William Lucas, Alan Chen, Krisham Prasai, Vinayak Deshpande

CS:3640 Intro to Networks and Applications

December 3, 2024

An Investigation of Privacy Practices and Compliance with CCPA in a Sample of 100 Websites

In this study, our group explored the state of privacy policies across a selected category of 100 websites, focusing on three key areas of compliance with privacy regulations. The category we chose was the top 100 most visited websites worldwide as of October 2024. We wanted to home in on this category because we believed it would be a good view of the internet today and how a very large set of privacy data is currently being handled. In order to analyze these websites, we devised three research questions that were to be answered.

Our first step was to examine the percentage of websites that adhere to the California Consumer Privacy Act (CCPA) requirements, particularly the presence and accessibility of the “Do Not Sell My Personal Information” (DNSMPI) link. Second, we evaluated the clarity and accessibility of privacy policies and DNSMPI links, investigating how easily users can locate these critical privacy features. Finally, we assess the prevalence of JavaScript-based fingerprinting mechanisms across these websites, identifying potential risks to user privacy. Our research aims to shed light on the current landscape of online privacy practices and the effectiveness of legal frameworks in protecting user data.

Compliance with CCPA: Analyzing DNSMPI Link Presence

The CCPA mandates that businesses disclose their privacy practices and provide a clear option for users to opt out of the sales of their personal information. A key component of this is the Do Not Sell My Personal Information (DNSMPI) link. Understanding the prevalence of this feature can shed light on how well websites comply with privacy laws and how well users are protected. CCPA compliance is crucial for ensuring transparency and trust in light of growing concerns about data privacy.

Our goal is to quantify how many websites in the top 100 most visited websites worldwide include a DNSMPI link, as required by the CCPA, and to assess their adherence to this legal requirement.

Methodology and Justification

We used a web scraping approach to extract privacy policy links and DNSMPI links from websites in our sample. This was done by developing a custom crawler using Playwright, which has been chosen due to its ability to interact with dynamic web pages and handle complex JavaScript-based websites. Playwright allows us to automate browsing and interaction with websites in a way that mimics real user behavior, ensuring we can access elements linke links embedded in dynamic or JavaScript-heavy pages. By using heuristics based on common keywords (e.g., “do not sell,” “my personal information,” “DNSMPI,” “privacy policy,” “legal policy,” “data protection,” etc.), we can locate DNSMPI links reliably using Python code.

Our crawler uses Playwright to visit each URL in the set and searches for links containing these keywords associated with DNSMPI links. The ‘fetch_link’ function searches and locates relevant links on each page. It checks for the visibility of these links and extracts their ‘href’ attributes. If a DNSMPI link is found, it is stored. If not, the entry is marked as “Not Found.” The results are stored in a structured format (JSON), which includes the URL of the website and whether the DNSMPI link was found or not. Once the data was collected, we counted how many of the 100 websites have a DNSMPI link and proceeded to calculate the percentage of compliance with CCPA.

Before we move onto our conclusions, let’s look at four potential limitations of our research. The first is that some websites may have DNSMPI behind dynamic content, or they could use non-standard terminology, which could prevent our scraper from detecting them. The next limitation could be that certain websites that heavily rely on JS for rendering content might not display the DNSMPI links in a way that is easily detected by the scraper. Finally, we must consider legal interpretation: our crawler identifies links based on keywords, but there may be variations in the language used across websites that could lead to missed links.

Conclusion

In our analysis of the top 100 most visited websites worldwide as of October 2024, we found that only 17% (17 out of 100) of these websites include a DNSMPI link and 48% included a privacy policy page. This finding suggests that a significant portion of the top websites may not be fully adhering to the CCPA’s privacy requirements. However, it is important to note several limitations that could affect the accuracy of this result. False negatives could arise due to any of these limitations, and any others that we missed. This would lead to an underestimation of compliance. Additionally, it is important to understand sampling bias, as these results should not be extrapolated to a broader set of websites to the non-diverse and non-random dataset. As a result, while our findings offer valuable insight into CCPA compliance among highly-visited websites, they should be interpreted with caution when considering broader trends across the internet.

Accessibility and Clarity of Privacy Policies

The accessibility and clarity of privacy policies are critical for allowing users to understand and control how their personal information is collected, used, and shared. Despite legal requirements like CCPA mandating such information, if the user cannot easily find or comprehend these documents, they might not be aware of how their personal data is being used. Evaluating these aspects helps determine whether the top 100 most visited websites genuinely help users make decisions about their privacy. Our goal is to find how easily users can locate privacy policies on the top 100 U.S. American websites and assess the readability of these documents to understand if it's written in user-friendly language.

To achieve this, we have outlined two primary measures: accessibility (how many clicks it takes to find the privacy policy) and clarity (the readability of documents).

Methodology and Justification

Our approach to assessing accessibility focuses on how many interactions (clicks) it takes for a user to find the privacy policy link from the homepage. This measure captures the ease of navigation, with fewer clicks indicating greater accessibility. By simulating user behavior through our custom web scraper, we can determine how many steps it takes for a user to locate the relevant links. This is important because more clicks may indicate that the link is buried within less accessible pages, which can hurt the user in terms of making informed privacy decisions.

To evaluate clarity, we used the Flesch-Kincaid Grade Level score, which calculates the readability of the text based on sentence length and word complexity. This score is an industry-standard for measuring the difficulty of a given document, with a lower score indicating more readable content. We applied the 'textstat' library in Python to compute the Flesch-Kincaid Grade Level of the privacy policy text, allowing us to quantify how easy it is for a general user to understand the language in the privacy policy.

We enhanced our crawler to not only detect and collect privacy policy but also to simulate clicks to locate these links and analyze the readability of the content. To count the number of clicks, we implemented a counter that tracks the number of clicks for each website that increments on each link clicked until we reach the privacy link. Once we reach a privacy link, we store the number of clicks it took to reach that link in the same JSON file that stored the data for each of the 100 websites. For websites that the crawler previously did not find a privacy link for, we make sure to not check that website. After locating the privacy policy, we use the 'textstat' library to extract the text and calculate the Flesch-Kincaid Grade Level score. Our final method will be to calculate the average number of clicks to find the privacy policy and the average Flesch-Kincaid Grade Level score.

Conclusion

Our analysis of the top 100 most visited websites revealed that the average number of clicks required to get to the privacy policy page is 1.07, indicating that most websites provide relatively easy access to their privacy policies. However, the average Flesch-Kincaid Grade was 13.95, suggesting that these policies are written at a level of complexity equivalent to college-level reading. This indicates a potential barrier to comprehension for the general public. Improving the clarity and simplicity of these documents could enhance transparency and user understanding.

JavaScript Fingerprinting: Prevalence and Impact on User Privacy

Fingerprinting has energed as a sophisticated tracking technique that collects detailed information about a user's device, browser, and behavior without relying on traditional identifiers like cookies. Unlike cookies, fingerprinting is difficult to detect and even harder to block, making it a growing concern for user privacy. This method bypasses many existing privacy safeguards, including cookie blockers and private browsing modes. Investigating the prevalence of fingerprinting scripts is crucial, as it reveals the extent to which websites adopt these invasive practices. This insight can help policymakers and developers develop more robust countermeasures.

Our goal is to identify and quantify the number of websites employing scripts that match a signature like those used for JavaScript fingerprinting. Specifically, we aim to identify scripts that collect unique device attributes such as screen resolution, installed fonts, GPU capabilities, or hardware configurations. By identifying patterns and signatures associated with known fingerprinting libraries, we can gauge the prevalence of this privacy-invasive behavior in our selected sample of websites.

Methodology and Collection

We analyzed the JavaScript execution environment of each website, looking for key indicators of fingerprinting. These indicators include the use of specific API calls (e.g. CanvasRenderingContext2D), timing attacks (performance.now()), and enumeration of fonts of plugins. To achieve this, we used a headless browser to monitor network requests, script execution, and function calls in real-time.

In the crawler code, we extended the browser context to include monitoring of JavaScript function calls. Using Playwright, we intercepted API calls and network requests to log suspicious behavior.

The data collection involved running the crawler on the selected sample. For each website, we will capture:

- API calls relevant to fingerprinting.

- Network requests to external fingerprinting libraries.
- Detailed logs of JavaScript execution

Conclusion

Our analysis of the top 100 most visited websites revealed that the proportion of websites in our sample that actively engage in fingerprinting is relatively high, shedding light on how pervasive this practice is in the modern web ecosystem. However, it is important to note the inherent limitations of this approach. False positives may occur, as legitimate scripts can exhibit behavior similar to fingerprinting. Conversely, some fingerprinting scripts may evade detection by obscuring their behavior, leading to false negative. Despite these challenges, our methodology provides a foundational understanding of JavaScript fingerprinting prevalence and offers a pathway for further investigation.