Poster: GDPR and Online-Advertising

Tobias Urban[1,2], Dennis Tatang[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

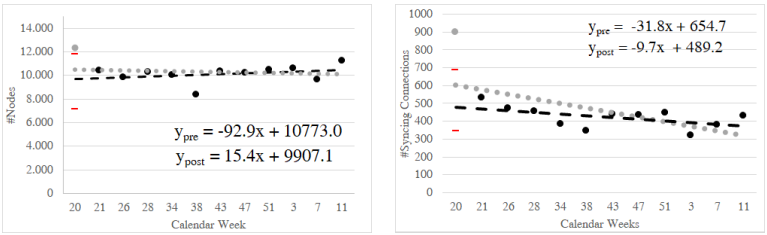
Introduction

Advertising remains one of the primary sources of income for many websites, apps, and online services. To individually target website visitors with ads, ad-tech companies gather and use personal data, mostly without users' explicit consent. The European General Data Protection Regulation (GDPR), which went into effect on May 25, 2018, introduced significant changes that affect how personal data can be collected and shared. Besides other things it requires companies to gain explicit consent before collecting or sharing personal data. In this work, we seek to provide insights into the effects of the GDPR on the information sharing by cookie syncing between ad services. Previous studies described the technology, but there is a lack of knowledge about its extent, the networks behind it, and changes over time. More specifically, we measure the relations of websites and third parties, as well as links between multiple third parties regarding ID syncing before and after the GDPR took effect. We used different browser profiles to visit more than 2.6 million websites over ten months to identify cookie syncing.

Method & Results

To measure the extent of tracking and cookie syncing, we used the *openWPM* [2] platform. For our study, we deployed the platform on two computers at a European university to ensure a European origin of our generated web traffic. We conducted twelve measurements over the course of ten months. The first measurement started a few days before (19th of May 2018), the second on the day the GDPR went into effect (25th of May 2018).

We build a graph representation of our measurements in which each node represents an observed third party and each edge cookie syncing activities. To measure the syncing relations of third parties, it is necessary to identify URLs—that contain user IDs—inside a request (e. g., **foo.com/sync?partner=https://bar.com?/id=abcd-1234**). For each observed URL, we check if this URL has `GET` parameters that might be an ID, according to the definition of Acar et al. [1], and that include a syncing partner. Figures 1a and 1b show the number of nodes and edges per measurement. The y-axis represents the number of nodes or connections, and the x-axis represents the calendar weeks (CW). The light gray dot on the left is the first measurement *before* the GDPR came into effect and the further darker gray (black) dots represent the corresponding other



(a) Number of third parties (b) Number of syncing connections

Fig. 1: Regression lines of our measurements including the pre-GDPR measurement (gray) and excluding it (black). The red dashes represents the confidence interval for the prediction for the pre-GDPR measurement point.

measurements. We performed two types of linear regression analysis, including the measurement before the GDPR took effect y_{pre}(gray line) and excluding it y_{post} (black line). To confirm that the amount of embedded third parties over all websites between M#2-M#12 are statistically significantly different from M#1, we calculate the confidence interval (99% confidence) for the prediction of the previous curve for the pre-GDPR measurement based on the values without the value of the measurement before introducing the GDPR. If the value of our measured pre-GDPR measurement is outside the confidence interval, we confirm that by the time of the introduction of the GDPR, the number of nodes reduced. We see evidence that the number of parties used in M#1 is independent of the number of parties observed in the remaining EU measurements.

We analyzed the effects of the mean betweenness centrality between the pre-GDPR measurement and the post-GDPR measurements. The betweenness centrality is an index to measure how many shortest paths in a graph include a node. The higher the betweenness centrality of a node, the higher the amount of information that flows through this node. Similar to our syncing connection regression, we performed a linear regression of the mean betweenness centrality and found a statistically significant ($\alpha = .01$ with p-value $< .001$) decrease in the betweenness centrality. The number of well-connected nodes and connected nodes decreases, which also means that fewer companies sync ids with each other. These observations are in line with the results of our previous observations of the ecosystem that the general structure within the ecosystem did not change, but we have shown that ID syncing dropped in a statistically significant way.

References

1. Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., Diaz, C.: The Web Never Forgets. In: Proceedings of the 21st ACM Conference on Computer and Communications Security. pp. 674–689. CCS'14 (2014)
2. 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 (2016)