers) include logging statements with description, variables, and verbosity levels (Figures 2 and 3) into the source code. Then, while the software is running, the logging statements are logged, if appropriate configurations (e.g., verbosity level) are enabled. In the simplest case, log messages are written to a single log file. However, in a distributed system, there can be multiple log files in different formats. The focus of our survey is on this type of logs which are also called execution logs or event logs [44]. Event logs are the outcome of logging statements that software developers insert into the source code. Event logging and log files are playing an increasingly important role in computer systems and network management [2], [16], [44], which we will review later in this survey.

2.1 Transaction Logs

It is also worth mentioning briefly the difference between the execution logs and transaction logs. A transaction log (also called journal) is a record file of the transactions between a system and the users of that system, or a data collection method that naturally captures the type, content, or time of transactions made by a user from interacting with the system. In a database server, a transaction log is a file in which the server stores a record of all the transactions performed on the database [45]. The transaction log is an important component of database servers and cryptocurrency protocols (e.g., blockchain [46]) when it comes to recovery. If there is a system failure, transaction logs are used to revert the database back to a consistent state. In summary, transaction logs act as a ledger to accurately record the transactions in the system which are agreed, shared, and synchronized among all the parties involved, whereas execution logs capture events of interests with different verbosity and severity levels, e.g., DEBUG, INFO, ERROR, etc., which do not necessarily require sharing, agreement (i.e., consensus), or synchronization between different software modules.

2.2 Log Example

Software developers utilize logging statements inside the source code to gain insight into the internal state of applications amid their execution. In the simplest form, logging statements are print statements utilized in different programming languages. In this case, the logging statement may contain a textual part indicating the context of the log, i.e., the description of the log, and a variable part providing contextual information about the event. Figure 2 shows an example of logging statements in C programming language.

```
printf("Cannot find BPService for bpid=%d", id);
```

Fig. 2: Log example in C.

Logging statements may utilize logging libraries to improve the organization of the logged information. For example, in Java, libraries such as Log4j [31] and SLF4J [47] provide a higher degree of flexibility to the developers. Logging libraries, also called logging utilities (LU) [48] or logging libraries and utilities (LLUs), provide extra features, such as log level, which indicate the verbosity and the severity of the logging statement. Log levels help to better distinguish the importance of runtime events and control the number of logs collected on the storage device [49]. For example, less verbose levels, i.e., FATAL, ERROR, and WARN, are leveraged to alarm the user when a potential problem happens in the system, and more verbose levels such as INFO, DEBUG, and TRACE are utilized to record more general system events and information or detailed debugging. In practice, INFO and more verbose levels are utilized during the software development phase by programmers, and INFO or less verbose levels are, by default, for the software deployment phase, as the end user observes. In case more insight about the internal state is needed, end users might enable more verbose logging. An example of a logging statement with library usage for WARN verbosity level is shown in Figure 3.

```
log.warn("Cannot find BPService for bpid=" + id);
```

Fig. 3: Log example with Log4j library.

Logging statements are generally saved in log files. Figure 4 shows ten lines of logs from Apache Spark [50] execution logs collected in our execution of k-means clustering algorithm [51] on a standalone cluster. In real-world cases, a cloud computing system can generate millions of such
log messages per minute [52]. For example, for an online store with millions of customers worldwide, it is common to generate tens of terabytes of logs in a single day [53, 54].

| Log file example |
|------------------|
| 2020/02/21 13:47:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in java classes where applicable. |
| 2020/02/21 13:47:49 INFO SparkContext: Running Spark version 2.4.4 |
| 2020/02/21 13:47:48 INFO SparkContext: Submitted application: JavaKMeansExample |
| 2020/02/21 13:47:48 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(10); groups with view permissions: Set(); users with modify permissions: Set(); groups with modify permissions: Set(); |
| 2020/02/21 13:47:48 INFO EventLoggingListener: Logging events to file:///tmp/spark-events/app-20200221134749-0004 |
| 2020/02/21 13:47:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... |

Fig. 4: A log file example from Apache Spark.

As observed in Figure 4, a major part of logging messages is unstructured text. Thus, in order to make log files useful and avoid the hassle of manually processing a plethora of log files, the first and foremost step of log processing is the automatic “parsing” of log messages, which transforms unstructured logs into structured events. Not to be confused with syntactic parsers in programming languages, which parses source code and confirm whether it follows the rules of the formal grammar, log parser, on the other hand, transforms unstructured raw log files into a sequence of structured events, to enable automated analysis of logs. We review the log parsing process later in Section 6.4.3.

2.3 Program Traces vs. Logs

The term log is often used to represent the way a program “TRACES” log level in logging libraries such as Log4j) is used to capture the temporal sequence of events during a particular execution of a program [55], in contrast to logs which are generally the consolidation of continuous execution of software systems. Tracing is typically performed by an external program/tool that instruments the runtime environment, such as network traffic traces, whereas logs are the direct output of logging statements’ execution inside the software. Moreover, while traces are typically structured data, logs are free-form and unstructured text. For example, a trace can contain the software execution paths, the events triggered during the execution, and the value of variables, which are used for debugging and program understanding. Stack traces are common examples that are used for function call tree tracing during development and postmortem debugging. Some of the well-known tracing systems include Google’s Dapper [56], X-Trace [57], Microsoft’s MagPie [58], the black-box approach [59], and Casper [60].

3 STUDY DESIGN

We perform a systematic literature review (SLR) and mapping study (MS) following Kitchenham et al. [43], [61] and Petersen et al. [62] guidelines. In the following, we elaborate on the details of this process.

3.1 Research Questions

While we review the prior literature, we aim to pursue and answer the following research questions:

- RQ1: How to systematically review and categorize prior logging research into different topics?
- RQ2: What are the publication trends based on venues, topics, and years?
- RQ3: How the research in each topic can be systematically compared with their approaches, pros, and cons?
- RQ4: What open problems and future directions are foreseeable for logging research?

With the pursuance of the aforementioned RQs, we ensure to follow the established evidence-based software engineering (EBSE) paradigm for our literature review. As a contributing improvement, our survey combines and benefits from the advantages of both systematic literature review (SLR) and systematic mapping (SM) paradigms. Prior research indicates the main differences between SM and SLR are that SM methodology is broader and more based on qualitative measures, while SLR focuses on narrower research questions and quantitative measures [63]. While being comprehensive, i.e., SM, our survey also provides details on the experimentation and results of each primary study, i.e., SLR. In summary, RQ1, RQ3, and RQ4 qualitatively assess the prior literature into different research categories base on different topics, i.e., systematic mapping (SM) [5], whereas RQ2 quantitatively measures the publications based on venues, topics, and years, and RQ3 provides details aligned with SLR data extraction methods for each study, e.g., the aim, experiments, results, and findings [63].

3.2 Search Strategy and Paper Selection

Studying log mining and logging analysis techniques in software systems is a challenging and widespread topic, and there have been numerous prior studies that focus on log analysis. Table 2 summarizes the databases and the search query used in our survey, and Figure 5 provides the flowchart of the steps in the database selection and reference analysis. We start with the established research databases listed in the table and search for the list of log-related keywords. If the database allows for metadata search (i.e., title, abstract, and keywords), we limit our search to metadata, to avoid the inclusion of numerous unrelated research that has the keywords in their main text. However, for Springer, we could not use meta-search, and thus, the reason behind the high number of returned publications. We use reference management software, such as Zotero 1 and JabRef 2 to facilitate and automate our process of reference consolidation from different sources and duplicate removal. After duplicate removal, our process resulted in

1. https://www.zotero.org/
2. https://www.jabref.org/
4,906 papers in the first revision of this survey. In the second revision (June-October 2021), we gathered additional 633 publications.

---

### Research databases

| Research databases |
|-------------------|
| ACM Digital Library, IEEE Xplore, ScienceDirect (Elsevier), Scopus, SpringerLink, and Google Scholar (snowballing). |

### Search query

| Search query |
|--------------|
| ‘software AND log AND (statement OR file OR record OR event)’ AND ‘Publication Date: (01/01/2010 TO 05/31/2021)’ |

---

### Exclusion criteria

**C1** Primary study is in English language.

**C2** Final (and long) version of an study is selected.

**C3** The date of publication is ≥ 2010.

**C4** It paper is published in very good to flagship venues.

**C5** The search query exists in the title, abstract, or keywords of the publication.

**C6** Additional studies that are included through single iteration of backward and forward snowballing search.

---

**TABLE 1:** Literature review databases and keywords. We searched in various online research databases for different combinations of log-related keywords.

**TABLE 2:** Selection criteria for primary studies.

---

### 3.3 Inclusion and Exclusion Criteria

Table 1 summarizes the selection criteria in our study. For exclusion criteria, our focus has been on the full (i.e., not short), recent papers in English in the last decade in the established venues, ranging from very good to flagship ones. We focus on the last decade as our goal is to provide a fresh picture of the recent trends in logging research to enable relevant directions for future work. In addition, other surveys, such as Salfner et al. [65], provided a comprehensive overview of failure detection approaches up to 2010, which also leverage log analysis. Moreover, Candido et al. [16] reviewed and found very few studies earlier than 2010 in their systematic literature review of software monitoring through logs. Laterally, we observe that another factor that has triggered a momentum in logging research is the recent rise in artificial intelligence (AI), machine learning (ML), and in particular deep learning (DL) approaches [66]. A significant portion of logging code automation and automated log analysis approaches leverage ML and DL algorithms.

After we gather the publications with the search query, we manually investigated the results and included relevant publications that directly tackle issues associated with log statements or log files, and excluded the ones which had a weak association with logs. Due to the significant number of curated papers, for manually reading the papers, we follow the prior work’s suggestion [67] to read the title and abstract first and then other parts of the papers to decide on inclusion in an efficient way. A significant number of papers are excluded as we read the beginning of the papers. Also, if a research project has multiple variations, i.e., a conference paper followed by a more comprehensive journal paper, we only include the more comprehensive version. Additionally, once we find an influential paper (i.e., highly-cited based on the absolute number of citations > 100 [68]), we also check all of its references and its citations, i.e., a single round of backward and forward snowballing search [67] with Google Scholar. Followed by exclusion criteria, we are narrowed down to 92 publications, and at last, 11 additional references are added with snowballing, bringing the total count to 103 publications in the first revision of the paper. In the second revision, based on the reviewers’ feedback, we added three additional primary studies [69], [70], [71]. We also reviewed the databases and list of accepted papers in venues in Table 5 for new publications, for June-October 2021 interval, and found six recent studies [72], [73], [74], [75], [76], [77]. This brings the total number of primary studies to 112.

---

3. https://www.core.edu.au/conference-portal
3.4 Data Extraction and Collection

After the primary studies (112) are selected, from each study, we extract the data required for analyzing and answering each research question. For example for RQ1 and RQ2, from each paper, we extract its primary topic and other secondary topics (i.e., subtopics) which are discussed in the study. For RQ3, for each primary study, we provide its 1) aim, i.e., the problem it is trying to address, 2) experimentation, 3) results and findings, 4) advantages, i.e., pros, and 5) disadvantages, i.e., cons. We extract pros and cons based on 1) authors explanation of the main benefit and limitation of their work, 2) findings of the follow-up studies that have performed comparisons, or finally, 3) our gained knowledge during the survey and by reviewing the research in comparison to other related studies. The data extraction is initially performed by the first author of this survey, and then reviewed by the second author. In cases where there exists a disagreement, both authors discuss the issue until reaching a consensus. Table 3 lists the data extracted from primary studies for each RQ.

Based on the extracted data, we categorize the selected publications (112) into twelve categories after carefully studying them. Our methodology for categorizing the publications has been based on a top-down approach. Meaning that, we first were able to draw categories that either focus on logging statements or the ones that focus on log files. We then further narrowed down each category based on its primary focus. Next, we also extracted subtopics for each paper, as usually, publications also partially cover some other related topics in their research. For example, Zhao et al. work [33] primarily focuses on log statement automation, but it also covers topics related to logging cost analysis. The details for each category is available in Table 19 in Section 9.

3.5 Summary of Differences

Our survey differs from other related surveys in various ways as listed in Table 4. We summarize how our study differs from the other recent surveys that cover parts of the software logging research domain. Not only our approach is different in a variety of ways from the compared surveys (e.g., we extract different types of data from primary studies), our work also differs significantly in the selected primary studies. In comparison, out of 112 primary studies that we have reviewed, 57, 70, and 80 of them are not included in our work. For example, Zhao et al. [33] primarily focuses on log statement automation, but it also covers topics related to logging cost analysis. The main differences for each category is available in Table 4.

**TABLE 3:** This table summarizes the data extracted for each RQ from each primary study.

| RQs       | Extracted data                                                                 |
|-----------|--------------------------------------------------------------------------------|
| RQ1       | General categories of main topics and subtopics of the primary studies.         |
| RQ2       | Details associated with its year of publication, venue, and the topic that it belongs too. |
| RQ3       | Five items for each primary study: aim, experiments, results, pros, and cons.   |
| RQ4       | Open problems and opportunities for the categories of the primary studies if available, i.e., if explicitly or implicitly can be extracted. |

**TABLE 4:** This table summarizes the difference between this survey and other recent survey studies.

| Survey               | Online databases | Type   | Type                     | RQs                                                                 | Main differences (Us vs. them)                                                                 |
|----------------------|------------------|--------|--------------------------|----------------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| He et al. [78]       | IEEE, ACM, Springer, Elsevier, Wiley, and ScienceDirect | SLR    | ‘log OR logging OR log parsing OR log compression OR (log + anomaly detection) OR (log + failure prediction) OR (log + failure diagnosis)’ | No clear-cut RQs; the survey discusses logging mechanism, compression, and mining for the goal of reliability analysis. | Main differences in the search query and RQs. Our work is different on logging cost, log statement automation, evaluation metrics, and taxonomy chart. |
| Chen and Jiang [79]  | IEEE, ACM, and DBLP | SLR    | ‘logging OR instrumentation OR tracing’ | No clear-cut RQs; the goal of the survey is to techniques used in all three software log instrumentation approaches: conventional logging, rule-based logging, and distributed tracing, and uncovers four categories of challenges. | Main differences in the databases, search query, and RQs. Our work is different on various topics covered in automated logging. |
| Candido et al. [80]  | Google Scholar, ACM, IEEE, Scopus, and Springer | SM     | ‘log AND (trace OR event OR software OR system OR code OR detect OR mining OR analysis OR monitoring OR web OR technique OR develop OR pattern OR practice)’ | RQ1: trends in log-based monitoring; RQ2: Different scopes of log-based monitoring | Main differences in the search query, RQs, and focus of the survey; it focuses on monitoring for studies up to the end of 2019. |
| Gholamian and Ward (This study) | IEEE, ACM, Springer, ScienceDirect, Scopus, and Google Scholar | SM/SLR | ‘software AND log AND (statement OR file OR record OR event)’ | Four RQs listed in Section 3.1 on the categorization of subtopics, publication trends, systematic comparison of each topic, and future directions. | In sum, we differ in the search query and more comprehensive approach in searching, the proposed RQs, and one-of-a-kind discussion on taxonomy (Figure 6), findings for each RQ, and directions on research opportunities (Table 32). |
reviewed by He et al. [78], Cândido et al. [80], and Chen and Jiang [79], respectively. Additionally, 39 primary studies that we reviewed are not examined in the three aforementioned surveys combined, i.e., they are exclusive to this survey.

4 RQ1: HOW THE PRIOR LOGGING RESEARCH CAN BE CATEGORIZED TO DIFFERENT TOPICS?

In this section, we provide our findings for the first research question by reviewing the available literature and categorizing the logging research into its topics (and subtopics), which enables us to explain the survey scope that follows next.

4.1 Survey Scope

Based on our methodology, the scope of our survey is developed as follows. To the best of our knowledge, there is no prior work that provides a systematic and comprehensive coverage on log mining and automation techniques in software systems, covering different aspects of logging such as mining source code and log files, automating log printing statements in the source code, their evaluation techniques, as well as a comprehensive review of log mining and log analysis approaches. As mentioned before, there are a few existing surveys on the application of execution monitoring [16], reliability engineering [78], and instrumentation [79].

Thus, we cover the following sections in this survey:

- **Logs and log files.** We explain what are log files, logging statements, and log messages, and what sort of applications and analyses they are leveraged for.
- **Logging cost.** We point out the quantitative and qualitative costs and benefits associated with logging.
- **Logging statement mining and automating.** Logging research aims to understand current logging practices and use the findings to improve the log printing statements with automatic log insertion and learning to log techniques. Thus, we review:
  - **Logging code practices.** This section includes studies that empirically or automatically investigate how developers insert logging statements into the software’s source code and how the logging evolution and improvement can benefit the usage of the logging code.
  - **Automatic log insertion and learning.** We cover the studies that leverage static code analysis, heuristics, and machine learning techniques to automatically add (or improve) logging statements in the source code to make them more effective in failure diagnosis.
  - **Evaluation.** As we study automatic ways of addition and learning of log statements, we introduce several metrics to measure the effectiveness of the proposed approaches.
- **Mining logs.** This part provides insight into methods and tools to analyze log messages and log files. This can be further divided into log management, log parsing, and their applications. Thus, we review:

  - **Log management and maintenance.** Management and collection of logs are important as a pre- or post-step of log analysis.
  - **Log parsing.** To enable log message analysis, we first require to parse the log messages and extract their templates.
  - **Application of logs.** We review a wide range of applications that leverage automated log analysis for various software engineering tasks, such as anomaly detection and failure diagnosis.
  - **Emerging applications of logs.** We review the recent special interest in applications of logs in other domains, such as mobile devices and big data.
  - **Opportunities for future work.** Based on the current log-related research, we comment on the future directions and opportunities for each category of logging research.

As our aim is to provide a comprehensive and end-to-end survey of the logging in software, we have covered the topics of mining and automation of logging statements and mining of log files side-by-side in this survey. We have seen that these topics go hand-in-hand and the synergy between them has resulted in more effective logs and logging practices. In fact, the ultimate goal of a mountain of studies for logging statement automation is to improve various log mining tasks (Figure 1). For example, ErrLog [81] and LogEnhancer [82] automatically introduce new logging statements or add additional variables to the logging statements (i.e., automation of the logging statements) to improve the quality of logs, and, subsequently, improve log mining tasks such as error detection and program diagnosability.

In another example, authors of Log20 [33], an automated log placement tool, explain that the main objective of their log placement tool is to “disambiguate” the execution paths, and consequently, improve the effectiveness of log mining methods in “debugging real-world failures”. Furthermore, we observe that ignoring the cross-cutting concerns of logging has resulted in log-related issues in the past, such as stale and confusing logging statements, and has hindered effective log file mining [83], [84].

It is also important to mention that there exist additional industry products, with an aggregate market cap beyond $125 billion USD, that perform log analysis for various goals such as performance evaluation, cloud monitoring, and data analytics, to name a few: Datadog [85] ($49.7B), Splunk [22] ($27.3B), Elasticsearch [24] ($15.5B), Loggly [86] ($5.7B), Dynatrace [87] ($22.1B), New Relics Inc. [88] ($4.9B), and XpoLog [89]. However, this paper mainly focuses on surveying academic works, or peer-reviewed publications from the industry. We acknowledge that there is a mountain of work in the industry that performs log analysis, however, it is mainly outside the scope of this survey. Additionally, logging is also used in other computing systems, such as embedded or hardware devices/sensors, which are generally outside the scope of this survey, as we mainly focus on software systems. In addition, we mention some of the mobile devices studies that are closely related and have

4. Values are collected at the time of this survey from Google Finance. For example for Datadog: https://www.google.com/finance/quote/DDOG:NASDAQ
Fig. 6: This figure provides a taxonomy of the present-day logging research. We initially provide a background on logs and log messages in §2, our study design in §3, and the categorization of research to subtopics in §4. We then provide the publication trends in §5 and explain logging costs and benefits in §6.1. Log printing statement research covers two sections, §6.2-6.3, followed by log files’ research in §6.4. §6.5 and §7 provide emerging applications and opportunities for future work, respectively. We conclude our survey in §8.
aimed to replicate the efforts in software logging research in Section 5.5. Lastly, Figure 8 summarizes a taxonomy of modern-day logging research in our survey, and we leverage the classification in this figure to divide the logging research into subtopics and study them in the upcoming sections.

5 RQ2: WHAT ARE THE PUBLICATION TRENDS BASED ON VENUES, TOPICS, AND YEARS?

For this RQ, after categorizing the 112 selected publications, we organize the publications based on their venues and publication years and draw high-level trends. Based on our findings from the trends, we summarize some of the log-related challenges, such as automating the logging code and automated log analysis, that the prior research has aimed to address at the end of this section.

5.1 Venue Trends

Table 5 provides a breakdown of surveyed publications per venue. Although looked for, we could not find a related work in TOSEM. The majority of the research in this field is published in EMSE, ICSE, TSE, ASE, and JSS. We suggest that the authors of future publications consider the following venues in Table 5 for submitting their works, and consider the number of related references that their work aligns with in that venue. This ensures that their works will receive higher visibility and a thorough comparison with the prior work.

5.2 Topic Trends

Figure 7 shows the percentage of publications per each topic. Overall, we have divided the logging research into twelve subcategories (i.e., topics), which we will review throughout this survey. Table 6 (located at the end of the survey) lists the topics and provides the related references for each one. The table serves as a quick-access guide to review the research happening in each topic. Based on our analysis, the top-5 active and popular research topics in the field of logging based on the number of publications are: 1) log mining for anomaly detection, 2) log printing statement automation, 3) log mining for performance and failure diagnosis, 4) log maintenance and management, and 5) log parsing.

5.3 Year Trends

Figure 8 illustrates the number of conference (blue), journal (orange), and archived (gray) publications per year until October 31, 2021. The upward trend in the plot suggests a continuous and growing interest of the research community to explore various dimensions of logging research. The publications are from both academia and industry such as Microsoft [14], [30], [35], Twitter [15], Huawei [90], RIM [83], and others [40]. Additionally, there exists valuable research from the synergy between academic researchers and industry teams, which further emboldens the efforts by bringing real-world industry experiences [54], [90], [91]. As such, we foresee the research in this area will continue to grow and foster in the upcoming years as there are interesting and promising trends for future research, explained in Section 7.

5.4 Logging Challenges

Based on the knowledge gained throughout the survey, we summarize the challenges that the prior literature has aimed to tackle in the following. Although current research advances have made logs more useful and effective, there are still multiple remaining challenges and avenues for future work and improvement. Categories of challenges remain in various aspects of logging source code and log file analysis and mining.

5.4.1 Logging Code

As a result of a lack of well-accepted standards and guidelines for logging practices [29], [35], [40], currently, developers mostly rely on their personal experience or intuition to perform their logging decisions. However, for this manual process, i.e., developers inserting logging statements into the source code, to lead to effective logging practices, we are facing four main challenges:

1) The first challenge is where-to-log, which is the decision of selecting appropriate logging points. Logging statements can be placed in different locations of interest in the source code, such as inside try-catch block, function return value, etc. Although log statements provide valuable insight into the running system’s state, they are I/O intensive tasks and excessive logging can incur performance and maintenance overhead [33], [76], [91]. Consequently, developers are often faced with the challenge of making an informed decision for where-to-log in order to avoid introducing unjustified performance degradation and maintenance overhead.

2) The second challenge, what-to-log, concerns with what information to include in the log message. As explained in Figures 2 and 3 the log statement description provides a brief context of the execution and the internal variables provide more insights into the dynamic context of the current execution state. Therefore, the logging description and logged variables should satisfy their purposes and be clear and informative about the current state of the program. The logging description should also stay up-to-date and in-sync with the feature code updates, as some developers fail to update the logging statements as feature code changes [29], [35].

3) The third challenge is how-to-log, which concerns with how the logging code, as a subsystem, combines with...
Fig. 7: Percentage of publications in each topic.

The rest of the software system. As the logging code is intertwined across different source code modules, some prior researchers have suggested modularizing the logging code, as an independent subsystem, which becomes compiled into the feature code in the later stages of the system release [84], [85]. Nevertheless, many industrial and open-source software projects still tend to mix the logging code with the feature code [28], [29], [40], [96]. As a result, maintaining and developing high-quality logging code as the feature code evolves remains challenging and crucial to the overall quality of the software.

4) The fourth challenge, and also the recent [49] and mostly overlooked in prior research [30], discusses the question of whether-to-log, which concerns with dynamically adjusting the degree of logging in response to the runtime requirements. For example, if a suspected anomaly is detected, the logging platform can enable more detailed logging, and if the system is acting normal, it minimizes the number of logs to lessen the overhead.

Figure 9 summarizes the challenges regarding the logging code. In the following, we briefly review the related research concerning each challenge. Later on, we will revisit these research efforts in more detail in their related section in the survey.

Where to log. The research in this area is interested in finding appropriate logging points. One approach is...
to analyze the source code and look for specific types of code blocks, i.e., unlogged exception code blocks, and insert logging statements inside them [81]. Other log placement objectives, such as disambiguating execution paths [33], minimizing the I/O and performance overhead [91], and feature extraction and learning approaches [30], [35] also exist in the literature.

**What to log.** The research in this area is concerned with what the content of logging statements should be to make the logs more effective for future system observation purposes, such as debugging and failure diagnosis. Possible approaches include automatically adding variables that can clarify execution paths to the LPSs [82], [117]. Additionally, concerns on the content of logging statements arise when developers neglect to update the LPSs as the related feature code is updated [92], which is also a common problem in source code comments [166].

**How to log.** This avenue is concerned with the problem of how to develop and maintain a high-quality logging code. Additionally, it pays attention to the interaction and integration of the logging code as a subsystem with the rest of the software system. Paradigms such as Aspect Oriented Programming (AOP) [94] aim to look at log statements as a submodule of the software which is separate from the feature code, and unifies with the rest of the software at later time during the development process. However, the current logging practices in the industry and opensource projects commonly fuse the logging code inside the feature code. Logging libraries and utilities [48] also take part in organizing the logs and improving their formatting and quality. As such, enhancement of logging libraries can positively impact on ensuring how we log.

**Whether to log.** One approach to tackle the challenge of the number of logs is to enable dynamic filtering of logs during runtime [40], [71]. Enablement of this paradigm would allow to take the pressure off the developers and enable more conservative addition of log statements in the source code, without being concerned about the overhead. Thus, depending on the program state, logs are dynamically discarded or collected if they serve any online or postmortem analysis purpose.

In sum, the challenges associated with the logging code have initiated a variety of research to analyze and mine logging statements, their progression, and their issues. In addition, this has also triggered the development of tools and approaches to automatically predict and suggest logging statements and their associated details, such as log verbosity levels and variables, inside the source code.

### 5.4.2 Log File Analysis and Mining

The challenges involved in this category stem from the large scale of log files as the computer systems become larger and further distributed. In addition, because log files are unstructured (or semi-structured), due to the intermixing of both static (e.g., log description) and dynamic (e.g., variable values) runtime content, this further complicates the automated analysis of logs.

**Logging cost & benefit.** As the size of computer systems increases, the voluminousness and heterogeneity of logs, which turns it into a big-data problem, imposes additional challenges. Collecting, processing, and storing of logs become more challenging as logging can infer additional computation, storage, and network overhead. This calls for further quantitative [76] and qualitative [37] cost analysis of logging.

**Automated log analysis.** Due to the voluminousness and heterogeneity of generated logs, and in some cases, the need for real-time processing of logs, the development of efficient, scalable, and real-time log analysis tools becomes challenging [55]. These tools aim to achieve a variety of goals through logs such as log parsing, anomaly detection, runtime behavior monitoring, and failure detection. Moreover, other less-investigated software territories, such as mobile systems and big-data applications, face additional or sometimes different log analysis challenges stemming from their different use cases and operating conditions, such as performance and power limitations.

**Log maintenance and management.** As the scale of logs increases, their maintenance and management also become more challenging. These challenges revolve around efficient collection, organization, compression, and storage of logs. In addition, for various machine learning-based analyses of logs (e.g., anomaly detection), labeled log collections are hard to find. This challenge may be addressed by proposing efficient and scalable automated (or semi-automated) labeling techniques that can infer the labels for large-scale data from a small sample of labels data [110].

**Finding 2.** The top five research topics for logging research are: 1. log mining for anomaly detection, 2. log printing statement automation, 3. log mining for performance and failure diagnosis, 4. log maintenance and management, and 5. log parsing.

**Finding 3.** The top five publications venues for logging research are: 1. Empirical Software Engineering (EMSE), 2. International Conference on Software Engineering (ICSE), 3. Transactions on Software Engineering (TSE), 4. Automated Software Engineering Conference (ASE), and 5. Journal of Systems and Software (JSS).

**Finding 4.** Due to the challenges associated with logging code, i.e., what, where, whether, and how to log, and log file analysis and mining challenges, i.e., logging cost, automated log analysis, and log maintenance and management, there exists a continuously growing interest in log-related research and the prior work is published in top-ranked venues in yearly basis (Figure 6 and Table 5).

### 6 RQ3: HOW THE RESEARCH IN EACH TOPIC CAN BE SYSTEMATICALLY COMPARED WITH THEIR APPROACHES, PROS AND CONS?

In this section, we review the available literature in each category of logging research and provide a comparative analysis.

#### 6.1 Category A: Logging Cost and Benefit Analysis

Although logs are useful and provide insight into the internal state of the running software, they also impose inherent costs on different subsystems of a computer system. We can assess the costs and benefits of logging both quantitatively and qualitatively, which we review in the following.
As the logging BW increases, OS might slow down, and other processes that require disk space and BW may crash, and even the logging subsystem could crash. Additionally, more logging requires more storage space.

Survey of 66 developers and users [91] interviewed engineers in Microsoft and applies a constraint solving-based method to calculate an optimal logging placement, i.e., when compared with the original logging statements, logs, as they contributed to the costs and benefits associated with logging.

This study categorized seven generic patterns of error sites based on the study on 250 failures, such as exceptions, function return errors, etc., which, however, should have been ≤200KB/s.

TABLE 6: System’s performance overhead associated with logging.

| Reference - Approach | Results | Type | Pro | Con |
|----------------------|---------|------|-----|-----|
| Yuan et al. [81]     | This study categorized seven generic patterns of error sites based on the study on 250 failures, such as exceptions, function return errors, etc. | Quan. | Errorlogging provides three different levels of configurable logging overhead. | Focuses on Error Logging Statements (ELS). |
| Zeng et al. [161]    | The authors measured up to 5% performance overhead when the audit logging is enabled. | Quan. | Proposes an adaptive approach to reduce the overall system overhead from 5% to 1.5%. | Reduced auditing might result in a lower level of security protection for the system. |
| Ding et al. [91]     | Maximized extracted runtime information and, concurrently, minimizes the I/O and performance overhead. | Quan. | Two levels of filtering, i.e., local and global filters, to discard less-informative logging messages and simultaneously keep important messages. | Curtailed to performance analysis of logs, and fails short for logs recording error and failure information. |
| Li et al. [37]       | Main benefits of logging communicated by developers include: diagnosing runtime failures, using logs as a debugger, user/customer support and system comprehension. | Qual. | Survey of 66 developers and a case study of 223 logging-related issue reports from the issue tracking systems. | Limits to open-source projects and closed-source projects might evaluate differently on their logging costs and benefits. |
| Yao et al. [38]      | If the performance model’s prediction error is noticeably impacted, it implies that the modified log helps to model the CPU usage property. | Qual. | The approach finds and suggests removing insignificant log statements. | Logging statements that are not covered by the performance tests cannot be identified by this approach. |
| Gholamian and Ward [76] | It presents nine findings for logging cost in different scenarios with and without system failures. The research observes 8.01% and ~2680X overhead in the execution time and storage when the trace log level is enabled versus the info log level. | Quan. | The research measures the impact of different types of failures in a distributed environment on the generated logs and performs a case study for anomaly detection from OpenStack logs with entropy values. | The evaluation is performed on a small cluster of four nodes. The findings should be confirmed on a large-scale distributed cluster. |

TABLE 7: Logging cost and benefit research - Topic (A). ‘Type’ shows qualitative, quantitative, or both.

6.1.1 Quantitative assessment

Quantitative assessment for benefits of logging measures to what extent logging improves a specific debugging task. For example, Yuan et al. [51] observed the benefits of improved logs, as they contributed to ~60% faster diagnosis time when compared with the original logging statements, i.e., prior to the enhancement. Log associated overheads can also be evaluated quantitatively [91], [161]. Table 6 summarizes logging cost breakdown on various subsystems of a computing system [91], including I/O, storage, CPU, and memory. For example, one approach [33] to simplify and measure the slowdown caused by logging statements is to calculate the number of times (n) each log statement is being executed and multiply that with the overhead of a single log statement execution (l), i.e., n × l. Other research efforts in this area have measured the overhead of Linux security auditing through log messages by enabling and disabling audit logging [161], and statistically mapped logging statements to the performance of the system, i.e., CPU usage [98], [160]. Gholamian and Ward [76] evaluated the computation and storage cost associated with logging in a distributed environment. In addition, they associated the logging in different verbosity levels with the amount of information gained while synthesizing various failure scenarios.

6.1.2 Qualitative assessment

In contrast to quantitative metrics for measuring logging overhead, e.g., system slowdown or I/O cost, qualitative approaches aim to understand the underlying trade-offs from developers’ perspectives through surveys or questionnaires. A developer survey at Microsoft [91] uncovered main overheads associated with logging, from developers’ perspective, as listed as “Details” in Table 6. Developers were also inquired about the methods they use for containing the logging overhead for large-scale online service systems. They mentioned a variety of methods to limit the logging overhead such as adjusting the logging verbosity level (93% of developers have applied this approach), manual removal of unnecessary logs (64%), and periodic archiving of log files to save disk space (43%). Additionally, this study observed the lack of a cost-awareness guideline during
log instrumentation. Some developers often had little idea about the logging overhead when they planned to add new logging statements to the source code. Thus, developers require to be more mindful in adding logging statements in scenarios such as for-loops, which iterate a large number of times and could cause high overhead, especially on CPU, I/O, and storage throughput. A recent study [37], qualitatively examined logging cost and benefits from developers’ perspectives. One qualitative measure of the logging cost, i.e., too much logging, causes noisy log files which hinders program comprehension and results in strenuous log file analysis. In contrast to costs associated with logging, main qualitative benefits of logging communicated by developers include the capability to diagnose runtime failures with logs, using logs as a debugger, user/customer support, and system comprehension. Table 7 summarizes the research in Category A.

Finding 5. In sum, although logs provide insight into the internal state of the running software, they also impose inherent costs on different subsystems of a computer system. Developers should pay close attention to both quantitative and qualitative costs and benefits of logging while making logging decisions.

6.2 Mining Log Printing Statements
There has been a significant body of research aiming to mine, understand, and characterize various source code logging practices [28], [29], [35], [96]. Because, intuitively, understanding previously applied logging practices is the gateway to help developers improve their current logging habits. Thus, to derive the in-the-field LPS practices, the first step is to mine the source code’s logging statements and extract useful insight and observable patterns. Consequently, there are two broad classes of prior studies that have sought after understanding and mining of the logging practices in both industry and open-source projects: 1) mining logging code, and 2) mining log files. We review the research for mining logging code in this section and mining of log files in Section 6.3.

Mining log printing statements (LPSs) in the source code principally focuses on understanding how developers insert LPSs into the source code and how logging code evolves over time to gain insight into the common logging practices. Figure 10 shows the categorization of research for logging code mining of software projects which is divided to research on logging practices (Category B), logging code progression (Category C), and logging-code issues (Category D). In the tables that follow, our convention is that pro signifies an advantage, or an improvement, and con signifies a limitation, room for improvement, or an avenue for future work. This section includes research in Categories B, C, and D, that we review in detail.

6.2.1 Category B: Logging Practices
Mining logging practices aims to gain insight into the current logging habits of developers both in open-source and industrial proprietary software projects.

Open-source projects. Prior work in this category includes empirical studies to characterize current logging practices in open-source projects such as Apache Software Foundation (ASF) projects [28], [96]. Other works [96], [99] aimed to find recurrent mistakes in the logging code and its relationship to overall source code quality. Another tread of research [48], [168] has examined logging configurations, libraries, and utilities.

Industry projects. Similar to the open-source software, logging is a widely adopted practice in industry software projects. Fu et al. [35] conducted a study on logging practices of two software systems at Microsoft, and Pecchia et al. [40] examined application-critical software logging practices at Selex ES.

Table 8 provides a high-level comparison of primary studies concerning logging practices. In sum, prior research has considered logging practices in logging source code (LSC) and logging libraries and utilities (LLU), for different programming languages (Lang.) and for different number of projects (# Proj.) in both open-source (OS) and closed-source (CS) software. Table 8 provides additional details for each study and compares the research on mining of the logging source code for open-source and industrial projects.

Finding 6. In sum, logging is a pervasive convention in various software domains (e.g., server, client, and support applications) and developers utilize various logging practices and spend a significant amount of time updating logging statements.

6.2.2 Category C: Logging Code Progression
Thus far, we have discussed the research investigating logging practices in both open-source and industry projects. Prior research has also studied the progression (i.e., evolution) of the logging code in software projects. Progression means that how logging code changes over time. Prior studies have concluded that logging code evolves significantly
### Reference - Aim

| Reference | Category | Lang. | Source | # Proj. |
|-----------|----------|-------|--------|--------|
| Yuan et al. [28] | Study and characterizing logging practices in four open-source C/C++ based projects. | C/C++ | OS/CS | 3 |
| Fu et al. [35] | Conducts source code analysis on two software systems at Microsoft, to categorize logged and unlogged snippets. | C/C++ | OS/CS | 3 |
| Pecchia et al. [40] | Studies the logging practices on a critical industrial software at Selex ES. | C/C++ | OS/CS | 3 |
| Shang et al. [42] | Explores the relationship between logging characteristics and the code quality. | C/C++ | OS/CS | 3 |
| Chen and Jiang [46] | A replication work of Yuan et al.'s work on 21 Java projects. | C/C++ | OS/CS | 3 |
| Zhi et al. [168] | Conducts an exploratory study on the logging configuration practices and how they evolve over time. | C/C++ | OS/CS | 3 |
| Chen et al. [48] | Studies logging utilities (LUs) usage in Java projects. | C/C++ | OS/CS | 3 |

### Table 9: Logging practices research - Topic (B).

#### Finding 7.
In sum, prior studies have investigated the evolution of the logging code and libraries from different angles as the feature code evolves. These angles include

1. addition, deletion, and modification of logs
2. logging library migrations
3. impact of log context and owner on log changes
4. log revision analysis
5. intentions and concerns in log evolution

#### 6.2.3 Category D: Log-related Issues

The extensive usage of logs comes with mistakes, improper, and not well-thought logging practices, which results in logging issues and low-quality logging statements. Some of the research in this thread overlaps with logging practices and logging code progression, as some of the logging issues are uncovered during the examination of logging practices and their evolution. Yaun et al. [141] presented a characteristic study on real-world failures in distributed systems, and observed that the majority of failures print explicit failure-related log messages which can be used to replay (i.e., recreate) the failures. However, recorded log messages are noisy, which makes the analysis of logs tedious. Several efforts have aimed to identify and reduce log-related is-
TABLE 11: Logging code progression research - Topic (C).

Reference - Aim | Experiments | Results | Pro | Con
--- | --- | --- | --- | ---
Shang et al. [83], [169] | Two open-source (Hadoop and PostgreSQL) and one industrial (EA) software projects. | The logging code changes at a high rate across versions, which might break the functionality of log processing applications (LPA). | Developers could avoid the majority of the logging code modifications through better logging designs. | The broad range of the avoidable logging code changes raises the concern of if the observed values are software system dependent.
Kabina et al. [92] | Studies 223 ASF projects with their issue tracking systems in JIRA. | As more flexible logging libraries with additional features emerge, many ASF projects have undergone logging library migrations or upgrades. | Although adding more flexibility and performance improvement are cited as the primary drivers for logging library migrations, performance after library migration is rarely improved. | A questionnaire survey from developers involved in logging migration efforts can bring additional value and more insight into their rationale behind the logging updates and best practices.
Kabina et al. [84], [170] | Studies the logging library migrations in Apache Software Foundation (ASF) projects. | A significant portion of logging statements change throughout their lifetime, and factors such as file ownership can affect the stability of logging statements. | Developers of LPAs should rely on more stable logging statements for designing their tools. | The research considers only the first change of logging statements. However, the already changed logging statements might become more stable after going through modifications and prior fixes.
Li et al. [102], [171] | Studies the co-evolution process of logging statements as bug fixes and feature code updates are committed. | Proposes LogTracker, a tool that proactively predicts log revisions by correlating the rules learned from historical log revisions, e.g., the logging context, and the feature code. | Utilizes code clones to learn log revision rules with the insight that semantically similar codes will likely require similar logging revisions. For reasons such as lack of supporting facilities and the version evolution of source code, the developers’ I&C are poorly reflected in the log statements. | The tool can only guide log revisions for codes that share similar logging context, and the percentage of these revisions is not substantial.
Rong et al. [172] | Investigates the status of developers’ intention and concerns (I&C) on logging practices. | Developers’ interviews and code analysis on three industrial software projects. | Major gaps and inconsistencies exist between the developers’ I&C and real log statements in the source code. | Only missing log statements are considered as inconsistencies, and unnecessary log statements are not evaluated.

TABLE 12: Comparison of log-related issues research - Topic (D). We provide a link under ‘# Proj.’ if the data or implementation is available.

Reference - Category | Experiments | Lang. | Source | # Proj.
--- | --- | --- | --- | ---
Yuan et al. [141] | et noisy logs, Aspira- tor checker | C/Java | OS | 5
Shang et al. [83] | et JIRA tickets, log intentions | Java | OS | 3
Chen and Jiang [29] | et anti-patterns, LC-Analyzer | Java | OS | 3
Hassani et al. [101] | et inappropriate and missing log statements | Java | OS | 2
Li et al. [114], [173] | et repetitive logging statements, DLFinder | Java | OS | 5
Gujral et al. [74] | et logging questions semantics | Q&A websites | OS | 6 [174]
Chen et al. [75] | et challenges and benefits of analyzing user logs | bug reports | OS | 10 [175]

TABLE 13: Comparison of log-related issues research - Topic (D). We provide a link under ‘# Proj.’ if the data or implementation is available.

Finding 8. In sum, prior studies have investigated the log-related issues form different perspectives with providing some automated tools to detect log issues: 1) noisy logs interfere with failure diagnosis, 2) development knowledge (JIRA tickets) can help with log intention discovery, and 3) anti-patterns, duplicates, and missing LPSS are among log-related issues.
6.3 Category E: Log Statement Automation

As mentioned earlier, execution logs, which are the output of logging statements in the source code, are a valuable source of information for system analysis and software debugging. Thus, high-quality logging statements are the precursor of effective log file mining and analysis. Conversely, low-quality LPSs result in log-related issues (Section 6.2.3) and they hinder the understanding of software problems whenever they happen. Currently, due to the ad-hoc nature of logging, lack of general guidelines, and because developers mostly insert logging statements based on their personal experiences, the quality of log statements can hardly be guaranteed [163]. Therefore, automated logging which aims to add or enhance log statements inside the source code either proactively or interactively is a well-motivated effort and can improve the quality of logging statements and, ultimately, result in more effective log mining tasks.

Fig. 11: Logging research with the emphasis on logging code automation research, Category E.

Figure 11 presents the log statements’ research with emphasis on automated logging. As per this figure, we review the theory behind automated logging, automated logging approaches, and how they are evaluated. Prior studies have suggested creating and utilizing statistical models from common logging practices, and learning logging heuristics
from experience, and using them to provide new logging suggestions or enhance the already existing LPSs.

6.3.1 Log Automation - Motivation and Theory

One of the common approaches for log automation is the application of machine learning methods to predict whether a code snippet needs a logging statement by training a model on a set of logged code snippets, and testing it on a new unlogged code set, i.e., supervised learning. In this section, we first review the background and theory for machine learning methods and continue with automated logging approaches.

**Motivation.** With the ever-increasing size of software systems, it is most likely that a single developer is in charge of developing only a small subsystem of the whole software system. Under this situation, making wise logging decisions becomes quite challenging as developers do not have full knowledge of the whole system [30]. As logs are quite pervasive and useful for system maintenance [178], if the logging decisions can be learned automatically, a log suggestion tool can be constructed to help developers make better decisions. Ultimately, such a tool can increase the quality of logs and save developers time.

**Learning to Log.** The idea of learning to log is to construct a machine learning (ML), or deep learning (DL), model that can learn common logging practices and provide logging suggestions to the developers or directly make logging decisions and insert logging statements into a newly-developed source code snippet. A typical learning to log tool [30] is outlined in Figure 12. The learning steps are: 1) code collection from repositories, 2) labeling the collected source code, 3) feature extraction and selection, 4) feature vectors and model training, and finally 5) logging enhancement, suggestion or automatic insertion. Based on this platform, once the training phase is completed, during the testing phase, the learning model decides whether a new code snippet requires a logging statement by extracting its features and feeding it to the ML model, and observing the model’s output. Learning algorithms apply a wide range of techniques such as: pattern or rule-based [1], [33], [81], [82], machine learning such as Naïve Bayes, Bayes Net, Logistic Regression, SVM, and Decision Trees [30], Random Forrest [102], [112], Ordinal Regression [103], and most recently, Deep Learning [36], [109], [117], [120].

**6.3.3 Automated Logging Categories**

Figure 14 highlights the research in this area categorized into three subtopics: log enhancement, log suggestion, and log insertion. These approaches are primarily concerned with log location prediction, i.e., where to log, and secondarily the content to include in the logging statements, i.e., what to log. Log enhancement aims to improve the quality of existing logging statements, such as adding more runtime context [82]. Log suggestion aims to provide suggestions for logging locations that might have been missed, and log insertion aims to proactively insert logging statements into the source code [33]. The approaches can also be categorized based on the targeted LPSs. Some approaches focus on error logging statements (ELS), such as logging when an error happens inside a catch clause [81], [126], and others focus on normal logging statements (NLS), e.g., such as method-level logging [35]. Although the goals of these approaches are similar, the intended way of their implementation can be different. For example, one practical scenario of implementing log suggestion approaches is as IDE plugins that provide just-in-time suggestions [35]. However, the log insertion techniques are implemented as post-processing tools that scan the source code and insert logs for various criteria of interest, such as disambiguating execution paths [33] or logging catch clauses [81]. This categorization is not
mutually exclusive and some of the prior work overlap in their approaches.

**Log Verbosity Level (LVL) and Description Predictions (LSD).** In addition to predicting log location, prior research has also investigated approaches for prediction of the appropriate logging `verbosity level` for newly composed logging statements [49, 101, 103, 109, 159]. The approaches either apply some type of learning to predict the verbosity level [109], or perform dynamic adjustment of the log level during the runtime [49]. Other research also aims to predict the description [122], or variables included in the logging statements [82, 117]. In Table 14, we provide a high-level comparison of automated logging research, and Table 17 provides a more detailed comparison of each research.

![Confusion Matrix](image)

**TABLE 15: Confusion matrix for log prediction.**

### 6.3.4 Evaluation Metrics

After training the learning model, its performance should be evaluated during the testing phase by applying new code instances as input and finding out the prediction outcome that whether or not this new code snippet requires a logging statement, e.g., Figure 12. This is an example of a binary classification problem [184]. Differently, log verbosity level prediction is evaluated by first extracting the size of n-grams that comes in the general formula for BLEU [182]. Similarly, for ROUGE, refer to [183].

![Evaluation Metrics](image)

**TABLE 16: Evaluation metrics for automated log prediction.**
| Reference | Aim | Experiments | Results | Pro | Con |
|-----------|-----|-------------|---------|-----|-----|
| Zhang et al. [163] | Proposes AutoLog, which generates additional informative logs to help developers discover the root cause of a software failure. | Performs a proof-of-concept case study on Apache Hadoop Common. | AutoLog embeds a two-stage process of log slicing and log refinement to the program to narrow down the execution paths that could have led to the system’s failure. AutoLog is targeted for interactive in-house development. | The approach narrows down the execution paths that could have led to a system’s failure. AutoLog is targeted for interactive in-house development. | The program needs to be re-executed every time new log statements are added, which is time-consuming. |
| Yuan et al. [42] | Proposes a tool, LogEnhancer, to find and add useful variables to log statements. | Evaluated on a total of 9,125 log messages from eight different domains, including apache httpd, postgresql, and etc. | LogEnhancer is effective in automatically adding a high percentage of log variables, on average, 95.1%, that programmers manually included. | The tool performs static analysis on the source code starting from the log state- ment and navigates backward to find variables that are causally along the path that results in the execution of the log statement. | As LogEnhancer’s improvement is limited to the existing log statements, its effectiveness diminishes if the logging statements are missing. |
| Zhu et al. [30] | Proposes LogAdvisor, which aims to provide logging suggestions for exception and return-value-check code blocks. | Two industrial software systems from Microsoft and two open-source software systems from GitHub (SharpDevelop and MonoDevelop). | LogAdvisor achieves a high balanced accuracy, ranging from 84.6% to 93.4%, to match developers’ logging decisions, and the decision tree model achieves the highest scores. | Trains a machine learning model (e.g., SVM and decision trees) to predict whether a focused code snippet requires a logging statement. | It is focused and limited to two categories of code snippets: 1) exception snippets and 2) return-value-check. |
| Lal et al. [165] | Introduces LogOptPlus, a tool for automated catch and if code block logging prediction. | Two open-source projects: Apache Tomcat and CloudStack. | The prediction model with random forest achieves the highest F1-score 80.70% (Tomcat) and 92.25% (CloudStack) for if-block logging prediction. | Applies five different learning techniques, e.g., AdaBoost, Gaussian Naive Bayesian, and Random Forests achieve the highest Precision and Recall. | It is limited to specific code blocks, i.e., if-block and catch clause. |
| Zhao et al. [33] | Introduces Log20, a tool that finds a placement of logging statements to minimize execution path ambiguity. | Evaluated on four open-source Java projects: HDFS, HBase, Cassandra, and ZooKeeper. | Log20 achieves a lower logging overhead with the same level of informativeness (i.e., entropy) compared to existing logging statements by developers. | It applies Shannon’s information theory equation to measure the entropy of the program by approximately considering all of the possible execution paths. | The approach does not consider developers’ concerns and practices, does not explain the static content of LPSs, and change of workload can cause extra logging overhead. |
| Li et al. [102] | Analyzes log changes in open-source projects and proposes commit-time logging suggestions. | Two open-source projects: Hadoop, Directory Server, Commons HttpClient, and Qpid. | Performs a manual analysis on four software systems and categorizes the changes to logging statements into four major groups: 1) block change; 2) log improvement, 3) dependence-driven change, and 4) logging issues. | Proposes a random forest (RF) classifier for each code commit to suggest whether a log change is required. The classifier’s balanced accuracy for within-project suggestions is 0.76 to 0.82. | As the model is trained on prior log changes, it might miss scenarios that there are no prior logging changes to learn from. |
| Li et al. [103] | Determines the appropriate log verbosity level for the newly-developed logging statement. | Analyzes four open source projects: Hadoop, Directory Server, Hama, and Qpid. | Collected five categories of quantitative metrics that play important roles in determining the appropriate log level: logging statements metrics, containing block metrics, file metrics, change metrics, and historical metrics. Achieves AUC in the range of 0.75 to 0.81 for log level prediction. | Metrics from the block which contains the logging statement, i.e., the surrounding block of a logging statement, play the most important role in the ordinal regression models for log levels. | The results show that the ordinal regression models for log level prediction are project-dependent. |
| Jia et al. [126] | Proposes SmartLog, which is an intention-aware error logging statement (ELS) suggestion tool with two intention models: IDM and GIDM. | Experiments on six open-source projects in C/C++: Httpd, Subversion, MySQL, PostgreSQL, GIMP, and Wireshark. | It improves on the Recall values (average of 0.61) and achieves higher scores compared to LogAdvisor [30] (average of 0.45) and Errol [81] (average of 0.18). | This work improves on prior work by going beyond code patterns and syntax features, and considers source code intentions, i.e., semantics. | This work is limited to ELS prediction, i.e., exception and function return-value logging. |
| Ana et al. [159] | Proposes a method to make the logging level decisions by understanding the logging intentions. | Four open-source software projects: Hadoop, Tomcat, Qpid, and ApacheDS. | It reaches AUC values higher than 0.9 in log level prediction. The approach extracts the contextual features from logging code snippets and leverages a machine learning model (i.e., a random forest model) to automatically predict the verbosity level of logging statements. | As a proof of concept, the authors also implement a prototype tool, VerbosityLevelDirector, to provide guidance on log verbosity level selection in focused code blocks. | The work is limited to focused code blocks: exception handling blocks and condition check blocks. |
| Liu et al. [117] | Presents an approach to recommend the variables to include in logging statements. | Evaluates on nine open-source Java projects: ActiveMQ, Camel, Cassandra, CloudStack, DirectoryServer, Hadoop, HBase, Hive and Zookeeper. | The approach first learns “rules” from existing logged code snippets by extracting contextual features with deep learning recurrent neural networks (RNN). The approach outperforms five baselines, including random guessing and IR methods in log variable prediction. | The tool provides a ranked list of variables that probably are required logging to the developer. | The method only considers the code preceding the logging statement. As such, extending this approach to include the code succeeding the logging statement can improve on logging variable recommendation. |

TABLE 17: Log printing statement automation research - Topic (E).
statements as *ground truth*. To create a set of training and testing data for the machine learning process and have a proper *ground truth* to compare with, one approach is to collect all of the code snippets with logging statements, and some samples of unlogged code, to include both positive and negative cases. Then, after deciding the train-test split and training the ML model, prior work removes the log statements from the test data. During the testing phase, the model’s performance is evaluated on the test code snippets with their logging statements being removed. This way we measure how well the model can remember which code snippets should have and which ones should not have logging statements, compared to the developers’ originally-inserted LPSs. Multiple iterations of the training-testing can be applied, e.g., cross-validation [187], to confirm the results. From Table 15, TP means that the model correctly predicted a code snippet that requires a logging statement, and FN denotes that the model incorrectly predicted that a code snippet does not require a logging statement.

Based on the confusion matrix, we can define some of the common metrics for evaluating the performance of log learning models in Table 16. The definitions for *Precision* and *Recall* are straightforward. In order to ensure a prediction model benefits from equally good or comparable Precision and Recall values, *F-Measure* is defined as the harmonic mean of Precision and Recall. Qualitatively, good performance of *F-Measure* implies good performance on both Precision and Recall. *Accuracy* represents correctly identified logged instances to the total number of cases. *Balanced Accuracy (BA)* is the average of the proportion of logged instances and the proportion of unlogged instances that are correctly classified. In case there is an imbalance in the data, e.g., in Table 15 if TN is much larger than TP, *Balanced Accuracy (BA)* is widely used to evaluate the modeling results [30, 102, 188], because it avoids the over-optimism that *accuracy* might experience. Receiver Operating Characteristic (ROC) plots *true positive rate* against *false positive rate*. AUC (area under the curve) is the area under the ROC curve. Intuitively, the AUC evaluates how well a learning method can distinguish logged code snippets and unlogged ones. The AUC ranges between 0 and 1. A high value for the AUC indicates a high discriminative ability of the learning model; an AUC of 0.5 indicates a performance that is no better than random guessing [102]. BLEU [182] and ROUGE [183] scores are equivalent to Precision and Recall and are leveraged to evaluate the auto-generated text compared to the original text developed by developers. These scores have applications in evaluating the auto-generated log statement descriptions (LSDs), which are sequences of tokens, i.e., words. For example, for a candidate LSD, C and the reference LSD, R, BLEU measures the ratio of tokens of C that also appear in R (analogous to Precision), and ROUGE measures the ratio of tokens of R that have appeared in C (analogous to Recall). The range of values for BLEU and ROUGE is [0,1], with 1 being the perfect score. These two measures combined explain the quality of the auto-generated LSDs. *Edit distance* can be also used for checking the distance between auto-generated text and the developer inserted LSD.

**Examples of Metrics Used.** Li et al. [173] utilized *Precision* and *Recall* to calculate the performance of DLFinder in detecting logging code smells. Zhu et al. [30] used *BA* to evaluate the accuracy of LogAdvisor, which advises the developer if logging statements are required for a focused code snippet. Li et al. [27, 102, 103] used ROC and AUC to evaluate their methods in *log verbosity level prediction*.

| Reference - Aim | Experiments | Results | Pro | Con |
|-----------------|-------------|---------|-----|-----|
| Kim et al. [142] | 22 open-source projects from three different domains: message queuing, big data, and web application server. | Applies semantic and syntactic features and recommends a new log level in case the current level is deemed inappropriate. It reaches 77% precision and 75% recall in log level validation. | Creates domain word model from all of the log messages in application domains, which enables knowledge sharing between different projects. | In some cases, the appropriateness of log levels is dependent on developers’ opinions and is quite arguable. |
| Gholamian and Ward [56] | Three open-source Java projects: Tomcat, Hadoop, and Hire. | Performs an experimental study of logging characteristics of source code clones and observes that code clones match in their logging behavior. Achieves 90% accuracy in log location prediction. | It applies source code features and machine learning methods to detect log-aware code clones for log statement prediction. | The approach can only suggest logs for code snippets that can find their clone pairs in the software code base. |
| Li et al. [120] | Seven open-source systems: Cassandra, ElasticSearch, Flink, HBase, Kafka, Wicket, and ZooKeeper. | The authors discover six categories of logging locations in different types of code blocks from developers’ logging practices. It achieves balanced accuracy of 80.1% using syntactic source code features. | Utilizes a pipeline of word embedding, RNN layer, and a dropout layer in its deep learning model for log location prediction. | The approach can be applied, e.g., cross-validation for BLEU and ROUGE is 0.1, with 1 being the perfect score. These two measures combined explain the quality of the auto-generated LSDs. Edit distance can be also used for checking the distance between auto-generated text and the developer inserted LSD. |
| Cândido et al. [112] | An enterprise software, Adyen, and 29 Apache projects. | The authors extract source code metrics from methods and evaluate the performance of five different learning approaches on log suggestions. The best performing model achieves 72% of balanced accuracy on Adyen’s log statements set. | Performs a study on 29 Java projects and leveraged learning transfer to generalize to an industry project. | The approach can only suggest logs for code snippets that can find their clone pairs in the software code base. |
| Li et al. [109] | Nine large-scale open-source projects: Cassandra, ElasticSearch, Flink, HBase, HBase, HBase, HBase, HBase, HBase, and ZooKeeper. | The authors initially perform a manual study and categorize five different logging locations. The model trained with syntactic features achieves an average AUC of 80.8%. | Their findings infer that the log levels that fall far apart on the verbosity scale manifest different characteristics. | Log levels that are closer in order, e.g., warn and error are more difficult to distinguish with this approach. |

**TABLE 17: Log printing statement automation research - Topic (E) (continued)**
and logging commit change suggestion. Kim et al. [142] used F-Measure to evaluate their log verbosity level recommendation approach, and Gholamian and Ward [36] utilized Accuracy to evaluate the performance of their log-aware clone detection approach. He et al. [122] leveraged BLEU and ROUGE scores to evaluate the effectiveness of the candidate log statement descriptions when compared to the developer-inserted log descriptions. Edit distance [189] also has applications in finding similar code snippets for enabling logging suggestions [122].

Finding 9. In sum, prior studies have investigated different learning paradigms, such as clustering, random forest, deep learning, and transfer learning. The approaches aim to predict the location (LSL), verbosity level (LVL), variables (VAR), and description (LSD) of the logging statements.

6.4 Mining Log Files

Priorly, we mentioned the purpose of logging statements added by developers is to expose valuable runtime information. The output of logging statements is written to log files, which are used by a plethora of log processing tools to assist developers and practitioners in different tasks such as software debugging and testing [3], [190], performance monitoring [160], and postmortem failure detection and diagnosis [6], [7], [8], [152]. We review log mining techniques and approaches in the following.

6.4.1 Category F: Log Maintenance and Management

As the size of logs increases, the job of methods and tools which manage and maintain logs becomes more crucial, cumbersome, and of value. For example, FLAP [152] provides an end-to-end platform for log collection, maintenance, and analysis. In the following, we review log collections, log compression, log rolling, and log removal, and Table 19 summarizes the research in this category.

Log Collections. The aim of maintaining a log collection is for auditing [191] or enabling benchmarking for different types of log analysis [124], [125], [157]. For example, Loghub [90] provides a repository of logs from various software systems, and Cotroneo et al. [192] have released an OpenStack failure dataset containing injected faults. The logs are used in various prior works for evaluating tasks such as compression techniques [104], [193], failure analysis [6], [132], [137] and bug detection [194]. We observe that although execution logs of different systems are conveniently available, it is difficult to find large-scale collections of labeled datasets, which are especially crucial for supervised learning of log mining tasks [2]. This is because manually labeling large datasets of logs is quite cumbersome. Thus, we see significant value in curating a database of labeled logs (e.g., normal vs. anomaly/failure log records), and the development of automated log labeling techniques [110].

Log Compression. With the continued growth of large-scale software systems, they tend to generate larger volumes of log data every single day, which makes the analysis of logs challenging. As such, to cope and contain this challenge, developers and practitioners apply tools for compression and continuous archiving of logs [51], [104]. Hassan et al. [195] applied log compression to extract common usage scenarios. Yao et al. [104] studied the performance of general compressors on compressing log data relative to their performance on compressing natural language data. Their work reviews 12 widely used general compressors to compress nine log files collected from various software systems. Because log files generally benefit from higher repetition than natural text [104], there is an avenue of outstanding work to develop log-aware compression techniques, that consider log characteristics in their algorithms and their parameter selections and achieve a higher compression/decompression performance.

Log Rolling. As log data generally grows rapidly during the system’s execution [104], [202], logging libraries such as Logback [203], Log4j/2 [31], and SELF4J [47] often support the continuous rolling i.e., archiving of log files as new logs become available. For example, as the size of the generated log goes beyond a user-defined value on the storage or a specific time interval has passed (e.g., daily, weekly), a Log4j Appender [39] can zip, rename, and store the log with a timestamp, e.g., “logs/app-%dMM-dd-yyyy.log.gz”. Log archiving helps with the long-term maintenance and organization of logs. There is certainly room for further research on improving and automating log archiving policies and techniques.

Log Removal. Although logs are useful, but due to the large volume of them, they can become noisy, hard to analyze, and cause inaccuracy in log analysis. As such, prior research has aimed to detect and remove duplicate logs.
| Reference | Aim | Experiments | Results | Pro | Con |
|-----------|-----|-------------|---------|-----|-----|
| Li et al. | Proposes a data-driven management framework by knowledge acquisition from historical log data. | Log files collected from several Windows machines in a university network. | Performs experiments on categorizing dependent and independent log events, and applies text mining techniques to categorize log messages, mines temporal data, and performs event summarization. | Provides a graphical representation of temporal relationships among events as an event relationship network (ERN) \[201\]. | Common categories of log messages are manually determined, which can be automated from historical data. |
| Marty | Proposes a proactive logging guideline to support forensic analysis in cloud environments. | N/A (the research does not provide experimentation). | Discusses the challenges of logging in the cloud environments such as decentralization and volatility of logs. | Outlines the guidelines for when the logging is required: business relevant, operational, security, compliance. | The guideline can be expanded to include forensic timeline analysis of logs, log review, and log correlation. |
| Li et al. | Introduces FLAP, a web-based integrated system to utilize data mining techniques for log analysis and knowledge discovery. | Network X event logs at Huawei Technologies. | Performs a case study and the results show the approach's applicability for different tasks, such as event summarization (graph) and root cause analysis. | It provides learning-based log event extraction and provides event summarization and visualization. | Some of the tasks, e.g., root cause mining, rely on domain knowledge to manually diagnose possible problems. |
| Amar et al. | Investigates the usage of finite-state models for log differencing. | Mutated logs from FSA models and a user study of 60 participants. | The proposed approaches \(i.e., 2KDiffs nKDiff\) can expedite the process of identifying behavior differences between logs. | The proposed models present sound, complete, and concise comparisons for log differencing, and presents two algorithm: 2KDiffs that compares two and nKDiff that compares multiple logs at once. | The presented work is limited to identifying k-differences between a set of comparing logs and does not consider temporal invariants. |
| Bao et al. | Proposes a statistical log differencing approach, which calculates the frequencies of behaviors found in the logs. | Controlled user study with 20 participants and log traces generated from 13 publicly available FSA models. | The proposed approaches achieve the required guarantees of the defined statistical test with an acceptable overhead. | It is a follow-up work to \[71\] and improves by enabling developers to control the sensitivity of log differencing by setting the statistical significance value \(s2KDiff\) and \(nKDiff\). | The approach is limited to highly structured logs, which might not be the case in practice. Thus, helper tools are required for preprocessing and structuring of logs. |
| Liu et al. | Proposes a new log compression method, logzip, to allow for more effective log compression. | Five log datasets: HDFS, Spark, Windows, Android, and Thunderbird. | Achieves higher log compression ratios compared to general-purpose compressors, \(i.e.,\) bzip2, and can generate compressed files around half of the size of general-purpose compressors. | Performs iterative clustering with template extraction and parameter mapping and can compress in three incremental levels: L1: field extraction, L2: template extraction, and L3: parameter mapping. | The performance of the decomposition should be also evaluated and compared with other compressors. |
| Yao et al. | Studies the performance of general purpose log compressors on compressing log data relative to their performance on compressing natural language data. | Nine system logs, such as HDFS and Linux Syslogs, and two natural language (NL) data, Wiki and Gutenberg. | Reviews twelve widely used general compressors to compress nine log files collected from various software systems. The observation is that log data is more repetitive than natural language, and log data can be compressed and decompressed faster than NL with higher compression ratios. | The performance of Loghub datasets have been widely utilized for research both in academia and industry. | The findings and implications of this research have not been utilized to propose a log-aware compressor. |
| Shin et al. | Introduces LogCleaner, which performs periodicity and dependency analyses for removing repetitive logs. | Two proprietary and eleven publicly available log datasets including: CVS, RapidMiner. | The approach can accurately detect and remove 98% of the operational messages and preserve 81% of the transactional log messages, and reduces the execution time of the model inference task from logs. | Segregates and only keeps transactional messages, which record the functional behavior of the system from operational messages of the system. | The performance of LogCleaner is heavily dependent on the quality of upstream log parser and template extraction, and requires manual analysis and domain knowledge. |
| He et al. | Provides a repository, Loghub, of logs from various software systems. | Provides 17 log datasets from various application domains, including distributed systems, supercomputers, and operating systems. | Provides a framework for AI-powered log analysis and applies a practical usage scenario of Loghub for anomaly detection for supervised and unsupervised approaches. | Loghub datasets have been widely utilized for research both in academia and industry. | There is still a shortage of labeled datasets to facilitate the evaluation of supervised log analysis tasks. |
| Chen and Jiang | Performs a survey on log instrumentation techniques. | N/A (the research does not provide experimentation). | Focuses on the three log instrumentation steps: logging approaches, logging utility integration, and logging code composition. | Defines four categories of challenges for instrumentation: usability, diagnosability, logging code quality, and security compliance. | The research can be improved by providing a connection between logging source code and its corresponding log messages in the log files. |
| Locke et al. | Proposes LogAssist to assist practitioners with organizing and summarizing logs. | Logs from one enterprise (IS) and two open source systems: HDFS and ZooKeeper. | Groups logs into event sequences to extract workflows and illustrate the system's runtime execution paths. LogAssist shrinks the logs by 75.2% to 93.9% and the unique workflow types by 70.2% to 89.8 in HDFS and Zookeeper logs. | LogAssist is able to reduce the number of log events of interest in a large quantity of logs by 75.2% to 93.9%, thus it saves time and improves the practitioners' experience in log analysis. | In some cases, the searched keywords \(e.g.,\) "error" or "exception" for problematic log lines result in a large quantity of logs for practitioners to manually review. |

**TABLE 19: Log maintenance and management research - Topic (F).**
6.4.2 Automated Log File Analysis - Challenges and Motivation

Logs record system runtime information and are widely used and examined for assessing the systems’ health and availability [178]. Traditionally, developers or operators often inspect the logs manually with keyword searches (e.g., “fail”, “exception”) and rule matching (e.g., “grep <RegEx>”) to find any potential problems in case of a system failure. However, manual or keyword inspection of log files becomes impractical with the ever-increasing volume of logs, making it unmanageable for a single developer to pin down an issue from concurrent and massive log files.

Parallel and distributed software systems generally apply various methods of fault-tolerant and performance optimization techniques in order to recover from a hardware failure or perform load-balancing and scheduling. For example, a resource manager daemon, e.g., on a Hadoop or Spark cluster, may intentionally terminate a running application and move it to another node in the system in order to expedite the execution of that task. As a result, the traditional and manual way of searching in the log files for keywords such as killed, terminated, failure might not be useful and can lead to multiple false-positive cases [207] and further muddle the manual inspection.

Moreover, although automatic log analysis helps developers and practitioners significantly to speed up the process (e.g., [9], [132], [138]), the automatic log analysis itself is still very challenging because log messages are usually unstructured free-form text strings, and application behaviors are often very heterogeneous and complicated [8]. As a result, effective automated log analysis methods are well sought after. To enable automatic log analysis, the very first step is log parsing, Category G, followed by applications of automated log analysis, i.e., anomaly detection, Category H, runtime behavior, Category I, and performance, failure, and fault diagnosis, Category J, that we review in the following.

### Table 19: Log maintenance and management research - Topic (F) (continued).

| Reference - Aim | Experiments | Results | Pro | Con |
|-----------------|-------------|---------|-----|-----|
| Yao et al. [72] | Proposes LogBlock a pre-compression approach to preprocess small log blocks to achieve a higher log compression ratio. | Evaluation on logs of 16 systems in different domains, such as ProxyServer, Android, BGL, and HDFS. | LogBlock’s preprocessing approach improves the relative compression ratios by a median of 5% to 21%, while it also achieves a faster compression time. | LogBlock results in higher compression ratios for logs by reducing the log repetitiveness through log header preprocessing and rearranging the log content. |
| Wang et al. [73] | Proposes DLOG, an approach to assist monitoring of big data applications. | 1000 randomly sampled spark-related questions on StackOverflow, six Spark benchmarks, and a user study of 20 developers. | DLOG introduces less than 10% overhead while reducing the big data application debugging time by 63%. | The approach provides the intermediate information of big data benchmarks, which allows the developers to better identify an cause for the chained methods and lazy evaluations. |
|               |             |         |     |     | Although DLOG performs well for small log blocks, other compression approaches, e.g., Logzip and gzip, outperform LogBlock for larger block sizes. |     |

#### Finding 10.
In sum, log collections, log compression, log rolling, and log removal are the pillars of log maintenance and management. Prior studies have investigated the approaches to compress, clean, and summarize logs to help developers and practitioners achieve more efficient interactions with logs.
Fig. 15: Log parsing for a raw log message to a parsed log from HDFS logs [2].

TABLE 20: Comparison of log parsing research - Topic (G). Systems: number of systems evaluates and some examples; Org.: Organization or link if available.

| Reference | Tool | Approach | Systems | Org. |
|-----------|------|----------|---------|------|
| Qu et al. [153] | Syslog Digest | word frequency | Syslogs | AT&T |
| Taerat et al. [149] | Baler | token attributes | 4 (BGL, Liberty) | N/A |
| Tang et al. [151] | LogSig | term pair and k partition | 5 (FileZilla, Thunder-Bird) | N/A |
| Vaarandi et al. [13] | LogCluste | associate rule mining and clustering | 6 OS +1 CS | 208 |
| Du et al. [108] | Spell | longest common subsequence | 2 (HPC, BGL) | 209 |
| He et al. [157] | Drain | fixed-depth parse tree | 5 (BGL, HDFS) | 210 |
| Messoudi et al. [138] | MoLI | search-based evolutionary approach | 6 (HDFS, Proxifier) | 211 |
| Dia [116] | Logram | n-grams dictionaries | 16 (Andriod, Apache) | N/A |

Finding 11. In sum, prior studies have applied various heuristics and approaches for automated parsing of numerous systems logs. These approaches include: 1 associate rule mining, 2 token frequency and k partitioning, 3 parse trees, 4 search-based evolutionary, and 5 n-gram dictionaries. The higher performance of log parsing, as the precursor step, can enable more effective downstream log mining tasks.

6.4.4 Category H: Anomaly Detection

Execution logs are extensively leveraged to monitor the health of software systems, identify abnormal situations, and detect anomalies that can lead to system failures. As such, the goal of anomaly detection from logs is to find cues in the log records that are tied to the identification of abnormal system behavior. If anomaly detection is used in online fashion, early detection of abnormal log lines could result in stopping anomalies before they turn into a partial or complete failure, e.g., slower system or shutdown, respectively. Role in incident management of large-scale systems. Table 22 provides an overview of anomaly detection from logs, and Table 26 provides a detailed comparison and pros and cons of various research for anomaly detection with logs. As a reference for further reading, He et al. [178] performed a quantitative comparison of various log-based anomaly detection approaches.

Finding 12. In sum, anomaly detection methods include various approaches, such as: 1 creating a state machine of normal execution and comparing the failure runs with normal models [3], 2 PCA-based approach which projects event logs to normal and abnormal subspaces [2], 3 deep learning approaches which learn an LSTM model from normal execution workflows [5], and unstable logs [138], 4 semi-supervised deep learning approaches with probabilistic label estimation [110], 5 a statistical approach using probabilistic suffix trees [123], 6 cloud deployment by correlating logs and resource metrics [124], and transformer-based approaches [77].

6.4.5 Category I: System’s Runtime Behavior

Researchers have also utilized logs for monitoring the system’s runtime behavior. This category of research aims to...
| Reference | Experiments | Results | Pro | Con |
|----------|-------------|---------|-----|-----|
| Qiu et al. [153] | Designs SyslogDigest that extracts log events from the router’s syslogs. | Syslog data from two large operational networks: a tier-1 ISP backbone network and a nationwide commercial backbone network. It combines an offline and online domain knowledge learning component automatically extracts relevant domain knowledge from raw syslog data. The authors showcase the applications in network troubleshooting, and health monitoring and visualisation. | Applies associated rule mining, then transforms and compresses raw syslog messages into their prioritized high-level events. The work is limited only to event template extraction from a specific system, i.e., syslog messages. | |
| Taerat et al. [149] | Introduces Baler, a token-based log parsing tool. | Logs of four supercomputers: BGL, Liberty, Spirit, and Thord. For clustering, Baler relies on token attributes rather than frequency or entropy of token positions that are applied in other log parsers. Baler handles large datasets better than compared tools and more efficiently, i.e., faster execution time. Requires only one pass to cluster log messages based on their event templates. Baler relies on the user to provide a dictionary of words. | |
| Tang et al. [151] | Proposes LogSig, a message signature based algorithm to generate events from textual log messages. | Logs of five real-world systems, including FileZilla, PVFS2, and Hadoop. It searches for the most frequent message signatures and then categorizes them into a set of event types. LogSig performs better in the quality of the generated log events (F-Measure) and scalability when compared to prior work. LogSig converts each log line into a set of ordered token pairs and then partitions log messages into k different groups based on the extracted term pairs. LogSig has a prolonged execution time on large log datasets and reaches low accuracy on the BGL data [272]. | |
| Vaarandi et al. [15] | Presents LogCluster, a data clustering and pattern mining algorithm for textual log lines. | A set of six system logs from a large national institution, including database systems, mail servers, and firewall logs. LogCluster improves SLCT [44] such that each Cluster Ci is uniquely identified by its pattern Pi, and each pattern consists of words and wildcard, which makes it insensitive to word shifts. LogCluster performs more accurate clustering and finds fewer groups compared to SLCT. The support of a cluster is calculated as the number of elements in that cluster: \( sup(P_i) = sup(C_i) = |C_i| \). Finally, clusters with support of equal or higher than a threshold value, s, are selected. The LCS approach achieves a faster template searching process by enabling subsequent matching and prefix trees. The algorithm requires a two-pass process to categorize the list of frequent patterns. | |
| Du et al. [108] | Proposes Spell, an online log parsing method based on longest common subsequence (LCS). | Two supercomputer logs: Los Alamos HPC log and BlueGene/L log. Parses unstructured log messages into structured events types and parameters in an unsupervised streaming manner with linear time complexity. Spell with pre-filtering has a faster computation time and achieves a higher accuracy compared to prior work. The prefix tree depth can grow arbitrarily without limitation, which can lead to lengthy computation time on large datasets. | |
| He et al. [157] | Proposes Drain, a fixed-depth tree-based online log parsing method. | Five real-world data sets: BGL, HPC, HDFS, ZooKeeper, and Praxisler. Constructs a tree data structure and groups the logs that belong to similar log events (i.e., templates) into the same log node of the tree. The approach achieves higher or equal accuracy and obtains 51.85% to 81.47% faster runtime compared to other online log parsers. Drain is specifically useful for web services as it enables log parsing in a streaming manner, and evaluation shows that Drain outperforms prior offline and online log parsing approaches. It appears that Drain does not handle positional shifts in the log templates, and log messages that belong to the same log event but have different lengths will be grouped separately. | |
| Messaoudi et al. [158] | Introduces MolFI (Multi-objective Log message Format Identification), which leverages an evolutionary algorithm for log message format identification. | Six real-world datasets: HDFS, BGL, HPC, ZooKeeper, Praxisler, and one industrial software logs. MolFI achieves a higher performance than the compared alternative algorithms in detecting the correct log message templates. Formulates the log template identification task as a multi-objective optimization problem and proposes a search-based solution based on the NSGA-II algorithm [213], i.e., a sorting genetic algorithm. MolFI suffers from low efficiency (i.e., high execution time) on large datasets as its iterative genetic algorithm, NSGA-II, is computationally intensive [98]. | |
| Dia [116] | Introduces Loggram, which uses n-grams dictionaries to perform log parsing. Logram initially calculates the number of appearance of each n-gram (i.e., token) in the log file. | Eventuated on 16 publicly available logs including Android, BGL, HDFS, Spark, and Zookeeper logs. Achieves a higher parsing accuracy than the prior work, and it is 1.8X to 5.1X faster than in its calculation when compared to prior work. Based on the threshold of an n-gram occurrence, Loggram decides dynamic and static parts of the log messages and extracts the log templates. One caveat for this log parser is the threshold selection for n-gram appearance, and if a dynamic n-gram occurs frequently, it might be mistakenly picked as a part of the log template. | |

TABLE 21: Log parsing research - Topic (G).
6.4.6 Category J: Performance and Failure Diagnosis

In many cases, log messages are one of the most important clues and often the only available resource for the system's failure diagnosis and performance degradation, as it might be difficult and undeterministic to reproduce a failed scenario by replaying (i.e., rerunning). Developers often have to diagnose a production run performance degradation or failure based on logs collected in the field and returned by customers without having access to the in-field user's inputs. Although some of the research is shared with 'anomaly detection', but for performance and failure diagnosis, prior research usually aims to diagnose a set of real-word known failures. Similarly, cloud-based distributed systems and supercomputers are the main categories that prior work has opted in for their evaluation. Table 24 provides an overview for the research on mining of log files for performance and failure diagnosis, and Table 29 compares and summarizes the research in this category with additional details.

Finding 14. In sum, performance and failure diagnosis approaches include: 1) probable program execution paths investigation with logs [6], 2) machine learning to compare and classify logs with good and bad performance [7], 3) correlating performance counters (e.g., CPU/memory usage) and logs [8], 4) automation-based workflow modeling [131], 5) process-oriented dependability analysis [132], 6) control and data flow analyses to extrapolate causal relations among logs [9], 10, 140, and 7) fault diagnosis with logs [147].

6.4.7 Category K: User, Business, Security, and Code Coverage

Other applications of log file mining include: analyzing user statistics and behavior [15], application security [134], duplicate issue identification [107], code coverage with logging statements [121], and business analytics [14]. For example, Figure 17 shows how recurrent issues are first classified through historical log records, and once a new record is available, it is analyzed and compared against the historical issues [107]. Table 25 provides a comparative summary of the research in this category, and Table 30 further explains each research effort.

| Reference     | Category            | Systems                      | Org. |
|---------------|---------------------|------------------------------|------|
| Tang et al.   | service composition | 144                           | ACA  |
| Oliner and    | component interaction| in large scale systems      | ACA  |
| Aiken         | LogMaster, event correlation from logs | Hadoop, HPC, BlueGene | ACA  |
| Fu et al.     | log analytics for cloud environment Analytics | three Hadoop-based applications | ACA  |
| Shang et al.  | statistical log analysis | four FSA from industrial logs | ACA  |
| Busany and    | performance model extraction through logs | Apache Tomcat access logs | ACA  |
| Masci         | clustering-based approach to detect potential system issues | an online large-scale Service X from Microsoft | ACA  |
| Di et al.     | potential correlation detection among system events | BlueGene/Q Mira supercomputer logs | ACA  |
| He et al.     | online large-scale Service X | from Microsoft | ACA  |
| Zhu et al.    | pattern extraction | HDFS, OpenStack              | ACA  |
| Zhao et al.   | pattern mining      | HDFS, BGL                    | ACA  |
| Yang et al.   | semi-supervised with probabilistic labeling | HDFS, BGL                   | ACA  |
| Li et al.     | transformer-based anomaly detection model | HDFS, BGL, Thunderbird, Spirit | ACA  |

TABLE 23: Comparison of system’s runtime behavior research - Topic (I).

| Reference     | Category            | Systems                      | Org. |
|---------------|---------------------|------------------------------|------|
| Xu et al.     | Principal Component Analysis | HDFS & Darkstar              | ACA  |
| Fu et al.     | finite state automaton (FSA) | Hadoop& SILK                 | IND  |
| Lou et al.    | invariant mining    | Hadoop, Microsoft            | ACA  |
| Chuah et al.  | resource usage and log correlation | ratlogs, syslogs             | IND  |
| Dai et al.    | online LSTM-based approach | HDFS, OpenStack              | ACA  |
| Bartero et al. | NLP and word2vec | Syslogs                      | ACA  |
| Bai et al.    | probabilistic suffix tree | HDFS, CloudStack            | ACA  |
| Farshidi et al. | regression-based statistical approach | AWS | ACA  |
| Meng et al.   | NLP and template2Vec | BGL, HDFS                    | ACA  |
| Zhang et al.  | semantic vectors of logs | HDFS, Service X              | ACA/IND |
| Zhang et al.  | numerical workflow relations | BGL, HDFS                   | ACA  |
| Huang et al.  | transformer-base DL model | BGL, HDFS, OpenStack        | ACA  |
| Zhou et al.   | pattern extraction | HDFS, OpenStack              | ACA  |
| Gao et al.    | transfer learning and word embedding (GloVe) | HDFS, Hadoop, proprietary log | IND/ACA |
| Yang et al.   | semi-supervised with probabilistic labeling | HDFS, BGL                   | ACA  |
| Le et al.     | transformer-based anomaly detection model | HDFS, BGL, Thunderbird, Spirit | ACA  |

TABLE 22: Comparison of log anomaly detection research - Topic (H). ACA: academic project, IND: industrial project; Systems: studies systems’ logs. Systems: number of systems evaluates and some examples; Org.: Organization or link if available.

Finding 13. In sum, the approaches for monitoring system’s runtime behavior include: 1) using logs to customize operational profiles for industry software [195], 2) web service composition [155], 3) detecting inter-component interaction [133], 4) mining system events correlation [127, 144], 5) assisting developers in cloud deployments [111], 6) performance model derivation [154], 7) statistical approach [69], 8) big-data analytics for cloud deployment [111], and 9) detecting impactful system problems [137].

Finding 14. In sum, performance and failure diagnosis approaches include: 1) probable program execution paths investigation with logs [6], 2) machine learning to compare and classify logs with good and bad performance [7], 3) correlating performance counters (e.g., CPU/memory usage) and logs [8], 4) automation-based workflow modeling [131], 5) process-oriented dependability analysis [132], 6) control and data flow analyses to extrapolate causal relations among logs [9], [10], [140], and 7) fault diagnosis with logs [147].
Finding 15. In sum, because the logs are readily available, prior studies have expanded the usage of logs to adjacent domains, such as code coverage analysis and business decision making. We presume the list of possible applications of the logs will continue to grow, as logs have proven to be a rich source of information, which is non-intrusive and

Fig. 17: Duplicate and recurrent issue detection tool.

### 6.4.8 Implementation and Evaluation

It is worth mentioning that many of the concepts with regards to feature selection, machine learning implementation, and evaluation metrics that are utilized for designing and assessing the log statement automation approaches (Sections 6.3.1, 6.3.3) are also leveraged for designing and evaluating log file mining tasks. For example, DeepLog [5], proposes a machine learning algorithm for log file mining. LogRobust [138] extracts feature vectors from the log files and implements an LSTM deep learning approach, and evaluates its anomaly detection approach with Precision, Recall, and F-Measure. Drain [157], a log parsing tool, evaluates its performance with F-Measure.

### 6.5 Category L: Emerging Applications of Logs

Thus far, we discussed different logging practices and log applications, mainly in large-scale software systems. However, most recently, there has been a special interest in applications of logs in other domains such as mobile devices [105], [106], embedded [236], [237], [238], and big data [55], [125], [164]. We summarize the key findings here:

- Prior studies have proposed the application of logs for emerging areas such as mobile and big data systems.
- For mobile, developers should be aware of different logging practices that might apply to alternative platforms with different design criteria and requirements. For example, because mobile devices operate on battery with limited storage space, the cost and overhead of logging (e.g., continuously flushing logs) become more exorbitant and unfavorable than software that is running on a workstation.
- Developers of log analysis tools have considered big-data platforms to scale and speed up log analysis.
- Natural language attributes of logs [113], [122] open up a new avenue for log statement automation, e.g., log statement description, and automated log analysis of logs, e.g., anomaly detection.

Table 27 provides an overview of emerging applications of logs, and Table 25 provides additional details on research in Category L.
| Reference - Aim | Experiments | Results | Pro | Con |
|-----------------|-------------|---------|-----|-----|
| Xu et al. [2]   | Logs of the Darkstar online game server and HDFS. | PCA extracts $k$ principal components by finding the axes with the highest variance among high-dimensional data. The approach can detect anomalous logs with high accuracy and few false positives while being efficient in its computation time. | For anomaly detection with PCA, two subspaces, i.e., normal, $S_n$, and abnormal, $S_a$, are created. $S_n$ is created with the first $k$ principal components, and $S_a$ is constructed with the remaining components. The approach cannot detect anomalies that are associated with the extra system resource usage that lead to system failures. | Relies heavily on log parsing step to extract log structure from the logs and detect event sequences, and will fail if log messages do not follow the preferred structure. |
| Fu et al. [113] | Introduces a method to detect anomalies by converting unstructured log files to log keys. | The research learns a finite state automaton (FSA) from the training set log keys to model the normal behavior of the system. The results show that the approach can detect system issues, such as workflow errors. | For anomaly detection with FSA and a performance model, the authors can identify anomalies in newly generated log files. The work detects three types of anomalies: 1) work flow error, 2) transition time low performance, and 3) loop low performance. | The approach does not work properly for loop low performance detection, results in false positives, and it is workload dependent. |
| Lou et al. [136] | Two distributed systems: Hadoop and SILK. | It detects anomalies if the new logs break certain invariants, e.g., if an “open file” log message appears without observing a “close file” log, the invariant is violated and an anomaly is detected. Generally, produces fewer false-negative cases compared to the PCA-based approach. | Improved upon the PCA-based method, this approach provides the operators with intuitive insight (i.e., what invariant is breached) on anomalies, and, hence, facilitates faster anomaly track down. | This approach is not able to detect anomalies that no invariant is broken, e.g., many files are opened and closed continuously. |
| Bertero et al. [17] | Experimental evaluation on Hadoop and Microsoft CloudDB. | Ranger supercomputer logs in two formats: syslog and ratlogs (rationalized logs). Evaluates the effectiveness of three different algorithms, PCA, ICA, and Mahalanobis distance. The results reveal a list of events with a strong correlation with system problems, such as soft lockup. | Performs anomaly and correlation analyses to detect the cluster nodes and jobs that are associated with the extra system resource usage. | The approach cannot detect system problems that are not manifested as extra resource usage. |
| Bao et al. [122] | Two minute-time-window (2 mTW) metric observation prior to the anomaly achieves the highest F-measure. | DeepLog decodes the log message, including timestamp, log key, and parameter values, and applies both deep learning and density clustering approaches. The approach outperforms PCA and IM methods and produces a lower number of false-negative and positive cases for anomaly detection. | Learns log patterns from the normal execution and constructs workflows, and detects anomalies when running log patterns deviate from the normal execution. | DeepLog is evaluated on systems with highly regulated logs and with a limited log key space, i.e., Hadoop (29 keys) and OpenStack (40 keys). For feature extraction, the approach only takes into consideration the number of occurrences of an event and does not consider the sequential relationship of the traces. |
| Zhang et al. [15] | Experiments on Amazon Web Services (AWS) logs. | For anomaly evaluation, the work injects faults in 22 iterations of a rolling upgrade task and utilizes the learned model for fault prediction. Two-minute-time-window (2 mTW) metric observation prior to the anomaly achieves the highest F-Measure. | The authors utilize a regression-based method to detect the most statistically significant metrics for anomaly detection and observe cloud metrics changes and signal anomalies in case of divergence. | The evaluation is performed with synthetically injected faults. Further evaluation on actual system faults can help to better validate the approach. |
| Meng et al. [15] | Two datasets: BGL and HDFS. | LogAnomaly achieves a better F1-Score (0.96 on BGL and 0.95 on HDFS) compared to DeepLog (0.90 on BGL and 0.88 on HDFS). It leverages a new word embedding approach, template2vec, to model the sequential and quantitative patterns of logs and extract the semantic information of log templates. | It leverages a new word embedding approach, template2vec, to model the sequential and quantitative patterns of logs and extract the semantic information of log templates. For feature extraction, the approach only takes into consideration the number of occurrences of an event and does not consider the sequential relationship of the traces. | The approach does not take into account the runtime parameter values. |

**TABLE 26: Log anomaly detection research - Topic (H).**
TABLE 26: Log anomaly detection research - Topic (H) (continued).

| Reference | Category | Systems | Impl. |
|-----------|----------|---------|-------|
| Miransky et al. [129] | challenges of big data log analysis | N/A | ACA |
| Salman et al. [164] | PhelkStat, log analyt ics for Apache Spark | supercomputers and network data, e.g., Spirit | ACA |
| Mavridis et al. [125], Gourley et al. [105], He et al. [122] | various log file analysis tasks in cloud environment, energy consumption of logging for mobile applications, natural language attributes of log statement descriptions | Hadoop and Spark, 24 Andro id apps, C++ and Java projects | ACA, ACA, ACA |
| Zeng et al. [130], Ghoshal and Ward [133] | logging practices for mobile devices, naturalness and loc alness of software logs | 1,444 Andro id apps, system logs and natural corpora, e.g., Spark and Wikipedia | ACA, ACA, ACA |

TABLE 27: Comparison of emerging log research - Topic (L).

Finding 16. Researchers have contributed to logging research in various categories, and research continues to progress in the existing categories and also grows to the emerging domains. Emerging application of logs include mobile and big data, and natural language processing of logs.

7 RQ4: CHALLENGES AND OPPORTUNITIES FOR FUTURE WORK

In the previous sections, we reviewed and discussed the state-of-the-art logging research by providing an introduction to log messages and log files, costs and benefits associated with logging, automated logging approaches and their evaluation metrics, mining of log files, and finally, the emerging areas of log application. In this section, as we revisit each section, we put emphasis on answering RQ4 and specify future directions and opportunities for each category. We point out the missing pieces of the puzzle for each area, followed by our intuitive approaches for tackling those issues, which are inspired by the collective knowledge of the prior work. In the following, we include opportunities for future work based on the research categories.

7.1 Category A: Logging Cost

7.1.1 Adaptive and Constraint-Based Logging

An imperative trend that we foresee as future research is the need for adaptive logging. On the one hand,
| Reference - Aim | Experiments | Results | Pro | Con |
|-----------------|-------------|---------|-----|-----|
| Tang et al. [155] | Proposes a log-based approach to identify service composition patterns by finding associated services using Apriori algorithm. | A case study on 74 service-oriented applications. | The approach can detect service composition patterns from control flow with a high accuracy. | The approach first starts with collecting and preprocessing of execution logs, and continues with identification of frequent web services, and then extrapolates the control flows. | The approach fails to extract service patterns in cases that control flow has alternative branches. |
| Oliner and Aiken [133] | Proposes an approach to address the interactions among the components of large-scale systems by analyzing logs. | Log of eight systems: four supercomputers (Blue Gene/L, Thunderbird, Spirit, and Liberty), two data clusters (Mail Cluster and Junior), and two autonomous vehicles (Stanley and SQL Cluster). | Log data signal compression allows for the scalability of log correlation, and with minimal loss, this approach identifies system’s behavioral model. | Performs a two-stage analysis: 1) PCA compression to summarize the anomaly signals, and 2) log correlation to identify if the signals relate to each other with a time lag. | The extracted signals show correlation and not a causal relationship, and in addition, manual analysis of a system administrator is required to make sense of the data. |
| Fu et al. [127] | Proposes LogMaster, a tool to mine correlations of events in log files of large-scale cloud and high-performance computing (HPC) systems. | Experimented on three system logs, namely: Hadoop, HPC cluster and BlueGene/L. | Results show the approach is successful in correlating events related to failures with acceptable precision scores but with lower recall rates. | LogMaster parses the log lines into event sequences where each event creates an informative nine-tuple, and then uses an algorithm, named Apriori-LIS, to mine event rules from logs, and measures the events correlations. | Experiments on cloud and HPC systems show LogMaster can predict failures with high Precision, however, the Recall scores are low and require improvement. |
| Shang et al. [111] | Suggests using execution logs from the cloud environment to assist developers of Big Data Analytics (BDA) applications. | A case study on three Hadoop-based apps: WordCount, PageRank, and JACK (industrial). | The approach reduces the verification effort and reaches comparable precision with traditional keyword search methods in verifying cloud deployment procedures. | This approach exposes the differences between pseudo and large-scale cloud deployments and it points the developers’ to examine the inconsistencies, and therefore, facilitates the deployment verification effort. | The approach suffers from a high number of false positives in flagging repetitive problematic log sequences, and results in low precision. |
| Busany and Maoz [69] | Proposes an approach for behavioral analysis of logs. | Logs generated from four state-of-the-art application models. | For logs of 2000 traces, reduces the analysis requirement to between 60 to 800 traces and unobserved log ratio of 0.01% and 2.45%. | The approach leverages sampling and statistical inference to provide scalable behavioral analysis of large logs. | The approach assumes that the log is sampled independently and adequately reflects the behavior of the system under review. |
| Awad and Menascé [154] | Proposes an approach to use system logs and configuration files to automatically extract performance models of the system. | Experiments on Apache Tomcat access logs from a multi-tier server. | The results show the method is effective in extracting the workloads and system model by parsing the system configuration files and log files. | The approach extracts the interaction pattern between servers and the probability associated with each interaction. | The approach assumes the log templates are known for the systems under analysis. |
| Di et al. [144], [221] | Proposes LogAider, an analysis tool that mines potential correlations between various system events for the diagnosis purpose. | Logs of BlueGene/Q supercomputer [232]. | The approach shows effective correlation between fatal system events and job events, with both high precision and recall values (99.9-100%). | LogAider can reveal three types of potential correlations between log events: acrossfield, spatial, and temporal. | It uses a threshold to find correlation candidates, and the evaluation scores, such as Precision and Recall, appear to be threshold dependent and vary significantly with the threshold. |
| He et al. [137] | Proposes Log3C, a clustering-based approach to detect system problems. | Three datasets of an online large-scale service system X from Microsoft (confidential). | Log3C achieves F1-measures values 0.91, 0.86, and 0.868 for three datasets, and outperforms PCA and Invariant Mining approaches. | It applies cascading clustering to cluster and match the log sequences efficiently, and then correlates the log sequence clusters with KPIs to identify the impactful problems. | The research only considers a single KPI, i.e., failure rate, to correlate with the log sequences. The research can be enriched with the inclusion of additional KPIs. |

TABLE 28: System’s runtime behavior research - Topic (I).

continuously logging in details (e.g., in trace verbosity level) can incur performance overhead, and on the other hand, logging very little might degrade the effectiveness of the logs. Therefore, we anticipate further research that will work on dynamically adjusting the amount of logged data from the least verbose to the most verbose level in order to help with detailed postmortem analysis, if the system is in a detected anomaly state, and on the other side, minimize the performance overhead of logging while the system operates normally and as expected.

7.1.2 Whether to Log?

We mentioned that the prior research has explored various challenges such as ‘what to log?’, ‘where to log?’, and ‘how to log?’. We see potential for further research on all challenges of logging and with more emphasis on ‘whether or not to log?’. With the emergence of adaptive logging and logging less when not needed and log more details whenever necessary, the idea of whether or not to ultimately print an existing logging statement becomes of importance. We foresee future research that explores different scenarios of whether logging statements are eventually printed or filtered based on the goal of the logging analysis tasks, i.e., performance evaluation or failure diagnosis, and the operating state of the system, i.e., normal state vs. when a system anomaly is detected.
| Reference - Aim | Experiments | Results | Pro | Con |
|-----------------|-------------|---------|-----|-----|
| Cinque et al. [15] - Proposes a software fault injection approach to assess the effectiveness of logs in the recording of software faults in the deployed environment. | Three open-source systems: Apache Web Server, TAO Open Data Distribution System, and MySQL Database Management System. | Approximately, only 40% of the injected faults are covered by the existing logging statements in the three studies systems. | Faults are intentionally injected into the experimented software systems to determine the most common failure sequences and identify logging deficiencies and improve them. | The faults are synthetic (might not necessarily match real faults) and it requires access to the source code of the software. |
| Chuah et al. [16] - Presents an approach to reconstruct event order and establish correlations among log events to discover the root causes of a given failure. | Syslogs of Ranger and Turing supercomputers, and BlueGene/L RAS logs. | The authors received positive feedback from system admins that they have found the tool analysis useful in facilitating their diagnosing efforts. | Introduces a Fault Diagnostics tool lprof, to discover faults, which comprises three components: a Message Template Extractor (MTE), a Statistical Event Correlator (SEC), and an Episode Constructor. | The approach depends on the availability of event-specific keywords as domain knowledge for correlation, and does not provide causality. |
| Yuan et al. [17] - Proposes SherLog, a tool that leverages runtime log information and source code analysis to infer the probable execution paths during a failed production run. | Evaluated on eight real-world software failures collected from different applications such as rmdir, Squid, and ln. | The experiments show the information inferred by SherLog is useful to assist developers in failure diagnosis. | By accepting the execution log of a failed run and the source code, SherLog aims to identify what must or may have happened along the execution path. | SherLog relies on the amount of information available in log messages to perform its analysis. As such, log messages that lack the necessary debugging information will significantly limit SherLog's effectiveness. |
| Pecchia et al. [18] - Conducts an experimental study to examine factors from event logs that help with the detection of failures. | Performs experiments on a set of 17,837 instances of injecting faults into three systems: Apache Web Server, TAO Open DDS, and MySQL DBMS. | Features such as system architecture, placement of the logging statements, and support provided by the execution environment can have an impact on the accuracy and effectiveness of the logs at runtime. | DISTALYZER is able to uncover undiagnosed performance issues for the experimented systems. | This research additionally investigates the logging improvement that can potentially increase the usefulness of the execution logs. |
| Nagaraj et al. [19] - Presents DISTALYZER, a tool which utilizes log data to assist developers in diagnosing performance problems. | Case studies on three systems: TritonSort, HBase, and BitTorrent. | Results show that DISTALYZER can be used to discover undiagnosed performance issues for the experimented systems. | DISTALYZER uses machine learning techniques (e.g., Welch’s t-test [23]) and dependency networks [24] to compare log files with acceptable and unacceptable performance. | DISTALYZER leverages ad-hoc approaches (e.g., thread id) to group log messages, which limits its application for less-structured logs. |
| Fronza et al. [20] - Proposes an approach to perform log-based prediction by applying Random Indexing (RD) and Support Vector Machines (SVMs). | Experimented on log files of a large European manufacturing company (anonymous). | According to the findings, weighted SVMs achieve the best performance by slightly shrinking specificity (true negative rate) scores to improve sensitivity or recall, and specificity stays greater than 0.8 in the majority of the experimented applications. | It applies weighted SVM, which utilizes cost-sensitive learning to achieve balanced TPR and TNR values, and makes the method more reliable in classifying both failures and non-failures. | SVM classification performs well in classifying non-failure instances, but poor in identifying failures, i.e., low true positive rate or recall. |
| Syer et al. [21] - Proposes an approach to reconstruct event order and establish correlations among log events to diagnose memory-related issues in load tests. | A case study of WordCount application on Hadoop. | The approach flags less than 0.1% of the execution logs with a high precision of $\geq 90\%$. | After clustering the events, authors apply scoring techniques to identify clusters that are abnormal and can be associated with a performance issue. | The approach has limited applications to memory performance issues, such as memory leaks, spikes, and blots. |
| Zhao et al. [22] - Proposes lprof, a log profiling tool that recreates the execution flow of distributed applications. | Evaluated on four distributed systems: HDFS, Yara, Cassandra, and HBase. | lprof’s reaches 88% accuracy in attributing log messages to requests, and 65% of the diagnostics are helpful for the operators. | lprof performs control-flow (CF) and data-flow (DF) analyses, and infers if log messages are causally related and what variables are unmodified between multiple log printing statements, and then groups the logs and use them for diagnosing performance issues. | lprof’s static analysis is limited to a single software component and needs to be readjusted for different languages (bytecode), which is cumbersome in practice. |
| Xu et al. [132], [162] - Utilizes system logs to provide rolling updates in a cloud environment for process oriented dependability (POD) analysis. | Experiments with rolling upgrade on AWS with injecting 8 different types of faults into the cloud-based clusters. Faults include machine image (MI) change during upgrade, key pair management fault, and security group configuration fault. | The evaluation results show acceptable performance (90%+) in injecting, recall, and accuracy scores in diagnosing the injected sporadic faults. | It creates a process model of the desired provisioning activities through log data with added annotation and checkpoints. The deployment logs are checked and assertions are raised in case there has been a deployment violation. | The approach heavily relies on the specific information in the logs and the absence of this information severely impacts the performance of error detection. Following research [23] aims to combine logs with system metrics for a more robust analysis. |

**TABLE 29: Performance, fault, and failure diagnosis research - Topic (J).**
Experiments on log sequences of 25 different applications of a software system for telemetry and performance of cars. The evaluation achieves 78% recall and 95% precision. Uses three popular support vector machines (SVMs): multilayer perceptron, radial basis function, and linear kernels to learn and predict defective (i.e., faulty) log sequences. The study is performed on the logs of a single system. Thus, the result of the applicability of the proposed approach to other systems and other software domains remains unknown.

The approach is accurate enough to check and infer workflows for most interleaved log sequences. CloudSeer reaches 92+% accuracy in checking interleaved logs for six experimented groups with a satisfactory checking efficiency (i.e., computation time). CloudSeer constructs automations for the workflow of management tasks based on their normal execution scenarios, and later checks log messages against these automations to detect workflow discrepancies and divergences in an online approach.

Additional tools that can automatically detect log-related issues, which combines console logs with resource usage data to improve the error detection accuracy in distributed systems. The approach is able to identify 80% of errors leading to failures, and achieves f-measure over 70%.

Evaluated both through a controlled user study and lab experiments on Hive, Spark, and OpenStack. On average, Stitch achieves 96% and 95% accuracy for object and edge detection, respectively, for workflow reconstruction.

Performs experiments on logs of components (e.g., disk, I/O, memory) of a cloud environment, StrongCloud, collected over a year period. Twelve categories of faults are detected, and fault detection precision is maxing out at 88% when the length of logs is more than 200 words.

Evaluated on 18 randomly sampled real failures on four systems: HDFS, HBase, ZooKeeper, and Cassandra. Pensieve is able to reproduce 72% of the sampled failures within ten minutes of analysis time. Pensieve leverages event chaining and extrapolates a chain of causally dependent events leading to the failure while using partial trace observation technique, which significantly limits the execution paths to observe.

TABLE 29: Performance, fault, and failure diagnosis research - Topic (J) (continued).

### 7.2 Categories B, C, D: Logging Practices, progression, and Issues

#### 7.2.1 Improved Logging Practices

Although logging practices in the software development process have been reviewed and improved over the past decade [28, 29], there is still room for betterment [101]. Additional tools that can automatically detect log-related issues are required. Moreover, because the majority of current logging practices and decisions are ad-hoc and decided by developers on the spot, the introduction of systematic logging practices that can provide suggestions to developers while composing the code can ensure a higher quality of logs. Also, further research that can provide directives and insights for developers with regards to good versus poor logging practices, and hence help to improve their logging practices and make better use of logging, is of interest. For example, more effective logging can enable the customers of the software systems to solve problems themselves using the logs without relying on developers or avoid unnecessary logging costs, such as exposing users’ sensitive information in the logs [57]. Another angle for logging practices improvement includes studies that investigate cost-aware logging, which can help developers to estimate and optimize the cost of logging while benefiting from the logs. Although efforts such as Log20 [33] have aimed to address this issue, there is still a sizable room to improve upon. This avenue of future research can be also expanded to other platforms such as mobile devices as the logging practices can be different depending on the applications and the system requirements [106]. Prior research has shown other areas such as mobile systems, that are not in the research community’s spotlight, are even more in dire need of systematic guidance and automated support tool for assisting in logging practices [106].

#### 7.2.2 Representation of Log Files

It is a safe assumption that the log analysis methods require, or at the very least, perform better on logs with good quality to conduct meaningful analyses. Therefore, we foresee future research in improving the formatting and defining universal structure for log messages, which will directly help in achieving more symmetric organization of log files, and consequently, more effective log analysis with higher Precision and Recall values. This goal is also partially realized
with proper selection and improvement of logging libraries and utilities, e.g., Log4j, SLF4J, and Logback.

### 7.2.3 Logging Libraries and Utilities

Logging libraries and utilities (LLU) provide additional functionality, structure, and flexibility in logging for developers such as log verbosity levels and thread-safety [48]. Although LLUs facilitate logging, there has been insufficient research on this topic. Further research that aims to improve the performance of logging libraries by performing some of the logging tasks during the compile time is necessary [240]. Furthermore, the development of application-specific logging libraries will provide higher logging flexibility and better API for developers in a specific domain, similar to Log++ [241] for cloud logging, to perform workload-related logging. For example, in a cloud deployment and provisioning process, users are further interested in logging the machine image initialization and termination steps in more details to enable better debugging in case of failures. Additionally, LLUs can improve to bring in new configurability, such as supporting different log verbosity levels for separate parts of a logging statement [37]. Another angle that LLUs can improve is to provide checks on the format of the developer’s provided logging text and ensure that the provided content passes a minimum set of standards in order to make logs more useful and organized. Additionally, logging libraries can help to reduce the overhead of logging. The way that some of the LLU work is that, during the runtime, all of the logging statements are executed but based on the verbosity level of the logging statements, some of the logs are filtered from being written to the log file. This approach can still introduce a considerable overhead if logging statements make calls to other methods and variables. To cope with this situation, developers include log statement guarding (e.g., putting the log statement inside an if-clause) to avoid the logging statement being executed based on whether or not the level is enabled. Therefore, this type of log guarding improvement would be beneficial to be implemented inside the LLUs [101].

### 7.2.4 Application Specific Logging

As logging messages can provide valuable information with regards to the different aspects of the running software, it is also evident that different tasks and applications that rely on logs require different types of information from the log files. Therefore, we anticipate application-specific logging research in the future to grow. That is to say, depending on the application, we need to log different categories of runtime information. For example, for a security log, certain values need to be printed while not compromising sensitive

| Reference | Aim | Experiments | Results | Pro | Con |
|-----------|-----|-------------|---------|-----|-----|
| Cheng et al. [121] | Presents an approach, called LogCoCo (i.e., Log-based Code Coverage), to estimate and measure the source code coverage using the readily-available execution logs. | Five commercial proprietary projects from Baidu and one open-source project, i.e., HBase. | Measures the accuracy and usefulness of LogCoCo, and it achieves high accuracy for different types of code coverage (Must and May have been executed). Additionally, the tool's results are useful to evaluate and improve the test suites for code coverage. | The research highlights how event log data can be used in decision-making processes in recurrent and unknown issues. | The approach cannot accurately infer whether a May executed code region is actually covered in a test. |
| Lim et al. [107] | Addresses the problem of automated identification of recurrent and unknown performance issues. | Evaluated on two datasets: Transaction Processing Performance Council Benchmark (TPC-W) and System X, a real production system (confidential). | The approach shows the approach achieves higher AUC when compared to approaches such as Fingerlog, Signature, K-means, and Hierarchical approaches in recurrent and unknown issues identification. | The approach works based on the mining and metric extraction of historical log records, and utilizes a Hidden Markov Random Field (HMRF) based approach for the clustering of recurrent issues. | The proposed statistical analysis only works if a large amount of monitoring data over a long period of time is available. |
| Oprea et al. [134] | Proposes a framework that analyzes log data collected at the enterprise network borders on a regular basis (e.g., daily). | Experimenter with DNS logs released by Los Alamos National Lab (LANL), and AC dataset of web proxies logs generated at the border of a large enterprise network (confidential). | The approach can detect malicious web domains with high accuracy and low false-negative rates. The work also detects new malicious domains that are not previously reported/detected by other tools in the enterprise. | Creates a bipartite graph \( G = (V, E) \), such that hosts and domains are vertices on each side of the graph. There will be an edge between a host and a domain if the host connects with the domain. | The approach cannot detect regular connections to malicious domains which happen in the training phase. |
| Bank et al. [14] | Performs a case study of log utilization in Microsoft for business decisions and analytics. | Performed interviews with 26 engineers at Microsoft, and followed that up with a survey of 1,823 respondents to confirm their findings. | Use of log event data is pervasive within the organization and the usage primarily falls into eight categories, among them engineering the data pipeline, instrumenting for event data, troubleshooting problems, and making business decisions. | This research highlights that event log data surely plays an important role in the company's decision-making process as the industry makes a transition towards a data-driven decision-making paradigm. | The study is performed on a single software company, and the findings may not generalize to other software and institutions. |

| TABLE 30: User, business, security, and code coverage research - Topic (K). | | | | |

## Experiments

### Pro

- “Client events” within Twitter logging framework.
- Discusses a variety of applications for the proposed approach, such as summary statistics, event counting, funnel analytics, and user modeling.
- The research applies techniques from natural language processing (NLP) to process the user’s behavior on the website, the user’s behavior right now is strongly influenced by immediately preceding actions.
- Currently, user session sequences only capture event names and do not provide enough details for more sophisticated types of analyses.

### Con

- Log++ libraries will provide higher logging flexibility and performance of logging libraries by performing some of the research on this topic.
- Further research that aims to improve the way that some of the LLU work is that, during the runtime, all of the logging statements are executed but based on the verbosity level of the logging statements, some of the logs are filtered from being written to the log file. This approach can still introduce a considerable overhead if logging statements make calls to other methods and variables. To cope with this situation, developers include log statement guarding (e.g., putting the log statement inside an if-clause) to avoid the logging statement being executed based on whether or not the level is enabled. Therefore, this type of log guarding improvement would be beneficial to be implemented inside the LLUs [101].

## References

- Lim et al. [107] - Addresses the problem of automated identification of recurrent and unknown performance issues.
- Oprea et al. [134] - Proposes a framework that analyzes log data collected at the enterprise network borders on a regular basis (e.g., daily).
- Bank et al. [14] - Performs a case study of log utilization in Microsoft for business decisions and analytics.
- Cheng et al. [121] - Presents an approach, called LogCoCo (i.e., Log-based Code Coverage), to estimate and measure the source code coverage using the readily-available execution logs.

---

[48]... additional functionality, structure, and flexibility in logging for developers. Log4j, SLF4J, and Logback.

[240]... necessary. Further research that aims to improve the performance of logging libraries by performing some of the logging tasks during the compile time is necessary.

[241]... Log++ for cloud logging, to perform workload-related logging.

[37]... that supports different log verbosity levels for separate parts of a logging statement. Another angle that LLUs can improve is to provide checks on the format of the developer’s provided logging text and ensure that the provided content passes a minimum set of standards in order to make logs more useful and organized.

[101]... certain values need to be printed while not compromising sensitive
information; however, this might not be an issue in post-mortem analysis of logs. Additionally, developers of other platforms, such as mobile apps, should be aware of the differences between desktop/server and mobile practices as it comes to logging, as for mobile, there is energy overhead concern that should be taken into account. Therefore, different platforms also might end up logging different information with varied frequencies of outputting logging statements.

7.2.5 Maintenance of the Logging Code

As the software systems continue to grow, maintaining the logging code becomes more challenging. Previous studies have observed that the logging code is not maintained as well as the feature code, as there is no straightforward way to test the correctness of the logging code. Therefore, we emphasize that further research should consider the systematic maintenance and testing of logging code alongside the feature code evolution. Additionally, there are interesting opportunities for developing automated tools that can read the context of the feature code changes and suggest logging code maintenance and updates concerning the feature code updates while the new feature code is being checked in.

### TABLE 31: Emerging log research - Topic (L).

| Reference - Aim | Experiments | Results | Pro | Con |
|-----------------|-------------|---------|-----|-----|
| Miransky et al. [5] | Discusses the challenges of event log analysis for big data systems, as the logs generated by a BDS can be big data themselves. | Categorizes seven challenges of log analysis for big data systems. | Highlights currently available solutions to each challenge and discusses unanswered questions, based on the authors’ and industrial experience. | The authors categorize the challenges of big data log processing into seven classes, including scarce storage, unsalable log analysis, inaccurate capture & replay, and inadequate tools for instrumenting BDS source code. Accurate mapping and case studies and examples of challenges in real-world big-data software can further illustrate the current issues. |
| Salman et al. [164] | Proposes PhipStat, a tool for analysis of system event logs of large-scale data centers on Apache Spark’s big data platform. | A set of public (e.g., Spirit, Thunderbird, and Liberty) and private (e.g., Cray and dartmouth/campus) log data. | Performs evaluation on a set of log analysis tasks such as arrival rate distribution, anomaly content, and runtime analysis. | The authors utilized a set of attributes, i.e., temporal and spatial metrics such as arrival rate and byte count, to characterize system event logs and then correlate the metrics with the runtime performance of the system. Log analysis tasks are partially correlated with system events, but are not analyzed meaningfully to draw actionable steps for admins and users of the system. |
| Mavridis et al. [125] | Evaluates various log file analysis tools with two cloud computational frameworks, Apache Hadoop and Apache Spark. | Experiments on log files of an Apache HTTP server, and implemented useful log analysis tasks such as checking for denial of service (DoS) attacks from the available logs. | Compared performance of Hadoop and Spark by evaluating their execution time, scalability, resource utilization, and cost and power consumption. | The research showed the potential of utilizing distributed big-data platforms for facilitating log analysis. The analysis needs to be expanded to designing tools that can leverage the parallelism of distributed big-data platforms to accomplish a faster and scalable analysis of logs. |
| Chowdhury et al. [105] | Performs an exploratory study to investigate the energy impact of logging in Android applications using GreenMiner [239], an automated energy test-bed for mobile applications. | Studies approximately a thousand versions of 24 Android applications (e.g., CALCULATOR, FEDEEX, FIREFOX, and VLC) with logging enabled and disabled, accompanied by a controlled experiment on a synthetic application. | There is little to no energy impact when logging is enabled for most versions of the studied applications. However, about 79% (19/24) of the studied applications have at least one version that exhibits a noticeable impact on energy consumption. | The authors found that the rate of logging, the size of messages, and the number of log buffer flushes are significant factors of energy consumption attributable to logging on mobile devices. More accurate models that consider mobile log events with the amount of energy consumption are required. |
| He et al. [122] | Characterizes natural language attributes of log statements' descriptions. | Experiments on ten Java and seven C++ open-source projects and answers four research questions. | Findings confirm the natural characteristics of logs, such as endemic and specific. | Proposes an automated approach for log description prediction based on source code similarity and edit distance. The dynamic part of the log statements is left out, and the approach for log description automation is limited to cases that a similar code snippet is found. |
| Zeng et al. [106] | Replicates the work of Yuan et al. [25] and investigated the logging practices in Android applications. | Performs a case study on 1,444 open-source Android applications in the F-Droid repository. | Although mobile app logging is less pervasive than server and desktop applications, logging is leveraged in almost all studied mobile apps, and there are noticeable differences between the logging practices applied in mobile applications versus the ones in server and desktop applications, as observed by prior studies. | The majority of the logging statements in mobile apps are in debug and error verbosity levels, while info level logging statements are the prevailing level in server and desktop applications. The research can be expanded by providing developers with guidelines for mobile apps. |
| Gholamian and Ward [113] | Performs an experimental study on natural and local characteristics of log files. | Experiments on eight system logs (e.g., HDFS and Spark), and two natural language data (e.g., Gutenberg and Wiki). | Six findings confirm that log messages are natural and local, even more or so than common English text. | Applies the findings and proposes an NLP-based anomaly detection approach from log files, which utilizes n-gram models. More advanced NLP models (e.g., deep learning and BERT) need to be investigated to improve the anomaly detection task. |

7.3 Category E: Log Printing Statement Automation

7.3.1 Automatic LPS Generation

In contrast to developer-inserted logs, LPS automation aims to auto-generate or suggest new logging statements or enhance the quality of currently available logs inside the source code based on various source code and applica-
tion criteria. Although this topic has been of interest recently [30, 33, 36, 81, 82, 109, 126], considering the continuous advancement and birth of new AI and learning methods, we anticipate future research in the development of machine learning methods to implement and automate logging, with statistical modeling, supervised, unsupervised, and deep learning approaches will continue to foster. These methods should consider automating different aspects of the logging statements, such as the location, content, and verbosity level. The automated methods can also consider different criteria for automation, such as diagnosability versus cost-awareness [37]. The ultimate goal is to achieve an automated approach that can introduce high-quality log suggestions or enhancements for various applications. Subsequently, assuming the development of different approaches, a comparative study of different approaches and the areas that each one performs better becomes necessary, similar to the comparison of different log parsing techniques in [52].

7.3.2 Constraint-based Logging
The majority of the log automation tools have aimed to mimic developers’ logging habits [30, 36, 126]. In other words, the log learning approaches work to learn developers’ logging habits to decide if a new unlogged code snippet requires a logging statement. However, prior work [101] has also shown that developers make mistakes, and in some places, they even forget to log in the first place. Thus, one remaining important challenge is to develop constraint-based automated logging approaches to guarantee a particular logging goal, e.g., at minimum, one iteration out of 100 iterations of method MtdM() is logged, or a particular execution path is fully disambiguated with logging, i.e., we can accurately determine which code segments ’must’ have been executed. Another example can be ensuring the beginning and the end of all methods of interest are logged. By doing so, we can guarantee that at least a minimum quality of logs is granted.

7.3.3 Golden Quality LPSs for Benchmarking
To the best of our knowledge, there is no prior work that quantitatively measures the quality of logging statements in each software project. Many of the prior work consider the developers’ inserted logging statements as ground truth to evaluate their automated logging approach [30, 36, 126]. However, prior research has shown that there is no general guideline for logging and developers mostly rely on their intuitions and insert logging statements in an ad-hoc manner [17]. As such, defining a set of quantitative metrics that can be applied to evaluate the quality of logs on various software projects and give them scores can be highly beneficial. This allows to find projects with high-quality logs, learn from them, and use them as a golden benchmark for comparison with auto-generated logs.

7.4 Category F: Log Maintenance and Management
Log collections and compressors. Log collections, similar to LogHub [90], are useful for evaluating various log analyses. Additionally, datasets that are labeled and differentiate normal against abnormal log records are well sought after, as they enable the application of supervised and deep learning approaches for log analysis [138, 244, 245]. As such, we see value in further research to collect log data from various software and application domains, and develop automatic and accurate probabilistic methods [110] to label the data to facilitate log analysis and logging practices research. For log compression, because logs generally benefit from higher repetition than natural text, future research can benefit from designing and evaluating log-tuned compressors, which not only can result in more effective compression but also more efficient and streamlined decompression, for later auditing and analysis.

7.5 Categories G, H, I, J, K: Automated Log Analysis Applications
7.5.1 Log Parsing
With the recent advancements of NLP models [227, 246], we foresee the development of transformer-based log parsers that can potentially improve the performance of log parsing, and consequently, the downstream log mining tasks. In addition, pre-trained language models [247] can potentially be fine-tuned on log files to enable higher performance for log parsing.

7.5.2 Log Analysis and Tools
Prior research has proposed plenty of log analysis methods and tools for different applications, such as anomaly and problem detection [2, 4], performance and failure diagnosis [9, 10, 31, 140], system’s runtime behavior [111, 137, 195], system profile building [134], code quality assessment [99], and code coverage [121]. Log analysis, starting with log parsing, plays an essential role in extracting useful information from the log files. As logs can be viewed in different ways, such as events, time series, and feature vectors, this enables different types of analyses. Complementary to the available research, because logs are non-intrusive and readily available, we anticipate new methods of log analysis or improvement of the current methods will be sought after for different applications. The quality of log analysis can directly impact the amount of actionable information that we can extract from the log files. Therefore, we expect new logging analysis approaches will emerge that utilize and combine a variety of algorithms to achieve a more accurate analysis. The approaches might also assume a specific format of logs, e.g., log messages following a specific template within the log files, to achieve a more personalized analysis. For example, the research can benefit from considering multiple factors, such as the content of each log message, the frequency, and the sequencing of log messages in log analysis tasks, e.g., anomaly detection, in order to achieve a deeper understanding of what happens in the logs. Lastly, we foresee that future research will benefit from utilizing AI approaches in understanding and leveraging the hidden semantics of the log messages, rather than solely focusing on learning log patterns and templates. This will enable a more sophisticated log analysis.

7.5.3 Scalable and Online Log Processing
In order to keep pace with the massive amount of growing logs in size (at the rate of approximately multiple terabytes per day [245]) and various formats, which is the
by-product of the software growth as well as the number of software users’ growth, we anticipate further research will be conducted to develop and update the current logging processing tools and platforms. Thus, future research should consider leveraging distributed and parallel processing platforms (e.g., Apache Spark) in conjunction with efficient machine learning approaches to implement scalable log analysis tools for all stages of the process, i.e., real-time collection, processing, and storage of voluminous logs [16]. In addition, as many of the enterprise software platforms require 24/7 up-time and availability, the need for online tools that can perform the log analysis simultaneously as the system generates logs becomes more apparent. We require the tools to be efficient enough to perform analysis at the same speed or faster than the log generation rate.

7.6 Category L - Emerging Logging Research

7.6.1 Natural Language Processing of Logs

Prior work [249], [250] in software engineering has utilized natural language processing (NLP) for software tasks such as source code next token suggestion. Recently, there has been a thread of research on analyzing logging statements as natural language sequences. He et al. [122] characterized the NLP characteristics of LPS descriptions in Java and C# projects, and Gholamian and Ward [113] showed software execution logs are natural and local, and these features can be leveraged for automated log analysis, such as anomaly detection. We hypothesize that further research is required to confirm the NLP characteristics of software logs, and eventually, leveraging NLP characteristics of logs will further benefit automated log generating and analysis of log files. Moreover, the recent advancements in NLP models, e.g., BERT models [246], calls for further investigation and application of them in improving the performance of log mining tasks. The intuition is that these models can embed and learn a higher degree of log semantics, and thus, can better enable actionable diagnosis from logs.

7.6.2 Log Summarization and Visualization

Prior works [113], [251], [252] have proposed approaches to summarize and visualize console and security logs. Log summarization and visualization is a natural response to the ever-growing scale of logs to gain high-level insight into the logs. In large-scale distributed software systems, as the scale of logs continue to grow, and various subsystems continue to generate logs in heterogeneous formats and rates, we foresee the development of approaches and solutions, both in academia and industry, that aim to make high-level sense of logs and to gain big-picture insight with visualizing logs. In addition, log summarization will help developers and practitioners to focus their troubleshooting efforts on a smaller set of relevant logs. Knowledge graph representation is a potential candidate for this aim [253]. Lastly, we provide a digest of the avenues for future of logging research in Table 32.

8 Conclusions

Logging statements and log files are the inevitable pieces of the puzzle in analyzing and ensuing various aspects of correct functionality of software systems, such as debugging, maintaining, and diagnosability. The valuable information gained from logs has motivated the research and development of a plethora of logging practices, logging applications, and log automation and analysis tools.

In this survey, we initially started with the basics of log statements and log files and the involving challenges in extracting useful information from them in Sections 1 and 2. As we conduct the survey, we aim to answer four crucial research questions related to software logging: (RQ1) categories of logging research, (RQ2) publication trends based on topics, years, and venues, (RQ3) available research in each category, and finally (RQ4) challenges and opportunities for future logging research. We next reviewed the costs and benefits associated with logging in Section 6.1 and followed that up with research that mines logging statements to derive logging practices in Section 6.2. In Section 6.3, we reviewed the proposed methods for automated logging, and we mentioned evaluation methods and metrics for auto-generated logs and learning-to-log platforms in Section 6.3.4. In section 6.4, we reviewed log file mining and log analysis research which aims to expedite and scale up the log processing, and apply logs for different system maintenance tasks such as anomaly and failure detection/diagnosis, performance issues, and code quality assessment. We also reviewed the emerging domains and applications for logging, such as in NLP, mobile, and big data in Section 6.5. Finally, we discussed the opportunities for future research in different aspects of logging statements and log files, their practices, and their analyses in Section 7. Overall, we reviewed 112 primary studies published between 2010 and 2021 (inclusive), and we included our findings for each subsection of the survey, which tallied up to 17 succinct and quick-grasping findings in total.

Although current research advances have made logs more useful and effective, there are still multiple remaining challenges and avenues for future work and improvement. Categories of challenges remain in various aspects of automated log analysis, LPS auto-generation, scalable logging analysis and infrastructure, cost-aware logging, log maintenance and management, and improved logging practices. We foresee future research in multiple directions for logging as follows:

- As the size of computer systems increases, we anticipate the voluminousness and heterogeneity of logs, which turns it into a big-data problem, will demand further quantitative cost analysis for collecting, processing, and storing of logs, as logging can infer computation, storage, and network overhead. Additionally, due to the voluminousness and heterogeneity of generated logs, and in some cases, the need for real-time processing of logs, we anticipate the development of efficient, scalable, and real-time log analysis tools [55].
- We anticipate continued research on current logging practices and log-statement-related issues, as this will enable improvements of future practices and help to create guidelines for developers when making logging
decisions. We also predict the evolution of learning and AI-based log recommender tools and IDE plugins, which utilize the readily available code repositories of open-source projects to provide just-in-time logging practice suggestions to developers [36], [102]. Additionally, we expect further work on logging libraries to collaborate with emerging logging practices and bring in the development of application-specific logging practices [242].

- We foresee automated log file analysis techniques continue to evolve and become more effective and sophisticated (with machine learning and AI-based techniques) in their information extraction from log files and log statements. We also see an emerging trend of new applications that utilize log analysis recently, such as log analysis for code coverage [121]. Moreover, we predict further research will be performed in enabling the analysis of logs for other platforms, such as mobile systems and big-data applications. Log collections will also continue to grow to help with log analysis.

- With regards to log statement prediction, we anticipate that future research on supervised, unsupervised, and deep learning techniques will continue to benefit logs and their analyses.

In this study, we aim to systematically summarize, discuss, and critique the state-of-the-art knowledge in the logging field for experienced researchers, and simultaneously, help new researchers to get a quick and critical grasp of the available research in this area. Additionally, we envision the uncovered research opportunities in this survey serve as a beacon for advancing the logging research. Lastly, we provide a link to the data used in this survey, available at: https://github.com/sgholamian/comprehensive-software-log-granulation.

### TABLE 32: Summary of avenues for future work in logging research.

| No. | Avenue                             | Rationale                                                                 | Selected research |
|-----|-----------------------------------|---------------------------------------------------------------------------|-------------------|
| 1.  | Adaptive logging                  | Dynamic adjustment of the logging level from the least to the most verbose level helps with detailed postmortem analysis. | 49                |
| 2.  | Whether to log?                   | Different scenarios of whether LPSs are printed or filtered based on the goal of the logging analysis tasks should be studied. | 91                |
| 3.  | Logging practices                 | Future research can improve on logging practices in the software development process and reduce the ad-hoc and forgetful developers' logging habits. | 28, 29, 101       |
| 4.  | Representation of the log files   | Further research can improve the formatting and standardization of log messages, which directly results in more organized log files and more accurate automated analysis. | 32, 248           |
| 5.  | Logging libraries and utilities   | Logging libraries and utilities (LLU) can provide additional functionality, structure, and flexibility in logging for developers. | 48, 240, 241      |
| 6.  | Application-specific logging      | Research can investigate and ensure that how different tasks and applications that rely on logs can record application-specific information (i.e., based on the application needs) into the log files. | 106, 242          |
| 7.  | Maintenance of logging code       | As maintenance of the logging code becomes more challenging, future research requires to develop automated approaches to ensure up-to-date and issue-free logging code. | 29, 173, 243      |
| 8.  | Automated and constraint-based log generation | Research requires to improve the quality of auto-generated LPSs and, also, enhance the quality of the developer-inserted ad-hoc logs by adding additional variables, etc. | 30, 33, 36, 81, 82, 90, 104, 117, 126 |
| 9.  | Golden quality log statements     | High quality logs are required to learn from, and use them as a golden benchmark for comparison with auto-generated logs. | 30, 36, 117, 126  |
| 10. | Log collections                   | There is a need for labeled log data collections and development of automated log labeling approaches to facilitate automated log analysis. | 90, 192, 194      |
| 11. | Log compression                   | Compressors which are designed for logs are needed to improve the compression/decompression efficiency and enable efficient long-term storage and backup of logs. | 54, 104, 195      |
| 12. | Log analysis for various objectives | Research will actively continue to propose and improve approaches for more accurate log analysis for different log mining tasks and postmortem debugging. | 2, 1, 9, 10, 110  |
| 13. | Scalable and online log processing | Scalable and real-time log processing is required to keep pace with the massive amount of growing logs in size, e.g., multiple terabytes per day. | 131, 140          |
| 14. | Natural language processing of logs | Leveraging NLP characteristics of logs will benefit automated log generating (e.g., log description generation) and NLP processing of log files. | 251, 252, 253     |
| 15. | Log summarization & visualization | Development of approaches that aim to elicit and condense big-picture insights from logs with visualizing and summarization are in demand, and this will enable practitioners to only focus on significantly smaller but most important portion of logs. | 36, 122, 249, 250 |

[36], [102], [242], [121].
### 9 List of Papers

Table 33 provides the list of papers per each category of logging research.

| Topic | Paper Title | Year | Venue | Subtopic(s) |
|-------|-------------|------|-------|-------------|
| (A) Costs and benefits of logging (6) | Be conservative: Enhancing failure diagnosis with proactive logging [81]. Linux auditing: Overhead and adaptation [161]. A qualitative study of the benefits and costs of logging from developers’ perspectives [37]. Log4JPerf: Suggesting and updating logging locations for web-based systems’ performance monitoring [96]. What distributed systems say: A study of seven spark application logs [76]. | 2012 | OSDI (J) (E) | |
| | | 2015 | ICC (C) | |
| | | 2015 | ATC (C) (E) | |
| | | 2020 | TSE (J) | |
| | | 2020 | EMSE (J) (E) | |
| (B) Logging practices (7) | Characterizing logging practices in open-source software [28]. Where do developers log? An empirical study on logging practices in industry [35]. Industry practices and event logging: Assessment of a critical software development process [40]. Studying the relationship between logging characteristics and the code quality of platform software [59]. Characterizing logging practices in Java-based open source software projects - A replication study in Apache Software Foundation [96]. An exploratory study of logging configuration practice in Java [168]. Studying the use of Java logging utilities in the wild [16]. | 2012 | ICSE (C) (C) | |
| | | 2014 | ICSME (C) (B) | |
| | | 2015 | ICSE (C) (A), (E) | |
| | | 2015 | EMSE (J) (D) | |
| | | 2017 | EMSE (J) (C) | |
| (C) Logging progression (5) | An exploratory study of the evolution of communicated information about the execution of large software systems [85]. Logging library migrations: A case study for the Apache software foundation projects [92]. Examining the stability of logging statements [84]. Guiding log revisions by learning from software evolution history [100]. Can you capture information as you intend to? A case study on logging practice in industry [172]. | 2014 | ICSME (C) (C) | |
| | | 2018 | EMSE (J) (B), (D) | |
| | | 2019 | EMSE (J) (E) | |
| | | 2020 | ICSME (C) (C) | |
| (D) Log-related issues (7) | Simple testing can prevent most critical failures: An analysis of production failures in distributed data-intensive systems [141]. Understanding log lines using development knowledge [53]. Characterizing and detecting anti-patterns in the logging code [29]. Studying and detecting log-related issues [101]. Studying duplicate logging statements and their relationships with code clones [114]. Demystifying the challenges and benefits of analyzing user-reported logs in bug reports [75]. | 2014 | ICSME (C) (B) | |
| | | 2014 | OSiD (C) | |
| | | 2015 | ICSE (C) (B) | |
| | | 2016 | MSR (C) (B) | |
| | | 2017 | ICSE (C) | |
| | | 2018 | EMSE (J) (E) | |
| | | 2019 | EMSE (J) (D) | |
| | | 2021 | TSE (J) (C) | |
| | | 2021 | SP&E (J) (I) | |
| (E) Log printing statement automation (13) | AutoLog: Facing log redundancy and insufficiency [163]. Improving software diagnosability via log enhancement [52]. Learning to log: Helping developers make informed logging decisions [30]. LogOptPlus: Learning to optimize logging in catch and if programming constructs [165]. Log20: Fully automated optimal placement of log printing statements under specified overhead threshold [53]. Towards just-in-time suggestions for log changes [102]. Which log level should developers choose for a new logging statement? [103]. SMARTLOG: Place error log statement by deep understanding of log intention [126]. An approach to recommendation of verbosity log levels based on logging intention [159]. Which variables should I log? [177]. Automatic recommendation to appropriate log levels [142]. Logging statements’ prediction based on source code clones [36]. Where shall we log? Studying and suggesting logging locations in code blocks [120]. An exploratory study of log placement recommendation in an enterprise system [112]. DeepLV: Suggesting log levels using ordinal based neural networks [159]. | 2011 | APsys (C) | |
| | | 2012 | TOCS (J) (J) | |
| | | 2015 | ICSE (C) (B) | |
| | | 2017 | CCF/CS (C) | |
| | | 2018 | SOSP (A) (C) | |
| | | 2017 | EMSE (J) (C), (D) | |
| | | 2018 | EMSE (J) (B) | |
| | | 2019 | EMSE (J) (E) | |
| | | 2020 | SANER (C) (A) | |
| | | 2019 | ICSE (C) (C) | |
| | | 2019 | TSE (J) | |
| | | 2020 | SP&E (J) | |
| | | 2020 | SAC (C) (B) | |
| | | 2020 | ASE (C) (B) | |
| | | 2021 | MSR (C) (B) | |
| | | 2021 | ICSME (C) (B) | |
| (F) Log maintenance and management (13) | An integrated data-driven framework for computing system management [145]. Cloud application logging for forensics [139]. FLAP: An end-to-end event log analysis platform for system management [152]. Using finite-state models for log differencing [70]. Statistical log differencing [73]. Effective removal of operational log messages: an application to model inference* [143]. Logcomp: Extracting hidden structures via iterative clustering for log compression [54]. A study of the performance of general compressors on log files [104]. Log20: A cost-aware logging mechanism for performance diagnosis [161]. Loghub: A large collection of system log datasets towards automated log analytics [90]. A survey of software log instrumentation [79]. Where shall we log? Studying and suggesting logging locations in code blocks [120]. An exploratory study of log placement recommendation in an enterprise system [112]. DeepLV: Suggesting log levels using ordinal based neural networks [159]. | 2010 | TSMCA (J) (I) | |
| | | 2011 | SAC (C) | |
| | | 2017 | KDD (C) (I), (J) | |
| | | 2018 | ESEC/FSE (C) | |
| | | 2019 | ASE (C) (G) | |
| | | 2019 | ASE (C) (G) | |
| | | 2020 | EMSE (J) | |
| | | 2020 | arXiv (A) (I) | |
| | | 2020 | arXiv (A) (H) | |
| | | 2021 | CSR (J) (B), (C) | |
| | | 2021 | TSE (J) | |
| | | 2021 | TSE (J) | |
| | | 2021 | ESEC/PG (C) | |

*Snowballing

**TABLE 33:** A full list of reviewed publications. ‘Subtopic’ column shows what other topics are discussed in the research, if applicable.
### TABLE 33: A full list of reviewed publications (continued). ‘Subtopic’ column shows what other topics are discussed in the research, if applicable.

| Topic | Paper title | Year | Venue | Subtopic(s) |
|-------|-------------|------|-------|-------------|
| (G) Log parsing (8) | What happened in my network: Mining network events from router sylogs [153] | 2010 | IMC (C) | (J) |
| | Logsig: Generating system events from raw textual logs [151] | 2011 | CSRD (I) | |
| | LogCluster - A data clustering and pattern mining algorithm for event logs [13] | 2015 | CNSM (C) | |
| | Spell: Streaming parsing of system event logs [108] | 2016 | ICDM (C) | |
| | Drain: An online log parsing approach with fixed-depth tree [157] | 2017 | ICWS (C) | |
| | A search-based approach for accurate identification of log message formats [158] | 2018 | ICPC (C) | |
| | Logram: Efficient log parsing using n-gram dictionaries [116] | 2020 | TSE (J) | (L) |
| (H) Anomaly detection (16) | Detecting large-scale system problems by mining console logs* [3] | 2009 | SOSP (C) | (J) |
| | Execution anomaly detection in distributed systems through unstructured log analysis* [4] | 2009 | ICDM (C) | (G) |
| | Mining invariants from console logs for system problem detection* [136] | 2010 | ATC (C) | (G) |
| | Linking resource usage anomalies with system failures from cluster log data [130] | 2013 | SRDS (C) | (I) |
| | Experience report: Log mining using natural language processing and application to anomaly detection [17] | 2017 | ICSE (C) | |
| | Execution anomaly detection in large-scale systems through console log analysis* [123] | 2018 | JSS (J) | (J) |
| | Metric selection and anomaly detection for cloud operations using logs and metric correlation analysis* [124] | 2018 | JSS (J) | (J) |
| | LogAnomaly: Unsupervised detection of sequential and quantitative anomalies in unstructured logs [150] | 2019 | JICAI (C) | (G) |
| | Robust log-based anomaly detection on unstable log data [138] | 2019 | ESEC/FSE (C) | (C) |
| | Anomaly detection via mining numerical workflow relations from logs [125] | 2020 | SRDS (C) | (J) |
| | HitAnomaly: Hierarchical transformers for anomaly detection in system logs [146] | 2020 | TNMS (J) | |
| | LogSayer: Log pattern-driven cloud component anomaly detection with machine learning [136] | 2020 | IWOQ (S) | (J) |
| | LogTransfer: Cross-system log anomaly detection for software systems with transfer learning [113] | 2020 | ISSRE (C) | |
| | Semi-supervised log-based anomaly detection via probabilistic label estimation [110] | 2020 | ICSE (C) | |
| | Log-based anomaly detection without log parsing [77] | 2021 | ASE (C) | |
| (I) Runtime behavior (8) | An approach for mining web service composition patterns from execution logs [155] | 2010 | WSE (C) | |
| | Online detection of multi-component interactions in production systems [133] | 2011 | DSN (C) | (H) |
| | Logmaster: Mining event correlations in logs of large-scale cluster systems [122] | 2012 | SRDS (C) | (J) |
| | Assisting developers of big data analytics applications when deploying on Hadoop clouds [111] | 2013 | ICSE (C) | (J) |
| | Behavioral log analysis with statistical guarantees [69] | 2016 | ICSE (C) | |
| | Performance model derivation of operational systems through log analysis [154] | 2016 | MASCOTS (C) | (J) |
| | Exploring properties and correlations of fatal events in a large-scale hpc system [144] | 2018 | TDPS (J) | (J) |
| | Identifying impactful service problems via log analysis [137] | 2018 | ESEC/FSE (C) | |
| (J) Performance, fault, and failure diagnosis (15) | Assessing and improving the effectiveness of logs for the analysis of software faults [135] | 2010 | DSN (C) | (I) |
| | Diagnosing the root-causes of failures from cluster log files [148] | 2012 | HPDC (C) | |
| | ShenLog: Error diagnosis by connecting clues from run-time logs [6] | 2010 | ASFLOS (C) | (I) |
| | Detection of software failures through event logs: An experimental study [119] | 2012 | ISSRE (C) | (B) |
| | Structured comparative analysis of systems logs to diagnose performance problems* [7] | 2012 | NSDI (C) | |
| | Leveraging performance counters and execution logs to diagnose memory-related performance issues [8] | 2013 | JSS (J) | (I) |
| | lprof: A non-intrusive request flow profiler for distributed systems* [140] | 2013 | ICSE (C) | |
| | POD-Diagnosis: Error diagnosis of sporadic operations on cloud applications* [132] | 2014 | DSN (C) | (J) |
| | Mining system logs to learn error predictors: A case study of a telemetry system [17] | 2015 | EMSE (J) | |
| | Cloudscreer: Workflow monitoring of cloud infrastructures via interleaved logs* [151] | 2016 | ASFLOS (C) | (I) |
| | CRUDE: Combining resource usage data and error logs for accurate error detection in large-scale distributed systems [129] | 2016 | SRDS (C) | (H) |
| | Non-Intrusive performance profiling for entire software stacks based on the flow reconstruction principle [9] | 2016 | OSDI (C) | (I) |
| | Ulrog: Improving log-based fault diagnosis by log analysis [147] | 2016 | JCST (J) | |
| | Pensive: Non-intrusive failure reproduction for distributed systems using the event chaining approach* [10] | 2017 | SOSP (C) | |
| (K) User, business, security, and code coverage (5) | The unified logging infrastructure for data analytics at Twitter [15] | 2012 | VLDB (C) | |
| | Identifying recurrent and unknown performance issues [107] | 2014 | ICMD (C) | (J) |
| | Detection of early-stage enterprise infection by mining large-scale log data [134] | 2015 | DSN (C) | (H) |
| | The bones of the system: A case study of logging and telemetry at Microsoft [12] | 2016 | ICSE (C) | (B, I) |
| | An automated approach to estimating code coverage measures via execution logs [121] | 2018 | ASE (C) | |
| (L) Emerging applications (7) | Operational-log analysis for big data systems: Challenges and solutions [55] | 2016 | IEEE Softw (J) | (A), (B), (I) |
| | Designing PhelkStat: Big data analytics for system event logs [164] | 2017 | HICSS (C) | (H) |
| | Performance evaluation of cloud-based log file analysis with Apache Hadoop and Apache Spark [125] | 2017 | JSS (J) | (A) |
| | An exploratory study on assessing the energy impact of logging on android applications [108] | 2018 | EMSE (J) | (A) |
| | Characterizing the natural language descriptions in software logging statements [122] | 2018 | ASE (C) | (E) |
| | Studying the characteristics of logging practices in mobile apps: a case study on F-Droid [190] | 2019 | EMSE (J) | (A), (B) |
| | On the naturalness and localness of software logs [113] | 2021 | MSR (C) | (H) |

*Snowballing*
M. Kubacki and J. Sosnowski, “Holistic processing and exploring event logs,” in International Workshop on Software Engineering for Resilient Systems. Springer, 2017, pp. 184–200.

B. A. Kitchenham, D. Budgen, and P. Bretenolle, Evidence-based software engineering and systematic reviews. CRC press, 2013, 2013.

R. Vaarandi, “A data clustering algorithm for mining patterns from event logs,” in Proceedings of the 3rd IEEE Workshop on IP Operations & Management (IPOM 2003)/IEEE Cat. No. 03EX764). IEEE, 2003, pp. 119–126.

T. Davis and G. Shaw, SQL Server Transaction Log Management, ser. Stairway handbook. Red Gate Books, 2012. [Online]. Available: https://books.google.ca/books?id=TLcVmAECAAJ

A. Beck, J. Stenum Czeplich, N. Lollike, and S. Malone, “Blockchain—the gateway to trust-free cryptographic transactions,” 2016.

“The Simple Logging Facade for Java (SLF4J),” www.slf4j.org, 2020.

B. Chen and Z. M. J. Jiang, “Studying the use of java logging utilities in the wild,” 2020.

T. Mizouchi, K. Shimari, T. Ishio, and K. Inoue, “Padla: a dynamic log level adapter using online phase detection,” in Proceedings of the 27th International Conference on Program Comprehension. IEEE Press, 2019, pp. 135–138.

“Apache Spark,” https://spark.apache.org/, 2019.

J. MacQueen et al., “Some methods for classification and analysis of multivariate observations,” in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

J. Zhu, S. He, J. Liu, P. He, Q. Xie, Z. Zheng, and M. R. Lyu, “Tools and benchmarks for automated log parsing,” in 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE, 2019, pp. 121–130.

D. Logothetis, C. Trezzo, K. C. Webb, and K. Yocum, “In-situ mapreduce for log processing,” 2011.

J. Liu, J. Zhu, S. He, P. He, Z. Zheng, and M. R. Lyu, “Logzip: Extracting hidden structures via iterative clustering for log compression,” in 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2019, pp. 863–873.

A. Miransky, A. Hamou-Lhadj, E. Ciallini, and A. Larsson, “Operational-log analysis for big data systems: Challenges and solutions,” IEEE Software, vol. 33, no. 2, pp. 52–59, 2016.

B. H. Sigelman, L. A. Barroso, M. Burrows, P. Stephenson, M. Plakal, D. Beaver, S. Jaspan, and C. Shanbhag, “Dapper, a large-scale distributed systems tracing infrastructure,” 2010.

R. Fonseca, G. Porter, R. H. Katz, and S. Shenker, “X-trace: A pervasive network tracing framework,” in 4th (USENIX) Symposium on Networked Systems Design & Implementation ((NSDI) 07), 2007.

P. Barham, R. Isaacs, R. Mortier, and D. Narayanan, “Magpie: Online modelling and performance-aware systems.” in HotOS, 2003, pp. 85–90.

M. K. Aguiler, J. C. Mogul, J. L. Wiener, P. Reynolds, and A. Muthukacharoen, “Performance debugging for distributed systems of black boxes,” ACM SIGOPS Operations Systems Review, vol. 37, no. 5, pp. 74–89, 2003.

R. Wu, X. Xiao, S.-C. Cheung, H. Zhang, and C. Zhang, “Casper: An efficient approach to call trace collection,” ACM SIGPLAN Notices, vol. 51, no. 1, pp. 678–690, 2016.

S. Keele et al., “Guidelines for performing systematic literature reviews in software engineering,” Citeeseer, Tech. Rep., 2007.

K. Petersen, S. Vakkalanka, and L. Kuzniarz, “Guidelines for conducting systematic mapping studies in software engineering: An update,” Information and Software Technology, vol. 64, pp. 1–18, 2015.

B. Napoleo, K. R. Felizardo, É. F. de Souza, and N. L. Vijaykumar, “Practical similarities and differences between systematic literature reviews and systematic mappings: a tertiary study.” in SEKE, 2017, pp. 85–90.

C. A. Kohlin, P. Olofsson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, Experimentation in software engineering. Springer Science & Business Media, 2012.

F. Salfner, M. Lenk, and M. Malek, “A survey of online failure prediction methods,” ACM Computing Surveys (CSUR), vol. 42, no. 3, p. 10, 2010.

T. Greene, “2010–2019: The rise of deep learning,” The Next Web. Available online at: https://thenextweb.com/artificial-intelligence/2020/01/02/2010-2019-the-rise-of-deep-learning. 2010.

C. Wohlin, “Guidelines for snowballing in systematic literature studies and a replication in software engineering,” in Proceedings of the 18th international conference on evaluation and assessment in software engineering, 2014, pp. 1–10.

V. Garousi and J. M. Fernandez, “Highly-cited papers in software engineering: The top-100,” Information and Software Technology, vol. 71, pp. 108–128, 2016.

N. Busany and S. Maoz, “Behavioral log analysis with statistical guarantees,” in 2016 IEEE/ACM 38th International Conference on Software Engineering, IEEE, 2016, pp. 877–887.

H. Amar, L. Bao, N. Busany, D. Lo, and S. Maoz, “Using finite-state models for log differencing,” in Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018, pp. 49–59.

L. Bao, N. Busany, D. Lo, and S. Maoz, “Statistical log differencing,” in 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2019, pp. 851–862.

K. Yao, M. Sayagh, W. Shang, and A. E. Hassan, “Improving state-of-the-art compression techniques for log management tools,” IEEE Transactions on Software Engineering, 2021.

Z. Wang, H. Zhang, T.-H. Chen, and S. Wang, “Would you like a quick peek? providing logging support to monitor data processing in big data applications,” in Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 516–526.

H. Guijar, S. Lal, and H. Li, “An exploratory semantic analysis of logging questions,” Journal of Software: Evolution and Process, vol. 33, no. 7, p. e2361, 2021.

A. R. Chen, T.-H. P. Chen, and S. Wang, “Demystifying the challenges and benefits of analyzing user-reported logs in bug reports,” Empirical Software Engineering, vol. 26, no. 1, pp. 1–30, 2021.

S. Gholamian and P. A. Ward, “What distributed systems say: A study of seven spark application logs,” in 2021 40th International Symposium on Reliable Distributed Systems (SRDS). IEEE, 2021, pp. 222–232.

V.-H. Le and H. Zhang, “Log-based anomaly detection without log parsing,” 2021.

S. He, P. He, Z. Chen, T. Yang, Y. Su, and M. R. Lyu, “A survey on automated log analysis for reliability engineering,” ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1–37, 2021.

B. Chen and Z. M. Jiang, “A survey of software log instrumentation,” ACM Computing Surveys (CSUR), vol. 54, no. 4, pp. 1–34, 2021.

J. Cándido, M. Ariche, and A. van Deursen, “Log-based software monitoring: a systematic mapping study,” PeerJ Computer Science, vol. 7, p. e489, 2021.

D. Yuan, S. Park, P. Huang, Y. Liu, M. M. Lee, X. Tang, Y. Zhou, and S. Savage, “Be conservative: enhancing failure diagnosis with proactive logging,” in Presented as part of the 10th (USENIX) Symposium on Operating Systems Design and Implementation ((OSDI) 12), 2012, pp. 293–304.

D. Yuan, J. Zheng, S. Park, Y. Zhou, and S. Savage, “Improving software diagnosability via log enhancement,” ACM Transactions on Computer Systems (TOCS), vol. 30, no. 1, p. 4, 2012.

W. Shang, Z. M. Jiang, B. Adams, A. E. Hassan, M. W. Godfrey, M. Nasser, and P. Flora, “An exploratory study of the evolution of communicated information about the execution of large software systems,” Journal of Software: Evolution and Process, vol. 26, no. 1, pp. 3–26, 2014.

S. Kabinna, C.-P. Bezemer, W. Shang, M. D. Syer, and A. E. Hassan, “Examining the stability of logging statements,” Empirical Software Engineering, vol. 23, no. 1, pp. 290–333, 2018.

“Datadog: Cloud Monitoring as a Service,” https://www.datadog.com/ 2021.

“Fast, powerful searching over massive volumes of log data,” https://www.loggly.com/ 2021.

“Dynatrace: The Leader in Cloud Monitoring,” https://www.dynatrace.com/ 2021.

“NewRelic: Real-time insights for the modern enterprise,” https://newrelic.com/ 2021.
“XpoLog: Log management & analysis automations,” http://www.xplg.com. 2021.

S. He, J. Zhu, P. He, and M. R. Lyu, “Loghub: a large collection of system log datasets towards automated log analytics,” arXiv preprint arXiv:2008.06448, 2020.

R. Ding, H. Zhou, J.-G. Lou, H. Zhang, Q. Lin, Q. Fu, D. Zhang, and T. Xie, “Log2: A cost-aware logging mechanism for performance diagnosis.” in USENIX Annual Technical Conference, 2015, pp. 139–150.

S. Kabrina, C.-P. Bezemer, W. Shang, and A. E. Hassan, “Logging library migrations: a case study for the apache software foundation projects,” in Mining Software Repositories (MSR), 2016 IEEE/ACM 13th Working Conference on. IEEE, 2016, pp. 154–164.

W. Shang, M. Nagappan, A. E. Hassan, and Z. M. Jiang, “Understanding log lines using development knowledge,” in Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on. IEEE, 2014, pp. 21–30.

G. Kiczales, “J. lamping, a. mendhekar, c. maeda, c. lopes, j.-m. lointiert, and j. irwin. aspect-oriented programming,” Lecture notes in computer science. ECOOP, pp. 220–242, 1997.

“The AspectJ Project.” https://eclipse.org/aspectj/ 2019.

B. Chen and Z. M. J. Jiang, “Characterizing logging practices in java-based open source software projects—a replication study in apache software foundation,” Empirical Software Engineering, vol. 22, no. 1, pp. 330–374, 2017.

B. Chen, G. S. Iosif, and W. Pedrycz, “Mining system logs to learn error predictors: a case study of a telemetry system,” Empirical Software Engineering, vol. 20, no. 4, pp. 879–927, 2015.

K. Yao, G. B. de Padua, W. Shang, C. Sporea, A. Toma, and S. Sajedi, “Log4perf: suggesting and updating logging locations for web-based systems’ performance monitoring,” Empirical Software Engineering, vol. 25, no. 1, pp. 488–531, 2020.

W. Shang, M. Nagappan, and A. E. Hassan, “Studying the relationship between logging characteristics and the code quality of platform software,” Empirical Software Engineering, vol. 20, no. 1, pp. 1–27, 2015.

S. Li, X. Niu, Z. Jia, X. Liao, J. Wang, and T. Li, “Guiding log revisions by learning from software evolution history,” Empirical Software Engineering, pp. 1–39, 2019.

M. Hassani, W. Shang, E. Shiab, and N. Tsantalis, “Studying and detecting log-related issues,” Empirical Software Engineering, vol. 23, no. 6, pp. 3248–3280, 2018.

H. Li, W. Shang, Y. Zou, and A. E. Hassan, “Towards just-in-time suggestions for log changes,” Empirical Software Engineering, vol. 22, no. 4, pp. 1831–1865, 2017.

J. Li, W. Shang, and A. E. Hassan, “Which log level should developers choose for a new logging statement?” Empirical Software Engineering, vol. 22, no. 4, pp. 1684–1716, 2017.

K. Yao, H. Li, W. Shang, and A. E. Hassan, “A study of the performance of general compressors on log files,” Empirical Software Engineering, pp. 3043–3085, 2020.

S. Chowdhury, S. D. Nardo, A. Hindle, and Z. M. J. Jiang, “An exploratory study on assessing the energy impact of logging on android applications,” Empirical Software Engineering, vol. 23, no. 3, pp. 1422–1456, 2018.

Y. Zeng, J. Chen, W. Shang, and T.-H. P. Chen, “Studying the characteristics of logging practices in mobile apps: a case study on f-droid,” Empirical Software Engineering, pp. 1–41, 2019.

M.-H. Lim, J.-G. Lou, H. Zhang, Q. Fu, A. B. J. Teoh, Q. Lin, R. Zhine, and D. Zhang, “Identifying recurrent and unknown performance issues,” in 2014 IEEE International Conference on Data Mining. IEEE, 2014, pp. 320–329.

M. Du and F. Li, “Spell: Streaming parsing of system event logs,” in 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016, pp. 839–846.

Z. Li, L. Li, T.-H. P. Chen, and W. Shang, “Deeply: Suggesting log levels using optimal based neural networks,” in 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). IEEE, 2021, pp. 1461–1472.

L. Yang, J. Chen, Z. Wang, W. Wang, J. Jiang, X. Dong, and W. Zhang, “Semi-supervised log-based anomaly detection via probabilistic label estimation,” in 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). IEEE, 2021, pp. 1448–1460.

W. Shang, Z. M. Jiang, H. Hemmati, B. Adams, A. E. Hassan, and P. Martin, “Assisting developers of big data analytics applications when deploying on hadoop clouds,” in 2013 35th International Conference on Software Engineering (ICSE). IEEE, 2013, pp. 402–411.

J. Cândido, J. Haesen, M. Aniche, and A. van Deursen, “An exploratory study of log placement recommendation in an enterprise system,” in 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR), 2021, pp. 155–166.

R. Li, T.-H. P. Chen, J. Yang, and W. Shang, “Studying duplicate logging statements and their related code clones,” IEEE Transactions on Software Engineering, 2021.

S. Locke, H. Li, T.-H. P. Chen, W. Shang, and W. Liu, “Logassist: Assisting log analysis through log summarization,” IEEE Transactions on Software Engineering, 2021.

H. Dai, H. Li, C. S. Chen, W. Shang, and T.-H. Chen, “Logram: Efficient log parsing using n-gram dictionaries,” IEEE Transactions on Software Engineering, 2021.

Z. Li, X. Xia, D. Lo, Z. Xing, A. E. Hassan, and S. Li, “Which variables should i log?” IEEE Transactions on Software Engineering, 2019.

R. Chen, S. Zhang, D. Li, Y. Zhang, F. Guo, W. Meng, D. Pei, Y. Zhang, X. Chen, and Y. Liu, “Logtransfer: Cross-system log anomaly detection for software systems with transfer learning,” in 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2020, pp. 37–47.

A. Pecchia and S. Russo, “Detection of software failures through event logs: An experimental study,” in 2012 IEEE 23rd International Symposium on Software Reliability Engineering. IEEE, 2012, pp. 31–40.

Z. Li, T.-H. Chen, and W. Shang, “Where shall we log? studying and suggesting logging locations in code blocks,” in 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2020, pp. 361–372.

B. Chen, J. Song, P. Xu, X. Hu, and Z. M. J. Jiang, “An automated approach to estimating code coverage measures via execution logs,” in Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. ACM, 2018, pp. 305–316.

P. He, Z. Chen, S. He, and M. R. Lyu, “Characterizing the natural language descriptions in software logging statements,” in 2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2018, pp. 178–189.

L. Bao, Q. Li, P. Lu, J. Lu, T. Ruan, and K. Zhang, “Execution anomaly detection in large-scale systems through console log analysis,” Journal of Systems and Software, vol. 143, pp. 172–186, 2018.

M. Farshchi, J.-G. Schneider, I. Weber, and J. Grundy, “Metric selection and anomaly detection for cloud operations using log and metric correlation analysis,” Journal of Systems and Software, vol. 137, pp. 531–549, 2018.

I. Mavridis and H. Karatza, “Performance evaluation of cloud-based log file analysis with apache hadoop and apache spark,” Journal of Systems and Software, vol. 129, pp. 133–151, 2017.

Z. Jia, S. Li, X. Liu, X. Liao, and Y. Liu, “Smartlog: Place error log statement by deep understanding of log intention,” in 2018 IEEE 25th International Conference on Software Analysis, Evolution and Reengineering (SANER). IEEE, 2018, pp. 61–71.

X. Fu, R. Ren, J. Zhan, W. Zhou, Z. Jia, and G. Lu, “Logmaster: Mining event correlations in logs of large-scale cluster systems,” in 2012 IEEE 31st Symposium on Reliable Distributed Systems. IEEE, 2012, pp. 71–80.

B. Zhang, H. Zhang, P. Moscato, and A. Zhang, “Anomaly detection via mining numerical workflow relations from logs,” in 2020 International Symposium on Reliable Distributed Systems (SRDS). IEEE, 2020, pp. 195–204.

N. Gurumdimma, A. Jhumka, M. Liakata, E. Chuah, and J. Browne, “Crude: Combining resource usage data and error logs for accurate error detection in large-scale distributed systems,” in 2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS). IEEE, 2016, pp. 51–60.

E. Chuah, A. Jhumka, S. Narasimhamurthy, J. Hammond, J. C. Browne, and B. Barth, “Linking resource usage anomalies with system failures from cluster log data,” in 2013 IEEE 32nd International Conference on Dependable Systems and Networks (DSN). IEEE, 2013, pp. 63–72.
tional Symposium on Reliable Distributed Systems. IEEE, 2013, pp. 111–120.

[113] X. Yu, P. Joshi, J. Xu, G. Jin, H. Zhang, and G. Jiang, “Cloudseer: Workflow monitoring of cloud infrastructures via interleave logs,” vol. 44, no. 2, pp. 489–502, 2016.

[114] X. Yuan, L. Zha, L. Weber, L. Bass, and D. Sun, “Pod-diagnosis: Error diagnosis of sporadic operations on cloud applications,” in 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. IEEE, 2014, pp. 252–263.

[115] A. J. Oliner and A. Aiiken, “Online detection of multi-component interactions in production systems,” in 2011 IEEE/IFIP 41st International Conference on Dependable Systems & Networks (DSN). IEEE, 2011, pp. 38–40.

[116] A. Oprea, Z. Li, T.-F. Yen, S. H. Chin, and S. Alrwais, “Detection of early-stage enterprise infection by mining large-scale log data,” in 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. IEEE, 2015, pp. 45–56.

[117] M. Cinque, D. Cotroneo, R. Natella, and A. Pecchia, “Assessing and improving the effectiveness of logs for the analysis of software faults,” in 2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN). IEEE, 2010, pp. 457–466.

[118] J.-G. Lou, Q. Fu, S. Yang, Y. Xu, and J. Li, “Mining invariants from console logs for system problem detection.” in USENIX Annual Technical Conference, 2010, pp. 23–25.

[119] S. He, Q. Lin, J.-G. Lou, H. Zhang, M. R. Lyu, and D. Zhang, “Detecting impactful service system problems via log analysis,” in Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018, pp. 60–70.

[120] X. Zhang, Y. Xu, Q. Lin, B. Qiao, H. Zhang, Y. Dang, C. Xie, X. Yang, Q. Cheng, Z. Li et al., “Robust log-based anomaly detection on unstable log data,” in Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2019, pp. 807–817.

[121] R. Marty, “Cloud application logging for forensics,” in proceedings of the 2011 ACM Symposium on Applied Computing, 2011, pp. 178–184.

[122] 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 11th [USENIX] Symposium on Operating Systems Design and Implementation (OSDI) 14, 2014, pp. 629–644.

[123] D. Yuan, Y. Luo, X. Zhuang, G. R. Rodrigues, X. Zhao, Y. Zhang, P. U. Jain, and M. Stumm, “Simple testing can prevent most critical failures: An analysis of production failures in distributed data-intensive systems,” in 11th [USENIX] Symposium on Operating Systems Design and Implementation (OSDI) 14, 2014, pp. 249–265.

[124] T. Kim, S. Kim, S. Park, and Y. Park, “Automatic recommendation to appropriate log levels,” Software: Practice and Experience, vol. 50, no. 3, pp. 189–209, 2020.

[125] D. Shin, D. Bianculli, and L. Briand, “Effective removal of operational log messages: an application to model inference,” arXiv preprint arXiv:2004.07194, 2020.

[126] S. Di, H. Guo, R. Gupta, E. R. Pershey, M. Snir, and F. Cappello, “Exploring properties and correlations of fatal events in a large-scale hpc system,” IEEE Transactions on Parallel and Distributed Systems, vol. 30, no. 2, pp. 361–374, 2018.

[127] T. Li, W. Peng, C. Perrig, S. Ma, and H. Wang, “An integrated data-driven framework for computing system management,” IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 40, no. 1, pp. 90–99, 2010.

[128] S. Huang, Y. Liu, C. Fung, R. He, Y. Zhao, H. Yang, and Z. Luan, “Hitanomaly: Hierarchical transformers for anomaly detection in system log,” IEEE Transactions on Network and Service Management, vol. 17, no. 4, pp. 2064–2076, 2020.

[129] D.-Q. You, H. Qin, and H. Jin, “Ulog: Improving log-based fault diagnosis by log analysis,” Journal of computer science and technology, vol. 31, no. 5, pp. 1038–1052, 2016.

[130] E. Chuah, S.-h. Kuo, P. Hiew, W.-C. Tji, G. Lee, J. Hammond, M. T. Michalewicz, T. Hung, and J. C. Browne, “Diagnosing the root-causes of failures from cluster log files,” in 2010 International Conference on High Performance Computing - HC2010. IEEE, 2010, pp. 1–10.

[131] N. Taerat, J. Brandt, A. Gentile, M. Wong, and C. Leangsuksun, “Baler: deterministic, lossless log message clustering tool,” Computer Science-Research and Development, vol. 26, no. 3, pp. 285–295, 2011.
