# FACEBOOK AND GOOGLE TECHNICALLY PROVEN TO BE RIGGING POLITICS WITH DIGITAL TRICKS

A look at the landscape of third-party inclusions for  websites with left and right bias in lead-up to the midterms.

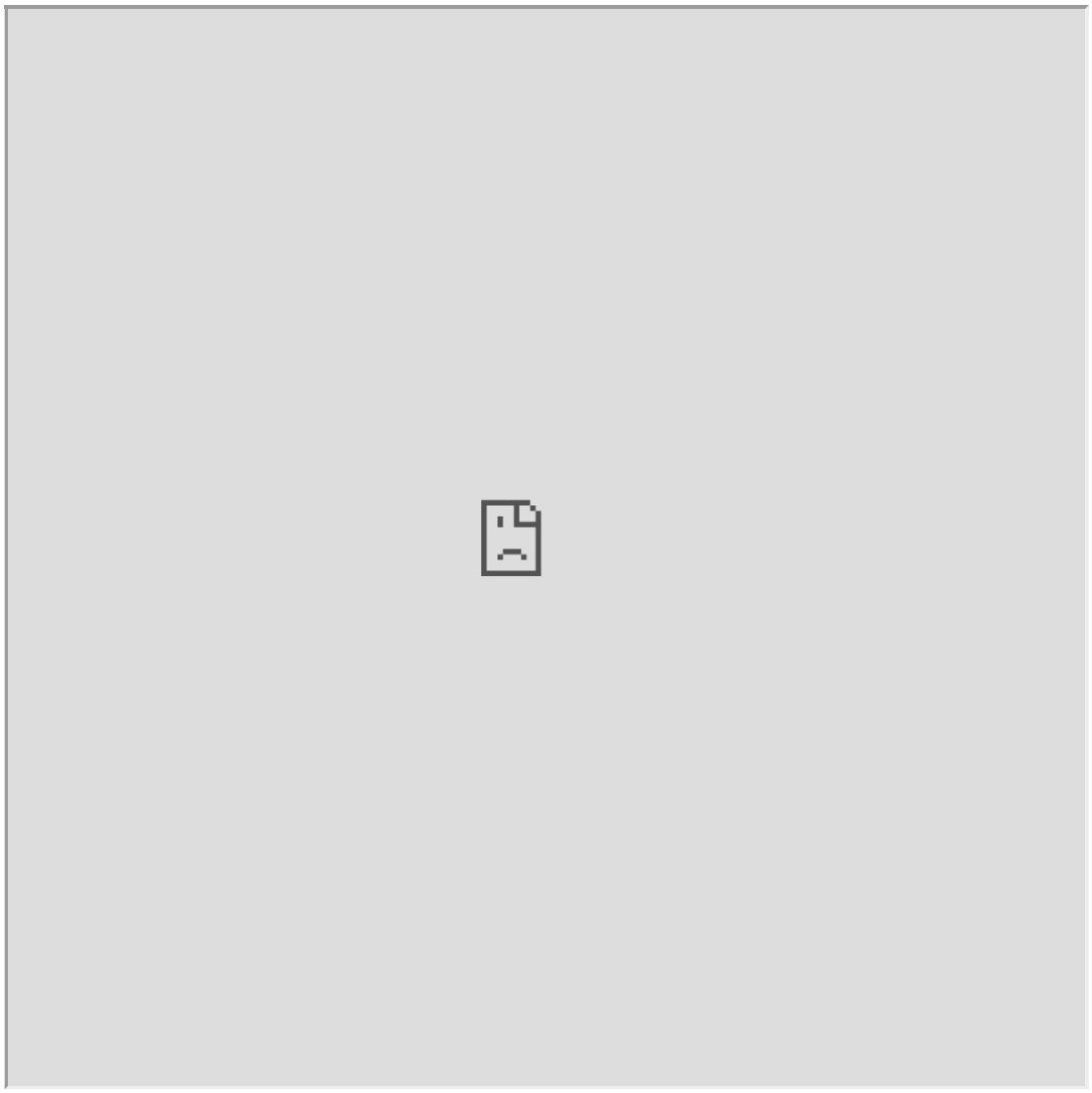
Netograph ingests a real-time stream of URLs from social media and our own crawling engine. Each URL is loaded in an instrumented browser, and its behaviour is captured in minute, exhaustive detail. All of this data is then fed into a vast database, and indexed for querying and analysis. This means that we have an ever-growing history of the underlying structure of the web: the rise and fall of third party trackers, Javascript inclusions, certificate authorities and other behaviours that are not visible to users at first sight. There are many stories to tell from this fascinating dataset - we plan to publish about one post a fortnight on this blog looking at some facet of our work.

For our first post, we will take a high-level look at website behaviour on the left and right of the political spectrum during the period approaching the US mid-term elections. To do this, we needed a reasonable independent classification of the political stance of online media, and after some searching, we eventually settled on Media Bias Fact Check. The site doesn't reveal much about its creators, but the methodology seems fair, and the classifications pass the sniff test. We retrieved the domains Media Bias Fact Check lists as left biased and right biased, and used Netgraph's query API to fetch a sample of captures prior to the midterms for each site. This gave us data for 149 sites on the left, and 146 on the right.

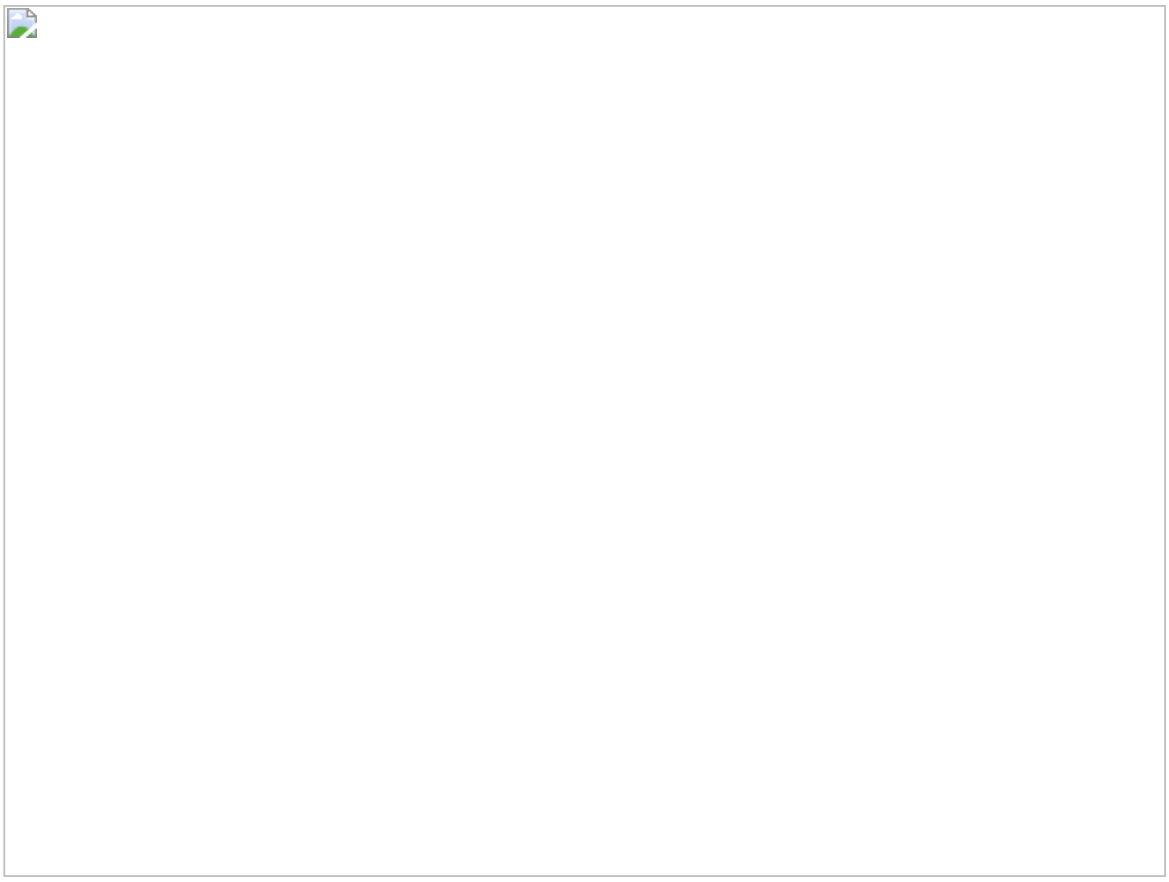
## The third party landscape

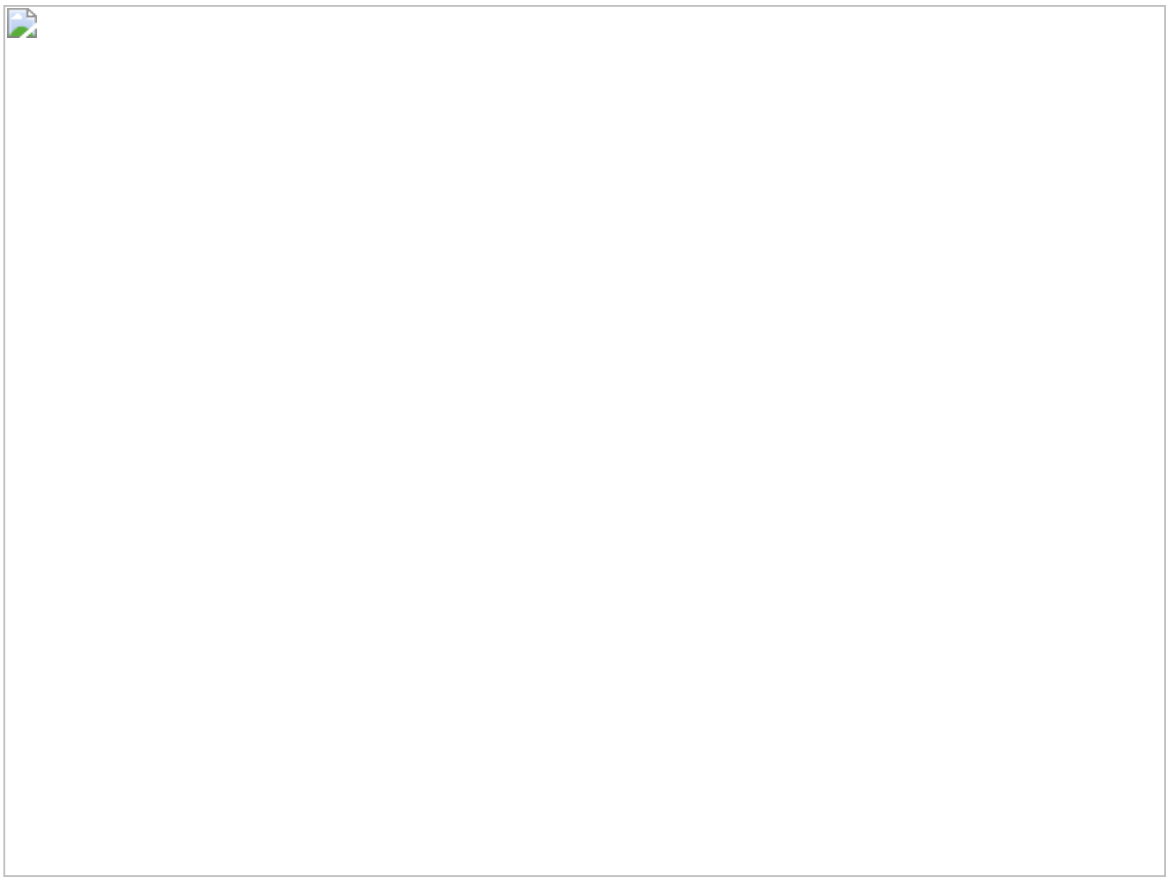
We define a **third party** as any unrelated domain that is contacted while viewing a URL. Third party connections can be innocuous - for instance, when loading assets from a CDN. However, third-party inclusions - usually in the form of JavaScript - are also an essential hook for the vast industries of user tracking, ad brokering, and profile building that underpin much of the web.

It might be surprising to some just how chatty most websites are. In this dataset, the **median** site by number of included third-party domains is the right-biased lifesitenews.com, and it contacted 32 third-party domains. We can see what this looks like visually using Netograph's interactive starmap visualisation:



To compare the behaviour between left and right, we graphed the distribution of the number of third-party domains contacted.





Distribution of the number of third-party domains connected to

So, we can see that nearly 50 websites on the left contacted between 0 and 20 third-parties. We can also see that there is a huge outlier on the left, which contacted more than 160 third-party domains on load. This is bust.com, which describes itself as a "groundbreaking, original women's lifestyle magazine". You can check out the details in our capture report (don't worry, this link doesn't take you to the site itself).

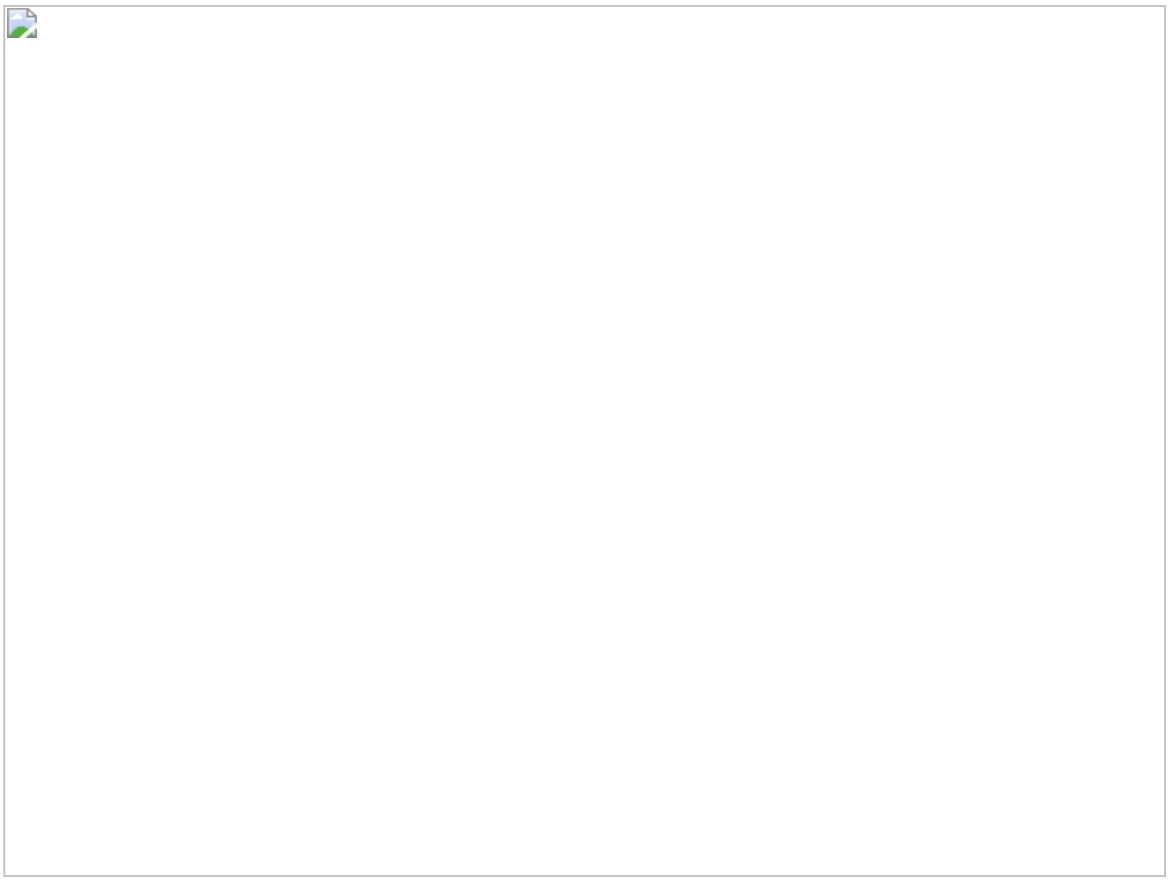
Comparing the left and right graphs, we can see that **websites on the right tended to be less discriminating about third-party inclusions than websites on the left**. The median left-biased site contacted 31 domains, and the median right-biased site contacted 37.

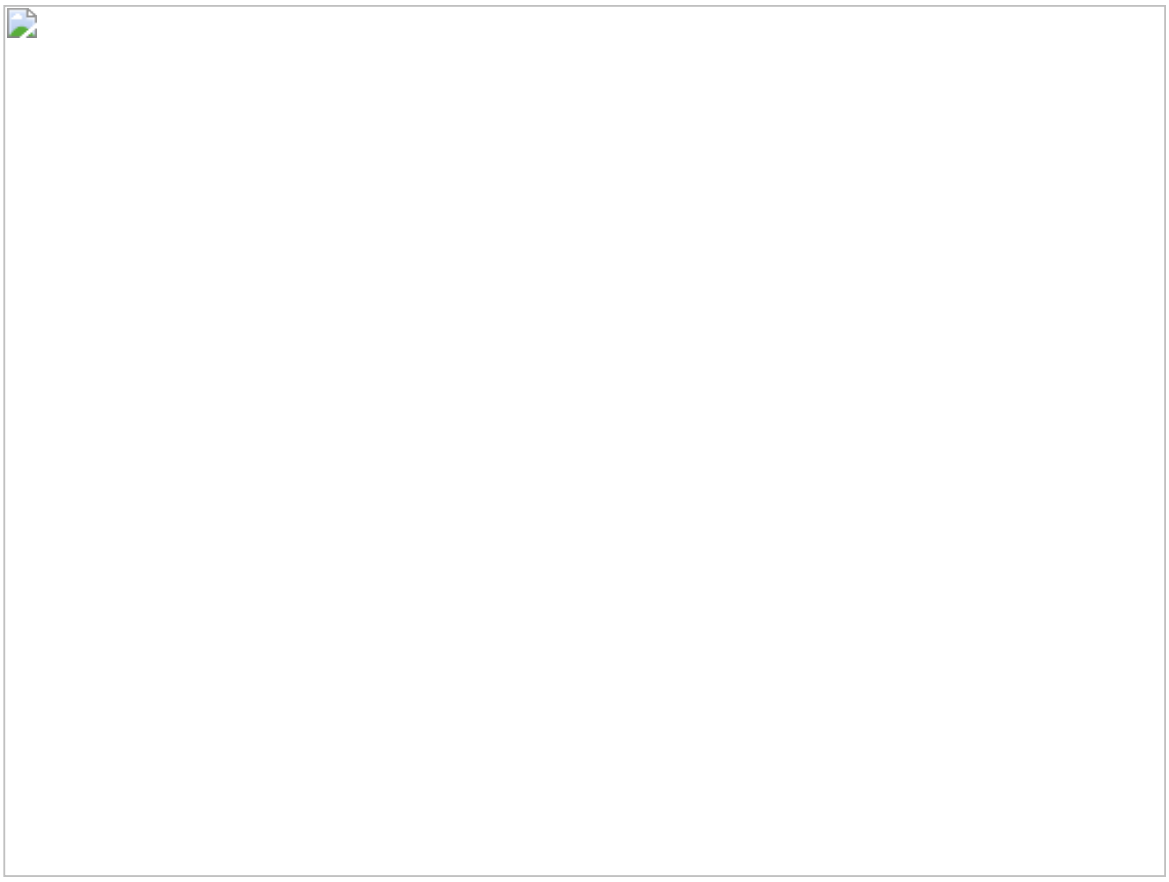
## Trackers

Another aspect of third-party behaviour is user tracking, which often takes the form of persistent state set in the user's browser by third-party domains. The most common type of persistent state is cookies, but Netograph also monitors other types, like IndexedDB, WebSQL and local storage.

This is a rapidly evolving part of the web landscape, with Firefox now joining Safari in taking active measures to prevent third-party domains from setting tracking state without permission. This has spurred something of an arms race, which we're monitoring closely and will have much to say about in future posts.

We graphed the number of third party domains that set persistent state for each capture in our sample, in the same way as we did for third-party connections. We kept the scale and graph bounds constant to make comparison easier.





Distribution of the number of third-party domains that set persistent state

The story here is less clear than for connections. The difference is small, and this is reflected in the fact that the median number of domains that set state for left-biased websites was 18, while the median for right-biased websites was 19. **The takeaway is that left-biased and right-biased sites included about the same number of third party domains that set state.**

There is another way to look at this data that highlights some clearer distinctions. The visualisation below considers the top third-party domains that set state on our sample sites, and looks at the proportional split in the occurrence of each domain between the left and the right. A domain that occurred equally on both sides of the spectrum would lie on the black center line at 50%.

> The percentage share of observations for each domain on the left and right

The domain that occurs most disproportionately on the left is bluekai.com, a data collection outfit owned by Oracle. Bluekai builds profiles of users and their interests, which is then used to target ads.

The site that skews most to the right is more interesting - **Facebook was substantially more likely to set persistent state in browsers visiting right-biased sites**. It's not immediately clear why this might be the case - we suspect that right-biased sites are more likely to integrate certain of Facebook's social features than left-biased sites. We plan to dig into these differences in more detail in future posts.

# Google and Facebook, the behemoths that bestride the web

In this post we've looked at the data with an eye to teasing out differences between the left and the right leading up to the midterms. However, perhaps the most remarkable figure in our dataset is to be found in the domains that unite both sites. The graph below shows the percentage of sites in our sample on which we observed each domain.

Total percentage of sites in our survey on which we saw each third party domain

The most prevalent site is doubleclick.net, a domain that forms part of Google's online advertising network, which was seen on over 90% of the sites we surveyed. The next most prevalent is Google Analytics. Going down the list, we find something remarkable: **6 of the top 10 domains included on sites in our sample belong to Google**. In fact, if we collapse these domains together under one banner, then **Google was included on 97% of the sites in our sample**. Similarly, if we collapse facebook.com and facebook.net, we find that **Facebook traffic was seen on 76% of sites in our sample**. It's hard to over-state the immense reach of these two giants on today's web.

## Next steps

In this post, we looked at behaviour differences for one slice of data at a fixed moment in time - the period immediately before the US mid-term elections. An interesting next step would be to look at how websites evolved over time - are there differences between the period before the midterms, and, say, the week following? Another question left unanswered is why, exactly, Facebook is setting state more often on right-leaning sites. We have the data to answer both these questions, and will return for a closer look down the track.


**Aldo Cortesi**

Read more posts by this author.