On Vulnerability and Security Log analysis: A Systematic Literature Review on Recent Trends

Jan Svacina
Baylor University
Waco, TX
jan_svacina2@baylor.edu

Jackson Raffety
Baylor University
Waco, TX
Jackson_Raffety1@baylor.edu

Connor Woodahl
Baylor University
Waco, TX
Connor_Woodahl1@baylor.edu

Brooklynn Stone
Baylor University
Waco, TX
brooklynn_stone1@baylor.edu

Tomas Cerny
Baylor University
Waco, TX
Tomas_Cerny@baylor.edu

Miroslav Bures
Czech Technical University
Prague, Czech Republic
buresm3@fel.cvut.cz

Dongwan Shin
New Mexico Tech
Socorro, NM
dongwan.shin@nmt.edu

Karel Frajtak
Czech Technical University
Prague, Czech Republic
kfrajtak@gmail.com

Pavel Tisnovsky
Red Hat
Brno, Czech Republic
ptisnovs@redhat.com

ABSTRACT

Log analysis is a technique of deriving knowledge from log files containing records of events in a computer system. A common application of log analysis is to derive critical information about a system's security issues and intrusions, which subsequently leads to being able to identify and potentially stop intruders attacking the system. However, many systems produce a high volume of log data with high frequency, posing serious challenges in analysis. This paper contributes with a systematic literature review and discusses current trends, advancements, and future directions in log security analysis within the past decade. We summarized current research strategies with respect to technology approaches from 34 current publications. We identified limitations that poses challenges to future research and opened discussion on issues towards logging mechanism in the software systems. Findings of this study are relevant for software systems as well as software parts of the Internet of Things (IoT) systems.

CCS CONCEPTS

• **Software and its engineering → Software maintenance tools**; **Software infrastructure**; • **Information systems → Web log analysis**; **Query log analysis**; *Data mining*; • **Networks →** *Network experimentation*; *Cloud computing*.

KEYWORDS

Log Analysis, Log Mining, Anomaly Detection, Intrusion Detection, Machine Learning

RACS '20, October 13–16, 2020, Gwangju, Republic of Korea

ACM ISBN 978-1-4503-8025-6/20/10…$15.00
https://doi.org/10.1145/3400286.3418261

ACM Reference Format:
Jan Svacina, Jackson Raffety, Connor Woodahl, Brooklynn Stone, Tomas Cerny, Miroslav Bures, Dongwan Shin, Karel Frajtak, and Pavel Tisnovsky. 2020. On Vulnerability and Security Log analysis: A Systematic Literature Review on Recent Trends. In *International Conference on Research in Adaptive and Convergent Systems (RACS '20), October 13–16, 2020, Gwangju, Republic of Korea.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3400286.3418261

1 INTRODUCTION

Software systems produce large amounts of log data as the program executes. These logs are usually produced in order to help the developers and system administrators detect defects, runtime errors, and unexpected output. These logs can also be used for intrusion detection. Through the tracking of key events in log data, anomalous behavior and signatures can be identified as attack or intrusion of the system and mitigated. Unfortunately, most software systems produce an inordinate amount of log data, making it an infeasible task to analyze the logs manually. Moreover, it is often the case that threats cannot be detected from individual log entries, but from a pattern of log entries spread throughout the file [33]. To this end, the automation of analyzing the log data becomes necessary.

In the cybersecurity of software systems, most effort is put in the protection against external threats [43]. Internal threats and intrusion detection are often overlooked fields due to the difficulty of the task [43]. If left overlooked, a large monetary loss can be accumulated due to the lack of action in attacking these threats [15]. The longer these threats are allowed to fester within the system, the more damage is caused. Log analysis is one of the most popular methods of detecting these threats. Traditionally, one or more experts would manually analyze the logs for anomalies, but as systems grow exponentially, so does the number of experts required to efficiently and accurately do their job [18] and automation becomes a necessity.

The goal of this study is to identify relevant existing works and determine what the current fields of research are and their progress.

We performed a systematic approach [38] to collect, filter, and analyze the available research publications. In our research, we summarized numerous different strategies and methods in analyzing logs for security purposes. Anomaly detection, machine learning, and clustering were all dominant topics in the papers we identified and assessed, with a noticed trend in the increase in popularity of machine learning. Along with these methods, we also identified current limitations and problems in the subject, which have been addressed but require more research. Such fields include real-time analysis [4], multi-source analysis [37], speed and accuracy of the algorithms [47], and universal log formatting [44]. In this paper, we report in detail on these findings. Discussed trends and issues are relevant for a wide variety of software systems or components using logging mechanisms. These systems include enterprise and control software systems as well as software components of the Internet of Things (IoT) systems, in which various security vulnerabilities are reported as a significant issue in the recent period [2, 7].

The rest of the paper is organized as follows: Section 2 contains the motivation for this paper, as well as related work discovered. In Section 3, we discuss the method we used in our research and our established research questions. Section 4 discusses the results and findings to our research, as well as answers to our research questions. Threats to the validity of our research are addressed in Section 5. Finally, we conclude and summarize the results of our work in Section 6.

2 MOTIVATION AND RELATED WORK

Several factors incentivized the decision to carry out this research. First off, while many studies exist regarding the topic of log analysis, few if any focus on both the process of analysis and security aspects. Second, three similar studies dating up to as late as 2016 were found [1, 14, 20], so a more up to date study is needed. Third, a mapping study over the work and research about the topic of log analysis and security will benefit peer researchers trying to address research questions concerning this topic.

The most recent study by ElTayeby et al. [14] was similar in structure and methodology, but used logs for a different purpose. While their work was used to analyze logs, their research was to extract user interaction and insights from log analysis rather than anomalies and security aspects of a system. Therefore, the papers reviewed in their survey on the topic of log analysis would be a complement to our research, yet deviates from ours due to the difference in focus.

The research performed by Hussain et al. is in the form of a mapping study as well. It provides insights into how to pre-process web logs, effectively removing the majority of unneeded information. Since this mapping study was created in 2010, it only covers web usage mining techniques between the years 2000 and 2010. Therefore, our study provides more relevant research, which includes cloud-based systems, machine learning, and high-end security techniques, which were not as prevalent in the early 2000's.

Finally, the research performed by Agosti et al. [20] is similar in concept to our study, as it is a mapping study over the research performed over log analysis from the previous decade. Their research covers Web search engines, log analysis, and Digital Library Systems log analysis. However, this mapping study provides a similar issue to the previous study in that it presents outdated research. It also lacks a focus on security, which is the main focus of our mapping study.

3 METHOD

In this study, we followed a process for gathering data, analyzing the data, and delivering results based on our findings, similar to a methodology presented by Kitchenham et al. [26]. As a software engineering-focused research team, we took advantage of the systematic approach known as a mapping study [38].

In the first stage of a mapping study, we defined the research questions and over the time, refined them in order to represent a holistic approach to our topic: Security and Log Analysis. Once a rough draft of research questions was developed, we defined search terms that would be used in our search query across various databases. Once existing related works were identified, we filtered out papers that did not fit the scope of our study. After refining our research questions, search query, and related works, we assessed the remaining papers and found information that assisted us in answering our research questions.

After a discussion of possible research directions, we formed the following research questions:

- RQ1: What topics/subjects have been addressed in the log analysis research, and what is their distribution? (log mining, log security)?
- RQ2: What are the existing strategies, tools, and techniques for log analysis and security?
- RQ3: What kinds of benchmarks (industrial or otherwise) are used to evaluate log taxonomy/analysis techniques?
- RQ4: What are the possible directions for future research?
- RQ5: How much information can be extracted from logs?

We provide a summary of our results to these questions in the discussion portion of our results.

Based on practices used in [38], our research direction of log analysis and log type classification indicated that we would be using indexing servers ACM, IEEE, Science Direct, and SpringerLink. The resulting search string for ACM, IEEE, Science Direct, and Springer Link can be found under listing 1.

For our discussion regarding log analysis and security, a search string would need to combine the two to produce the desired paper topics. The search string then becomes a set of two phrases combined with an AND. The first part of the search string will pull papers related to log analysis and log mining. Combined with the second portion, which specifies topics related to threats, attacks, compromise, anomalies, and other security-related terms, the search results are narrowed down to log analysis and security/intrusion detection.

The results of our search query can be found in Table 1 [1]. The first step in narrowing down the results was to eliminate papers with unrelated or far off topics based on the title of the paper. We selected 34 relevant papers from 1374 publications for further analysis. Next, we read through the abstract of each paper, ruling out papers that went into topics, which were unrelated to log mining and security.

Listing 1: The Search Query for Research Indexing Sites

```
(log mining OR log analysis) AND (insider threats OR insider attack OR compromise
OR access control OR access policy OR SIEM OR intrusion* OR anomal* OR privac* )
```

Table 1: Search Query Results for Various Index Sites

Indexer	Found results	Used results
ACM DL	199	8
IEEE Xplore	130	11
SpringerLink	224	10
ScienceDirect	821[1]	5
Total	1374	34

Finally, once we narrowed down the related works to 34 papers, we studied the various topics covered by other researchers, analyzed their findings, and have presented our final analysis in the following portions of this paper (Although 37 related papers were found and used, we have chosen to cite 34 due to brevity and papers with related topics).

4 RESULTS

In this section, we present findings of this study split to existing strategies and discussion of current problems and possible future research directions.

4.1 Existing Strategies

The traditional method of detecting security issues through log files has required that an expert manually examine the log files themselves. However, this is often infeasible due to the large amounts of log data produced in modern applications. For this reason, numerous strategies to automate the process of analyzing logs for security threats and intrusion detection have been made.

Strategies for intrusion detection can generally be placed in one of two categories: *signature-based* detection and *anomaly-based* detection. Signature-based detection relies on a set of rules as input to determine which patterns of logs should be flagged. These rules are generally derived from past or known attack patterns. While this strategy works for attacks that match the set of rules and is simple, they fail to identify and flag novel types of attack. Anomaly-based detection instead determines which sort of behavior is abnormal by comparing it to other logs in the same stream. This strategy allows for all anomalous events to be captured, even those which signature-based detection would detect. However, this strategy tends to have numerous amounts of false-positive flags, as not all anomalous behavior are examples of intrusion. One paper claimed that anomaly-based detection works poorly on it's own, and works best when paired with signature-based detection [23], and another created a hybrid method using both signature and anomaly-based detection [35]. However, of the works we reviewed, a majority were focused solely on anomaly detection [5, 6, 15, 18, 19, 23, 27, 36, 40–42, 46].

[1]Since ScienceDirect limits the number of boolean connectors we combined the results using BibDesk tool [34])

In most of the analyzed studies, machine learning was a topic that held some importance in the paper [3, 6, 18, 23, 29, 32, 39–41]. The previously mentioned strategy, anomaly-based detection, use machine learning as the core of the algorithm. While anomaly-based detection is typically supervised learning, the degree to which it is supervised differs depending on the implementation. The use of machine learning also differs. While used in anomaly-detection, it is also used to determine the semantic meaning of log messages in order to classify and analyze them. One work [3] discussed using Natural Language Processing for this purpose.

Clustering is a specific branch of machine learning with a unique focus in log-analysis. Over time, it becomes difficult to detect attacks by analyzing individual log statements. Thus, clustering is often used to group log statements together and then compare clusters with each other to determine anomalies. A decent number of papers discussed clustering in use with log analysis [13, 18, 22, 31, 36, 41]. A very recent survey [30] goes in-depth into the applications of log clustering for security purposes.

4.2 Answers to RQ's and Discussion

From our research with related work, we can answer RQ1 (What topics/subjects have been addressed in the log analysis research, and what is their distribution?) with numerous aforementioned directions. Topics such as log mining and anomaly detection were present in the majority of the papers. Machine learning, log security, intrusion detection, and clustering were also present in many papers found in our related work.

The answer to RQ2 (What are the existing strategies, tools, and techniques for log taxonomy?) is a very long list of different techniques and tools found in our related work. The most common technique was machine learning found within four papers [3, 6, 39, 40]. We found three papers performing a comparative study over available log analyzers and numerous techniques [18, 22, 24]. The topic of clustering algorithms being used as method for log can be found within [13, 18, 22, 30, 36, 41]. Three papers created their own tool or process: Dilaf [5], Beehive[44], and UiLog [48]. Several other processes were used such as: process mining[36], text analysis[21], word frequency [27] and [47], data mining [16], event correlation[4], statistical analysis [17], and the use of tree structures[25]. Lastly, one paper tried to explore the idea of leveraging bioinformatics tools by using a method for re-coding log data into the alphabet used for representing amino acid sequences, which enables the application of high-performance bioinformatics tools for outlier detection in the domain of log data processing [43].

In regard to RQ3 (What kinds of benchmarks (industrial or otherwise) are used to evaluate log analysis techniques?), number of publications used logs from web based systems [32], [40], [16], [21], [23], [20], two systems used cloud based systems [48], [24] and finally two publications used IoT (Internet of Things) systems [15], [39]. Other publications did not specified further the source of their benchmarks.

Table 2: Papers

ID	Ref	Paper Title	Year
1	[17]	Abnormality analysis of streamed log data	2014
2	[19]	AD2: Anomaly detection on active directory log data for insider threat monitoring	2015
3	[33]	Investigating event log analysis with minimum apriori information	2013
4	[27]	Multidimensional Log Analysis	2016
5	[35]	Operational-Log Analysis for Big Data Systems: Challenges and Solutions	2016
6	[42]	An Approach of Anomaly Diagnosis with Logs for Distributed Services in Communication Network Information System	2017
7	[48]	UiLog: Improving Log-Based Fault Diagnosis by Log Analysis	2016
8	[24]	Cloud Log Forensics	2016
9	[22]	Towards structured log analysis	2012
10	[5]	Incremental Analysis of Large-Scale System Logs for Anomaly Detection	2019
11	[28]	Scalable intrusion detection systems log analysis using cloud computing infrastructure	2013
12	[18]	Experience Report: System Log Analysis for Anomaly Detection	2016
13	[44]	Beehive: large-scale log analysis for detecting suspicious activity in enterprise networks	2013
14	[43]	Discovering Insider Threats from Log Data with High-Performance Bioinformatics Tools	2016
15	[12]	Semantic Mediation for A Posteriori Log Analysis	2019
16	[13]	Taming the logs - Vocabularies for semantic security analysis	2018
17	[36]	Discovering process models for the analysis of application failures under uncertainty of event logs	2020
18	[3]	Detect and correlate information system events through verbose logging messages analysis	2018
19	[30]	System log clustering approaches for cyber security applications: A survey	2020
20	[41]	Graph clustering and anomaly detection of access control log for forensic purposes	2017
21	[4]	Insider Threat Detection Using Log Analysis and Event Correlation	2015
22	[6]	Execution anomaly detection in large-scale systems through console log analysis	2018
23	[21]	Normalizing Security Events with a Hierarchical Knowledge Base	2015
24	[23]	Online anomaly detection using dimensionality reduction techniques for HTTP log analysis	2015
25	[25]	Fast attack detection system using log analysis and attack tree generation	2018
26	[37]	Hercule: attack story reconstruction via community discovery on correlated log graph	2016
27	[15]	A log mining approach for process monitoring in SCADA	2012
28	[16]	Visualization of System Log Files for Post-incident Analysis and Response	2014
29	[32]	Improving the system log analysis with language model and semi-supervised classifier	2019
30	[39]	Mining system logs to learn error predictors: a case study of a telemetry system	2014
31	[40]	Anomaly Detection from Network Logs Using Diffusion Maps	2011
32	[47]	Improving Log-Based Fault Diagnosis by Log Classification	2014
33	[29]	Practical Machine Learning for Cloud Intrusion Detection: Challenges and the Way Forward	2017
34	[45]	Digging Evidence for Violation of Cloud Security Compliance with Knowledge Learned from Logs	2018

RQ4 (What are the possible directions for future research?) has multiple suggestions for future research. Most suggestions for future research centered around converting their tools or methods from post occurrence to real-time analysis [5, 6, 13, 30]. Other suggestions for future research were to improve the tools and methods introduced by the research paper [5, 28, 43, 44] to make the results of log analysis more efficient or accurate. Lastly, one suggestion for future research was to look for alternate sources for anomaly detection other than logs.

To answer RQ5(How much information can be extracted from logs?), we did an analysis to conclude that most, if not all, necessary information about a system for security and anomalies are available within logs if extracted properly [35]. This information can lead to system deviation [15], mitigation strategies [15], malicious behavior[36], system performance, and system health [24].

4.3 Current Problems and Future Research

A problem addressed by numerous papers is that most log analyses are performed post-mortem. While these analyses can alert system administrators of attacks, they cannot quickly address the attack while in action. As a result, more emphasis has been put on real-time approaches to intrusion detection. Multiple existing works [4, 25, 30, 47] look into real-time log analysis. One work [4] found that a real-time analysis was actually able to improve performance, as well as security, due to being able to stop threats in real-time.

Another issue addressed by a few studies is that as the scale of systems increase, the number of log files increases as well. With separated statements, attacks can become impossible to detect by analyzing only a single source. Multi-source log analysis then becomes necessary to correctly identify anomalies and threats. Research by Pei et al. proposed a fast and accurate intrusion detection system using multi-source log analysis, called HERCULE [37]. Their study

showed low false-flag rates and high accuracy. Their suggested future research idea was to experiment with log files distributed across multiple hosts.

Because machine learning is used so frequently, tackling the limitations of machine learning are also areas of future research. These algorithms tend to be very resource-intensive and slow. These limitations also pose problems with the above strategies. Real-time analysis is hindered by the slow speed of machine learning, and multi-source log analysis is bogged down even more. Another area to look into is to reduce the number of false flags from the results by being able to accurately identify log statements that are rare, but not anomalous [6].

Another discussed topic is the inconsistent formatting of logs across different files and systems[35, 44]. These different formats can make clustering, analysis, and anomaly detection much more difficult. One paper's solution to this problem and recommendation for future work from their research was to standardize log format [22].

5 THREATS TO VALIDITY

The main threat to validity for this mapping study is the omission of relevant research. To mitigate this threat, the aforementioned search strings used for finding research papers were made to be as broad as possible, and the resulting paper counts for each database confirm that. While our broad search string allowed for a vast amount of resulting papers, we only allowed for papers from 2010 or later, which could potentially omit valid yet older research. Lastly, we manually reviewed the resulting papers, reading the title, abstract, and keywords to filter from the hundreds of results, which might have lead to research wrongly left out. In the next step, we had multiple reviewers for each potentially applicable paper so as not to have a single researcher invalidate the research. If either reviewer believed it was applicable, we left it for the deeper analysis so as not to leave out potentially applicable research.

6 CONCLUSION

Security intrusions in a computer system lead to severe consequences and monetary losses if left undetected. Manual analysis of a log file for intrusion detection is ineffective and has limited resources. However, an automatic approach has challenges in terms of data volume and velocity. In the past decade, researchers conducted numerous studies in tackling the issue using various techniques. In this paper, we selected 34 relevant papers on security aspects of log analysis form 1374 publications between 2010 and 2020.

We analyzed papers based on the research questions and identified their relevant characteristics, and common techniques, such as machine learning, data mining, and text analysis. Next, we identified areas that pose significant challenges in the current research, for instance, real-time analysis, universal log formatting, and multi-source analysis. Our study contains publications relevant to web systems, cloud computing, and IoT (Internet of Things).

In future research, we are going to focus on categorizing types of log and their distribution. Next, we want to focus on detecting similar log items across multiple log files. Our long term goal is to create a complex approach that encompass security analysis together with failure prediction [11] and constraint checking [10]. Furthermore, we also aim to connect security log analysis together with source code [8, 9].

REFERENCES

[1] Maristella Agosti, Franco Crivellari, and Giorgio Maria Di Nunzio. 2011. Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery* 24, 3 (2011), 663–696. https://doi.org/10.1007/s10618-011-0228-8

[2] Bestoun S Ahmed, Miroslav Bures, Karel Frajtak, and Tomas Cerny. 2019. Aspects of quality in Internet of Things (IoT) solutions: A systematic mapping study. *IEEE Access* 7 (2019), 13758–13780.

[3] Flora Amato, Giovanni Cozzolino, Antonino Mazzeo, and Francesco Moscato. 2018. Detect and correlate information system events through verbose logging messages analysis. *Computing* 101, 7 (2018), 819–830. https://doi.org/10.1007/s00607-018-0662-1

[4] Amruta Ambre and Narendra Shekokar. 2015. Insider Threat Detection Using Log Analysis and Event Correlation. *Procedia Computer Science* 45 (2015), 436–445. https://doi.org/10.1016/j.procs.2015.03.175

[5] M. Astekin, S. Özcan, and H. Sözer. 2019. Incremental Analysis of Large-Scale System Logs for Anomaly Detection. In *2019 IEEE International Conference on Big Data (Big Data)*. 2119–2127.

[6] Liang Bao, Qian Li, Peiyao Lu, Jie Lu, Tongxiao Ruan, and Ke Zhang. 2018. Execution anomaly detection in large-scale systems through console log analysis. *Journal of Systems and Software* 143 (2018), 172–186. https://doi.org/10.1016/j.jss.2018.05.016

[7] Miroslav Bures, Tomas Cerny, and Bestoun S Ahmed. 2018. Internet of things: Current challenges in the quality assurance and testing methods. In *International Conference on Information Science and Applications*. Springer, 625–634.

[8] Vincent Bushong, Russell Sanders, Jacob Curtis, Mark Du, Tomas Cerny, Karel Frajtak, Miroslav Bures, Pavel Tisnovsky, and Dongwan Shin. 2020. On Matching Log Analysis to Source Code: A Systematic Mapping Study. In *International Conference on Research in Adaptive and Convergent Systems(RACS '20) (RACS '20)*. ACM, New York, NY, USA, 1–6. https://doi.org/10.1145/3400286.3418262

[9] Tomas Cerny, Jan Svacina, Dipta Das, Vincent Bushong, Miroslav Bures, Pavel Tisnovsky, Karel Frajtak, Dongwan Shin, and Jun Huang. 2020. On Code Analysis Opportunities and Challenges for Enterprise Systems and Microservices. *IEEE Access* (2020), 1–22. https://doi.org/10.1109/ACCESS.2020.3019985

[10] Tomas Cerny, Andrew Walker, Vincent Bushong, Dipta Das, Karel Frajtak, Miroslav Bures, and Pavel Tisnovsky. 2020. Mapping Study on Constraint Consistency Checking in Distributed Enterprise Systems. In *International Conference on Research in Adaptive and Convergent Systems(RACS '20) (RACS '20)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3400286.34182571

[11] Dipta Das, Micah Schiewe, Elizabeth Brighton, Mark Fuller, Tomas Cerny, Miroslav Bures, Karel Frajtak, Dongwan Shin, and Pavel Tisnovsky. 2020. Failure Prediction by Utilizing Log Analysis: A Systematic Mapping Study. In *International Conference on Research in Adaptive and Convergent Systems(RACS '20) (RACS '20)*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3400286.3418263

[12] Farah Dernaika, Nora Cuppens-Boulahia, Frédéric Cuppens, and Olivier Raynaud. 2019. Semantic Mediation for A Posteriori Log Analysis. *Proceedings of the 14th International Conference on Availability, Reliability and Security - ARES 19* (2019). https://doi.org/10.1145/3339252.3340104

[13] Andreas Ekelhart, Elmar Kiesling, and Kabul Kurniawan. 2018. Taming the logs - Vocabularies for semantic security analysis. *Procedia Computer Science* 137 (2018), 109–119. https://doi.org/10.1016/j.procs.2018.09.011

[14] Omar ElTayeby and Wenwen Dou. 2016. A Survey on Interaction Log Analysis for Evaluating Exploratory Visualizations. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV '16)*. Association for Computing Machinery, New York, NY, USA, 62–69. https://doi.org/10.1145/2993901.2993912

[15] Dina Hadžiosmanović, Damiano Bolzoni, and Pieter H. Hartel. 2012. A log mining approach for process monitoring in SCADA. *International Journal of Information Security* 11, 4 (2012), 231–251. https://doi.org/10.1007/s10207-012-0163-8

[16] John Haggerty and Thomas Hughes-Roberts. 2014. Visualization of System Log Files for Post-incident Analysis and Response. *Lecture Notes in Computer Science Human Aspects of Information Security, Privacy, and Trust* (2014), 23–32. https://doi.org/10.1007/978-3-319-07620-1_3

[17] A. N. Harutyunyan, A. V. Poghosyan, N. M. Grigoryan, and M. A. Marvasti. 2014. Abnormality analysis of streamed log data. In *2014 IEEE Network Operations and Management Symposium (NOMS)*. 1–7.

[18] S. He, J. Zhu, P. He, and M. R. Lyu. 2016. Experience Report: System Log Analysis for Anomaly Detection. In *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*. 207–218.

[19] C. Hsieh, C. Lai, C. Mao, T. Kao, and K. Lee. 2015. AD2: Anomaly detection on active directory log data for insider threat monitoring. In *2015 International Carnahan Conference on Security Technology (ICCST)*. 287–292.

[20] Tasawar Hussain, Sohail Asghar, and Nayyer Masood. 2010. Web usage mining: A survey on preprocessing of web log file. *2010 International Conference on Information and Emerging Technologies* (2010). https://doi.org/10.1109/iciet.2010.5625730

[21] David Jaeger, Amir Azodi, Feng Cheng, and Christoph Meinel. 2015. Normalizing Security Events with a Hierarchical Knowledge Base. *Information Security Theory and Practice Lecture Notes in Computer Science* (2015), 237–248. https://doi.org/10.1007/978-3-319-24018-3_15

[22] D. Jayathilake. 2012. Towards structured log analysis. In *2012 Ninth International Conference on Computer Science and Software Engineering (JCSSE)*. 259–264.

[23] Antti Juvonen, Tuomo Sipola, and Timo Hämäläinen. 2015. Online anomaly detection using dimensionality reduction techniques for HTTP log analysis. *Computer Networks* 91 (2015), 46–56. https://doi.org/10.1016/j.comnet.2015.07.019

[24] Suleman Khan, Abdullah Gani, Ainuddin Wahid Abdul Wahab, Mustapha Aminu Bagiwa, Muhammad Shiraz, Samee U. Khan, Rajkumar Buyya, and Albert Y. Zomaya. 2016. Cloud Log Forensics. *Comput. Surveys* 49, 1 (2016), 1–42. https://doi.org/10.1145/2906149

[25] Duhoe Kim, Yong-Hyun Kim, Dongil Shin, and Dongkyoo Shin. 2018. Fast attack detection system using log analysis and attack tree generation. *Cluster Computing* 22, S1 (2018), 1827–1835. https://doi.org/10.1007/s10586-018-2269-x

[26] B. Kitchenham and S Charters. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering. (2007).

[27] M. Kubacki and J. Sosnowski. 2016. Multidimensional Log Analysis. In *2016 12th European Dependable Computing Conference (EDCC)*. 193–196.

[28] M. Kumar and M. Hanumanthappa. 2013. Scalable intrusion detection systems log analysis using cloud computing infrastructure. In *2013 IEEE International Conference on Computational Intelligence and Computing Research*. 1–4.

[29] Ram Shankar Siva Kumar, Andrew Wicker, and Matt Swann. 2017. Practical Machine Learning forCloud Intrusion Detection: Challenges and the Way Forward. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec '17)*. Association for Computing Machinery, New York, NY, USA, 81–90. https://doi.org/10.1145/3128572.3140445

[30] Max Landauer, Florian Skopik, Markus Wurzenberger, and Andreas Rauber. 2020. System log clustering approaches for cyber security applications: A survey. *Computers & Security* 92 (2020), 101739. https://doi.org/10.1016/j.cose.2020.101739

[31] Marcello Leida, Paolo Ceravolo, Ernesto Damiani, Rasool Asal, and Maurizio Colombo. 2019. Dynamic Access Control to Semantics-Aware Streamed Process Logs. *Journal on Data Semantics* 8, 3 (2019), 203–218. https://doi.org/10.1007/s13740-019-00106-2

[32] Guofu Li, Pengjia Zhu, Ning Cao, Mei Wu, Zhiyi Chen, Guangsheng Cao, Hongjun Li, and Chenjing Gong. 2019. Improving the system log analysis with language model and semi-supervised classifier. *Multimedia Tools and Applications* 78, 15 (2019), 21521–21535. https://doi.org/10.1007/s11042-018-7020-3

[33] A. Makanju, A. N. Zincir-Heywood, and E. E. Milios. 2013. Investigating event log analysis with minimum apriori information. In *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*. 962–968.

[34] M McCracken, A Maxwell, and C Hofman. 2015. BibDesk. (2015).

[35] A. Miranskyy, A. Hamou-Lhadj, E. Cialini, and A. Larsson. 2016. Operational-Log Analysis for Big Data Systems: Challenges and Solutions. *IEEE Software* 33, 2 (2016), 52–59.

[36] Antonio Pecchia, Ingo Weber, Marcello Cinque, and Yu Ma. 2020. Discovering process models for the analysis of application failures under uncertainty of event logs. *Knowledge-Based Systems* 189 (2020), 105054. https://doi.org/10.1016/j.knosys.2019.105054

[37] Kexin Pei, Zhongshu Gu, Brendan Saltaformaggio, Shiqing Ma, Fei Wang, Zhiwei Zhang, Luo Si, Xiangyu Zhang, and Dongyan Xu. 2016. Hercule: attack story reconstruction via community discovery on correlated log graph. *Proceedings of the 32nd Annual Conference on Computer Security Applications* (May 2016). https://doi.org/10.1145/2991079.2991122

[38] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology* 64, Supplement C (2015), 1 – 18. https://doi.org/10.1016/j.infsof.2015.03.007

[39] Barbara Russo, Giancarlo Succi, and Witold Pedrycz. 2014. Mining system logs to learn error predictors: a case study of a telemetry system. *Empirical Software Engineering* 20, 4 (2014), 879–927. https://doi.org/10.1007/s10664-014-9303-2

[40] Tuomo Sipola, Antti Juvonen, and Joel Lehtonen. 2011. Anomaly Detection from Network Logs Using Diffusion Maps. *Engineering Applications of Neural Networks IFIP Advances in Information and Communication Technology* (2011), 172–181. https://doi.org/10.1007/978-3-642-23957-1_20

[41] Hudan Studiawan, Christian Payne, and Ferdous Sohel. 2017. Graph clustering and anomaly detection of access control log for forensic purposes. *Digital Investigation* 21 (2017), 76–87. https://doi.org/10.1016/j.diin.2017.05.001

[42] K. Sun, L. Meng, S. Guo, S. Xu, Y. Wang, and W. Li. 2017. An Approach of Anomaly Diagnosis with Logs for Distributed Services in Communication Network Information System. In *2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)*. 938–940.

[43] Markus Wurzenberger, Florian Skopik, Roman Fiedler, and Wolfgang Kastner. 2016. Discovering Insider Threats from Log Data with High-Performance Bioinformatics Tools. *Proceedings of the 2016 International Workshop on Managing Insider Security Threats - MIST 16* (2016). https://doi.org/10.1145/2995959.2995973

[44] Ting-Fang Yen, Alina Oprea, Kaan Onarlioglu, Todd Leetham, William Robertson, Ari Juels, and Engin Kirda. 2013. Beehive: large-scale log analysis for detecting suspicious activity in enterprise networks. *Proceedings of the 29th Annual Computer Security Applications Conference on - ACSAC 13* (2013). https://doi.org/10.1145/2523649.2523670

[45] Yue Yuan, Anuhan Torgonshar, Wenchang Shi, Bin Liang, and Bo Qin. 2018. Digging Evidence for Violation of Cloud Security Compliance with Knowledge Learned from Logs. (Oct 2018). https://doi.org/10.1007/978-981-13-5913-2_20

[46] Dongxue Zhang, Yang Zheng, Yu Wen, Yujue Xu, Jingchuo Wang, Yang Yu, and Dan Meng. 2018. Role-based Log Analysis Applying Deep Learning for Insider Threat Detection. *Proceedings of the 1st Workshop on Security-Oriented Designs of Computer Architectures and Processors - SecArch18* (2018). https://doi.org/10.1145/3267494.3267495

[47] Deqing Zou, Hao Qin, Hai Jin, Weizhong Qiang, Zongfen Han, and Xueguang Chen. 2014. Improving Log-Based Fault Diagnosis by Log Classification. *Advanced Information Systems Engineering Lecture Notes in Computer Science* (2014), 446–458. https://doi.org/10.1007/978-3-662-44917-2_37

[48] De-Qing Zou, Hao Qin, and Hai Jin. 2016. UiLog: Improving Log-Based Fault Diagnosis by Log Analysis. *Journal of Computer Science and Technology* 31, 5 (2016), 1038–1052. https://doi.org/10.1007/s11390-016-1678-7