Data Harvesting 2.0: from the Visible to the Invisible Web

Claude Castelluccia, Stéphane Grumbach, Lukasz Olejnik

The Twelfth Workshop on the Economics of Information Security,
Georgetown University, Washington, D.C., June 11-12, 2013

Data Harvesting 2.0: from the Visible to the Invisible Web

Claude Castelluccia
INRIA
claude.castelluccia@inria.fr

Stéphane Grumbach
INRIA
stephane.grumbach@inria.fr

Lukasz Olejnik
INRIA
lukasz.olejnik@inria.fr

Abstract

Personal data are fuelling a fast emerging industry which transform them into added value. Harvesting these data is therefore of the outermost importance for the economy. In this paper, we study the flows of personal data at a global level, and distinguish countries based on their capacity to harvest data. We establish a cartography of international data channels on the visible and invisible Web. The visible Web is composed of the sites that are available to the general public and are typically indexed by search engines. The invisible Web refers to tags, Web bugs, pixels and beacons that appear on Websites to track and profile users.

It is well known that the US dominate the visible Web with more than 70% of the top 100 sites in the world. We show that this domination is even stronger on the invisible Web.The largest proportion of trackers in most countries are indeed from the US. Apart from the US, two countries exhibit an original strategy. China, which dominates its visible Web with a majority of local sites, but surprisingly these sites still contain a majority of US trackers. Russia, which also dominates its visible Web, and is the only country with more local trackers than US ones.

1. Introduction

Sun Tzu famously wrote in the 6th century BC: *"If you know your enemies and know yourself, you will not be imperilled in a hundred battles; if you do not know your enemies but do know yourself, you will win one and lose one; if you do not know your enemies nor yourself, you will be imperilled in every single battle"* [1]. This statement might become more relevant than ever with the explosion of information now available from the Internet and Web 2.0 systems. Some countries are widely collecting data on the visible and/or the invisible Web about their citizens as well as sometimes about citizens of other countries. Some other countries on the other hand are relying mostly on foreign systems, thus letting a large amount of their data be handled outside their borders. This discrepancy between these two approaches to the management of personal data could result in information asymmetries between the players, whether industrial or governmental, in their capacity to access strategic data [23]. It might seem irrelevant to know which corporations are handling and concentrating data, and in what countries they are based, but it has tremendous consequences related in particular to the jurisdiction that applies, not to mention the economic impact, direct or indirect. Some complex fiscal issues have already been raised in Europe on the economy of the data, which at this stage is mostly invisible to European states[1].

Personal data have become an essential resource for the new economy of the information society, much like iron ore or crude oil were for the industrial economy. Developing tools to harvest personal data is therefore of strategic importance to catch up with the digital revolution. Harvesting data is mostly done by systems, such as search engines, social networks, clouds, etc. where people provide their personal data in exchange for a service, which most of the time is accessible essentially for free [2]. More precisely, the private data given by users, constitute the means by which users pay for services. It is therefore not only a resource, but can be seen in a sense as a virtual currency. Harvesting data can be done as well in more subtle ways, on the "invisible Web", by tracking people, which is carried on mostly by third party entities.

The main **objective** of this paper is to analyze the current world situation, and perform a geographical analysis of data harvesting on both the *visible* and the *invisible* Web. The visible Web is composed of the sites that are available to the general public and are typically indexed by search engines. The invisible Web refers to tags, Web bugs, pixels and beacons that appear on Websites to track and profile users.

This paper aims at answering the two following questions:

[1] Some countries, such as France [18], actually intend to tax the tracking and data-mining providers. The premise behind this assumption is that consumers pay for service with their private data and this good is exchanged without issuing a tax.

[2] Note that most people pay about 1000 dollars/year (ISP, cellular data subscription,...) to access the Internet. It is arguable whether these services are really free.

1. What are the most influential Web systems? In which countries are they based? What are their regional influence?
2. Which are the biggest trackers on the Internet? In which countries are they based? What are their regional influence?

In other words, this paper aims at analyzing the cyber-strategies of different countries in terms of data harvesting on both the visible and the invisible Web.

Interestingly, the answers to the previous questions reveal very different patterns, not necessarily correlated with the penetration rate. Numerous studies have measured the level of penetration of the information society such as the World Economic Forum [?]. Recently, the Web Foundation[3], led by Tim Berners-Lee, launched its Web Index, which as previous studies ranks top of the list North America and Northern Europe, as well as some countries of Asia. The index measures three key attributes of the Web: "Web readiness", for the communications infrastructure; "Web use", for the population online and the contents available; and "the impact of the Web", for its economic, social and political impact. The Web Foundation gives a high weight to the political openness. China ranks 29^{th} out of 61 countries in this index, a rather low rank, which reveals though an increased Web use, but a relatively slow evolution of Web content.

There is one measure that is little taken into account in these rankings, namely the local development of the Web 2.0 industry, which offers an indicator of the potential strategic activity on Big Data in the country[4]. The US would be ranked at the top position for such a criteria. They have indeed developed the strongest industry worldwide, with most of the first online social systems accessed in the world, such as Google, Facebook, YouTube, Yahoo!, Wikipedia, Windows Live, Twitter, Amazon, to name the most popular. With these corporations, USA harvests private data of people all around the World that can be analysed for an unpredictable set of purposes, with considerable economic impact. Of course the US have a dominant position in a number of strategic sectors of the information society, including the operating systems, the browsers, or the clouds that support the systems of the Web.

Technically speaking, tracking is made possible by the design of the HTTP protocol [16] itself. Third-party scripts can indeed be used for tracking purposes. Every inclusion of a third-party script on a visited site requires the browser to execute a request to this third-party server, download the script and execute it into the user's browser.

Browser cookies [2] are traditionally used to maintain a browser-state of a Web user. They allow to tie a given browser with the internal profile in a third-party database. This motivates to limit the use of cookies on the Web. Recently certain countries, notably the UK, passed laws requiring Websites to gain their users consent for the cookie usage [13]. Examples include the site of the BBC, recently extended with an information pop-up. Users are expected to consent though they mostly mechanically do so, without any special consideration or understanding.

It is important to note the existence of other tracking mechanisms. Of particular importance, in addition to Evercookies [10], are non-standard browser fingerprinting techniques such as browser configurations [3], history [17], host identifiers [25], pixels [15], allowing tracking of the browser across various sites. The discussion of the risks and protections against potential tracking by social buttons (i.e. Facebook's Like) is covered in the work of [21]. According to [17], most popular sites can be leveraged to basic history-based fingerprinting and using this intuition, we assumed this can also be the case of the most important sites for tracking.

New risks resulted in new reactions. Quite recently, a new and bold initiative attempts at limiting the tracking on the Internet. This initiative, known as Do Not Track [14], promotes the consensual and easy solution of opting out from tracking on the Web. It is technically achieved by a simple addition of an another HTTP header in the browser's request: DNT [24]. According to Mozilla, more than 11% of Firefox users have activated the DNT. The DNT initiative lead to difficult negotiations with the advertisement industry in the US [8].

Tracking can also be limited by using dedicated browser extensions, which can block unwanted tracker scripts and/or ads. We selected two of the most prominent ones, Ghostery and AdBlock Plus, and compared them to present actual metrics of performance. For a global analysis, we refer to [11], which shows how tracking has changed with time and acquisitions of various companies by others. Our work focuses on a global approach, in contrast with [17], where the situation of individuals was addressed. Geographic differences in the Twitter network have been studied as well [12]. While some analogies to our research can be found, trackers information are fundamentally different from information on Twitter's users.

Paper organisation: The paper is organized as follows. In the next section, we present the top harvesting sites of the visible Web, and their geographical influence. Section 3 is devoted both to the trackees and the trackers of the invisible Web. In Section 4, we analyse the correlations between the different harvesting techniques.

2. The visible Web

Our investigation focuses on the top Websites of 55 countries, which have been identified using the statistics provided by Alexa. For simplicity, in the sequel we use the country

[3] `http://www.Webfoundation.org/`

[4] In the sequel, we consider the most popular Websites, globally or in one specific country based on Alexa's ranking. Alexa is a subsidiary of Amazon. `alexa.com`

code top-level domain in CcTLD format[5]. Alexa maintains lists of over 500 sites per country, but we restricted our attention to the 25 to 100 most popular sites in these lists.

2.1 Top sites by country

Data on the Web 2.0 are produced by users everywhere in the World, but they are accumulated by corporations, which for most users worldwide are not in their own country. A first measure of this phenomenon can be estimated by measuring for each country the percentage of national Web corporations among **the top 25 sites** used in that country[6]. The results are striking. Table 1 presents for a few representative countries the percentage of national Web corporations among the top 25 in each country.

CC	Nat. ratio	Foreign Sites
US	100%	no foreign site
CN	92%	only foreign site: Google
RU	68%	mostly american sites
JP	36%	mostly american sites
KR	24%	half American half Chinese
FR	36%	Only american sites
NG	24%	mostly american sites

Table 1: Percentage of national Websites among top 25

In the US, there are no foreign sites among the top 25 Websites. For all other countries we considered, apart from China and Russia, the ratio of national sites amounts at best to around a third of the Websites. Both in Japan and France, only 36% of the top 25 Websites are national, but this number hides very different realities in the two countries. First, while in France all 64% of foreign sites are American, in Japan, there is more diversity. Two Chinese sites (search engine Baidu, instant messaging QQ) and one Korean site (search portal Naver) belong to the top 25 sites in Japan. Second, and more importantly, the French sites are mostly sites[7], such as newspapers, which do not gather as much personal data, while in Japan, national sites include very data intensive ones, such as Web portals, e-commerce, blogs, etc. Similar patterns would be found for other European countries. Italy for instance has only 28% of national sites. In Africa, Nigeria, has only 24% of national sites, mostly in the Press.

China is the only country which has developed systems with a number of users, in the hundreds of million, comparable with american systems. Both China and Russia have developed a very powerful industry which harvests most of the data produced by their citizens. In China most of the first 50 sites are Chinese [9]. As shown in the infography produced by Ogilvy[8], there is no area of the social media where a Chinese company cannot be found. Moreover, in some areas, several very large systems coexist, while only one dominates in the US, not to mention the rest of the World. It is the case for microblogging platform for instance, where Sina Weibo, and Tencent Weibo coexist with both around 300 million users, and both ahead of Twitter. The ratio of local sites in Russia is lower than in China. Most of the top sites of the US have predominant positions (e.g., Facebook) in Russia, while they are blocked in China. The relative size of the two countries though impact on the size of their first systems, but both also have their respective sphere of influence abroad.

South Korea has an extremely interesting pattern of diversity. Among the top 25 sites, there are only 24% of sites which are national, while there are 36% of both American and Chinese sites, a remarkable situation. A mongolian portal (zaluu) also belong to the list.

Let us consider now the **top 100 sites**. When looking beyond the top 25 sites, the ratio of national sites increases, in particular with most of the local newspapers and magazines, as well as some services such as banking institutions.

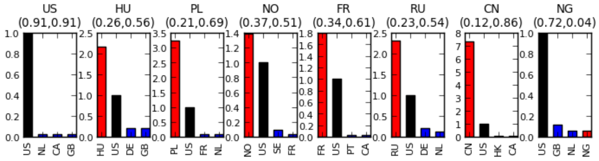

Figure 1: $Ratio_{us}$ for sites in Alexa lists, by origin.

Figure 1 shows for a selected list of countries, the proportion of sites from the US (in black), the country itself (in red), as well as the two others locally most active foreign countries (in blue), in the top 100 sites. The results are normalized to the number of sites from the US, that is they correspond to a ratio, where the per-country count is divided by the number of the US based sites.

Different patterns can be observed. For European countries, Hungary, Poland, Norway and France, the number of local sites is about twice the number of US sites in the top 100. Russia has a similar pattern. China on the other hand maintains more than 80% of local sites in the top 100 list. Nigeria shows a very different situation with a very small proportion of local sites in the top 100.

Note that the percentage of users that visit specific Websites decreases very fast from the top sites in a list to the subse-

[5] AT, AU, BE, BR, CA, CH, CN, CY, CZ, DE, DK, DZ, EC, ES, FI, FR, GB, GR, HK, HN, HU, IE, IL, IN, IT, JP, KR, KW, KZ, LY, MA, MY, NG, NL, NO, NZ, OM, PA, PE, PL, PT, PY, QA, RO, RU, SA, SE, SG, SI, SN, TH, TN, US, UY, VN.

[6] Unless otherwise specified, the numbers presented in the following tables are extracted from Alexa's ranking as of mid september 2012.

[7] National sites among the top 25 in France: leboncoin, Orange, Free, commentcamarche, lemonde, lequipe, lefigaro, pagesjaunes, sfr

[8] China social media equivalents: a new infographic `http://www.asiadigitalmap.com/2011/02/china-social-media-equivalents-a-new-infographic/`

quent sites. The first ten sites generally concentrate an important share of the total traffic. Countries whose ratio of websites is low in the first segment and increases then for the top 100, have in general a relatively small share of the global traffic.

2.2 Top sites globally

We now consider the **global impact** of Web corporations.
Table 2 presents the proportions of top 50 sites in the world that are owned by companies in a given country.
The US have more than two thirds of the top 50 sites worldwide. These sites have a real prominence worldwide as we have seen on the previous table. The only two countries that have more than one site in this club, are China and Russia, which have both developed their own industry for fundamental tools such as search engines, blogs, e-commerce, etc. Three countries, in the European sphere, have one site among the top 50.
Here again, China occupies a remarkable position after the US, which have the absolute supremacy. China has eight of the first fifty sites worldwide according to Alexa's ranking. If the size of the Chinese population impacts of course on the number of users of the Chinese systems, and therefore ultimately on the ranking of these systems, the most important reason for their success is the association of a clear political ambition, a strong appetence for social networking, and a dynamic industry. The size of the population is by no means an explanation by itself. India for instance has only a few national sites among its top 25 sites, which are almost all American.
Unlike their American counterparts, Chinese systems have currently most of their users in China. Most of them are of course also widely used in Hong Kong and Taiwan. Some, such as Taobao a popular online shopping site, are also used in South Korea and in Russia. Their international ambition will most probably grow in the coming years.

CC	Ratio	Top sites with their (rank)
US	72%	Google (1); Facebook (2); YouTube (3); Yahoo (4); ...
CN	16%	Baidu (5); QQ (8); Taobao (13); Sina (17); 163.com (28); Soso (29); Sina weibo (31); Sohu (43)
RU	6%	Yandex (21); kontakte (30); Mail.ru (33)
IL	2%	Babylon (22)
UK	2%	BBC (46)
NL	2%	AVG (47)

Table 2: The Top 50 Websites worldwide by country

Table 3 shows the percentage of global users who visited a site in the last three months. It also displays the number of countries in which the site appears in the 10 top sites locally. Note that for the most popular sites, namely Google.com, Facebook.com and Youtube.com, we show the number of countries in which the site is among the top 2 or 4 sites (as indicated in the table).
It is important to note that the traffic decreases very fast in the top 50 list. Google.com, which occupies the top position for instance drives more than 40% of global users (not to mention the local sites such as google.de, google.com.hk, etc.), and is among the first 4 sites in more than 30 countries, while Sohu.com, which occupies the 50^{th} position, drives about 2% of users and is among the top ten sites only in China.

Rank	Website	% users	# countries
1	Google.com	43.75	30 (1^{st}-4^{th})
2	Facebook.com	42.65	34 (1^{st}-2^{nd})
3	YouTube.com	33.43	35 (1^{st}-4^{th})
4	Yahoo.com	20.11	29 (1^{st}-10^{th})
5	Baidu.com	12.19	3 (1^{st}-10^{th})
8	Amazon.com	8.26	15 (1^{st}-10^{th})
18	Yandex.com	2.97	6 (1^{st}-10^{th})
32	Tumblr.com	2.57	0 (1^{st}-10^{th})
50	Sohu.com	1.84	1 (1^{st}-10^{th})

Table 3: Percentage of global users and top countries

2.3 Search engines

Let us consider more carefully particular segments, such as the **search engine**, which plays an essential role in the way people access knowledge. Here again, distinct patterns can be found. Table 4 presents the top search engines for a few countries.

CC	1^{st} SE	share	2^{nd} SE	share
US	Google	65%	Bing / Yahoo	15%
CN	Baidu	73%	Qihoo / Sogou	8-9%
JP	Yahoo! Japan	51%	Google	36%
RU	Yandex	60%	Google	25%
UK	Google	91%	Bing	5%
FR	Google	92%	Bing	3%
CZ	Google	53%	Seznam	37%

Table 4: The top two search engines by country

The US have developed major search engines. The three which dominate the American market, Google, Yahoo and Bing, are among the most popular worldwide. Google has a dominant position with 65% market share, while Bing and Yahoo have both 15% share in USA. Globally Google has 65% market share, Baidu, 8.2% market share, Yahoo, 4.9% market share, Yandex, 2.8% market share, and Microsoft, 2.5% market share[9].

[9] http://searchengineland.com/google-worlds-most-popular-search-engine-148089

China[10] and Russia[11] are in a very similar situation, where the dominant search engine is the local one, Google being the next most widely used engine. Baidu has a relatively bigger share in China than Yandex in Russia. More recently, Google lost shares in China, with the sudden raise of two other local search engines, which are approaching 10% shares of the Chinese market, Qihoo 360, and Sogou.

In Europe[12], there are no local search engines with strong positions, and the market is dominated by Google, which has a quasi monopolistic position. Only Seznam has a reasonable share for czech, but which is now decreasing with respect to Google's share.

2.4 Social networks

Other domains of the information society such as **social networks** would lead to very similar conclusion, with Facebook largely dominating in Europe, while alternative systems have been developed in Asia. The size of the Chinese social networks deserve some attention. The ranking of the Global Web Index based on the percentage of global Internet users are striking. Table 5 shows that 6 out of 10 of the most widely used social systems[13] are Chinese.

CC	Corporation	share
US	Facebook	41%
US	Google+	21%
CN	Qzone	19%
CN	Sina Weibo	18%
CN	Tencent Weibo	16%
US	Twitter	16%
CN	Renren	11%
CN	Kaixin	8%
US	LinkedIn	7%
CN	51.com	6%

Table 5: Percentage of global Internet users

China has developed a large industry on the net, with essentially all the usual services initially proposed by mostly American companies, such as online search engines, social networks, news, business, instant messaging, etc. Chinese companies have taken advantage in their development of the difficulties to access their foreign counterparts from Mainland China, but they would most certainly have succeeded without the censorship of foreign sites. The diversity in some other Asian countries such as Japan and Korea for instance shows their appetence for local systems. The strong focus in Western media on the censorship imposed on the Internet has often led to underestimate the strategy of China towards IT and the information society, and overestimate the importance of the control of the content.

3. The invisible Web

While Section 2 deals with the cartography of the visible Web, this section analyses what is often referred to as the invisible Web. The invisible Web refers to tags, Web bugs, pixels and beacons that appear on Websites to track and profile users. We first present the methodology used to track the trackers. We then consider the invisible Web from the trackees point of view, we aim to show how Internet users are tracked across the world. Finally, we analyze the trackers, and consider whether the trackers are distributed uniformly on the planet, or whether some countries are dominating the tracking business.

3.1 Tracking the trackers

In order to establish a cartography of global third-party resource utilization on the Web, we used PlanetLab's infrastructure [19]. PlanetLab connects many servers in different countries. In our experiment, 37 proxy servers from distinct countries have been used[14].

To retrieve information on trackers, we created dynamic tunnels to the relevant PlanetLab's servers. Subsequently, all sites from the respective lists were visited and the trackers detected on these sites were saved for further analysis. This process was automated with the use of a WebDriver together with a Firefox browser, equipped with modified plugins and Flash enabled. All our data has been obtained between the end of october and the beginning of december 2012.

For our tests we have chosen two popular tools enabling the detection and blocking of third-party resources: *Ghostery* and *AdBlock Plus* (ABP). They both work in a similar manner requiring the scanning of the visited Website, and searching for an offending resource or connection. If a resource is found to be present on the respective list of blocked resources (filter lists), this may either be reported to the user or blocked by the plugin.

Ghostery. Ghostery is a popular extension which detects trackers and display their names (such as "DoubleClick", "Omniture") [5]. Ghostery analyzes the requests made by a browser and compares them against a database of known trackers. It is important to note that Ghostery maintains a list of *confirmed* trackers: a tracker is not only added to the database, but also included on the project's Webpage (e.g. `http://www.ghostery.com/apps/omniture`), with their respective privacy policy. In our experiment, we saved the

[10] `http://www.chinainternetwatch.com/1444/china-search-engine-market-share-by-revenue-q1-2012/`

[11] `http://www.bloomberg.com/news/2012-04-02/yandex-internet-search-share-gains-google-steady-liveinternet.html`

[12] `http://theeword.co.uk/seo-manchester/google_tops_july_2012_search_engine_market.html`

[13] `http://globalWebindex.net/thinking/social-platform-report-series-september-2012-facebook-on-track-to-hit-2bn/`

[14] If a server from a specific country was unavailable, we used a local IP from Inria.

names of the trackers found for every visited site for further analysis.

AdBlock Plus (ABP). For comparison purposes, we also used AdBlock Plus extension [20], in the same environment as described previously. AdBlock is able to block third-party resources, mostly unwanted advertisements, and maintains a dedicated trackers list: EasyPrivacy[15], which includes many known trackers as well as Web bugs, which allows AdBlock Plus to block these resources. We use this later list in our analysis.

Although these two tools are different, they provide results that are consistent. A more detailled comparaison of these tools and results are provided in Appendix A.

Country of origin. Determining the **country of origin** of Web corporations is a challenging problem. One solution is to identify the location of their headquarters, but this is not always relevant. It is also possible to use Whois databases to identify the location of the site holding their domain name. However, domain registrars are sometimes located elsewhere than reported in Whois databases.
We instead propose a technique that can be automated.
ABP associates to each tracker a Web resource, for example `http://edge.quantserve.com/quant.js`. We extracted the top-level domain name of each tracker, i.e. `quantserve.com`. We then resolved the domain name into an IP address and use a geolocation database, to identify its location. This approach can correctly identify the country of origin in most of the cases.
Ghostery website contains a description of each tracker (see `http://www.ghostery.com/apps/`). This description contains the url of the tracker's company website and potentially the postal address of the company. We used the company website's url to geolocalize it. The results were then cross-checked manually. For example `http://www.ghostery.com/apps/digilant` mentions Digilant company's Website `www.digilant.com`, which resolves to United States. The "Privacy Contact" tab on Ghostery's Website confirms that Diligant headquarters are indeed located in Boston, USA."
In both cases, the geolocation has been done using Python GeoIP and geoiplookup command-line utility tools, which query geolocation databases. These tools take an IP address or an url as input, and output the location of the corresponding website.

3.2 The trackees

We first consider the **average number of trackers per site** for each of the 55 countries considered. The average is computed by connecting to the top 100 most popular sites of a country, summing the number of trackers on each of them, and dividing the final result by the number of retrieved sites.

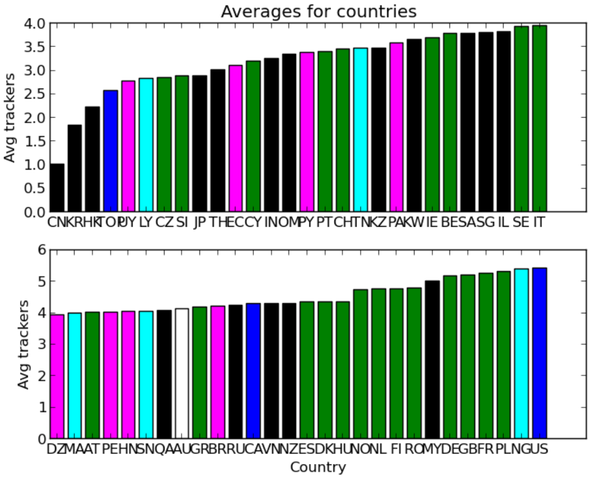

Figure 2: Average number of trackers (Ghostery)

Figure 2 displays the average number of trackers (Y axis) for each considered country (X axis). The top 100 sites are accessed from a local IP, and the trackers are detected using Ghostery data for the 55 selected countries. The results clearly show that users are tracked differently on the Internet according to their country. The different colors on the figure refer to the different continents[16]. As shown in Appendix B, these results do not depend in fact on the visitor locations. In other words, in most cases, a given site tracks its visitors independently of the visitor's IP address.
While results do not show big differences between continents, some countries seem to be much more tracked than others. For example, **US Internet users are tracked 5 times more than Chinese users** [17].
Figure 3 displays the same type of results as Figure 2, with the browser's User-Agent string set to a mobile device. Although the results are different, with a smaller average number of trackers, the trends are similar. Some countries are still much more tracked than others. This difference is likely due to the fact that many Websites redirect the mobile browser to a special version of the site, tailored towards devices with a smaller display. These sites, as it seems, include less third-party resources, probably to speed up loading. Furthermore many popular sites have a dedicated mobile application anyway, so the potential losses from tracking and/or advertising can be balanced with the use of in-app advertisements.
Figure 4 exhibits results from experiments similar to those of Figure 2, but while using ABP instead of Ghostery. The absolute numbers are slightly smaller, but the trends are very similar. Appendix A presents more experiments that analyse the consistency of the results obtained by ABP and Ghostery.

[15] copied on 30/11/2012

[16] Europe is in green, Asia in black, North America in dark blue, Africa in light blue, South America in purple, Australia in white

[17] We define as a US (resp. Chinese) user, a user that visits the Alexa 100 top, i.e. the 100 most popular, sites in the US (resp. in China)

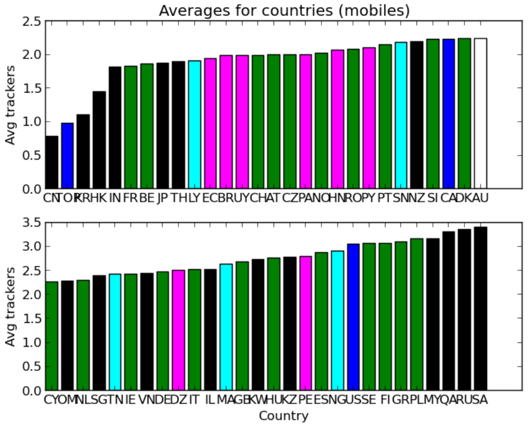

Figure 3: Average number of trackers (Ghostery, mobile)

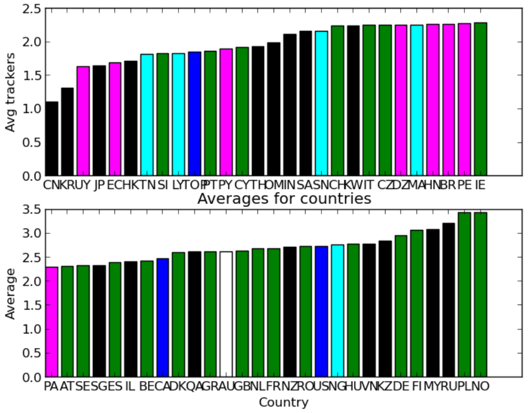

Figure 4: Average number of trackers (ABP)

We then considered the **distribution of trackers for selected countries**, which is shown on Figure 5. The distribution is obtained by counting, for a given country, all occurrences of a particular tracker, e.g., DoubleClick. These numbers are then ordered from the largest to the smallest, and plotted. The first value of a given curve shows the number of occurrences of the most popular tracker, the second value shows the number of occurrences of the second most popular tracker, and so on. The analysis is based on Ghostery.

The results clearly show that in China there are only 10 different trackers on the top 100 Chinese sites, and these trackers are not very active. Moreover, the most popular, CNZZ, only appears in 10 of the top 100 Chinese sites. In contrast other countries, such as the US, have a large number of different trackers (about 90 for the US), and some of these trackers, for example Google Analytics, are very popular in a large number of sites.

Does it mean that the US market is more fragmented? Indeed, it seems that the number of tracking companies is

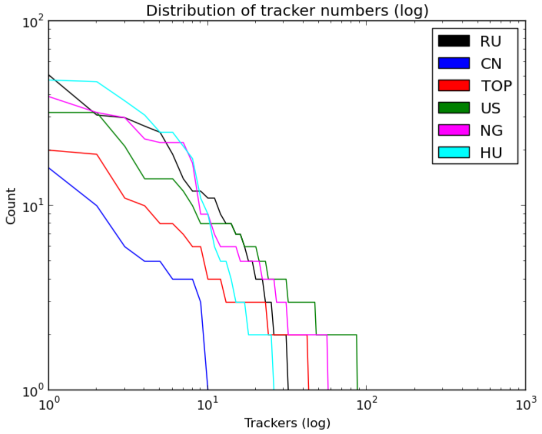

Figure 5: Distribution of trackers for selected countries

much larger in the US than in other countries, which naturally increases the average number of trackers per site in the US.

Finally, we considered the **distribution of different tracker types**. Ghostery divides its detected scripts into five categories [6], which we recall below:

1. *Ad*: advertisements provided by the ad-networks;
2. *Tracker*: scripts which actually perform tracking (often using very sophisticated behavioral analysis);
3. *Analytics*: utility scripts for Website creators allowing them to discover various statistical details about their visitors;
4. *Widget*: small Web applications such as clocks, weather tables, and others. Other examples include Facebook Social Plugins, Google +1, etc.;
5. *Privacy*: typically a script disclosing privacy policies and practices related to ads, such as Evidon Notice[18].

Figure 6 displays the distribution of trackers for the five categories of [6] for the 55 countries, based on Ghostery. It is interesting to note that the distribution of each type seems to be quite similar in different countries. The *Ad* trackers are the most common, followed by the *analytics* ones.

3.3 The trackers

We next consider the geographic origin of the trackers, and the way trackers proliferate depending upon the country they come from. We first start by analyzing the **origin of the trackers on the top 100 sites** of the global list, that is the list of the 100 most popular Websites worldwide.

Table 6 shows the distribution of the detected trackers using Ghostery. We computed the number of trackers, T, on the

[18] `http://www.evidon.com/about/`

[19] ROW: Rest Of World

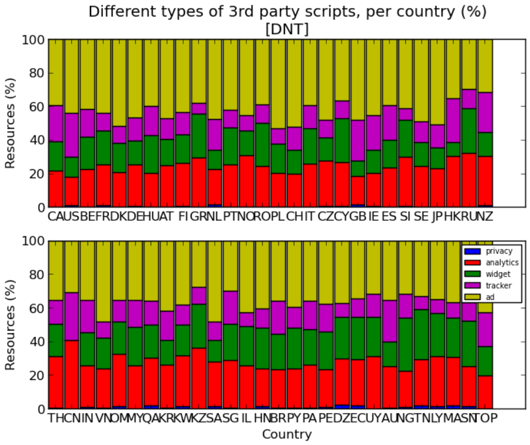

Figure 6: Distribution of trackers by types

CC	Share
US	87%
CN	3%
GB	3%
RU	2%
ROW[19]	4.2%

Table 6: Distribution of trackers on top 100 global sites

top 100 sites worldwide. We then counted for each country i, the number of trackers of origin i, C_i, and then computed, for each country, the percentage $p_i = C_i/T$.
The results show a clear domination of the US, with more than 80% coverage of the top 100 sites worldwide, with only a few countries, such as China, GB, and Russia, which have a small percentage of trackers.

Tracker (country of origin)	Count
Google Analytics (US)	25
DoubleClick (US)	20
ScoreCard Research Beacon (US)	19
Facebook Social Plugins (US)	11
Omniture (US)	10

Table 7: Top 5 trackers on top 100 global sites

Table 7 displays the top 5 most popular tracking corporations on the top 100 most popular sites worldwide. The results clearly confirm that the tracking business is largely dominated by US companies.

Let us push further the analysis of the origin of the trackers, and consider an **in-depth Analysis for Selected Countries**. For the sake of readability, we decided to display here the results for only 5 countries, namely Russia, China, USA, Nigeria, and Hungary, and analyze the origin of the trackers of the top 100 sites for each of these countries[20].

CC	Tracker (country of origin)	Count
RU	Google Analytics (US)	53
	LiveInternet (RU)	51
	Yandex.Metrics (RU)	31
	TNS (GB)	30
	Rambler (RU)	27
CN	Google Analytics (US)	20
	MarkMonitor (US)	16
	CNZZ (CN)	10
	ScoreCard Research Beacon (US)	6
	DoubleClick (US)	5
US	Google Analytics (US)	39
	DoubleClick (US)	32
	ScoreCard Research Beacon (US)	32
	Omniture (US)	21
	Facebook Connect (US)	14
NG	Google Analytics (US)	62
	Facebook Social Plugins (US)	39
	DoubleClick (US)	32
	Facebook Connect (US)	30
	Google Adsense (US)	23
HU	Google Analytics (US)	70
	Median (HU)	48
	Gemius (PL)	47
	Adverticum (HU)	37
	Facebook Social Plugins (US)	31

Table 8: Top 5 trackers for selected countries

Table 8 displays the top 5 most popular tracking companies on the 100 most popular Websites of these 6 selected countries. Once again, the results clearly indicate that the tracking business is largely dominated by US companies. Only few countries seem to resist this domination such as Russia, China, and Hungary. Interestingly, **Russia is the only country whose trackers are mostly local**.

We next consider the origin of the most prominent trackers in the following eight countries: USA, Hungary, Poland, Norway, France, Russia, China, and Nigeria. We performed the following analysis. We first collected all the trackers on the top 100 sites of a country, classifying them according to their origin, and counting their occurrences. Since the raw numbers, even averages, may not be the most informative in this analysis, we decided to plot the information with respect to the detected number of US-based trackers to ease the presentation. More specifically, we computed for all trackers of an origin N detected on the top 100 sites of a country C,

[20] The analysis of all 55 countries listed at the beginning of this paper is presented in Appendix C.

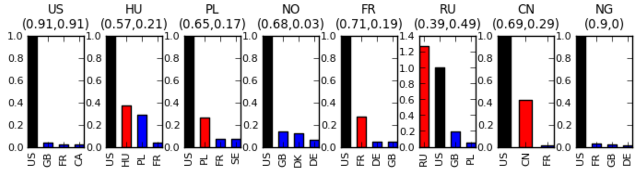

Figure 7: $Ratio_{us}$ for trackers by origin (Ghostery)

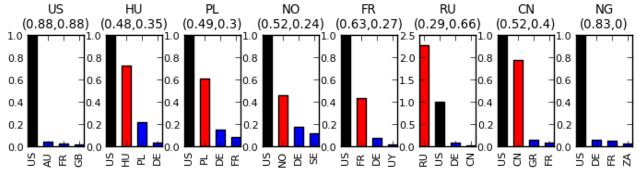

Figure 8: $Ratio_{us}$ for trackers by origin (ABP)

the following ratio:

$$Ratio_{us}(N) = \frac{\text{\# trackers of origin N}}{\text{\#US trackers}}$$

The US has been chosen as a reference because of the global prevalence of the trackers of this origin. Indeed, US trackers are present on almost every site.

Figures 7 and 8 show for each of these selected countries the origin of the top 4 most prominent tracking countries for both detection methods, Ghostery and ABP. In parenthesis for each country, the ratio of US trackers, and the ratio of local trackers. These figures clearly show that the US have almost always the largest ratio of trackers. The second largest tracker country, is apart from some exceptions, the country itself. Austria constitutes for instance an exception to this rule, with Germany as second tracker with a ratio of 0.24.

China and even more so Russia constitute the two exceptions, with remarkably strong ratios of local trackers over US trackers. In fact, China has a ratio of 0.78, and Russia a ratio of 2.3, that indicates that Russian sites contain more Russian trackers than US ones.

We also observed some variation in the number and the distribution of countries involved in tracking in a given country, beyond the US and the country itself. For example in France, there are third-party resource providers from 11 countries, whereas in China only 6 countries are "represented".

In addition, we observed that these trackers often come from neighboring countries. For example, Danish sites often contain trackers from Sweden or Finland. The same observation applies to Austria as we noticed already, as well as other countries in Central Europe, such as Hungary, Slovakia and the Czech Republic. Therefore, with the exception of the US, most trackers are regional.

4. Sites vs Trackers Analysis

In this section, we compare the harvesting activity on the visible and the invisible Web. Our objective is to understand whether the predominance of the US is larger on the visible or on the invisible Web. In order to achieve this goal, we compute and compare the proportion of US sites (resp. US trackers), with the proportion of local sites (resp. local trackers) for each of the top 100 sites of each of the 55 considered countries.

These proportions, P_{US} and P_{local}, are computed as follows:

$$P_{US}(C) = \frac{\text{\#US trackers (resp. sites) in country C}}{\text{\#All trackers (resp. sites) in country C}}$$

$P_{US}(C)$ is the number of US trackers (resp. US sites) divided by the total number of trackers (resp. sites) in the top 100 sites in country C. Quantitatively, it shows how trackers of a dominating country (e.g. US) track the world.

$$P_{local}(C) = \frac{\text{\#Local trackers (resp. sites) in country C}}{\text{\#All trackers (resp. sites) in country C}}$$

$P_{local}(C)$ is the number of local trackers (resp. local sites) divided by the total number of trackers (resp. sites) in the top 100 sites in country C. Quantitatively, it shows the strength of local trackers.

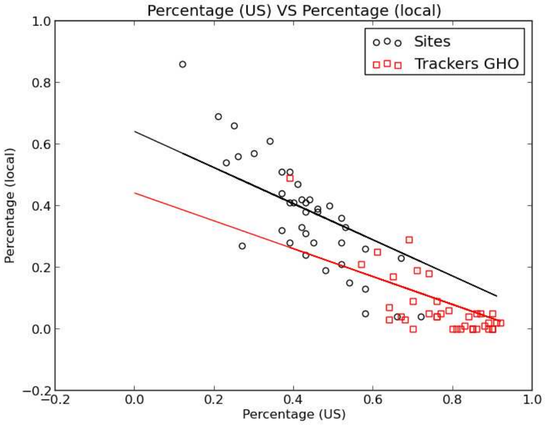

Figure 9: Ratio US vs local sites and trackers

The scatterplot for P_{US} and P_{local} of all the 55 countries considered for sites and trackers are shown on Figure 9. Every black square (resp. red circle) point corresponds to a specific country C, with coordinates $P_{US}(C)$ (X axis) and $P_{local}(C)$ (Y axis), for sites (resp. trackers). The trackers have been obtained using Ghostery, as for previous measures.

The results clearly show that **the US are even more present on the invisible than on the visible Web**. Most of the tracker points (in red) are located on the right lower corner of the plot. This indicates that the percentage of US trackers is large in most considered countries, while the percentage of local trackers is usually smaller, while the distribution of the proportion of the sites is more balanced between the US and local sites.

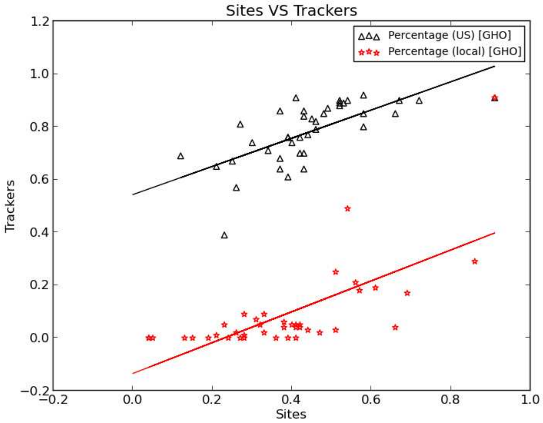

Figure 10: Sites vs Trackers (Ghostery)

This finding is further confirmed with the results shown on Figure 10, which presents the scatterplots for sites against trackers: the relationships between US sites (black triangles) (resp. local sites (red stars)) and the corresponding trackers for Ghostery. The red stars are on the lower part of the figure, corresponding to a small percentage of local trackers for most considered countries even with a large proportion of local sites. Whereas, the black triangles are on the higher part of the figure, corresponding to a large percentage of US trackers.

Let us now consider again an **in-depth Analysis of Selected Countries**. We consider the US and local proportion of sites and trackers for seven selected countries, namely Russia, China, Hungary, Poland, Norway, France and Nigeria.

CC	P_{US}	P_{local}
RU	0.23	0.54
CN	0.12	0.86
HU	0.26	0.56
PL	0.21	0.69
NO	0.37	0.51
FR	0.34	0.61
NG	0.72	0.04

Table 9: P_{US} and P_{local} for sites

Table 9 displays the respective percentages of US and local sites. It shows that local sites are often dominant, except for some developing countries such as Nigeria.

Table 10 displays the respective percentages of US and local trackers. The second value shows the results when the US sites are excluded. The results contrast with the results of the previous table, and clearly show that the percentage of US trackers is larger that the percentage of local trackers, except for Russia.

CC	P_{US}	P_{local}
RU	0.39/0.36	0.49/0.52
CN	0.69/0.68	0.29/0.31
HU	0.57/0.53	0.21/0.23
PL	0.65/0.63	0.17/0.17
NO	0.68/0.62	0.03/0.04
FR	0.71/0.67	0.19/0.22
NG	0.9/0.86	0/0

Table 10: P_{US} and P_{local} for trackers

Finally it is interesting to compare Russia and China with respect to the proportion of local sites and trackers.

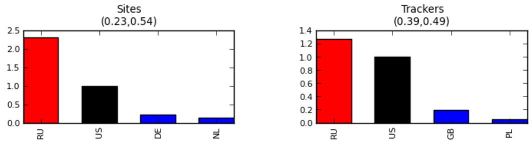

Figure 11: Sites and trackers in Russia

Figure 11 shows how Russia manages to have both more russian sites and trackers at home, although somehow with the same order of magnitudes as US sites and trackers, while Figure 12 shows that China has mostly Chinese sites, but mostly US trackers.

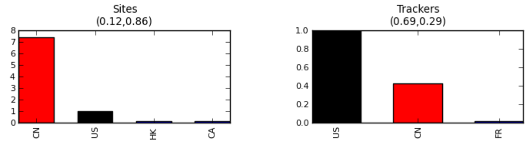

Figure 12: Sites and trackers in China

5. Conclusion

We studied the global distribution and proliferation of third-party resources on the most popular sites in various countries. Our research reveals very different strategies and/or capacities to harvest local as well as global data, both on the visible and the invisible Web.

- In Europe, most countries have twice as many local sites as US sites amongst their top 100 most popular sites, although the top sites are mostly US, and the trackers are mostly US as well, followed by local and regional trackers, thus leading to an important flow of data from Europe to the US.

- USA clearly dominates the visible Web with more than 70% of the top 100 sites in the world. The US domination is even stronger on the invisible Web with 87% of trackers on the top most popular 100 sites in the world, and the largest proportion of trackers in all countries (except Russia). This trend is even bigger in developing countries (such as in Africa).
- China dominates its visible Web with more than 80% of local sites [9], but these sites still contain a majority of US trackers. Another striking result of our experiment is that there are only 10 different trackers on the top 100 Chinese sites, and these trackers are not very active. For example CNZZ, the most popular tracker in China, only appears in 10 of the top 100 Chinese sites. In contrast, the US have a large number of different trackers (about 90), and some of these trackers, for example Google Analytics, are very popular in a large number of sites.
- Russia is the only country that contains more than twice as many local sites as US sites amongst its top 100 most popular sites, and more local trackers than US ones. This unique feature results probably from the different views that Russia has on the nature, potential and use of the cyberspace [7]. Russia raised serious concerns about the principle of uncontrolled exchange of information in cyberspace, which it considers as a threat to the society and the state.

The current non-uniform geographic distribution of data harvesting might result in strong information asymmetries between regions. If Big Data is now considered as an important economic issue, much less attention is devoted to data harvesting techniques and the data flows, which to our opinion constitute one of the economic and political challenges of the 21st century. To complete this study it would be interesting to measure the flows of data going from one country to another through the systems of either the visible or the invisible Web. This would require the use of widely deployed trackers, and was out of reach of the present investigation.

It is no surprise that China and Russia have developed powerful systems to handle most of their data locally. In Europe, there is an increasing concern about personal data handled abroad, particularly on issues such as privacy and taxation. The relations between governments and corporations handling personal data is of great concern in many places in the world [23]. One of the key issues is the legislation that applies to the data, and the capacity governments have to access the data [22]. In the US for instance distinct protection frames apply to residents' and foreigners' data [4]. We believe that these questions will raise considerable attention in the near future.

References

[1] *The Art of War*. MightyWords, Incorporated, 2000.

[2] A. Barth. Http state management mechanism. https://tools.ietf.org/html/rfc6265, 2011.

[3] P. Eckersley. How unique is your web browser? In *Privacy Enhancing Technologies*, pages 1–18, 2010.

[4] R. Gallagher. U.s. spy law authorizes mass surveillance of european citizens: Report. http://www.slate.com/blogs/future_tense/2013/01/08/fisa_renewal_report_suggests_spy_law_allows_mass_surveillance_of_european.html, January 2013.

[5] Ghostery. Ghostery. http://www.ghostery.com/.

[6] Ghostery. Knowyourelements.com - presented by ghostery. http://www.knowyourelements.com/.

[7] K. Giles. Russia's Public Stance on Cyberspace Issues. In *In Proceedings of the 4th International Conference on Cyber Conflict*, Tallinn, Estonia, June 2012.

[8] D. Goldman. Do not track is dying. http://money.cnn.com/2012/11/30/technology/do-not-track/index.html.

[9] S. Grumbach. The stakes of Big Data in the IT industry: China as the next global challenger? In *The 18th International Euro-Asia Research Conference, The Globalisation of Asian Markets: im- plications for Multinational Investors, Venezia, January 31 and February 1st, 2013*, Venise, Italie, Jan. 2013.

[10] S. Kamkar. Evercookie - virtually irrevocable persistent cookies. http://samy.pl/evercookie/.

[11] B. Krishnamurthy and C. Wills. Privacy diffusion on the web: a longitudinal perspective. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 541–550, New York, NY, USA, 2009. ACM.

[12] J. Kulshrestha, F. Kooti, A. Nikravesh, and K. P. Gummadi. Geographic Dissection of the Twitter Network. In *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, Dublin, Ireland, June 2012.

[13] D. Lee. Cookie law: websites must seek consent from this weekend. http://www.bbc.co.uk/news/technology-18194235.

[14] J. Mayer and A. Narayanan. Do not track - universal wev tracking opt out. http://donottrack.us/.

[15] K. Mowery and H. Shacham. Pixel perfect: Fingerprinting canvas in html5. *Web 2.0 Security and Privacy*, 2011.

[16] Network Working Group. Hypertext transfer protocol – http/1.1. http://www.w3.org/Protocols/rfc2616/rfc2616.html.

[17] L. Olejnik, C. Castelluccia, and A. Janc. Why johnny can't browse in peace: On the uniqueness of web browsing history pattern. In *Hot topics in Privacy Enhancing Technologies*, 2012.

[18] E. Pfanner. France proposes an internet tax. http://www.nytimes.com/2013/01/21/business/global/21iht-datatax21.html?pagewanted=all&_r=2&.

[19] PlanetLab. Planetlab: An open platform for developing, deploying and accessing planetary-scale services. http://www.planet-lab.org/.

[20] A. Plus. Adblock plus - for annoyance-free web surfing. http://adblockplus.org/en/firefox.

[21] F. Roesner, C. Rovillos, T. Kohno, and D. Wetherall. Sharemenot: Balancing privacy and functionality of third-party social widgets. *Usenix ;login:*, 2012.

[22] C. Savage. U.s. weighs wide overhaul of wiretap laws. http://www.nytimes.com/2013/05/08/us/politics/obama-may-back-fbi-plan-to-wiretap-web-users.html?nl=technology&emc=edit_tu_20130508&_r=1&, May 2013.

[23] B. Schneier. Do you want the government buying your data from corporations? http://www.theatlantic.com/technology/archive/2013/04, April 2013.

[24] W3C. Tracking preference expression (dnt). http://www.w3.org/TR/tracking-dnt/.

[25] T. Yen, Y. Xie, F. Yu, R. Yu, and M. Abadi. Host fingerprinting and tracking on the web: Privacy and security implications. In *Proceedings of NDSS*, 2012.

A. Comparing ABP and Ghostery Results

There are important differences between the behaviors of Ghostery and ABP. The most distinctive one is that Ghostery provides the tracker's names, while ABP identifies the actually detected resource, that is a script with a full domain name.

Let us see more carefully how ABP treats third-party script detection. If a site `www.X.com` has an external tracker which is served from `Y.X.com`, the latter is assumed to be a third-party script. But sites may serve third-party scripts as their *first-party* resource. For example, a site `www.X.com` can include a "*tracker.js*" script file, which belongs and provides information to a third-party tracker. This is why ABP may detect a file, present on the visited site, as an actual tracker. In certain cases, a file served from the Website the user visits, actually dynamically sends the tracking data to an external server. But eventually, it can be perceived as if the visited site was doing the tracking and this does not always make sense in our analysis. Consequently, whenever we identify a detected tracker from the accessed site's domain name, we count it as a tracker. If, however, the accessed site includes multiple scripts originating from the same third-party site, we treat it as a single tracker.

Our purpose is to identify the national origin of the trackers. We considered the tracking with respect to the **origin of third-source providers**. We thus had to discard all the trackers served from the visited sites, which in reality might belong to other entities. An illustrative example is the `http://www.index.hr/` site, which serves the `xgemius.js` tracker (file). If we had not done so, the origin analysis could have been biased towards local sites, as if the visited site would actually perform the tracking, and the rate of false positives for the origin analysis could have been significant. Being tracked by a visited site is theoretically possible and in fact any such Website is obviously capable of collecting the information required for tracking purposes. Subsequently that site could, behind the scenes, i.e. without knowledge of the user, transfer this information to tracker companies. In principle such tracking would be undetectable. We did not focus on this theoretical scenario, though, and only included this remark for completeness.

Moreover ABP supports full CSS selectors standard: it is possible to block particular nodes, which either are known for or are likely to contain a third-party resource. Example might be an HTML `div` tag with a specified `id` or `class` parameter. Since it is usually difficult to establish the national origin of such a tracker, we ignored all such appearances in the origin analysis as well.

In summary, we used Ghostery and ABP in the "third-party resources only" mode, which is sufficient, and although results obtained with both tools differ, their trends are very similar.

Figure 13 shows a scatterplot for P_{US} and P_{local} belonging to ABP and Ghostery (Pearson's $r > 0.9$ in both cases), which demonstrates a very high consistency between the results obtained with these two different tools.

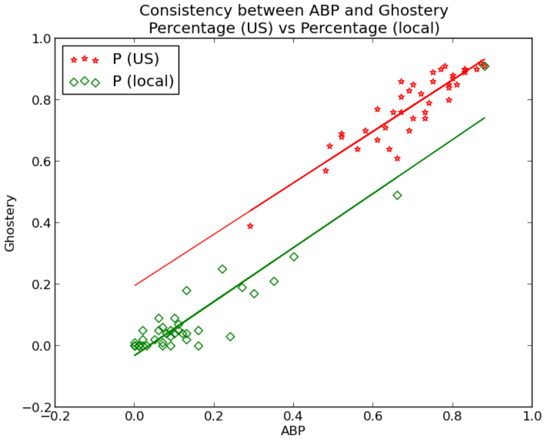

Figure 13: Scatterplot of P_{US} and P_{local}: ABP vs Ghostery. Every marker corresponds to a specific ratio belonging to ABP and Ghostery. The correlations are $r = 0.9$ (P_{US}) and $r = 0.94$ (P_{local}). This shows that ABP and Ghostery's results are consistent. (ABP)

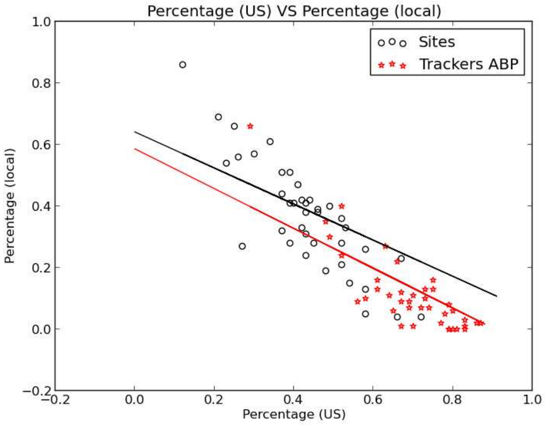

Figure 14: P_{US} (X) vs P_{local} (Y) for sites and trackers (ABP). Each marker corresponds to country. Data from Figure 1-like.

B. Tracker analysis according to visitor locations

One of our goals was to verify whether the physical presence of a user has influence on tracking. In other words, we intend to analyse whether the IP address of the visitor of a site affect how this visitor is tracked by that site.

For every country $x_1, ...x_i$ $(0 < i < 55)$, we visited all the sites $S_k, ...S_{100}$ belonging to the country x_i, using the source IP address from this country. We saved all the detected trackers and the results are presented in form of a heat map, which is seen on Figure 16. The vertical axis (Y) corresponds to the source country request, while the horizontal axis (X) shows the connection's target. Every column corresponds to the same country as a target for a visit, while every row is a visit from a respective country address. For example, the third row corresponds to a visit from IP address assigned to Belgium, while the fourth column is a French list of destination addresses. Variations within these columns are small and - if at all - correspond to the redirection from a general site to a local version. For example, a visit to "yahoo.com" may be a subject of redirection to "fr.yahoo.com", if the origin address is french.

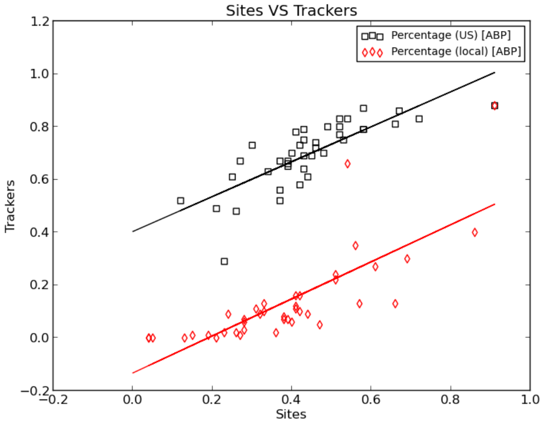

Figure 15: Sites vs Trackers (ABP)

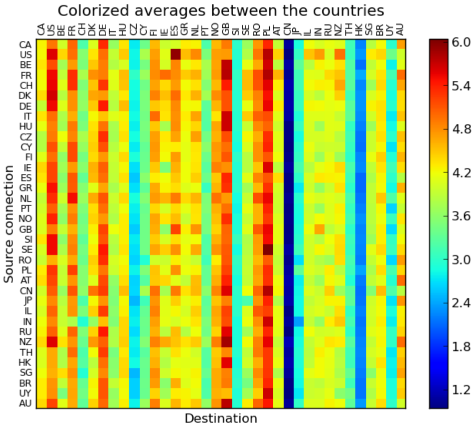

Figure 16: Heat map for average number of detected third-party scripts. The left column is the origin of request, and the top line is the destination of request, in a right place a color indication of the 'heat' resides.

C. Tracker heatmaps

In this section, we counted all observations of trackers as raw numbers on the top 100 sites of a country, and stacked them by their country of origin. We then drew heat-maps from these results. In these heat-maps, the X axis is the country of origin of a tracker, while Y axis shows the origin of specified list. For example X = US, Y = RU means the origin of a tracker is *US*, and the country-specific list is *RU*. Lighter color means more trackers.

Heat-maps on Figure 17 and 18 were made using 10 selected countries using ABP and Ghostery tools.

These results show a clear dominance of the US trackers. They also show that most of the 10 countries have local trackers. This is especially true for Russia.

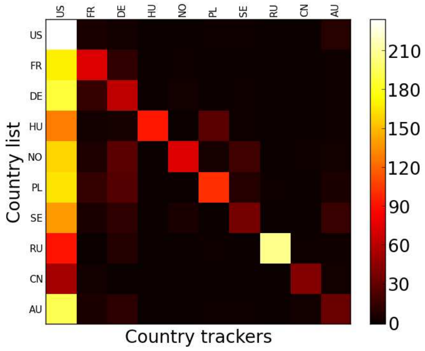

Figure 17: (ABP) Heatmap for trackers (origin, destination)

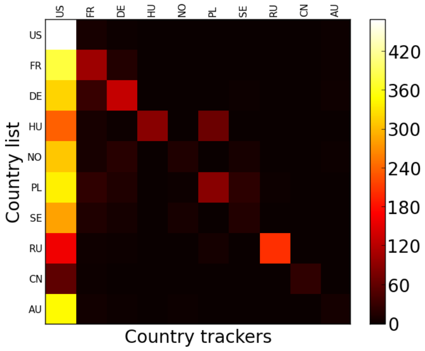

Figure 18: (Ghostery) Heatmap for trackers (origin, destination)

Heatmaps for all 55 countries considered in this study are shown on Fig. 19 and 20, respectively for ABP and Ghostery.

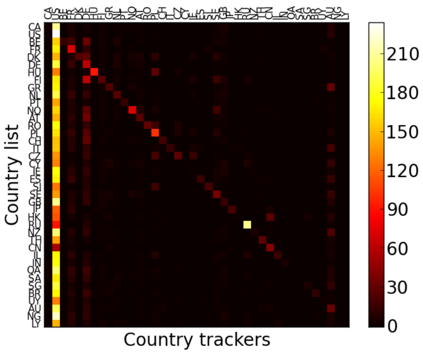

Figure 19: (ABP) Heatmap for trackers (origin, destination)

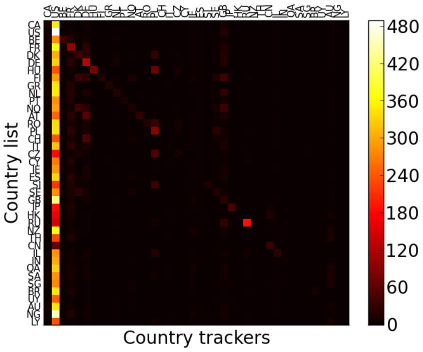

Figure 20: (Ghostery) Heatmap for trackers (origin, destination)