User Behavior Application Logging: A Solution to Log Analysis

(Technical Paper)

Cyber Forensics: Cybercrime Investigation through Logs

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Haris Saeed

Fall, 2022

Technical Project Team Members

Sven Brueckner

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Haris Saeed

Approved _____ Date _____

Capstone/Technical Advisor Name, Department

Approved ___ _____ Date 11/21/2022

STS Advisor: Richard D. Jacques, Ph.D., Department of Engineering & Society

Introduction

With the age of the Internet and the ongoing technological advancements in society, trying to log every action taken by users has become nearly impossible. In the past, companies used logs for troubleshooting problems by seeing where their product went wrong, but now they are used for many functions such as performance checking, recording user actions, and checking for malicious activity (Singh 2010). Logs are a collection of entries, and each entry contains information related to a specific event that has taken place. An enormous amount of operational information is created each time a user interacts with software. There is so much data that is separated amongst multiple logs that it becomes overflooded and difficult to decipher. As a result, when trying to analyze logs, there is either under-reporting of data by filtering out too much data that it becomes useless or an over-reporting of data by keeping too much data that it is still difficult to process and still has the initial problem of being difficult to decipher. The data that should be filtered out also depends on what is being searched for as some information may be helpful in one case but useless in another case. Many companies use logs as they are easy to capture at a large scale, quick at doing so, and to better understand customers and provide personalized experiences (Wang 2016). Additionally, companies and cyber forensics use logs to search for malicious behavior and find user information through behavior analysis (Singh 2010). Some companies have turned towards using behavioral-based application logging where each log has a specific purpose, is scoped to only commit to that purpose, and provides feedback mechanisms that quickly describe behaviors the log was tasked to analyze. In this prospectus, I will research whether user behavior application logging is a better option for log analysis compared to others based on filtering of information and readability as the technical paper and proceed to research how beneficial these logs can be in cyber forensics and cyberattacks as the STS paper.

Technical Discussion

Behavioral logs are footprints of human behavior that are collected through sensors to record user activity. Compared to other types of recording activity, logs are the most natural way to observe people as they use systems they would in everyday life without any external instructions. Since the 1930s, behavior has been recorded in psychology studies and it has become customary practice to do this and save them for future analysis (Dumais 2014). In recent years, the centralization of web-based computing has made it possible to capture these logs on a much larger scale. These logs are created by recording activities when people interact with different services. There are many distinct types of actions that can be recorded from a simple keystroke to browsing patterns and purchases on e-commerce sites. Client-side logging is implemented in operating systems, applications like one's browser, or third-party software/harder to capture comprehensive data usage, while server-side logging is commonly used by service providers such as web search engines corresponding to user requests (Neelima 2014).

Behavioral logs can provide insight into how people interact with different services and how skilled they are as well. For example, an analysis of over a million web searches found that queries were short, averaging 2.35 terms, and 80% of all queries did not include advanced operators (Dumais 2014). By diving deeper into this and looking at each log, you can determine the person the log is from and classify them as an advanced searcher or a novice searcher. There are two main ways to start partitioning logs, time, and user. Partition with time deals with the exact times that are timestamped in a log, and these can be used to split or classify logs into distinct groups and one can analyze when most logs are created with what is going on at a specific time. One example of this is researchers accurately predicting the strength of the

seasonal flu based on search engine log data with just a single day's logs (Dumais 2014). The second way logs are partitioned is the user characteristics, which was discussed earlier as differences between advanced users and novice users. The only drawback with logs is that there is limited information or none at all about the person generating the data and their intentions and emotions in those behaviors.

Initially, logging was used for testing purposes rather than user behavior analysis (Alspaugh 2014). As a result of this, there have been issues with logging and understanding user behavior. It is difficult to analyze logs with the excess amounts of unnecessary information that only a developer would need for debugging. Trying to alter current systems is difficult due to system infrastructures being complex, causing issues with both automated and manual logging (Maxwell 2021). Current logging systems contain an excessive amount of information when behavior analysis only needs the timestamp, user identification through IP address or username, and the user's action. Instead of receiving logs with all the information that a developer would normally deal with, analysts would prefer receiving logs that are already prefiltered through the system and easy to recognize, which is what behavioral logs can be useful for by only recording what is necessary, skipping the filtering step. The issue that I will be researching with using behavioral-based application logging as an alternative to other logging methods to resolve the problem of under-reporting and over-reporting information causing issues with data analysis as mentioned earlier and in the overview. The under-reporting and over-reporting issues stem from the filtering of data either removing too much or not enough information causing analysis to become difficult. Behavioral logs can be centered around a certain topic and collect logs for only that specific behavior. These logs will be researched in usage areas such as API session

management, user application navigation, and more. I will be working with LookingGlass Cyber, a company focused on cybersecurity, to research this topic over the next few months.

STS Discussion

Cyber forensics is a new, rapidly changing field of technology that uses investigative techniques to extract information and data from a person's computer. It is a systematic approach for collecting and preserving data that guarantees information accuracy, reliability, and the presentation of the data in a legal environment. The goal of cyber forensics is to aid in investigating crimes. Evidence in the form of electronic information can be useful in solving cybercrimes (Lu 2017). This is applied to not only cybercrimes but also password breaking, spamming, data recovery and analysis, tracking user activity, forensic imaging and verification, viruses, file types, encryption, and more (Singh 2010). By applying cyber forensics to crime investigations, it can reduce the time needed to investigate and decrease the scope of the crime by reducing the complexity and narrowing the investigation. It is not meant to solve the crime itself.

Cyber forensics uses the same logs that everyday companies use to watch actions that users take on their platform and track their behaviors. Just like these companies, crime investigators can use behavioral logs to help with their investigations as stated before. Forensic analysis of log data is necessary to deal with cybercrime (Lu 2017). Companies have a goal of preventing and stopping hackers from accessing protected information and crime investigators have the job of hunting these hackers down. Many of these attacks such as bombings could be investigated by looking into purchases made by the bomber. On the internet in e-commerce, the client access server will generate data that has log files and query data (Wang 2016). Investigators and the seller can use the logs and see information like where the person is buying

from, what they are buying, and more take action as needed. In a different scenario like a cyberattack, the firewall has logs that both the company and investigators can use to shut down the malicious attack and track down the attacker. It is not only companies that need logs to become much cleaner and filtered with only useful information being reported to them, but also crime investigators.

Crime investigators and cyber security analysts can identify users and their behaviors from server log data for various cybercrime including phishing (Ibrahim 2021). Analysts take server logs and analyze them to determine whether an action is an irregular behavior and mark it down as malware as is done in network security applications (Chen 2005). One way they can determine whether an action is irregular is by looking at patterns that show various user intents. The user's intent is what the user's goal is when using a program. Intents can differ between users and programs can be used for many different intents. When a person is using software with an intent, their intent strongly influences their behavior as they are more focused and engaged in a specific task and search through specific categories as compared to other users (Cheng 2015). By analyzing user actions with the timestamp, their intent can be discovered. Through intents, a user's gender, age, and categorical differences can be determined. Taking both the log information that contains timestamps, location through IP addresses, and more and using information gained through intent analysis by comparing everyday logs to suspicious actions, malicious action can be identified, stopped, and information about the suspect can be discovered. Researching behavioral logs and how it helps crime investigations can further support the technical research described in the previous section and support the need for better behavioral logging practices.

STS Research Question and Methodology

The STS research question being explored is whether logs can aid with crime investigations and improve the security of society by allowing analysts to uncover suspicious behavior quickly to resolve and prevent crime. This will be accomplished by first researching how logs are currently used in crime investigations and what types of crime they are beneficial for. Case studies and crime reports will be researched to analyze the current usage of logs and if they are even used at all. Next, log analysis for malicious behavior will be researched to see if logs can be incorporated into criminal investigations and whether or not logs will be efficient. This will incorporate the technical research stated previously involving behavioral logs to see whether a person can determine malicious/suspicious activity through the behavioral logging system efficiently. By researching these areas, we can see whether logs are already used or not in crime investigations, what types of logs are used, and whether they are useful or not, and then see whether malicious activity analysis is efficient with behavioral logs to determine if logs are beneficial to crime investigations and prevention.

Conclusion

The technical discussion topic is making logs more efficient through a behavioral-based application approach for readability and filtration of information. The STS discussion topic extends past the technical discussion topic by connecting it to real-world use involving cyber-crime investigations and tracking suspicious users that could be preparing to conduct a crime through their behaviors exposed by behavioral logs such as buying products online so that the security of society is improved. Understanding how behavioral logs can help with the current logging issues can further benefit society with how scalable they are from logs for minor projects to global investigations and tracking. Through the technical research, I expect to find that

behavioral logs are more efficient and beneficial than other types of logs and through the STS research, I expect to find the usage of behavioral logs to be beneficial in crime investigation and prevention, thus increasing the security of our society.

References

Alspaugh, S., Ganapathi, A., Hearst, M. A., & Katz, R. (2014). Better logging to improve interactive data analysis tools. *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA'14)* (pp. 19-25). Retrieved from https://core.ac.uk/download/pdf/34655536.pdf#page=19

Cheng, J., Lo, C., & Leskovec, J. (2017). Predicting intent using activity logs. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. https://doi.org/10.1145/3041021.3054198

Dumais, S., Jeffries, R., Russell, D.M., Tang, D., Teevan, J. (2014). Understanding User Behavior Through Log Data and Analysis. *Olson, J., Kellogg, W. (eds) Ways of Knowing in HCI. Springer, New York, NY*. https://doi.org/10.1007/978-1-4939-0378-8_14

Ibrahim, K., Obaid, A. (2021). Fraud usage detection in internet users based on log data. *International Journal of Nonlinear Analysis and Applications*, 12(2), (pp. 2179-2188). doi: 10.22075/ijnaa.2021.5367

Lu, W. (2017) Exploration and implementation of user behavior forensics analysis system of computer network based on system log1. *2017 Institute of Thermomechanics*, 62(2), (pp. 53-62). Retrieved from http://journal.it.cas.cz/62(2017)-2B/Paper%2006%20Wenzhe%20Lu.pdf

Maxwell, D., & Hauff, C. (2021, March). LogUI: Contemporary Logging Infrastructure for Web-Based Experiments. *European Conference on Information Retrieval* (pp. 525-530). Retrieved from https://chauff.github.io/documents/publications/ECIR2021-Maxwell.pdf

Neelima, G., & Rodda, S. (2016, March). Predicting user behavior through sessions using the web log mining. In *2016 International Conference on Advances in Human Machine Interaction (HMI)* (pp. 1-5). IEEE. doi: 10.1109/HMI.2016.7449167

Chen, S., et al. (2018), User Behavior Map: Visual Exploration for Cyber Security Session Data. *IEEE Symposium on Visualization for Cyber Security (VizSec)* (pp. 1-4). doi: 10.1109/VIZSEC.2018.8709223.

Singh, N. K., Tomar, D. S., & Roy, B. N. (2010). An Approach to Understand the End User Behavior through Log Analysis. *International Journal of Computer Applications*, 5(11), (pp. 27–34). https://doi.org/10.5120/953-1330

Wang, N., Zhang, Q., Yang, L., & Chen, M. (2016). A novel E-Commerce recommendation system model based on the pattern recognition and user behavior preference

analysis. *Information Science and Industrial Applications*. Retrieved from https://web.archive.org/web/20180602064635id_/http://onlinepresent.org/proceedings/vol138_2016/23.pdf