The Rise of Panopticons: Examining Region-Specific Third-Party Web Tracking

Marjan Falahrastegar, Hamed Haddadi, Steve Uhlig, Richard Mortier

The Rise of Panopticons: Examining Region-Specific Third-Party Web Tracking

Marjan Falahrastegar[1], Hamed Haddadi[1], Steve Uhlig[1], Richard Mortier[2]

[1] Queen Mary University of London
[2] University of Nottingham

Abstract. Today's web has a huge, diverse ecosystem of third-party websites collecting information about users and providing them with content such as targeted advertisements. In this paper we study this ecosystem of third-party websites. We sample every continent, targeting the 500 most popular websites in the US, UK, Australia, China, Egypt, Iran and Syria. This allows us to contrast the commonplace, western-dominated views of the web with less studied countries. We find 2,097 third-party web services, reflecting the diversity of services and types of application/content they involve, e.g., advertisement, ad trackers, CDNs, news, sport, and pornography. We find those third-party web services offering ad tracking services to be the most prevalent. In addition to the usual suspects (e.g., DoubleClick and Google), we find a rich ecosystem of *local* third-party websites that are country and language dependent.

1 Introduction

The role of the Internet in everyday life evolves continuously. Online Social Networks (OSNs), streaming videos, and online shopping are all now daily activities in the lives of most netizens. In addition, web interactions afforded by developments such as dynamic client-side interaction (e.g., Ajax [1]) and cloud-based services [2] have lead to significant changes in the Internet traffic [3] and website complexity [4].

One of the expanding family of new entrants in the Web ecosystem are the third-party tracking services and cookies. They provide features such as advertising, analytics, OSN plugins, and user tracking and profiling. Although some user interactions with these services may be conscious and explicit, e.g., sharing content or clicking *like* on various OSNs, most interactions users have with these services will not be explicit and indeed, users may often be unaware of the presence of the services at all. The increasing trade of personal information between these services and their increasing ubiquity and ability to track users from one page to the next, often on different websites, is a major source of concerns about privacy.

A number of recent works have investigated the effect of these third-party services in terms of the performance [4], privacy [5] and transparency [6] challenges they introduce. In this paper we study the presence of third-party services across the most popular websites in different regions of the world when viewed

from the vantage point of an individual user. We measure the prevalence of these services in the modern web, focusing on a few distinct but key countries in both the East and the West. Our measurements are taken from a single vantage point and are thus not based on user interactions with these websites in the countries in question. As DNS may redirect users from different regions to different servers, whose version of the webpage point to different third parties, one may observe different third parties than in our dataset. In this paper our specific contributions are as follows:

- We find a surprisingly large ecosystem of third-party sites: within just the 500 most popular websites, according to Alexa,[3] in the 7 countries we examine, we find 2,097 unique third party websites.
- We categorise these third-party websites based on the nature of the service they provide (ad trackers, Content Distribution Networks (CDNs), analytics services, OSNs, etc.) and we find that advertisement and ad trackers make the bulk of the third-party websites in their number. The prevalence of user tracking services (ad trackers, analytics and trackers) and their ubiquity across a broad portfolio of popular websites allow the corporations behind them to obtain a rich, detailed view of individuals' browsing behaviour, trends, interests and correlations between activities.
- Surprisingly, third-party CDNs are only the third category by number, despite capturing about half of all referrals from the origin websites. The CDN market therefore appears more consolidated than that for user tracking services. In addition to ad networks, trackers, and CDNs, we observe a rich ecosystem of other third-party sites that reflect specific user interests, activities and applications, e.g., sports, shopping, OSNs, porn, and gaming.

2 Method

We investigate the presence of third-party websites in a diverse set of countries by analysing the landing pages of popular websites in seven countries. We use the Alexa top-500 ranking per-country to determine the popular websites for each country. In this paper we present data from five countries representing each continent (the USA, the UK, Australia, China and Egypt), and from an additional two countries (Iran, Syria) as samples of less commonly-studied countries with a common language (Arabic).

We identify presence of a third-party site in a given landing page using a combination of the *domain* and *adns* approaches from Krishnamurthy & Wills [5]. Thus we identify a third-party site as one whose second-level domain and authoritative DNS server differ from the second-level domain name and authoritative DNS server of the origin site. Use of the authoritative DNS server in this way enables us to correctly classify cases such as `bbc.co.uk` and `bbci.co.uk`, where each belongs to the same company even though the second-level domain names are different.

[3] `http://www.alexa.com/`

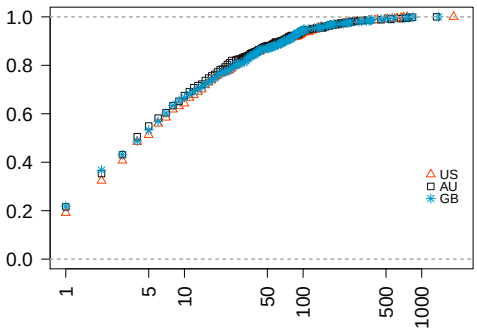
We automated this process with (*i*) a Python script that fetches the landing pages from a given list of sites, (*ii*) a Chrome extension that detected third-party websites based on the *domain* approach, and (*iii*) a Python script that carried out the *ADNS* approach. The initial script opens the landing page of the current target site in a Chrome browser a website in a browser tab. While the landing page is open on the Chrome, the extension monitors all HTTP requests sent by the browsed website and identifies third-party requests as those where the second-level domain name of the URL differs from the second-level domain name of the origin website. The list of third-party requests corresponding to each origin site is recorded in a log file. The ADNS server name of the obtained third-party requests and their origin site were fetched using `nslookup -type=soa` command available on Linux. Subsequently, the requests were compared against each other to refine those cases where two second-level domain names are different while one is an alias for the other one. Every 20 seconds the next site is fetched and its landing page is analysed as explained above.

We carried out this procedure once in October 2013 for each country, from a single PC, with Linux OS, located in the UK. We are currently extending the set of countries considered and investigating the possibility of running this process from within different countries to gain further insight into the prevalence of the local third-party websites.

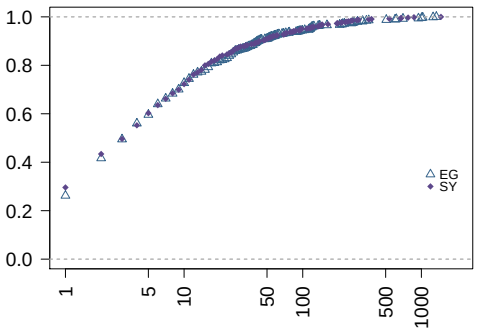
3 Third-party Presence in the WWW

In this section we describe our collected data and analyse the presence of third-party websites in the dataset. We retrieved a total of 2,104 unique websites that involved at least one third-party website. The origin site that referred to the most third-party websites was `free-tv-video-online.me`, a popular video hosting website in the US (Alexa rank 305), which issues 1,501 calls to 135 third-party sites. We identified 2,759 unique sites as third-party based on the *domain* approach, which reduced by 24% to 2,097 when the *adns* approach was applied. In total third-party sites were referred to 119,911 times across the entire dataset. A majority of third-party sites, 1,063 (51%), were referred to fewer than 5 times. However, a small number, 37 (1.8%), are very popular and referred to by more than 500 of the 2,104 websites we studied.

Figure 1, 2 provides the cumulative distribution of the identified third-party sites across different countries. Examining each country individually, we see that the English language countries (US, GB, Australia; Figure 1(a)) are all very close together, indicating similar prevalence of third-party services in these countries (although the specific third-party services may differ). However, the other countries studied (Egypt, China, Iran, Syria; Figure 1(b), 2) display quite marked differences, with only Arabic language countries (Egypt and Syria) being similar. This suggests that the prevalence of tracking does differ between different countries with different languages.



(a) US, GB and Australia



(b) Egypt and Syria

Fig. 1. The cumulative distribution of third-party websites in each country

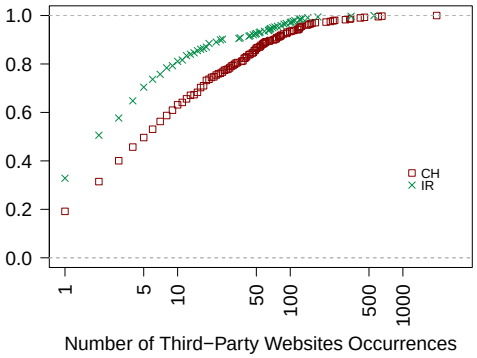


Fig. 2. The cumulative distribution of third-party websites in Iran and China

3.1 Role of Third-Party Websites

Third-party websites provide three main services: (*i*) advertisement, (*ii*) content delivery via CDNs and (*iii*) tracking user activity across websites. Given the prevalence of these third-party sites, we want to understand the relative importance of each category in this ecosystem.

We identified the category of each third-party website by combining three methods. First, we used information from the Abine website[4] to detect ad-trackers, trackers, and analytic sites. Second, we examined the third-party domain name itself for presence of keywords indicating the category. Finally, we manually categorised any remaining third-party websites in the top-300 (approximately 130 required manual categorisation).

	Category		Number		Referrals		
No.	Name	(Code)	Total	(%)	Total	(%)	Median
1	Advertising	(A)	192	(28%)	4073	(4%)	6.0
2	Ad trackers	(D)	141	(21%)	36768	(35%)	88.0
3	CDN	(C)	88	(13%)	17198	(16%)	28
4	Analysts	(Y)	68	(10%)	22081	(21%)	54.5
5	Web hosting	(W)	46	(6%)	2232	(2%)	8.5
6	Trackers	(T)	41	(6%)	7641	(7%)	35.0
7	News	(N)	29	(4%)	684	(0.7%)	3.0
8	Shopping	(S)	17	(2%)	1135	(1%)	64
9	Portal	(R)	12	(2%)	1094	(1%)	93.5
10	Sport News	(E)	9	(1%)	354	(0.3%)	4.0
11	Porn	(P)	8	(1%)	1811	(2%)	238.5
12	OSN	(O)	8	(1%)	6667	(6%)	67.5
13	Video hosting	(V)	7	(1%)	790	(0.8%)	120.0
14	Game	(G)	5	(0.7%)	354	(0.3%)	4.0
15	Other	(X)	10	(1%)	1213	(1%)	79

Table 1. Categorisation of third-party websites.

Table 1 presents a breakdown of the 28% (681 websites including top-300 from every country) of the third-party websites that were categorised.[5] For each category we give both the number of third-party websites and the number of referrals to them from the origin websites.

The largest category of observed third-party sites is *advertisement*. Websites in this category provide general advertisement services such as ad design. The second biggest category is *ad trackers*, sites that facilitate targeted advertisement

[4] `https://www.abine.com`

[5] This extends previous work [4] where just 200 third-party websites were categorised.

by profiling user online activities across different websites. Other categories that also profile user behaviour include *analysts* and *trackers*, in forth and sixth places respectively. Overall, we find that about 37% of categorised third-party sites track user behaviour.

It is worth noting that, despite the large number of advertisement third-party websites, this is not the most prevalent category in terms of referrals, i.e., in terms of the use by origin sites of third-parties within the category. *Ad trackers* and *analysis* are the most frequently referred categories amongst with CDNs, delivering content such as images or scripts on behalf of the origin websites, placed third. Although there are far fewer sites in categories such as OSNs or porn, the proportion of referrals is nearer the larger categories and, in the case of OSNs is larger than the top category, *advertising*.

3.2 Third-Party Websites Across Countries

In this section we investigate the distribution of third-party websites across the US, Great Britain (GB), Australia (AU), China (CH), Egypt (EG), Iran (IR) and Syria (SY).

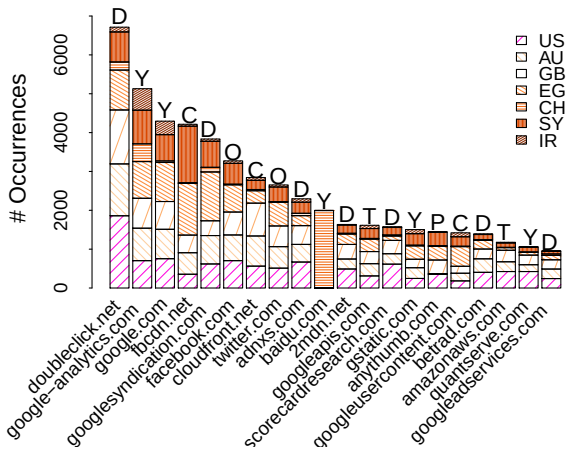


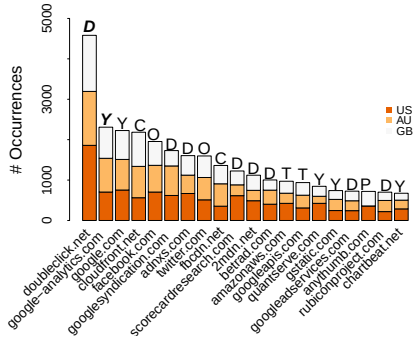
Fig. 3. Top 20 third-party websites with the most referrals from origin websites in the dataset for US, AU, GB, CH, EG, IR and SY. The categories of these third-party websites are indicated by the letter codes provided in Table 1.

Figure 3 shows the distribution of referrals to the top 20 third-party websites per country. The most popular third-party site is `DoubleClick.net` with 6,713 referrals from 2,104 origin sites. DoubleClick is a third-party ad tracker that records user activities across all sites subscribed to its service. This was the most frequently referred to third-party site in the US and was least frequently referred to in China and Iran. Google Analytics was the next most referred to third-party

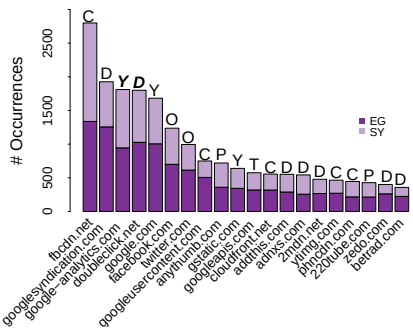
site, and actually obtained a larger share of referrals than DoubleClick from the Iranian and Chinese websites. Overall, Google properties dominate the top-20 by referrals occupying positions #1, #2, #3, #5, #13, #15, #17, and #20.

In general, the prevalence of third-party sites differs strikingly in China and Iran compared to the other countries studied. However, for example, the position of `Baidu.com` among Chinese websites is similar to DoubleClick among top 500 websites in the US. Baidu is a search engine specialised for Chinese-language queries. It was referred to by 207 unique websites from gaming sites (e.g., `61.com`) to news sites (e.g., `tiexue.net`).

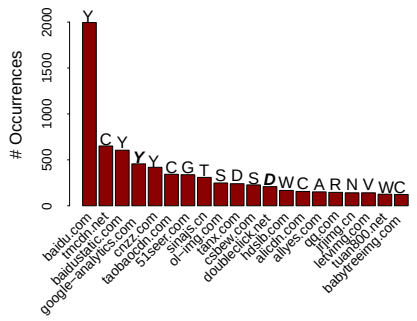
In general, the way that top websites in Arabic speaking countries (Egypt and Syria) refer to third-party sites is similar to the top websites in the English-speaking countries (US, Great Britain, Australia) apart from `fbcdn.net`, a Facebook property. This was referred to about 1,400 times from the popular sites of Egypt and Syria, about 3.8 times more than from the US. Indeed, Facebook is the most popular website in Egypt and Syria, while it ranks third in the US, Great Britain and Australia based on the Alexa records for October 2013.



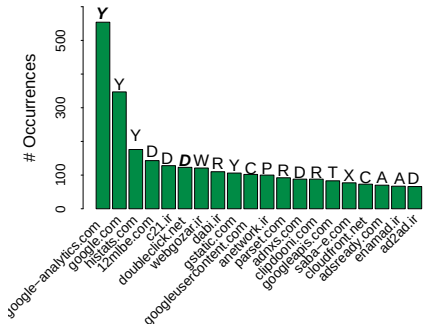
(a) English: US, Great Britain and Australia.



(b) Arabic: Egypt and Syria.



(c) Chinese: China



(d) Persian: Iran

Fig. 4. The top 20 third-party websites in the countries with different language. The categories of these third-party websites are indicated by the letter codes provided in table 1. Those common to all countries are in ***Italic Bold***.

We next look at the top 20 third-party sites across countries with a common language, i.e., English (US, Great Britain, Australia), Arabic (Egypt and Syria), Persian (Iran) and Chinese (China).[6] Figure 4 shows that there are just two common third-party websites among all the countries (DoubleClick and Google Analytics), while there are seven in common among the English, Arabic and Persian groups.

3.3 Regional Third-Party Websites

We now examine if some third-party websites appear exclusively in a specific group of countries. Table 2 shows the top-20 third-party websites in each country. In our data, 125 unique third-party websites appear in the top websites of English-language countries (US, AU, GB) but do not appear at all in the top-20 of other countries (CH, EG, IR, SY). The top-20 that fall in this situation are shown in Figure 5. These third-party websites were referred to by 133 unique sites such as `groupon.co.uk` (shopping website), `usnews.net` (news website). On the other hand, there are 30 unique third-party sites that appear only in Arabic-language countries (SY, EG). The top 20 of these third-party websites are provided in Figure 5.

4 Related work

A number of recent studies have explored modern web traffic, some of which specifically focused on analysing third-party trackers. Here we mention a few which are closely related to our work. Krishnamurthy & Wills [5] investigated the evolution of third-party trackers from 2005 to 2008. They identify the top-10 trackers using an approach based on the *adns* method. They had previously examined how different trackers work and collaborate with each other [7].

Roesner *et al.* [8] suggest a framework for classifying the behaviour of web trackers as well as showing the spread of the identified classes amongst top 500 websites in the world. Our work is closest to that done by Castelluccia *et al.* [9] which analyses the top 100 most popular sites worldwide across a number of countries to assess their tracker behaviours. Their work shows that despite heavy dominance of US-based websites across the visible and invisible web, some countries like China and Russia showing strong differences in local tracking behaviours.

All the above studies focus on specific types of third-party trackers on the web. A similar study was also carried out for mobile ads [10]. In contrast, our study examines the presence of all third-party websites among different sets of popular origin websites, when viewed from a single source.

Third-party websites, in the sense of any non-origin-site HTTP referral, have been studied by Butkiewicz *et al.* [4]. They categorise non-original sites based on

[6] We ignore differences in regional variants of English, as well as the many different regional languages in China.

USA	Great Britain	Australia	China	Egypt	Iran	Syria
(D) double click.net	**(D) double click.net**	**(D) double click.net**	(Y) baidu.com	(C) fbcdn.net	**(Y) google analytics .com**	(C) fbcdn.net
(Y) google.com	(C) cloudfront .net	**(Y) google analytics .com**	(C) tmcdn.net	(D) googlesyn dication.com	(Y) google.com	**(Y) google analytics .com**
(O) facebook .com	**(Y) google analytics .com**	(C) cloudfront .net	(Y) baidustatic .com	**(D) double click.net**	(Y) histats.com	**(D) double click.net**
(Y) google analytics .com	(Y) google.com	(Y) google.com	**(Y) google analytics .com**	(Y) google.com	(D) 12mlbe .com	(Y) google.com
(D) adnxs.com	(O) facebook- .com	(D) googlesyn dication.com	(Y) cnzz.com	**(Y) google analytics .com**	(D) c21.ir	(D) googlesyn dication.com
(D) googlesyn dication.com	(O) twitter.com	(O) facebook .com	(C) taobaocdn .com	(O) facebook .com	**(D) double click.net**	(O) facebook .com
(D) scorecard research.com	(D) adnxs.com	(C) fbcdn.net	(G) 51seer.com	(O) twitter.com	(W) webgozar .ir	(O) twitter.com
(C) cloudfront .net	(C) fbcdn.net	(O) twitter.com	(T) sinajs.cn	(C) googleuser content.com	(R) dabi.ir	(P) anythumb .com
(O) twitter.com	(D) googlesyn dication.com	(D) adnxs.com	(S) ol- img.com	(P) anythumb .com	(Y) gstatic.com	(Y) gstatic.com
(D) 2mdn.net	(D) 2mdn.net	(D) betrad.com	(D) tanx.com	(Y) gstatic.com	(C) googleuser content.com	(D) adnxs.com
(T) amazonaws .com	(C) opta.net	(T) googleapis .com	(S) csbew.com	(T) googleapis .com	(P) anetwork.ir	(D) addthis .com
(Y) quantserve .com	(P) anythumb .com	(Y) gstatic.com	**(D) double click.net**	(C) cloudfront .net	(R) parset.com	(T) googleapis .com
(D) betrad.com	(D) scorecard research.com	(D) rubicon project.com	(W) hdslb.com	(D) addthis .com	(D) adnxs.com	(C) googleuser content.com
(P) anythumb .com	(T) googleapis .com	(D) scorecard research.com	(C) alicdn.com	(C) ytimg.com	(R) clipdooni.c	(C) cloudfront .net
(C) fbcdn.net	(T) amazonaws .com	(D) 2mdn.net	(A) allyes.com	(D) 2mdn.net	(T) googleapis .com	(C) phncdn .com
(T) googleapis .com	(D) betrad.com	(T) amazonaws .com	(R) qq.com	(D) zedo.com	(X) saba- e.com	(P) 220tube .com
(Y) chartbeat .net	(Y) quantserve .com	(D) googlead services.com	(N) jrjimg.cn	(D) adnxs.com	(C) cloudfront .net	(D) 2mdn.net
(D) newsinc .com	(D) googlead services.com	(Y) chartbeat .net	(V) letvimg .com	(D) betrad.com	(A) adsready .com	(D) adsafeprot ected.com
(D) yieldman ager.com	(T) brightcove .com	(Y) google.co .uk	(W) tuan800 .net	(C) phncdn.com	(A) enamad.ir	(D) scorecard research.com
(Y) gstatic.com	(Y) gstatic.com	(T) twimg.com	(C) babytree img.com	(P) 220tube .com	(A) ad2ad.ir	(C) ytimg.com

Table 2. Top 20 third-party websites amongst popular websites of each country of our dataset. Those common to all countries are in **bold**. The categories of these third-party websites are indicated by the letter codes provided in table 1.

the type of services they offer, and strongly inspired our categorisation (Table 1). They categorised the top 200 trackers using regular expressions and automated lookups; in contrast, we categorised about 600 third-party websites, also partly relying on manual inspection.

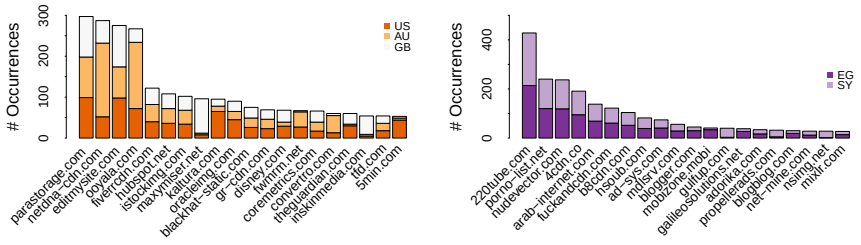
Ihm & Pai [11] report a longitudinal analysis on web tracking technological advances in various countries from 2011. They focus on the technology rather than providing detailed information on top individual ad-trackers per country. They find that ad traffic has been increasing, although they don't identify ad domains.

These studies attempt to explain the mechanisms of web trackers and their prevalence in today's Internet. In contrast, in this paper we have compared the presence of web-trackers across different countries and common language regions.

5 Conclusions and Future Work

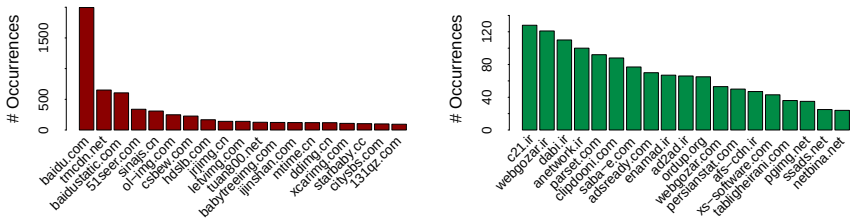
In this paper, we studied the presence and use of so-called third-party websites on the web. We sampled every continent, targeting the 500 most popular websites in the US, Great Britain, Australia, China, Egypt, Iran and Syria. We examined this data by country and by grouping countries with a common language. Perhaps unsurprisingly, our results show that Google properties dominate overall, occupying four of the top five positions, and eight of the top 20. Two specific Google properties sites, DoubleClick and Google Analytics, are the only ones that are in the top 20 in every country we examined.

Cultural/language effects are also significant, with a great deal of similarity between the top 20 lists and referrals for the three western English speaking countries (US, Great Britain, Australia), and between the two Arabic countries (Egypt, Syria), and notable differences between those each of the four language groups (English and Arabic plus the two singleton groups, Persian/Iran and Chinese/China). Google properties, and ad tracking and tracking sites generally dominate everywhere. There are also minor differences in the prevalence of categories in the different countries: web hosting, portals and advertising only appeared in Iran and China, and porn was slightly more prevalent in the Arabic



(a) English: US, Great Britain and Australia.

(b) Arabic: Egypt and Syria.



(c) Chinese: China

(d) Persian: Iran

Fig. 5. The top 20 third-party websites which appear exclusively in a specific group of countries.

countries (Egypt, Syria). Further analysis on the effect of language and culture, however, requires bigger groups of countries. For example in our dataset, the Chinese and Persian speaking groups each contained only one country but those languages are also spoken in several other countries. The English speaking group can be also expanded to contain other countries such as India which is culturally different from other countries in this group. Additionally, in our measurements we monitor those third-party websites which are visible through our single vantage point while monitoring the popular websites from a different vantage point (e.g., different countries) may reveal different third-party services. We plan to address these limitations in our follow-up studies of the personal data ecosystem.

Acknowledgements

This work was funded in part by Horizon Digital Economy Research, RCUK grant EP/G065802/1.

References

1. Crane, D., Pascarello, E., James, D.: Ajax in Action. Manning Publications Co., Greenwich, CT, USA (2005)
2. Popa, L., Ghodsi, A., Stoica, I.: HTTP as the narrow waist of the future internet. In: Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets), New York, NY, USA, ACM (2010) 6:1–6:6
3. Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., Jahanian, F.: Internet inter-domain traffic. SIGCOMM Comput. Commun. Rev. **41**(4) (August 2010)
4. Butkiewicz, M., Madhyastha, H.V., Sekar, V.: Understanding website complexity: measurements, metrics, and implications. In: Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC), New York, NY, USA, ACM (2011) 313–328
5. Krishnamurthy, B., Wills, C.: Privacy diffusion on the web: a longitudinal perspective. In: Proceedings of the 18th international conference on World Wide Web (WWW), New York, NY, USA, ACM (2009) 541–550
6. Mortier, R., Haddadi, H., Henderson, T., McAuley, D., Crowcroft, J.: Challenges & opportunities in human-data interaction. In: DE2013: Open Digital, MediaCityUK, Salford, UK (2013)
7. Krishnamurthy, B., Wills, C.E.: Generating a privacy footprint on the Internet. In: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement. IMC '06, New York, NY, USA, ACM (2006) 65–70
8. Roesner, F., Kohno, T., Wetherall, D.: Detecting and defending against third-party tracking on the web. In: USENIX Symposium on Networking Systems Design and Implementation (NSDI), USENIX (2012)
9. Castellucia, C., Grumbach, S., Olejnik, L.: Data Harvesting 2.0: from the Visible to the Invisible Web. In: The 12th Workshop on the Economics of Information Security, Washington, DC, USA (June 2013)
10. Vallina-Rodriguez, N., Shah, J., Finamore, A., Grunenberger, Y., Papagiannaki, K., Haddadi, H., Crowcroft, J.: Breaking for commercials: characterizing mobile advertising. In: Proceedings of the ACM Internet Measurement Conference (IMC). (2012)

11. Ihm, S., Pai, V.S.: Towards understanding modern web traffic. In: Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC). (2011)