Tracking the Trackers: Analysing the global tracking landscape with GhostRank

Sam Macbeth <sam@cliqz.com>

Cliqz GmbH, **https://cliqz.com**

July 2017

1 Introduction

In April 2016 we published our innovative anti-tracking method — based on the algorithmic detection of user identifiers in tracking requests — and included a large-scale study of online tracking in the wild, using data from 200,000 German users over a two week period [3]. This study covered 21 million page visits to over 350,000 different sites, the largest such study to-date. In this paper we rerun the same analysis, this time using data collected by the Ghostery [2] browser extension's GhostRank feature [1]. This study's scope therefore extends beyond those previous, as it:

- Covers all major browsers (Chrome, Edge, Firefox, and mobile), instead only desktop activity on the Cliqz browser;
- Captures regional differences in tracking, as the dataset is international;
- Increases the number of participants from 200,000 to 850,000, and number of page loads from 21 million to 440 million.

2 Data Overview and Processing

The Ghostrank data set is data gathered from users of the Ghostery browser extensions and mobile apps, who have opted-in to the collection of information about trackers on pages they browse to. It consists of two types of messages. Firstly, a *page* message is sent for each page visited by a participating user, secondly, the *tag* message is sent for each request during a page load which matches a pattern in the Ghostery tracker database. In order to complete an analysis per page load, these messages are combined into a single message which describes a page load and the tags within. This is done by creating a signature

[1] Ghostery was acquired by Cliqz in February 2017

for each page load against which we can match tag messages. If no tags match a page signature, we assume there were no tags for this page. Alternatively, if a tag cannot be matched to a page, we discard it.

Once we have combined *page* and *tag* messages to describe *page loads*, we must further clean the data. The Ghostery extension sends *page* messages for any page loaded in the browser, however several of these are not relevant to our analysis. We filter the following URLs:

- Chrome new tab page ($https://www.google.com/\_/chrome/newtab$)
- URL shorteners, or automatic redirects (e.g. $t.co$ and $https://www.google.com/url$)
- Non HTML pages (for example images or PDF documents)

Tag messages also contain information about whether the user had enabled blocking for the tracker in question. We can use this information to build up a picture of how many users enable blocking. Furthermore, comparing the pages loaded by users with blocking enabled vs. disabled tells us more about the tracking ecosystem — namely which trackers depend on others to load them into a page. Note, we cannot determine from this data if a user with Ghostery blocking disabled had another blocking system enabled to block trackers. This could cause the number of trackers be under-reported.

We take GhostRank data over two weeks, from May 1st to 14th 2017 (inclusive). After combination, this dataset contains 455M page loads, which is then reduced to 440M after the described filtering. Of this total, 112M page loads were done by users with blocking disabled, 33% of the total.

2.1 Regions

The GhostRank dataset reports a region for each page load, based on a lookup of the IP sending the message. We can therefore split the dataset into each regional set to compare the extent of tracking in each region. As can be seen in Figure 1, the US constitutes by far the largest proportion of the dataset, with 27% of the page loads. Russia, France and Germany each represent user groups, with around 10% each, and the remaining countries have under 5%.

2.2 Evaluation

As the GhostRank dataset is collected from real users' browsing, this dataset offers significant insight compared to the synthetic crawls of other studies such as those using OpenWPM [1]. Firstly, we can see deeper into the web and observe tracking where the crawler cannot: logged into private websites, during E-commerce checkouts, online banking, or after form submissions. Secondly, the data contains explicit weightings to allow us to quantify the popularity of sites and thus determine the impact of tracking there. This means that rather than having to estimate popularity to know the impact of tracking on a specific site, our data also contains the popularity in terms of number of page loads. Thus,

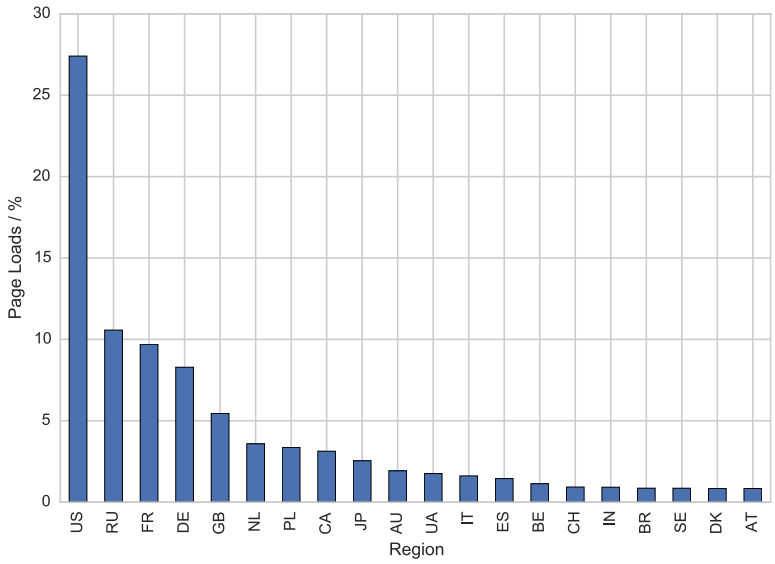

Figure 1: Top 20 Regions in the Ghostrank dataset, ranked by proportion of page loads

when we measure tracking reach in terms of page loads, this is a very accurate measure of the impact for the average browsing session.

As mentioned earlier, there are also some caveats to the results due to the nature of the data collection. Firstly, the messages collected for GhostRank make our calculation of page loads an approximation, as we have to group the *tag* messages back with the pages they came from. This could introduce errors, such as combining repeated page loads into a single one in some cases. Secondly, as data is only sent for URLs which match rules from the Ghostery database, trackers which are not in this database will be ignored by this analysis. This would cause tracking to be under-reported. Finally, other sources of blocking outside of our control will affect the results. Users may have adblockers, firewalls or other tools installed, which will reduce the number of trackers seen by the Ghostery extension.

3 Results

In this section we report and discuss the results of our analysis of the GhostRank data, and make comparisons to other similar studies, namely our own study [3] (henceforth Cliqz study) and that of [1] (henceforth OpenWPM) the current largest synthetic study of tracking.

Our analysis focuses on the following questions:

- How much tracking can be observed on any given page load?
- How much reach do specific tracking widgets and companies have over users' web browsing?
- How does the tracking ecosystem vary between different regions?

3.1 Trackers per page

The Cliqz study presented a figure depicting the number of trackers seen per page load. We split this into two measures: 1) equests to *potential* trackers, and 2) requests to *potential* trackers with *unsafe* data detected. We defined *unsafe* data as data elements included in a request, which have the potential to identify an individual user uniquely. As Ghostery utilises a blocklist to detect trackers, we assume that a match with the blocklist means that the request would have sent *unsafe* data, or led to it being sent. Therefore we compare to the second measure.

Furthermore, to maintain the equivalence of the measure, we only consider page loads from users with blocking disabled. This is because blocking will affect the number of trackers detected on a page.

Figure 2 shows the frequency distribution of trackers per page load, comparing the Cliqz study results to the Ghostrank data. Figure 3 shows the same comparison but by unique domain seen - taking the mean number of trackers for each first-party domain visited.

These results firstly reinforce the findings of the Cliqz study. Figures 2 and 3 are direct comparisons of the number of trackers per page load and for each unique domain respectively for the two datasets. The general trend is consistent, and notable differences can be attributed to variations in data collection and measurement. These differences are:

- The Cliqz dataset directly measured 'trackers sending unsafe values' by measuring the presence of a tracking cookie or unique identifier in a request. In the GhostRank data, we measure simply a match of a rule in the Ghostery database of trackers, under the assumption that this match would directly correspond to a request that would send an unsafe value. This assumption is not tested, however, and it may be that some rules could match when no tracking request would be made, leading to an over-reporting of tracking. Likewise, some third-parties may not be in the Ghostery database, but in the Cliqz study we may have measured tracking, leading to under-reporting.
- The demographics of the survey participants differ: users in the Cliqz study were based in Germany using the Cliqz browser extension, whereas this study uses global data with a regional distribution as shown in Figure 1. Furthermore, Ghostery users tend to be more privacy conscious and often install multiple privacy extensions [TODO: ref survey results], which can affect the number of reported trackers.

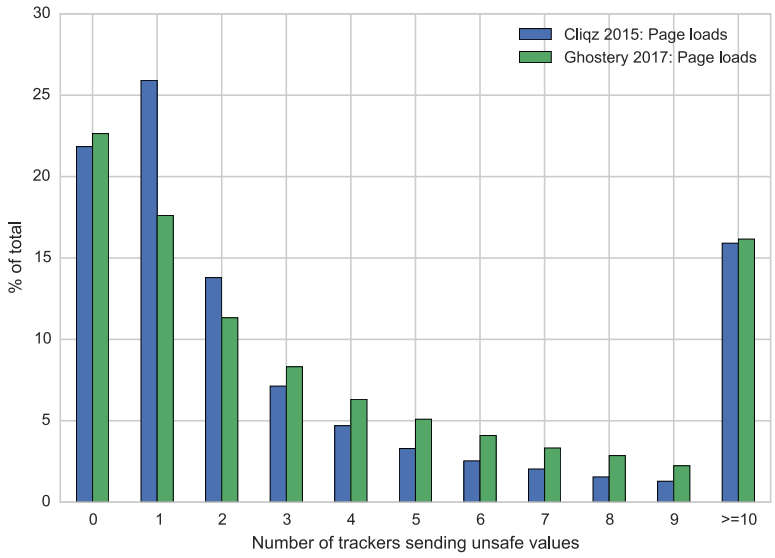

Figure 2: Number of unique tracker domains detected per page load in Cliqz and Ghostery datasets.

The results show pervasive tracking, with over 15% of page loads and 10% of all sites with 10 or more trackers seen. This shows that a large proportion of internet traffic has an extreme amount of tracking, and that this is more prevalent on popular websites. At the other end of the spectrum, 23% of web traffic has no tracking, and 21% of sites in this group. While this seems promising, further analysis of this traffic shows that a significant proportion is from certain web properties. Google's own sites constitute 15% of the page loads with no tracking, where they can obviously track the user as a first-party. Facebook constitutes 6%, and Wikipedia 5%. Overall, almost 50% of the page loads with no tracking come from only 25 web properties that drive large quantities of traffic.

3.2 Prevalence of tracking widgets

The GhostRank data also exposes which third-party services are most prevalent across the web. Each services is defined by a set of rules in the Ghostery database, which are then mapped to a user-understandable name shown in the user interface. In some cases, this name is a company, and in others, multiple services for a company are split into different names.

Table 1 shows the top 20 services in global reach, measuring the proportion of page loads where this service was detected. The dominance of Google and

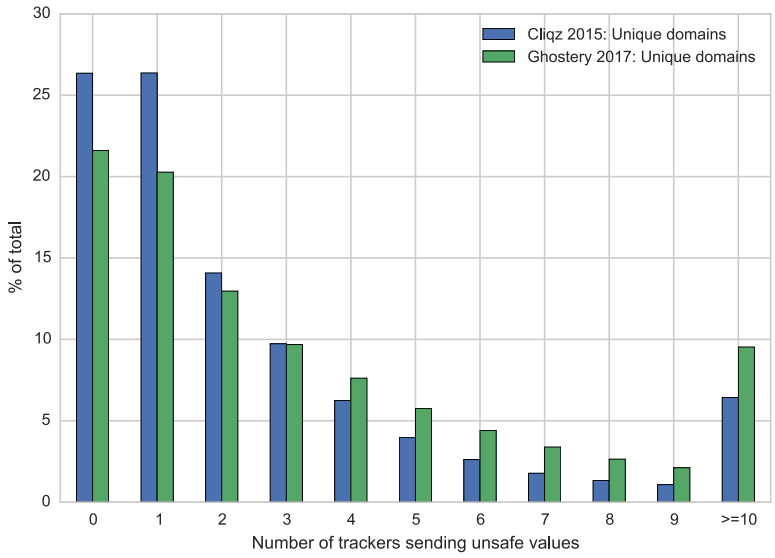

Figure 3: Number of unique tracker domains detected for each unique domain in Cliqz and Ghostery datasets.

Facebook can be seen here, with 8 Google[2] and 3 Facebook services present, respectively.

This result shows, as other studies have also demonstrated, the significant reach of certain trackers across the web, and in particular Google's dominance in this area. The reporting here, using service names rather than domain names, makes some comparisons with OpenWPM difficult, but we can still draw some interesting observations.

The OpenWPM study generally reported a higher reach for the same trackers as we report in Table 1. We can attribute this to a bias in the data collection: the OpenWPM data was collected by crawling the home pages of the top 1 million sites. This means that, firstly, there is no data from deep links inside sites, just homepages, and secondly, the reported reach figure weights all sites equally (i.e. the presence of a tracker on the most popular site is counted equality as its presence on the millionth most popular site). Our data addresses these issues by using data collected from real user browsing, thus these figures represent the probability of encountering these trackers during a normal browsing session.

[2]In addition to the services with Google in the name, DoubleClick is also a Google company.

Table 1: Reach of third-party services over all page loads

Rank	Service	Proportion of page loads %
1	Google Analytics	46.4
2	Facebook Connect	21.9
3	DoubleClick	18.5
4	Google Publisher Tags	15.1
5	Google Tag Manager	14.6
6	ScoreCard Research Beacon	12.1
7	Google Adsense	9.9
8	Twitter Button	9.0
9	Yandex.Metrics	7.5
10	Facebook Custom Audience	7.1
11	Facebook Social Plugins	6.7
12	Criteo	6.5
13	Google+ Platform	6.5
14	New Relic	6.0
15	Quantcast	5.8
16	Amazon Associates	5.7
17	LiveInternet	5.5
18	AppNexus	5.5
19	Google Dynamic Remarketing	5.4
20	Google AdWords Conversion	4.8

3.3 Reach of tracking companies

In the previous section there were several cases where different services belonged to the same company. In this section, we look at the reach of companies across the pages visited in the dataset. We can map the domains on which the services were detected to the companies that own them. We can further include the addresses of the first-party sites visited to create a picture of how much user browsing each company is a party to. The results for the top 20 trackers are shown in Table 2.

Here we see the dominance of Google: when their third-party services (analytics, advertising and social) are combined with first-party services (Search, Maps, Youtube, etc.), they are party to over 64% of all web-browsing worldwide. Similarly Facebook's reach is approaching 30% as their advertising tracking tools gain more reach.

3.4 Regional Differences

Using the GhostRank data, we can observe how the amount of tracking per page varies in different countries. Figure 4 shows the cumulative frequency of trackers per page load in several regions, compared to the global average. For this we can see:

Table 2: Reach of tracking companies

Rank	Company	Proportion of page loads
1	Google	64.4
2	Facebook	28.8
3	comScore	12.2
4	Twitter	11.0
5	Amazon.com	10.5
6	Yandex	8.0
7	Criteo	6.5
8	New Relic	5.9
9	Quantserve	5.8
10	LiveInternet	5.5
11	AppNexus	5.5
12	Adobe	4.8
13	AddThis	4.0
14	Microsoft	3.8
15	Mail.Ru	3.4
16	VKontakte	3.3
17	TNS	3.2
18	Automattic	3.2
19	Taboola	3.1
20	Chartbeat	2.9

- The US, Russia and UK have more trackers per page load than the global average, while Germany, France and India have fewer.
- The distribution in the US and UK is skewed to larger numbers of trackers (> 10), while Russia has more pages in the 1-10 region.

As only trackers in the Ghostery database are reported, if the coverage of this database is lacking in certain regions, the number of trackers will be under-reported. However, since the results for Germany match those from the Cliqz study, which consisted of only German users, it suggests that this is not the case in Germany at least. We cannot, however, make this same assertion about the results for India.

These results show that there are significant regional differences in the tracking ecosystem. This could be caused by several factors:

- Number of players in the advertising supply chain for a region/language: The advertising supply chain represents the vast majority of tracking companies, and the fact that many networks load each other into pages during ad loading contributes significantly to high tracker counts on pages. Therefore, in regions where there are fewer players in this space, there will be less scope for large numbers of trackers per page.

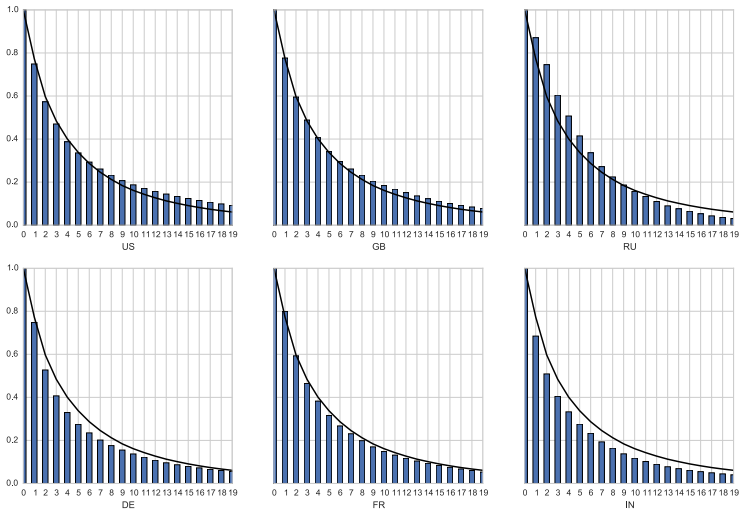

Figure 4: Comparison of proportion of pages with trackers between regions

- Different levels of acceptance of third-party tool usage versus in-house development: In different markets, companies may be faster to turn to third-party tools on their websites. This may be due to cultural difference, or concerns about data sharing regulations. For example, stricter data protection regulation in the EU may make companies more cautious about adding tracking beacons to pages.

- Different rates of external tracker blocking: If an external browser extension or blocklist is blocking tracking requests it will affect these results. Particularly if advertising networks are blocked, this will lead to a lower number of trackers per page. Different levels of adoption of these tools in the regions shown here could then explain some of the differences observed in the results.

3.5 Differences in Ecosystems

We can dig deeper into the results from the previous section to see how the presence of specific services varies between regions. As mentioned above, there are differences in the number of trackers seen by page, but are there simply fewer of the same group of trackers, or completely different sets of trackers in each region?

We can measure how different the tracking ecosystem is in each country using a distance metric to measure how much the observed reach for the service

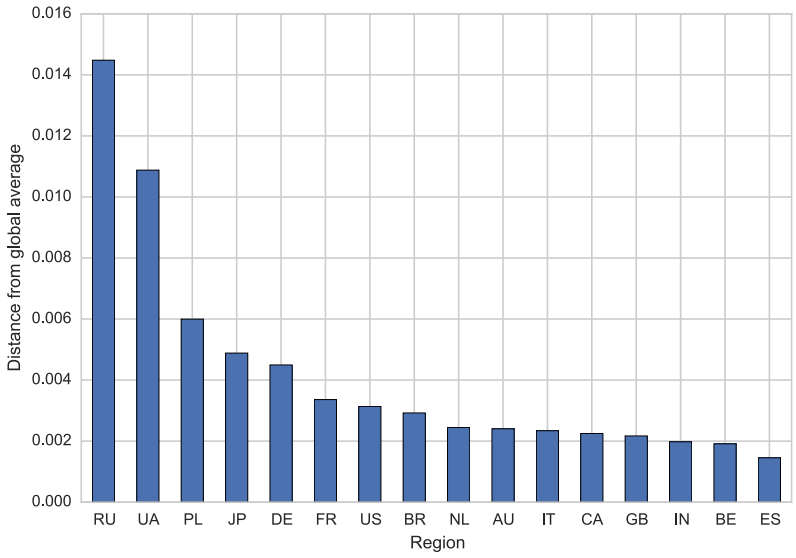

Figure 5: Differences in third-party services in top regions.

in a particular region differs from what we would expect from the global average. We analyse the top 20 regions from the dataset in terms of data quantity, and the top 50 trackers in each of these regions. For each region and tracker pair, we measure the difference between the proportional reach in that region versus the tracker's average over all 20 regions. We then take the sum of squares for the metric in each region to quantify the magnitude of the differences. Figure 5 shows this measure for the top regions in our dataset. We also show the absolute and proportional differences for selected trackers in a few regions in Table 3.

Our results show that:

- Russia and Ukraine have the most differentiated tracking ecosystems. Here, Yandex takes the place of Google, with their Google Analytics competitor at over 50% reach within Russia, an improvement almost 7 times their global average. Several other services have a similar relationship, with 6 to 10 times more traffic in Russia than elsewhere. We also see American companies have less scope in Russia, with Google, Facebook and AppNexus exhibiting between 25 and 70% less reach.

- Poland has a few region-specific advertising networks, including Gemius and BBelements who both have significant gains in this region. Interestingly, Google and Facebook also have better-than-average reach - an indication that there is generally high tracking on Polish sites.

- Japan is another outlier, with a selection of services that do not operate

Table 3: Regional differences in tracking reach for selected trackers

Region	Tracker	Global	Regional	Diff.	Change
RU	Yandex.Metrics	6.7	52.0	45.3	678.2
RU	LiveInternet	5.0	39.5	34.4	682.2
RU	TNS	2.5	25.7	23.2	927.1
RU	Mail.Ru Group	2.9	23.3	20.4	708.9
RU	VKontakte Widgets	2.6	19.0	16.4	628.2
RU	Google Publisher Tags	14.9	8.7	-6.3	-41.8
RU	Facebook Connect	22.8	16.6	-6.1	-26.9
RU	AdRiver	0.7	6.4	5.7	819.0
RU	AdFox	0.4	4.5	4.1	1015.7
RU	AppNexus	5.3	1.6	-3.7	-69.5
PL	Gemius	2.6	21.5	18.8	710.4
PL	Facebook Connect	22.8	36.6	13.9	60.8
PL	ScoreCard Research Beacon	11.5	4.0	-7.5	-65.1
PL	Criteo	6.9	13.5	6.6	94.6
PL	Google Tag Manager	15.4	21.7	6.3	41.0
PL	Facebook Social Plugins	7.2	13.3	6.1	83.9
PL	Google Analytics	47.2	53.0	5.8	12.3
PL	BBelements	0.3	4.4	4.1	1385.6
JP	Twitter Button	9.6	20.1	10.4	108.3
JP	Hatena	0.5	8.5	8.0	1479.1
JP	ScoreCard Research Beacon	11.5	3.5	-8.0	-69.5
JP	MicroAd	0.5	7.8	7.3	1417.1
JP	Google Publisher Tags	14.9	8.4	-6.5	-43.7
JP	Yahoo Analytics	1.5	7.1	5.6	381.3
JP	Yahoo! Retargetting	0.4	5.9	5.5	1428.3
JP	FreakOut	0.3	4.0	3.8	1447.6
DE	INFOnline	1.1	16.0	14.9	1302.3
DE	Google Analytics	47.2	39.5	-7.7	-16.3
DE	Facebook Connect	22.8	17.2	-5.5	-24.3
DE	ScoreCard Research Beacon	11.5	6.2	-5.3	-46.2
DE	Yandex.Metrics	6.7	1.8	-4.9	-73.5
DE	DoubleClick	18.9	14.3	-4.5	-24.1
DE	Twitter Button	9.6	6.1	-3.5	-36.3
DE	Adition	0.4	3.8	3.3	780.0
DE	Webtrekk	0.7	3.7	3.0	446.4
DE	emetriq	0.2	2.7	2.5	1339.0
DE	VG Wort	0.2	2.3	2.2	1289.7
DE	Piwik	1.1	2.9	1.8	168.1
US	Amazon Associates	4.1	11.8	7.8	191.9
US	Quantcast	4.5	10.6	6.1	134.6
US	ScoreCard Research Beacon	11.5	16.3	4.9	42.6
US	Taboola	2.6	4.9	2.3	85.7
US	Google Adsense	10.9	8.7	-2.2	-20.1
US	LiveRamp	1.4	3.6	2.2	158.6
US	Advertising.com	1.7	3.8	2.1	126.2
US	Moat	1.1	3.0	2.0	184.6
US	Google+ Platform	7.1	5.4	-1.7	-24.2

elsewhere, such as Hatena, MicroAd and FreakOut. Twitter Buttons are also twice as common in Japan compared to other regions, and Yahoo's platforms are much more prevalent, with 14 times more reach for Yahoo! Retargetting for example.

- Germany shows a general trend ofe less tracking, with many US services for example seeing reduced reach. We also see an increased usage of Piwik, indicating a preference for self-hosted analytics services. We can also see regional players in market research and Adtech, with INFOnline supplanting ScoreCard Research (a.k.a. ComScore), and Adition, Webtrekk and emetriq appearing with significant reach.

- The US tracking ecosystem is not as different as the other regions mentioned. Most of the major tracking companies originate from the US and are global in reach. There are still some interesting differences, with companies such as Amazon and Quantcast, who more than double their reach in this region.

These findings enhance the story from Section 3.4. First, this partly explains that result: the numbers of trackers per page varies between regions because different trackers are loaded on pages in different regions. This may be because of developer culture; whether one uses Google Analytics, Yandex.Metrics or Piwik for site analytics may depend on cultural perceptions of which tool, or company, is better. In this regard, tracking companies may perform better in their home countries because they can communicate with publishers and sell their services in their native language.

The other factor that will change the trackers loaded on pages are advertising networks. Online advertising networks are global, and targeted ads are based on who they think the user is, not on the content alongside which the ad will load. This can be seen when one opens a foreign-language site with advertising: the ads will likely be in your language, and for products from your country. The ad-system has determined your location - usually from the IP address - and served ads relevant to your region, often from region-specific ad suppliers. Therefore, the same website can have a different set of trackers by region, since ad networks in your region may bid higher for your ad impression than those from other regions.

4 Evaluation & Conclusion

These study results represent the largest study on tracking to-date - with an order of magnitude more data than our previous study. The results agree with previous studies, such as our own [3], and others [1], in that the extent of online-tracking is extraordinary, and that in addition to the dominance of major players such as Google and Facebook, there is a long tail of companies also hovering up significant quantities of user browsing data.

This study also adds new insights showing that tracking varies significantly between different regions, but also that it is remarkably global: Google and Facebook are omnipresent across the world, while other companies fortunes vary a great deal across markets. We can also see differences in the general level of tracking per page in different regions. Our results show that, for example, the average page load in the US will contain more tracking than in Germany.

Web tracking has become pervasive, and with Google and Facebook tracking 64% and 29% of pages loaded on the web, it is becoming almost impossible to avoid. Additionally, 15% of pages will send data to 10 or more different companies. As well as being a significant burden on resources (both CPU and network) to load scripts from all these different parties, there is little transparency about what is being shared and with whom. Users are increasingly turning to privacy tools like Ghostery to notify them about who is tracking on each page, and allow them to 'opt-out' as they wish. However, it remains to be seen if the increase in this behaviour will lead to a change in the pervasiveness of tracking.

References

[1] S. Englehardt and A. Narayanan. Online tracking: A 1-million-site measurement and analysis. [Technical Report], May 2016.

[2] Ghostery. Ghostery. `https://ghostery.com/`.

[3] Z. Yu, S. Macbeth, K. Modi, and J. M. Pujol. Tracking the trackers. In *Proceedings of the 25th International Conference on World Wide Web*, pages 121–132. International World Wide Web Conferences Steering Committee, 2016.