Research Paper Review – Device Fingerprinting

Gerald T. Rigdon

February 8, 2025

The paper "Searching for Indicators of Device Fingerprinting in the JavaScript Code of Popular Websites" (Good and Hoofnagle, 2014) addresses the topic of user privacy as it relates to the tracking of website visitations using device fingerprinting as opposed to the more ubiquitous cookie. Given the latter is a file that needs to be placed on a user's machine, it can be blocked or subsequently discovered and deleted. Conversely, device fingerprinting can be derived from various sources such as querying a web browser to create a user profile. This dynamic JavaScript code execution approach means fingerprinting is less constrained and thus harder to avoid, hence neutralizing a user's ability to prevent tracking.

Research Problem

As indicated by an appropriate title of the research effort, the paper attempts to determine the prevalence of device fingerprinting by analyzing the exposed JavaScript software running on popular websites. It takes a neutral position by stating the corporate rationale for wanting to identify internet users which includes analytics, marketing, and importantly, from a cybersecurity perspective, fraud protection. However, the traditional tracking methods using cookies had become less effective due to user awareness and their ability to effect both preventive and corrective actions. As a result, web-based device fingerprinting emerged as an alternative means to track users in a surreptitious manner. This kind of practice without user consent raises significant privacy concerns. As a result the core research focuses on device fingerprinting techniques, methods for identifying these, and ultimately a statistical measure of the prevalence derived from empirical data. In publishing a census, the intent was to provide insights into the impact of device fingerprinting on user privacy and the effects on the broader internet ecosystem.

Research Questions

The paper presents a comprehensive understanding of device fingerprinting and aims to address the importance of appreciating the implications of the practice by addressing the following key research questions:

- How prevalent is device fingerprinting on popular websites?
- What is the impact on the internet?
- What are the common techniques used for device fingerprinting?
- Can fingerprinting be detected?

- How effective is device fingerprinting technology?
- What are the implications and/or level of concern for user privacy?

While these key questions drive the research, other tangential questions naturally arise with respect to broader concerns such as the level of user awareness, the need for regulatory oversight, whether countermeasures be deployed, etc.

Hypothesis and Methodology

There is no stated hypothesis; the goal was to find the answers to the research questions. Hence, the implication is that an absence of preconceived notions was necessary so that the data could speak for itself and produce a census untethered to any expectations.

The research method was conducted in three phases.

- Phase 1: Understanding how JavaScript code is used by device fingerprinters by viewing the website source code of BlueCava, Iovation, and ThreatMetrix, companies who claimed to use the technique. This provided insight into identifying specific software function calls that could be used to fingerprint and to verify the tactics remained closely aligned to previous research in the "Cookieless Monster" paper.
- Phase 2: Deploying a custom crawler to visit popular websites to spot the use of JavaScript code. This included a clever strategy of presenting the crawler as a browser so that the website would serve up the typical content. Additionally, login or sign-in pages were visited assuming that the site owner would be more interested in fingerprinting users that were engaging with the security mechanisms including passwords.
- Phase 3: Analyzing the collected code. The heart of the analysis was using the fingerprinting tactics discovered in Phase 1 as the key to uncovering similar signatures from the data that was collected in Phase 2. This involved categorizing files and looking for keywords and suspect calls to JavaScript libraries.

Discussion

The paper starts by discussing the historical context of tracking technologies, focusing on traditional cookie-based methods and their limitations. It then transitions into the emergence and growing prevalence of device fingerprinting as an alternative tracking mechanism. It covers the various techniques employed in device fingerprinting, browser and operating system characteristics, and hardware configurations. It critically evaluates the effectiveness of these techniques in uniquely identifying users, comparing them with traditional tracking methods. Moreover, the paper presents a comprehensive and balanced overview of the state of device fingerprinting research already conducted. It sets the stage for the paper's contributions to understanding the ecosystem of web-based device fingerprinting.

Dataset and Sampling

The dataset for the study was limited to one thousand of the most popular websites. The sample size used, N = 1000, provides an adequate representation of a population with acceptable margins of error, supporting the results as reasonably accurate.

Ground Truth, Analysis and Results

The ground truth or baseline criteria by which the samples were judged was established from BlueCava, Iovation, and ThreatMetrix, who were known to engage in the device fingerprinting activity. Using the characteristics of the JavaScript code from these sources, the study found that of the one thousand websites visited, less than 6% contained the fingerprinting code signature.

The central theme of the analysis was to avoid false positives which would result in overstating the prevalence of device fingerprinting. Thus, after identifying three JavaScript software calls of interest that were associated with the fingerprinting criteria, positive identification required code that contained all three of these calls together. This involved categorizing files appropriately along with careful keyword searches. In essence, words and phrases thought to be sufficiently similar to the baseline or ground truth criteria passed through multiple filtering stages until a confidence threshold was reached.

Limitation and Gaps

- Lack of Previous Census Data: Although another paper published census data in 2013, there was no data available at the time the research activity was conducted. Without previous benchmarking in place, there was no stated hypothesis or expectations for the findings.
- Sample Weakness: While one thousand is a strong sample size, the websites chosen as the top thousand most popular were ranked by Quantcast. Choosing the most popular does not necessarily mean the best representative sample or that optimal randomness was achieved. Finally, while the sample size was technically one thousand the paper admits that ninety-eight (or roughly 10%) of the websites were unable to be visited by the custom crawler.
- Accuracy of Findings: The desire to avoid false positives may have resulted in false negatives which may mean soundness was sacrificed for completeness. The paper admitted the technical part of collecting the data was less challenging than analyzing the data to reach logical conclusions.
- Myopic Focus: The study revolved around the analysis of JavaScript code for identifying device fingerprinting instead of exploring other techniques.

- Detection Weakness: Obfuscated or minified JavaScript code could have resulted in missing the code used in the device fingerprinting.
- Skewed Results: The study admitted that the websites that fit the JavaScript code signature were overwhelmingly associated with Iovation versus the BlueCava and ThreatMetrix companies by more than a 10:1 ratio. Only a brief analysis was provided as to why this was the case and the implications of these differences.

Literature Review

The paper references sixteen sources which provide a respectable foundation for the research. During the time the research was conducted no papers on a census had been published and an acknowledgement is given to a 2013 paper in the ACM SIGSAC Conference as the first to do so. While most sources are referenced multiple times, the paper relies heavily on the "Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting" which is referenced eight times. The "Cookieless: Monster" paper is used to establish the JavaScript code analysis research approach for identifying device fingerprinting by relying on three companies, namely, BlueCava, Iovation, and ThreatMetrix. Given these companies were known or claimed to use device fingerprinting, a baseline for comparative analysis could be established by understanding how they did it.

Research Extension Possibilities

- Given the paper established a census at a given point in time, there would be an opportunity to use the same criteria and approach solely for the purpose of establishing a trend analysis.
- Since over a decade has passed since the paper was published one could leverage detection technology advances. Instead of crawling one thousand popular websites, collecting and analyzing JavaScript code for fingerprinting signatures, the research could simply identify how many browsers fitted with detection add-ons detect offensive fingerprinting during the website visits.
- Although the original study raised privacy concerns, more research could be done on the state of legislation and policy changes with respect to device fingerprinting in both the United States and the international community.

Conclusion

The findings of this study highlight that although the practice of device fingerprinting raises significant concerns regarding user privacy, the results suggest limited risk given only 6% of the websites visited and analyzed were found to be participating in the fingerprinting practice. The conclusions of this study are based on the results obtained given the JavaScript code identification approach and is feasible within the context of the overall reliability, validity, and sample size of the study. Further research would be necessary to explore this topic in broader contexts and to explore whether the prevalence of device fingerprinting was under-estimated.

Marymount Honor Pledge

I agree to uphold the principles of honor set forth by this community in the Marymount University mission statement and the Academic Integrity Code and Community Conduct Code, to defend these principles against abuse or misuse, and to abide by the regulations of Marymount University.

References

Rausch, M., Good, N., & Hoofnagle, C. J. (2014). "Searching for Indicators of Device Fingerprinting in the JavaScript Code of Popular Websites." In Proceedings, Midewest Instruction and Computing Symposium.