Privacy on the Web

Tianze Dong, Jaewon Jung
Department of Computer Science, Duke University

Abstract

Websites on the modern web use Hypertext Transfer Protocol (HTTP) to transfer files between the origin server and the user. Since HTTP is stateless, HTTP cookies (web cookie, browser cookie) are used to identify users and provide a seamless browsing experience. However, since the introduction of HTTP cookies, websites have been storing users' data unbeknownst to the user. While there is no encompassing set of user privacy laws, the General Data Protection Regulation passed in 2018 by the European Union as well as previous works provide a valid set of guidelines for websites regarding collecting and processing of user data. We set out to examine the state of user privacy on the modern web: how cookies are used and whether websites follow the mentioned guidelines.

Objectives

Our goal is to conduct a large-scale study on the top ten thousands most visited websites and analyze the data collected.

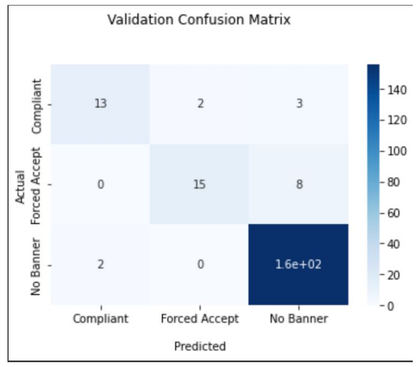
- Collect and analyze data on the different types of cookies and cookie banners
- Determine weather websites are complying with GDPR
- Identify possible correlation between the number of third party cookies and the tendency to include GDPR-compliant cookie banners
- Verify the validity of cookie banners

Methods

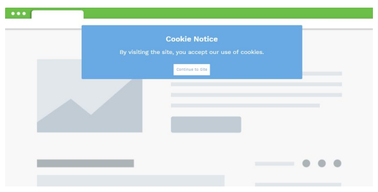
- Web scraping
  - Traverse through top 6,000 most commonly visited websites
  - For each website, gather all possible buttons and links
  - Match the text to a collection of parameters and categorize a website by its specific banner
- Validation
  - From an initial set of 500 websites, randomly select a subset of 200 websites and manually validate the result from the script
  - Create a confusion matrix and determine the accuracy of the script
- Cookie Data
  - Traverse through top 6,000 most commonly visited websites
  - Using cookie-script database, collect cookie data for each site

Results

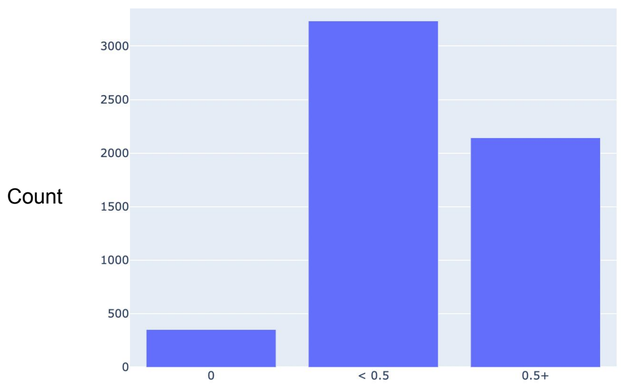
- Categories of cookie banners
  - Fully compliant banners
    - Able to decline and customize
  - Forced acceptance banners
    - Only able to accept but still informs user
  - No banner
- Results of banner script (6000 sites crawled):
  - 584 fully complaint sites
  - 768 forced acceptance sites
  - 3962 with no banner
  - Only 25.4% websites display a valid consent banner
- Validation Results
  - Accuracy: 0.9195



Fully Compliant Banner



Forced Acceptance Banner



Proportion of third-party cookies used

Conclusion

- Since the introduction of HTTP cookies, companies have been using them to track users' information. With the guidelines detailed in the GDPR, we ran a widespread measurement study on the web and found that a majority (74.6%) of sites do not inform their users or gather consent before using them. This is a serious privacy concern. Of that majority, we found that many websites use a dark pattern: companies often hide user consent prompts under a sign up process, forcing users to accept the privacy terms unknowingly and giving no alternative option to not be tracked.. Furthermore, after analyzing the cookies used on the most popular websites, we found that 94% of websites use third-party cookies. The huge prevalence of third-party cookies was surprising and it is important that users are more aware of how their internet activity is being tracked.

- Next measures
  - Conduct a user survey to determine the user perspective of privacy
  - Using the existing web scraping script, create a developer-side service to check compliance of the website

References

- Related Work
  - We Value Your Privacy…Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy, NDSS (2019)
  - On Compliance of Cookie Purposes with the Purposes Specification Principle, IEEE (2020)