Poster: Cookie Setting Practices in the Wild

Tobias Urban[1,2], Martin Degeling[2], Thorsten Holz[2], and Norbert Pohlmann[1]

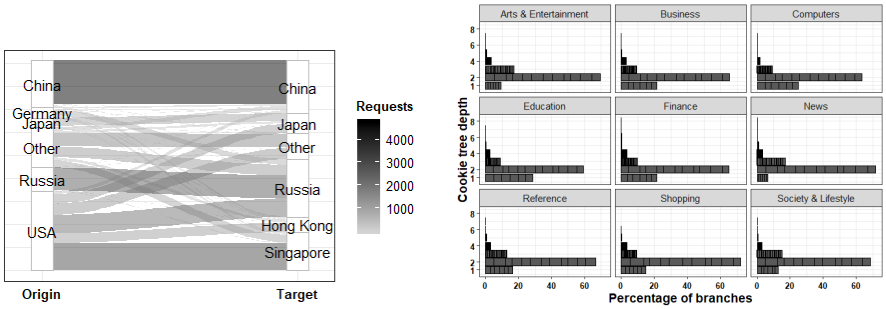
[1] Institute for Internet Security, Germany {lastname}@internet-sicherheit.de
[2] Ruhr-University Bochum, Germany {firstname.lastname}@rub.de

Introduction

Website today constantly increase in terms of complexitiy of their setup. To decrease time of development cycles, to decrease costs, and to provide users recognizable elements they use third party software. One artifact of this is that website providers are often unaware of implications when embedding third parties and sometimes provide additional attack surface to adversaries.Additionally, they can be held accountable for the actions of third parties with respect to the personal data they collect. Previous work has been focused on the analysis of website front pages. To get a comprehensive overview on third party use we introduce *cookie-trees*. A *cookie-tree* holds all third parties observed when visiting a websites and accounts who is responsible for loading each third party. For example *foo.com* embeds an iframe which loads content from *bar.com* (3rd party). The iframe again loads a script from *foobar.com* (4th party) which could again load another third party. Using *cookie-trees* we investigate to which extend services can control which third parties are actually embedded into their service and test if and at which point parties are emended that might handle user data but which might not be adequate regarding current privacy legislation.

Method & Results

In our analysis, we use the top 10k website ranking provided by Le Pochat et al. [2]. To measure the third party cookie behaviour of websites, we use the `openWPM` measurement platform [1]. We configured the platform to visit the landing page and 100 subsites of each website. The measurement platform is configured to store all cookies set or accessed via JavaScript and HTTP headers and we capture all JavaScript events that be used to access the local storage or HTTP cookie. We passively log all DNS responses in order to test if IP addresses are used which are associated with countries that do not offer a GDPR adequate privacy protection level. In total, we conduct the same measurement from three different locations (Japan, Germany, and the USA) to account for possible geographical differences of cookie usage. In this work, we evaluate the amount of third parties resulting from embedding a single third party object that itself can dynamically load code. To do so, we create a *cookie tree* for each visited URL (for each landing page and all subpages respectively) which includes all third parties loaded on the page visit. We build the trees based on two HTML object types (JavaScript and iframes) which can both be used to dynamically load more third-party code.



(a) Origins and targets of cookies set by services in non adequate countries.

(b) Distribution of tree depths based on the websites category.

Figure 1a shows the origins and targets of all requests that set a cookie and which are potentially in conflict with the new legislation. These numbers only refer to our EU measurement and, therefore, the results are not violations of the legislation but provide insights to potential data flow services providers might not be aware of. The origins/targets are based on the observed IP addresses in our measurements. Overall, 4.7 % of all cookies were set by a services outside an adequate country and only 7 % of the visited domains (TLD+1) do not set a cookie in a GDPR non-adequate country. Russia is the most prevalent target of such requests (29 %) in our dataset followed by China (21 %. Regarding absolute numbers, the US is the most prevalent origin of such requests (39 %) followed by China (18 %). However, percentage wise the amount of all requests that target a non-adequate country originate in Argentina and Australia (both around 3 %). We did *not* find statistical significant impact that the originating region has influence if a non GPDR adequate country is used. Thus, overall companies in all observed countries are equally well compliant to the GDPR and aware of the new legislation. Figure 1b shows the depth of the measured cookie trees. The average cookie tree has a depth of 2 (median 2) and the deepest tree we found has a depth of 8 (17.3 % $\geq$ 3). Each node of the tree has on average 0.9 (media 0) direct children (max 361). The the originating region has *no* statistical significant impact on the depth of a cookie tree. Furthermore, the depth of a tree is also impacted by the category of a website (ANOVA test p-value < 0.0001).

References

1. Englehardt, S., Narayanan, A.: Online tracking: A 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM Conference on Computer and Communications Security. pp. 1388–1401. CCS '16, ACM Press (2016)
2. Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., Joosen, W.: Tranco: A research-oriented top sites ranking hardened against manipulation. In: Proceedings of the 26th Symposium on Network and Distributed System Security. NDSS'19 (Feb 2019)