Nurse and Buckley *Hum. Cent. Comput. Inf. Sci.* (2017) 7:40
DOI 10.1186/s13673-017-0121-6

Human-centric Computing and Information Sciences

RESEARCH Open Access

Behind the scenes: a cross-country study into third-party website referencing and the online advertising ecosystem

Jason R. C. Nurse[1*] and Oliver Buckley[2]

*Correspondence:
jason.nurse@cs.ox.ac.uk
[1] Department of Computer Science, University of Oxford, Oxford, United Kingdom
Full list of author information is available at the end of the article

Abstract

The ubiquitous nature of the Internet provides an ideal platform for human communication, trade, information sharing and learning. Websites play a central role in these activities as they often act as a key point of interaction for individuals in navigating through cyberspace. In this article, we look beyond the visual interface of websites to consider exactly what occurs when a webpage is visited. In particular, we focus on the various web scripts that are often programmatically executed, to explore the extent to which third-party sites are referenced. Our aim is to study these references and the ecosystem that they create. To gain maximal impact while also allowing for a cross-country comparison, our study is scoped to an assessment of the top 250 sites in the UK, USA, Germany, Russia and Japan. From our analysis, there are various novel contributions of note. These include the empirical identification of a vast ecosystem of third-party information processing sites, especially advertisement networks, and the evidential discovery of a few significant players irrespective of country and locale. Through a user study, we also find that while individuals do have some knowledge of the prevalence of advertisements in websites, their understanding of the variety of activities that occur upon visiting websites, is not widely known. Going forward, we therefore advocate for increased transparency in such activities and the wider online advertisement ecosystem.

Keywords: Web browsing, Web scripts, Information studies, Online advertising, Advertisement ecosystem, Human information behavior, Online privacy

Introduction

Technology is the foundation of modern-day society. The Internet, for instance, is critical in facilitating large scale communication, trade, information sharing and learning. While there are numerous ways in which humans can interact with the Internet, websites play one of the most central roles. Current estimates indicate that there are more than 1 billion websites today [1], many of which act as the first point of contact for numerous individuals as they navigate through cyberspace. These sites offer a mixture of static and dynamic informational content, typically with the aim of encouraging user engagement and maintaining repeat visits.

Throughout the literature, the topic of websites has been researched in several contexts and for a variety of purposes. One of the most noteworthy areas of work has been

on enhancing user experience and building trust in sites. For instance, articles have identified and examined factors crucial to website design [2], defined frameworks for evaluating the quality of websites [3], and conducted user studies (e.g., via eye-tracking) to assess how websites are used and any innate complexities in perceptions [4].

Apart from usability, there is a significant strand of research exploring the behind the scenes activities, particularly as it relates to information processing and, the online tracking and profiling that occurs as users browse websites. These relate to the passive collection of user information often without user knowledge [5], the unsanctioned tracking of users and user web-surfing behaviour across websites by third-party trackers [6], and the proliferation of complicated privacy policies which make it difficult for users to understand conditions of website use [7]. Such work highlights the clear tension between users concerned about maintaining privacy and anonymity online, and advertising companies seeking to better track and profile users across the web.

This article engages in research in the area of online web references. Specifically, we consider the topic of websites in the context of third-party references and the online advertising ecosystem. Through an analysis of the top-accessed 250 websites in the UK, USA, Germany, Russia and Japan, we empirically examine the extent to which third-party sites are referenced by these popular domains. Our research objectives in this article are: (a) to investigate these 'behind the scenes' references; (b) to explore the size and characteristics of their ecosystem (i.e., viewing the ecosystem as an entity that can be studied), and (c) to consider the perception and understanding that users have of this ecosystem. While previous literature has explored this general topic (e.g., [6, 8, 9]), we believe our work is novel in several regards as highlighted below.

Firstly, as we analyse the links to these third-party websites, we also construct a model graph of the relationships between sites directly and the third-party organisations that run them; this allows detailed insight into the ecosystem and its connectivity. The second contribution is a comparison of the information on localised third-party references through assessing the most used sites in five of the world's leading economies. From this we can consider the top domains referred to and assess how these compare in prevalence across nations of different cultures and languages. Lastly, we conduct a user study to determine the extent to which users understand these 'behind the scene' activities of websites including the origin of third-party sites and how often they are referenced.

The remainder of this paper is structured as follows. The "State of the art" section reviews the current literature in online advertising and web tracking, most of which can be found in the online privacy field. In the "Methodology" section we detail the methodology that we adopt for our research study. This explains the sample of websites selected for analysis, the process through which we analyse them, and how we incorporate a user study to enrich our research. We then present and discuss our results in the "Results and discussion" section; these cover both the website analysis and the user-based study. Finally, we conclude and outline avenues for future work in the final section.

State of the art

Online advertising is an enormous industry, with advertising revenues reaching $27.5 billion in the first half of 2015 [10]. This figure climbed to $59.6 billion by the end of 2015, with both Google (with $30 billion) and Facebook (with $8 billion) being the biggest benefactors [11]. One of the key differences between online and offline advertising is the ease at which individual users online can be profiled, and then targeted by marketers based on their browsing history, personal information and location.

There are various approaches that have been used to gather user information and track their movements online. Traditionally, tracking cookies and scripts have been the main way of monitoring who is visiting a particular website. As sites are visited, browser cookies act as a unique identifier, or scripts—many of which are provided by third-parties—execute and attempt to collect user data or provide customised content (e.g., information or advertisements). Good examples of the proliferation of these trackers and how they work can be found in existing studies [6, 9], including analyses of various tracker properties.

Due to changing attitudes to user privacy, the increased usage of mobile phones, and recent legislation [12], other approaches to track users have surfaced. Browser fingerprinting [13, 14] is becoming an increasingly popular means of user identification in place of the more traditional cookie-based approaches. This is a technique that uses web scripts (often originating from third-party websites) to collect various information about a remote computer in order to try and (re)identify that particular machine (as a conduit to the user). The collected information can incorporate a wide variety of features including screen resolution, time zone or the availability of specific font sets.

The established field of information science [15] has been fundamental to the processes mentioned above. This is linked to the fact that companies have continued to pursue novel ways to gather information (and metadata) on prospective customers online, and to process and manage that information for optimum use. Common activities include leveraging big data to optimise digital marketing [16], and developing approaches to extract individual user profiles from online surfing information for purposes of behavioural targeting [17]. These concerns will almost certainly increase in the future as technology becomes more ubiquitous [18].

There have been various approaches proposed to elucidate and, in the interest of privacy, to block the range of external scripts and web tracking attempts. One stark reality however is that fully preventing third-party references is non-trivial, as often scripts are required in order for sites and services to function properly. For example, a number of social media sites will use widgets (e.g., the Facebook 'Like' button) to identify individuals but also as a means of the user interacting with content. If all references (via their respective scripts) are blocked then this will limit the usefulness of many core services and sites. Moreover, an increasing amount of websites (e.g., Forbes) are preventing site access if they detect that the user is using software (e.g., Adblock [19]) that blocks digital advertisements. Ad-blocking software also comes with its own risk as seen recently when a fake Adblock Plus extension made its way on to the official Chrome Web Store, and was downloaded by thousands of users [20].

Wu et al. describe a solution called 'DMTrackerDetector' which detects third-party trackers automatically using structural hole theory and supervised machine learning [9].

Nikiforakis et al. identify that one of the most significant issues is not that the generated fingerprints are unique but that they can be used to link the device across multiple visits [21]. Their work proposes 'PriVaricator', a solution which aims to add an element of randomisation to make the fingerprinting process non-deterministic and thus making it more difficult to link fingerprints across multiple user visits.

A salient point regarding privacy is that the way in which Internet users perceive privacy online is something that is still not fully understood. This is particularly related to attitudes towards online advertising and third-party website referencing and tracking. Melicher et al. interviewed Internet users to better understand their concerns about web tracking [22]. Interestingly, their results suggest that users want more control of tracking but are unwilling to invest the effort to control tracking and often distrust current tools. This point can be seen in much earlier works [23, 24], and typifies the existence of the well-researched Privacy Paradox [25].

In a recent TRUSTe study, 92% of British Internet users were concerned about their online privacy, but in contrast only 25% actually understand how organisations share their personal information [26]. In another context, Melicher et al. collected browsing histories and interviewed 35 people in USA about the benefits and dangers of online tracking [22]. The research found that, as with similar studies, individuals would like greater control over tracking but also that users were able to see benefits to controlled tracking. This study found that users were largely unwilling to dedicate the effort to control the tracking of their data.

In this article, we aim to build on the research of many of the articles mentioned above, to examine the size and characteristics of the current third-party website (including online advertising) ecosystem. This includes drawing out common networks and associations of sites, and any third-party networks.

Methodology

To briefly reiterate, our research objectives are threefold. Firstly, we aim to investigate the 'behind the scenes' references made in popular websites today. Secondly, this work explores the size and characteristics of the ecosystem in which these sites reside and the links that are made between other entities (i.e., viewing the ecosystem as an informational entity that can be studied). Finally we consider the perception and understanding that users have of this ecosystem, and thus explore the human-computer interaction perspective in more detail.

To structure our research, we adopted a methodology that involved a combination of technical and user-focused tasks. We felt that this was the most appropriate approach given the technical nature of analysing website source content, but also the importance of considering the understanding human users may have on third-party references and the online advertising ecosystem. There are four main steps in our approach.

The first step was the selection of websites that would be most appropriate for our research. Given our interest in also conducting a cross-country comparison of sites and ecosystems, we decided to focus on five of the major economies today and those with a significant number of citizens online. These were likely to be countries where third-party references were the most active. The countries that met our criteria were the UK, USA, Germany, Russia and Japan; China, India and Brazil were also viable, however we were

unable to reach other selection criteria as will be outlined below. We decided to scope our assessment to the top 250 sites in each of these countries, and used the rating index as as defined by Alexa [27], the commercial web traffic data and analytics company.

The second step comprised of analysing each website's HTML source and extracting the references to third-party scripts. These can typically be found within the `<script>` elements of the page, and included either as a `src` (source) attribute (e.g., `<script type="text/javascript"` *src="http://..."*`></script>`) or in the script body itself (e.g., `<script>..` *js.src="//..."* `..  url="//.."</script>`). We identified third parties as those references which did not match the current domain. As the number of sites and potential variety of external references could be quite high [28], we opted for a semi-automated approach to this task. Specifically, we developed a python script based on a combination of modules (e.g., urllib, beautifulsoup) and tailored regular expressions, to allow extraction of the desired external links.

To ensure that we captured all of the third-party references, we adopted an iterative method to development and testing, by comparing our script's output to that of manual analyses; we conducted our analysis on a site-by-site basis. Script extraction tools, such as LightBeam [29], were also very useful at supplying additional evidence and support. In situations where third-party websites or links made follow-on references to other sites, we recorded those as well; this allowed us to build a more complete picture of the range of third-parties involved.

Most importantly, and unlike in many other studies in the field, we ensured that each site that was visited was accessed as if we were physically in that country—Virtual Private Network (VPN) services and a physical presence in two cases, allowed for this. This acted to increase the authenticity of the third-parties referenced as initial sites were accessed; unfortunately, we were unable to find reliable VPN services for China, India and Brazil hence the exclusion of these sites. The output of this task was a list of third-party domains associated with each of the top websites.

Having gathered the listing of third-party website references of all sites, the next step concentrated on its analysis. We engaged in two levels of assessment. The first level sought to apply social network analysis (SNA) techniques to build a visual network that would better allow the identification of relationships within the site-domain listings. SNA can be described as a set of methods that allow the links between elements (e.g., people, devices, or things) to be studied [30]. We were keen on applying this technique to identify the third-party domains that were the most common across sites generally, or specifically when considering the site's locale. Moreover, we could identify the domains central to the created network, how cohesive the network is, and clusters of highly connected domains. This was our information processing and interpretation stage.

Based in part on the analysis from the first level, we then considered the companies behind the third-party website domains. This attempts to demystify the domains referenced and elucidate the organisations orchestrating them. As highlighted in other works (e.g., [8]), we expect that while there will be hundreds of domains, there will only be a small subset of core sources representing the giants in the online advertising and analytics ecosystem.

The last step in our methodology was to conduct a user study to determine the extent to which average web users understand the 'behind the scenes' activities of websites they

visit, including origins of online advertisements and the extent to which third-party scripts are referenced. We aim to build on previous work (e.g., [24]) that has examined the user's perspective and pose additional research questions to gather perceptions on topics such as which types of sites users think are the most private and estimations about third-party website references.

Results and discussion

Analysis of third-party website references and the online advertising ecosystem

Overview of websites and size of the ecosystem

Our approach to analysing the 'behind the scenes' ecosystem began with an assessment of the 1250 websites selected. The websites examined were of various types including search engines, news agencies, social media, commerce and finance, professional networks, video streaming, and cloud services. While each country had its unique set of websites, there were several common domains present, and with similar degrees of popularity. For instance, well-known sites such as `google.com`, `youtube.com` and `wikipedia.org` all featured in the top 10 sites of the five countries.

The largest similarity in web domains was between the UK and USA (at 46%), but this is to be expected given the likeness of the two countries (e.g., in society and language). This was followed by Europe's UK and Germany (at 35%), and then the USA and Germany (at 29%). Unsurprisingly, the most dissimilar countries were Russia and Japan (at 9%) and the USA and Russia (at 11%). Language spoken and location are likely to be reasons for these differences. It is important to note such similarities and differences as they may help explain differences in ecosystems across countries in the subsequent sections.

From our analysis of each country's websites, we were able to extract a comprehensive list of third-party sites that were referenced. We used a simple "`source-to-reference`" list format which also catered for third-party sites making their own references; where `source` was the initial site and `reference` was the site referenced. The real advantage of this format however, was the ease at which social network analysis (SNA) tools could be applied to allow the characteristics of the third-party ecosystem to be explored and visualised. As an example, we present Fig. 1, which was created using the SNA tool, Gephi.[1] This figure displays a network constituted of the UK websites, the third-party sites referenced, and the various connections between them.

The first point of significance regarding Fig. 1 was the substantial size of the network created based on only 250 initial sites; we do not intend this image to be legible but instead to use it to exemplify the network's vastness. Overall, the network contained 1078 unique domains (or in graph terms, nodes), with 4242 connections (also known as directed graph edges) between them. In this context, connections denote references that indicate third-party website calls, potential links to advertisement networks, or associations to legitimate sites (e.g., `bbc.com` making references to `bbci.co.uk` for images). From a user perspective two key aspects to highlight here are, the amount of additional domains referenced (4 times the initial number), many of which are likely to be unknown to users, and the large number of connections between these various sites.

[1] https://gephi.org/.

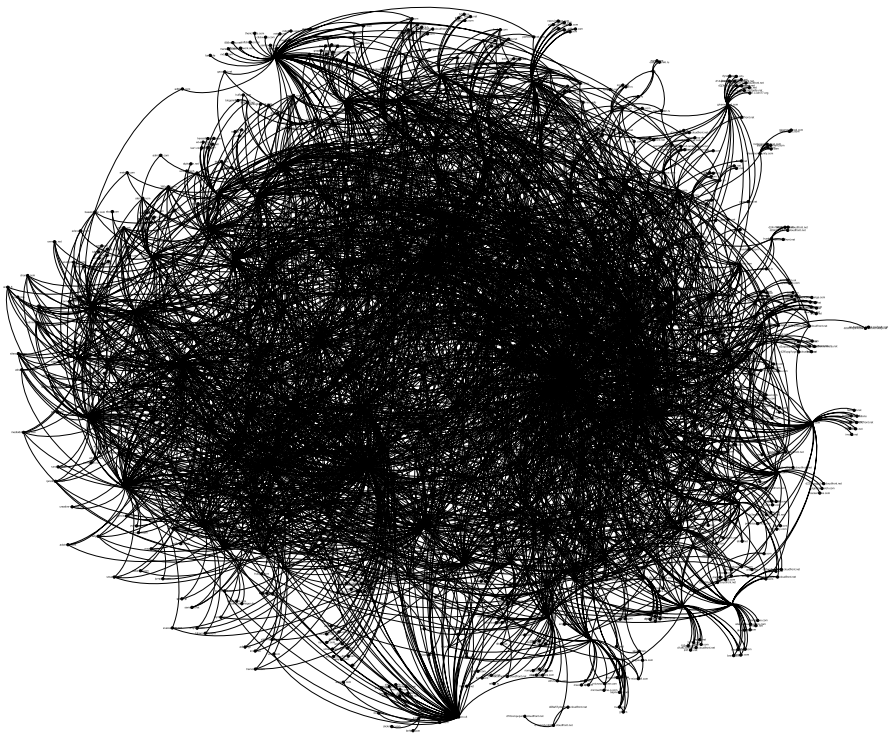

Fig. 1 The network of UK websites and all of the third-party references made by these sites. This illustrates the size and complexity of these networks, and while it is not possible to read the intricate details, this does highlight the significance of these networks

Moreover, as can be seen, some of the domains are very well connected and may act as hubs that gather user information from various sites—this point will be discussed further in the next section.

With regards to the characteristics of the other networks, the USA possessed the largest and most dense network with 1135 domains and 4455 references. It was followed by Germany with 1034 domains and 3808 references, Russia with 976 domains and 3582 references, and Japan with 796 domains and 2967 references. Each of these networks highlights a notable number of third-party sites that are referenced as individuals browse webpages.

Although the size of these networks was a noteworthy factor in itself, our most intriguing finding was the positioning of Japan given their large market size (126 million citizens) and high Internet adoption rate (at 91%) [31]. In particular, Japan has the 5th largest number of Internet users in the world (second only to USA in our study) and with a 91% Internet adoption rate is only behind the UK (at 92.6%) in our study. A market such as this would be ideal for web advertising networks, hence our expectation of a larger third-party website prevalence. Upon further investigation, we found that the reason for this disparity may not due to lack of attempts, but rather because Japan still primarily focuses on television advertising [32]. This is in contrast to markets such as the USA and UK where online advertisements—as might be inferred from the size of the networks discussed—are the dominant way to reach consumers.

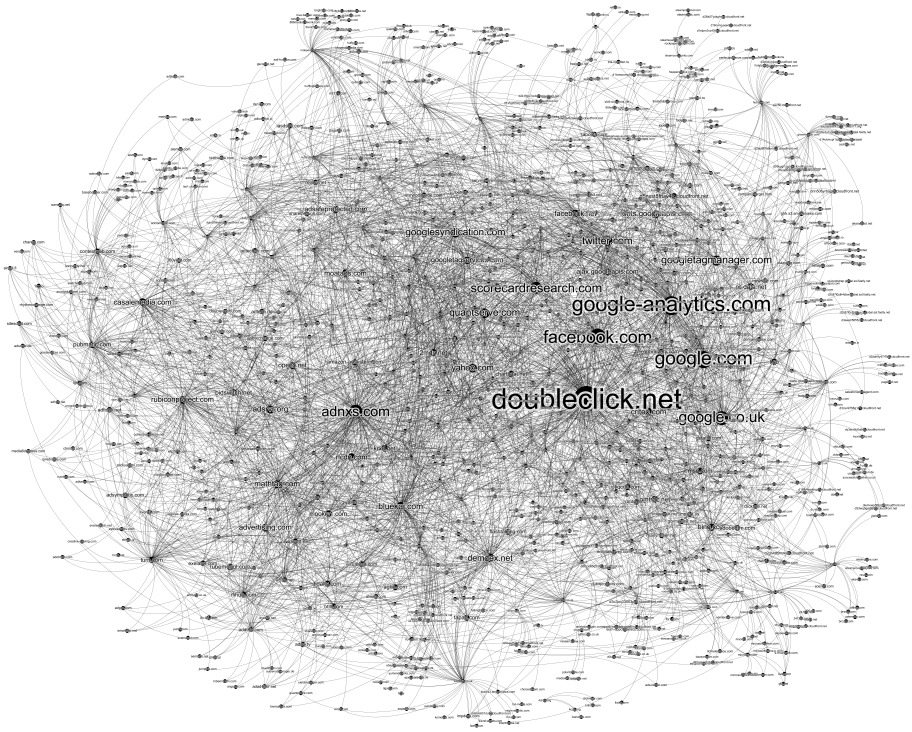

Fig. 2 The key sites that are referenced within the UK network, where the size of the domain name and node indicates the number of times that this site is referenced. While all of the finer details are not entirely clear, this figure is key in illustrating the sheer volume of referenced sites

The big players: third-party sites referenced the most

The networks discussed above provide a 'bird's eye view' of the website ecosystem and its complexities. Our aim in this section is to further understand that ecosystem, and in particular, the most referenced sites. This would allow us to empirically identify the central third-party websites in each country, the key players behind these domains, and how they may differ across locales.

To determine the most referenced websites in the country networks that were created, we calculated the in-degree network centrality (i.e., an SNA metric that measures the number of incoming connections) for all of the websites. We then used this in-degree measure to proportionally size the various nodes in each of the networks such that the larger the node the more times it was referenced. Figure 2 shows the visual of the UK network from Fig. 1 updated to account for this metric. For purposes of readability, we also present an alternative visualisation in Fig. 3 which focuses on the top 50 domains.

From an analysis of each website's in-degree, the top ten were: `doubleclick.net` (189 references to it), `google-analytics.com` (134 references), `google.com` (117 references), `facebook.com` (107 references), `google.co.uk` (90 references), `adnxs.com` (82 references), `scorecardresearch.com` (71 references), `google-tagmanager.com` (56 references), `twitter.com` (55 references) and `quantserve.com` (52 references).

Fig. 3 Word cloud displaying the key sites that are referenced within the UK network, where the size of the domain name indicates the frequency in which that site is referenced

The first observation that may be made from this list is the prevalence of domains that are attributable to the Internet giant, Google Inc. These include `doubleclick.net` (DoubleClick is a subsidiary of Google that develops and provides Internet advertisement-serving services), `google-analytics.com` (Google Analytics is a web analytics service offered by Google), `google.com` and `google.co.uk` (Google search engines or portals to other Google services), and `googletagmanager.com` (Google Tag Manager is a system that allows the management of tags or code snippets which send information to third parties). Given that most of these sites can be linked to online ads, this finding highlights Google's dominance of the online advertising industry in the UK and acts to re-emphasise the conclusions of earlier works [11].

While not as significant as Google, Facebook (`facebook.com`) and Twitter (`twitter.com`) also are referenced by several websites. These references may be due to embedded Facebook and Twitter services (e.g., 'Like' or 'Share' buttons on sites) or ads, considering that Facebook, for instance, is one of the lead companies in the online advertising industry [11]. Our analysis of the other top domain names uncovered that they are all generally linked to digital advertising services and platforms.

AppNexus (the owner of `adnxs.com`), for instance, regards itself as an independent technology company that provides trading solutions and powers marketplaces for Internet advertising—they partner with, and are backed by, companies such as Microsoft [33]. The aim of the next company, Scorecard Research (`scorecardresearch.com`), is to collect data (in this instance, information on users' website visitation patterns) to assists their clients in better meeting the needs of consumers.[2] Finally, Quantcast (`quantserve.com`) engages in website audience measurement for publishers and tailored online advertisement delivery.[3]

[2] http://www.scorecardresearch.com/about.aspx?&newlanguage=1.

[3] https://www.quantcast.com.

Broadly speaking, a central goal of these and other third-party organisations is to facilitate a better understanding of website users (including their profiles and browsing patterns across sites) to allow for more unique and targeted marketing. This is intended to benefit advertisers by enabling them to reach their ideal consumers, and consumers in ensuring they are shown the more suitable ads.

The networks of the other countries were similar to that of the UK with Google dominating in the top 10 most referenced sites. As shown in Table 1, Google domains, `doubleclick.net`, `google-analytics.com` and `google.com` were permanent features, with only `facebook.com` also appearing in all countries.

There were also a range of new domains found, many of which were associated with other digital advertising firms. These include: DemDex (`demdex.net`), now owned by the American multinational Adobe Systems Inc., which captures behavioural data on users to allow for better targeting of online ads [34]. InfoOnline (`ioam.de`) is a German-based organisation focused on digital audience measurement. In Russia, Yandex N.V. (`yandex.ru`) is a large technology firm which also engages in online advertising, while the site `tns-counter.ru` appears to be part of a project by a Russian enterprise (TNS) seeking to understand behaviour of Russian citizens on websites. Lastly, the technology company OpenX (`openx.net`), specialises in online advertising marketplaces.

For those domains not directly associated with advertising firms, one reason for their prevalence may be social plug-ins (as discussed with Facebook and Twitter earlier). For instance, VK (`vk.com`) is a Russian social networking service that has site widgets and plug-ins for sharing and 'liking' similar to Facebook. We also have found plug-ins for auctions and marketplaces, which may explain the popularity of sites such as `yahoo.co.jp` (though we should note that Yahoo! does also engage in the online advertising industry [35]).

While Table 1 is useful at highlighting the key domains in the ecosystem, we were also keen to explore the websites that were well referenced but not present in the initial 250 for each country. This would more clearly depict the prominent websites that were in

Table 1 The most referenced sites in the USA, Germany, Russia and Japan

USA	Germany	Russia	Japan
doubleclick.net (203)	doubleclick.net (163)	doubleclick.net (179)	doubleclick.net (168)
google-analytics.com (130)	google-analytics.com (125)	google-analytics.com (175)	google-analytics.com (144)
google.com (122)	google.com (90)	yandex.ru (173)	google.com (98)
facebook.com (118)	facebook.com (89)	yadro.ru (121)	facebook.com (78)
scorecardresearch.com (97)	adnxs.com (82)	google.com (97)	googletagmanager.com (72)
googlesyndication.com (69)	googlesyndication.com (72)	tns-counter.ru (83)	google.co.jp (71)
adnxs.com (68)	ioam.de (64)	facebook.com (80)	yahoo.co.jp (57)
demdex.net (59)	criteo.com (50)	vk.com (72)	googlesyndication.com (53)
rubiconproject.com (53)	googletagmanager.com (50)	google.ru (68)	twitter.com (48)
quantserve.com (51)	fonts.googleapis.com (47)	mail.ru (67)	openx.net (42)

Each list is ordered and next to the domain, the number of references is presented

the ecosystem only as a result of being referenced by other sites. Figure 4 presents our findings when considering the top 10 websites of each country. We also have noted the prevalence of these third-party sites in other countries to allow for comparison.

Looking beyond the websites that are well-known to us based on earlier findings, Fig. 4 uncovers several additional domains. These include `adition.com` (from the German-based, ADITION Technologies), `bluekai.com` (BlueKai, which was acquired by Oracle in 2014), `criteo.com` (Criteo) and `rubiconproject.com` (Rubicon Project). These are all companies that engage in digital advertising of some form. One noteworthy point here is the fact that although certain organisations (e.g., AppNexus) are established in multiple countries, others appear to concentrate only on local networks. This seems particularly relevant in the German and Russian cases with sites such as `adition.com`, `adriver.ru`, `ioam.de` and `yadro.ru`. While culture may generally be a factor, it could also be a business decision to focus on core markets and their dominance.

With respect to the US and UK, they have similar distributions across each of the various domains, with the US typically leading the UK. Figure 4, therefore, also mirrors the structure of those networks more broadly. Although Japan has the smallest network overall, it should be noted that several of the advertising domains are present which may suggest a growing interest by marketers generally.

The dominance of Google in Japan however, is not to be overlooked. We compared the percentage of references made to the main Google domains (i.e., those in the top 10 in Table 1) across the countries and found that 20.42% of all (2967) third-party references in Japan sites are to Google domains. For the other countries, the US, UK, Germany and Russia are at 11.76, 13.81, 14.36 and 14.48% of all references respectively. It will be interesting to monitor this trend in the future and examine whether Google's dominance expands, or other local advertising companies such as Dentsu and Hakuhodo establish a substantial online presence.

The inbetweeners: third-party sites acting as a bridge

In addition to analysing websites that were referenced the most, we also engaged in a brief assessment of the third-party sites that made the most references. In SNA terms, this is called the out-degree network centrality. Identifying these sites could indicate central points in the ecosystem which act as a connector or bridge for other sites, even if

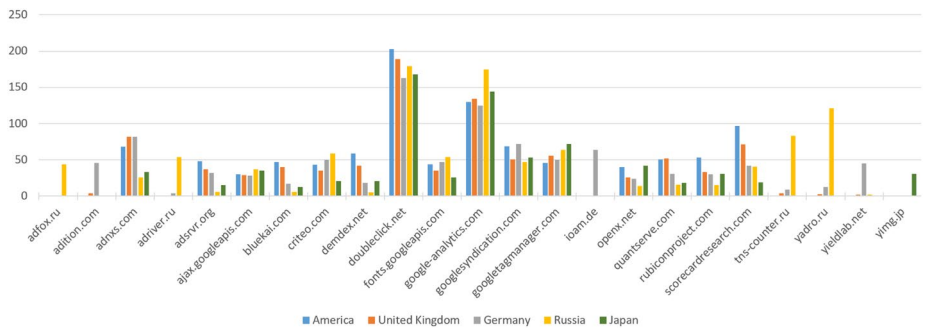

Fig. 4 A comparison of the most common third-party websites across the five countries that were studied (USA, UK, Germany, Russia and Japan). This shows the third-party sites that were the most prevalent across all of the countries

they themselves are not heavily referenced. In Fig. 5, we present a comparison of the top ten third-party websites that make the most references in each country (this is similar in concept to the depiction in Fig. 4).

The first point of note as indicated in Fig. 5 was that several of the domains (e.g., `adnxs.com`, `casalemedia.com`, and `doubleclick.net`) and organisations uncovered in the previous section, remain present. This means that not only were these domains heavily referenced, but that they were also responsible for several references to other third-party sites themselves. A reason for this could be partnering third-party services, shared advertising networks or advertising marketplaces. Moreover, although we found that Google domains were again prevalent, they were not as dominant as when examining sites by references made. For instance, other domains such as those by AppNexus (`adnxs.com`) and Rubicon Project (`rubiconproject.com`) featured reasonably well here.

Across countries, there were nuances in the domains present as might be expected. AdRiver (`adriver.ru`) for example, is a Russian company that specialises in Internet advertising technology; this explains their presence in Russia as opposed to in other nations. The US-based company AppNexus (`adnxs.com`), was responsible for a significant number of out-going references in the US, UK and Germany, but none in Russia or Japan. This finding echoes the low in-degree centrality measures for these countries in this domain (as shown in Fig. 4). In Japan, the Rubicon Project (`rubiconproject.com`) has the highest number of out-going references and could indicate a hub or main 'bridge' platform in the Japanese advertisement marketplace. This presence is the opposite in the US and Russia, where the Project is hardly prominent.

While understanding the domains with the highest out-degrees allows some insight into the main intermediate third-parties, we were also interested in the specific sites that they reference. This could enable for more insight into the ecosystem. We therefore analysed each of the top ten domains for each country to determine what domains they referenced. As an example of our findings, we present Fig. 6 which is a directed network graph of the top UK third-party domains.

In addition to re-emphasising the prominence of `adnxs.com` and `doubleclick.net`, Fig. 6 displays the range of domains which are referenced. A detailed inspection of

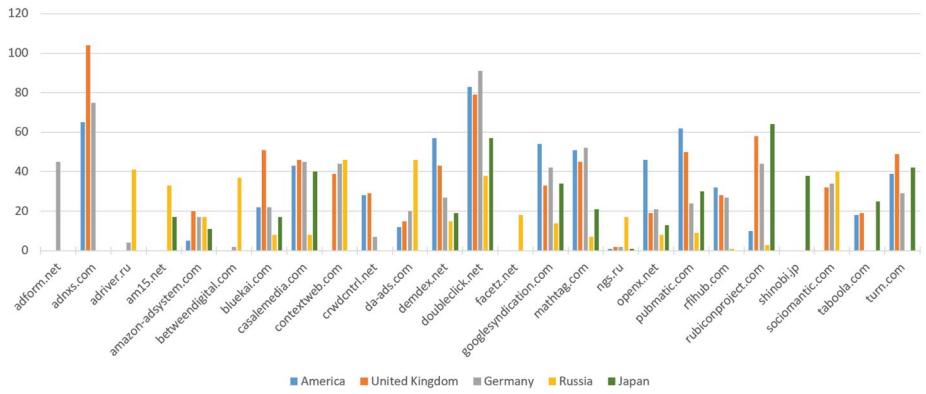

Fig. 5 A comparison to show the number of references made by each of the third-party websites, across the five countries that were studied (USA, UK, Germany, Russia and Japan)

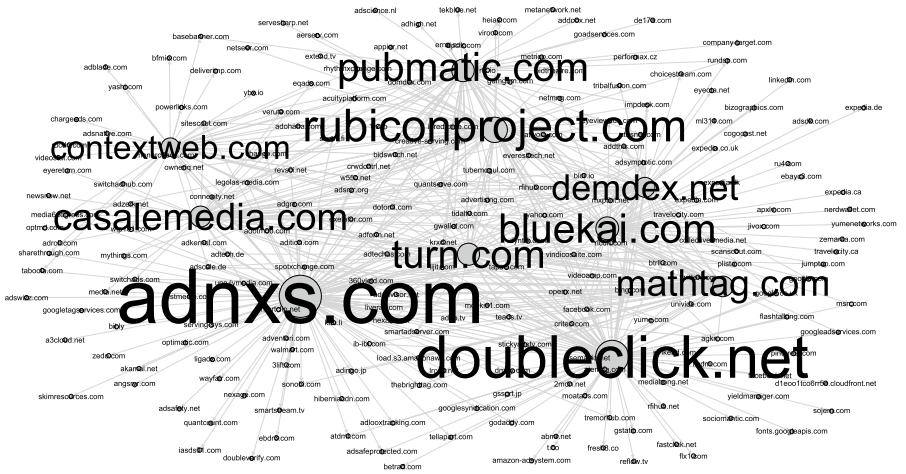

Fig. 6 A directed graph that shows the top ten third-party sites in the UK, which are ordered by out-degree. The size of the domain is proportionate to the number of references made by that site (i.e., the larger the text, the more references). Additionally, the other domains that are referenced by these third-party sites are also included

the graph highlights the existence of additional advertising networks and organisations such as Adblade (`adblade.com`), Optimatic (`optimatic.com`) and Smart AdServer (`smartadserver.com`). This may suggest some association or sharing across web domains to supply ads to users. We also identified a few well-known organisations such as Expedia (`expedia.co.uk`), Facebook (`facebook.com`), GoDaddy (`godaddy.com`), and Travelocity (`travelocity.com`). The links here could be justified by various means, but possibly the most likely is in the provision of ads. This, in addition to our earlier findings, exemplify the extent to advertising networks and organisations may work together in the website and third-party ecosystem.

To follow up on the potential association between these various third-party networks, we decided to conduct a brief exploratory study. For this, we returned to our initial network graphs of the full sites and references (e.g., as shown in Fig. 2). We then applied a clustering algorithm [36] to these graphs to determine the main associated groups and clusters that can be uncovered. Figure 7 depicts the clusters discovered in the UK network graph.

From Fig. 7, we can see that the clustering algorithm identified three large clusters as represented in purple, green and light blue. The purple cluster is the largest one and accounts for $\sim$ 25% of the sites. It contains many of Google's sites such as `google.com`, `doubleclick.net`, `google-analytics.com` and `googletagmanager.com`. An interesting addition to this cluster is `facebook.com`, given its own advertising ambitions, but this may be explained by the presense of these third-party references on many of the same sites.

The green cluster is the second in size at $\sim$ 20% of the network. The prominent sites within this cluster include `adnxs.com`, `rubiconproject.com`, `mathtag.com` and `yahoo.com`. Visually analysing the cluster, many of the second-tier third-party sites and advertising companies (e.g., AppNexus and the Rubicon Project) can be seen.

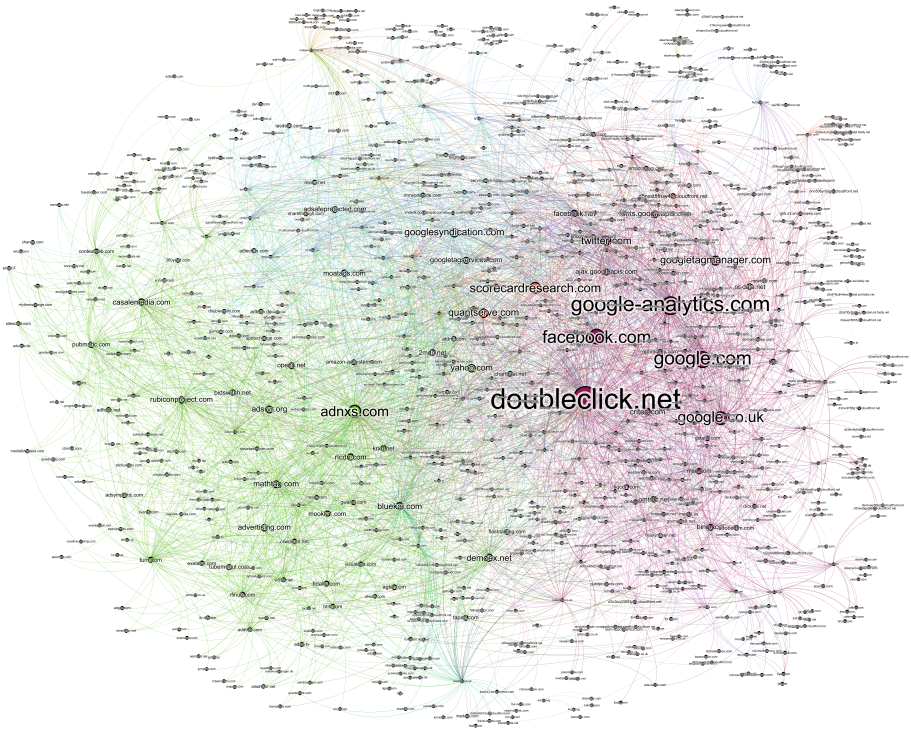

Fig. 7 A graph to highlight the different clusters within the third-party sites, which aims to draw out the distinct groupings within these sites in the UK network

This better hints to the reality that advertising networks and organisations, particularly smaller ones, may work together in the third-party ecosystem.

Finally, in the light blue and covering ∼ 13% of the network, is the third-largest cluster. Here there are more domains from advertisers (or enabling the provision of advertisements) such as `googlesyndication.com`, `moatads.com`, `adsafeprotected.com`, `googletagservices.com`. This cluster was somewhat understandable given the associations between the included sites, but it we might have expected a closer association with the first cluster (e.g., potentially the main Google domains being present in the same cluster). Possibly the most significant finding from this exploratory study was the close network (or grouping) between some sites that almost certainly alludes to a common advertising marketplace. This is an area we will seek to investigate more in our future work.

User perceptions and understanding

The results presented to date have highlighted the significance of advertising networks within popular websites across five countries. In fact it is clear to see that there are a number of networks that are prevalent across the world, with a core set of companies providing a large percentage of third-party content on websites. The next stage of our research involved understanding the perceptions of Internet users as it pertains to such networks.

To facilitate our investigation, an online survey was carried out over the course of 3 months. The specific aim was to consider topics such as personal privacy, and also, to understand how aware users are of the prevalence of advertising networks and other

third-party content on popular websites. The survey received ethical approval from our local university review board, and we also ensured to gain informed consent from participants. In total we received 109 responses, with questions covering standard biographic information (age, gender, education and technical understanding), perceptions of online privacy, and finally perceptions regarding the amount of third-party content that was linked to by a selection of popular websites (drawn from the Alexa Top websites). The majority of respondents to the survey were based in the United Kingdom, however, the survey was not limited by a user's location with a range of countries being represented.

The majority of responses to the survey came from females (54%) and 48% of the respondents were under the age of 34. The younger profile of the respondents suggested our sample had grown up with the Internet and associated technologies, therefore might be more knowledgeable in this area. This was further underlined by the participants' high knowledge of computers and the Internet: over 67% of participants ranked themselves at 7 or above, with no-one ranking themselves below a 3 (this is based on self-reports using a scale of 1 to 10, where 10 was extremely knowledgeable and 1 was very little knowledge).

After completing the biographic information, we queried participants about their own perceptions of online privacy and its importance to them; this would be especially important when considering knowledge of third-party networks. Specifically, they were asked to rate the importance of online privacy, using a 5-point Likert scale. In total, 80% of respondents expressed that online privacy was either very important (52%) or extremely important to them (28%). A Pearson's Chi-Squared test was used to determine any correlation between the importance of online privacy to a participant and their perceived level of technical knowledge. This test resulted in a p-value of 0.85, which indicates that there was no correlation between perceived technical knowledge and the importance of online privacy. This independence could be interpreted as online privacy being an important concept regardless of the level of technical knowledge and understanding of the individual, which in itself is a positive finding.

To assess participant's perceptions and understanding of online advertisements, we asked them to indicate the category of sites they regularly visit and which of these categories they thought contained the most advertisements. Figure 8 shows the comparison between the types of sites that are commonly used and those which participants thought contained the most advertisements. The results show that the participants were

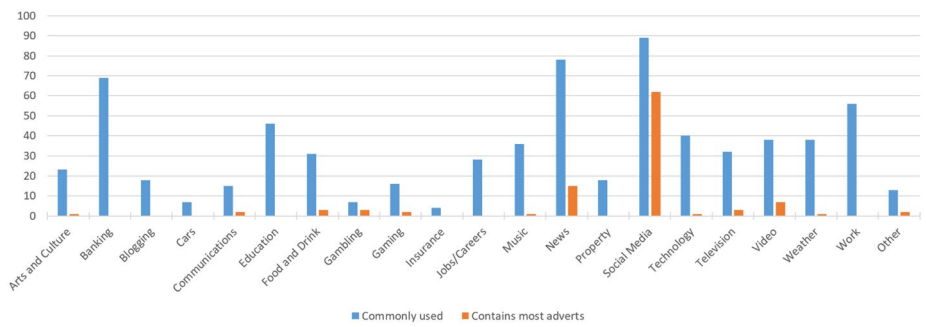

Fig. 8 A graph showing a comparison between the sites that were used most commonly by survey participants and the sites that they thought to contain the most advertising

of the opinion that news (15%) and social media (60%) sites were likely to contain the most advertisements. Interestingly, although that was felt to be the case, these were still the top two types of sites most commonly used by participants (with news at 75% of respondents and social media used by 86% of respondents). While there are several other variables that could be in play here, this could highlight the value of site utility over concerns of ads or online tracking.

We were also interested in understanding how often participants tended to access third-party links and advertisements online. The majority of respondents said that they 'never' (38%) or 'rarely' (53%) clicked on advertisements with only 2% of respondents claiming to 'always' click on such links. This result goes some way to explaining the responses to the previous question, in that users might be aware that certain sites contain the most advertisements (e.g. news or social media sites) but they simply choose to ignore these ads, for the most part. In further support of this point, we found that over half of the respondents (52%) claimed to use advertisement-blocking browser extensions 'frequently' (26%) or 'always' (26%). This potentially could account for why users were content visiting sites with advertisements and also, why they rarely clicked on them.

When comparing the importance of online privacy and the use of advertisement-blocking browser extensions, a Pearson's Chi-Squared test resulted in a p-value of 0.69. This suggests that there is little dependence between those people who use advertisement-blocking extensions and the importance of online privacy. This result is understandable because it may be argued that advertisement-blocking browser extensions are not necessarily a privacy preserving tool, with the main focus being to hide unwanted advertisements. For example, Adblock [19] is listed as the 'most popular Chrome extension with over 40 million downloads'. However, this extension is marketed as a means to block advertisements with no mention of preserving privacy. When comparing gender to the use of ad-blocking extensions it was found that there was a correlation (with a Pearson's Chi-Squared test resulting in a p-value of 0.01), meaning that there is a correlation between gender and the use of advertisement-blocking software, with males generally using ad-blockers more frequently.

Our next aim was to understand how much participants were aware of some components of the underlying advertising ecosystem. For this, participants were asked to identify where they thought that the majority of advertisements on web pages originated from. A large proportion of respondents (90%) correctly identified that advertisements typically originate from a third-party, as opposed to the company that owns the site (as demonstrated in our earlier website analysis section). We also found that over half of participants were either 'moderately aware' (42%) or 'extremely aware' (21%) that when a website was visited several third-party sites were also automatically referenced. While this was good to see as it demonstrated awareness, approximately a third of participants (37%) were in the lower bracket of 'not at all aware', 'slightly aware', and 'somewhat aware'. This is some cause for concern as it highlights that the activities that occur when users visit websites may not be transparent for all users involved.

In addition to exploring participants' awareness, we asked them to estimate the number of third-party sites that particular websites might make reference to. This would allow us to assess how well participants' perceptions match reality (as presented in earlier sections). We used a subset of the main-site (or home) pages from the Alexa Top

250 such that they would cover a range of different categories and would be generally recognisable by individuals. While there were a number of UK-based sites included they were those that have a global reach. As previously discussed, while the majority of participants were indeed based in the UK the sites were specifically chosen to be recognisable by a wider, global audience.

Table 2 contains a summary of the survey responses and for each site, it includes the minimum number of third-party sites stated, the maximum number of third-party sites stated, and the mean, median and mode number of sites stated across the entire set of responses. It should be noted here that we only asked individuals about the number of third-party website references on the main page of the site (e.g., `google.com`), and did not focus on sub-pages (e.g., `google.com/news`).

A point that is immediately obvious from the table is the extremely large estimates that participants expressed about third-party links. Overall, we found that in 38% of responses users estimated site references of above 500 third-party links (far above the actual cases as mentioned in earlier). This highlights a sizeable disparity between perceptions and reality (as will be discussed below). To take Google as an example, it was the site that participants felt would contain the most references to third-party sites. One factor in this ranking could be that Google is a company built on data. Therefore, as much of Google's business relates to collecting and analysing information, it is reasonable that participants would rank Google highly. In fact, from our earlier findings, we know that Google is one of the largest companies behind advertising networks.

The comparison between the number of third-party links on a website (the 'actual value' column in Table 2) and participants' perceptions of these links, is an interesting one. As can be seen, the estimates given by respondents are a long way from the values that were actually recorded in the course of our research. The mean values presented can be directly attributed to a number of outlying, and extremely high values for each of the websites as mentioned earlier. These values have acted to skew the mean and exemplify why the mean would not an appropriate centroid for comparison in such distributions. In Amazon's case for instance, there is a mean value of 71,974, as values ranged from 1 to 5,000,000. In reality however, there were only 3 third-party references made on the

Table 2 The number of third-party websites that participants estimate specific sites link to, and the actual number of links made by those sites

Website	Minimum	Maximum	Mean	Median (total %)	Modal group (total %)	Actual value
Amazon	1	5,000,000	71,974	50 (7%)	0–25 (45%)	3
Argos	0	5,000,000	58,611	15 (6%)	0–25 (63%)	29
BBC	0	500,000	5,884	10 (8%)	0–25 (67%)	5
Daily Mail	1	5,000,000	62,381	50 (9%)	0–25 (43%)	45
Expedia	0	2,000,000	41,788	30 (8%)	0–25 (42%)	59
Facebook	1	2,000,000	44,625	100 (15%)	0–25 (34%)	107
Google	2	10,000,000	306,946	100 (15%)	0–25 (30%)	3
Huffington Post	0	2,000,000	34,737	50 (10%)	0–25 (42%)	46
Independent	0	1,000,000	13,945	30 (8%)	0–25 (45%)	55
Just Eat	0	1,000,000	26,520	50 (10%)	0–25 (44%)	24
Spotify	0	10,000,000	128,892	30 (3%)	0–25 (48%)	22
TalkTalk	0	100,000	2630	40 (1%)	0–25 (48%)	24

site. Comparing these would be unhelpful, and would allow no insight into the difference between perceptions and reality.

To address the issue of the extreme values entered by participants and the resulting skewed mean, we also examined the median (this is a common use of the median [37]) and mode values for each site. In Table 2 we present the median values and mode values (after value grouping), along with the total number of times (and respective percentage that) those values occur. While the median identifies the central value in the list of values supplied by participants, the modal groups attempt to categorise the values (into 25 value groups, e.g., 1–25 links, 26–50 links, and so on) and give an indication of which value-groups appear the most in the sample.

Comparing the medians to the actual observed values, some were close (e.g., Facebook, Huffington Post and BBC), and overall, they were certainly within the correct order of magnitude—Google and Amazon being the main outliers. An interesting point to note here is that through the median, participants were largely correct at identifying the site with the highest number of links to third-party sites i.e., Facebook. Facebook generates a large percentage of their income through data and through advertising revenues, as alluded to earlier in our article. The fact that the survey participants were able to correctly identify this site as the one with the most third-party content demonstrates some understanding of which organisations are more likely to rely on third-party material. Conversely, Google was identified as a site that users would consider to reference numerous third-party websites. In reality, Google (at least google.com) ranks as one of the lowest sites in terms of links to third-party content.

From assessing the mode groups, we could see that the most common values supplied by participants were actually within the 0–25 range for all sites. This inclination was somewhat accurate for a few cases, with the BBC, Spotify and TalkTalk, but not for others, such as Facebook and Expedia. The total percentage (shown in Table 2) is useful here as it highlights the proportion of participants with a value within the modal range. From this we can see that for the BBC for instance, most participants (67%) were relatively close to the correct value. However, for Google this was not the case as less than 1/3 of individuals were within the modal range. To investigate this further, we examined the second most common value ranges and found that Google and Facebook were both in the range of 76–100 with 16 and 16% of participants respectively. This acts to support our other findings above, and emphasise that while participants do have some understanding of the 'behind the scenes' activities, there are still misconceptions about sites and how much third-party referencing is conducted.

Conclusions and future work

In this paper we have investigated the various activities that occur when an individual visits a webpage, specifically focusing on third-party references and their proliferation across the web. These references often automatically execute in the background of a website without the express knowledge of a user. This paper has explored the nature of these references and the degree to which a website will link to third-party content, as its information set for study. Further to this, a survey of user perceptions has provided insight into how aware users actually are about this third-party content and its potential privacy implications.

Our analysis of the top 250 websites in the UK, USA, Germany, Russia and Japan has shown, perhaps unsurprisingly, that the USA has the largest and densest advertising network. A more interesting observation was that Japan had the smallest advertising network even though it has a substantial market size of 126 million and a high Internet adoption rate of 91% [31]. When considering the different parties involved in advertising networks, our research found Google was the largest player thanks to DoubleClick, a Google owned company that specialises in developing and providing advertising services and was prevalent across all five of the countries studied. Similarly, Google Analytics, a web platform designed to offer website analytics was commonly seen across a wide-range of sites from all five countries.

To reflect on the companies engaged in advertising networks, it was not just Google who had a significant interest. Facebook are another company who had a notable presence across all of the countries that were included within our research. That is not to say that advertising networks are solely controlled by a few large organisations. There were a reasonable number of smaller organisations within the space, for example, ScoreCard and ADNXS. Our findings indicated that some of these smaller sites, such as ADNXS, could be thought of as bridging sites. That is to say, these sites were identified as part of the ecosystem that act as a bridge to other sites, with the original site perhaps not being heavily referenced.

The survey conducted on user perceptions of privacy on the web provided some interesting results and insights. One of the most notable findings was that comparing the median values to the observed values there were some very close results between the user estimates and the recorded values. The fact that participants were correctly able to identify Facebook as one of the sites with the most third-party link shows that there is some level of understanding of what types of organisations rely on third-party content. When assessing the modal groups of participant estimates, it became apparent that while there was some understanding of third-party references and the 'behind the scenes' website activity, there are still misconceptions about the true prevalence of third-party sites.

There are a number of avenues through which we intend to engage in future work. The first of these is to broaden our study to encompass more countries, and particularly those with different cultures or those considered as developing states. This would strengthen our cross-country research and enable us to better compare and contrast a variety of country types. China for instance, would be interesting to study given the state's control over the Internet, and India would be ideal to examine as the Internet is set to reach larger parts of the population.

With a better appreciation of the prevalence of third-party website references across various nations, we would then seek to thoroughly explore the topic of online privacy in the wider context. As a society we are becoming increasingly reliant on the Internet and as such privacy should be an increasingly important consideration. The research discussed in this paper highlights the extent to which third-party organisations are being referenced often without the express permission of the user. Another important question that should be raised going forward is the extent to which user data is shared with such third-parties, especially if users are not made aware of the implications of this data being shared.

The issue of online privacy is only going to become more pressing as our reliance on the digital world increases, and as such there are a number of questions about the extent to which gathered user data (and metadata) is passed between these third-party sites, and just how this data is used. The work in this paper provides an initial investigation into the extent that this data is gathered and shared without the explicit knowledge or consent of the user.

In addition to this, we are keen to undertake a deeper user study by presenting the full third-party networks to the users—ideally in an interactive format-to help better understand their impressions of them, especially as it relates to privacy. While we note that many advertising companies offer 'Opt out' options on their websites, it would be useful to see if there was some effective way at creating a tool that would: firstly, identify sites where this option is possible; and secondly, streamline the process from the stage of visualising a third-party referencing network to selecting and opting out from sites.

Authors' contributions
All parties assisted significantly in the research work and the manuscript drafting stage. Both authors read and approved the final manuscript.

Author details
[1] Department of Computer Science, University of Oxford, Oxford, United Kingdom. [2] Centre for Electronic Warfare, Information & Cyber, Cranfield University, Defence Academy of the United Kingdom, Shrivenham SN6 8LA, United Kingdom.

Competing interests
The authors declare that they have no competing interests.

Funding
There are no funding bodies to acknowledge for this research.

Publisher's Note

Received: 17 March 2017 Accepted: 24 October 2017
Published online: 02 November 2017

References

1. InternetLiveStats (2017) Total number of Websites. http://www.internetlivestats.com/total-number-of-websites/. Accessed 18 Sept 2017
2. Lin CJ, Hsieh T-L (2016) Exploring the design criteria of website interfaces for gender. Int J Ind Ergonom 53:306–311
3. George JF, Mirsadikov A, Mennecke BE (2016) Website credibility assessment: an empirical-investigation of prominence-interpretation theory. AIS Trans Hum Comput Interact 8(2):40–57
4. Wang Q, Yang S, Liu M, Cao Z, Ma Q (2014) An eye-tracking study of website complexity from cognitive load perspective. Decis Supp Syst 62:1–10
5. Falahrastegar M, Haddadi H, Uhlig S, Mortier R (2016) Tracking personal identifiers across the web. In: Proceedings of the international conference on passive and active network measurement, Springer, Berlin, pp 30–41
6. Li T-C, Hang H, Faloutsos M, Efstathopoulos P (2015) Trackadvisor: taking back browsing privacy from third-party trackers. In: Proceedings of the international conference on passive and active network measurement, Springer, Berlin, pp 277–289
7. Wilson S, Schaub F, Ramanath R, Sadeh N, Liu F, Smith NA, Liu F (2016) Crowdsourcing annotations for websites' privacy policies: Can it really work? In: Proceedings of the 25th international conference on world wide web, International world wide web conferences steering committee, pp 133–143
8. Metwalley H, Traverso S, Mellia M, Miskovic S, Baldi M (2015) The online tracking horde: a view from passive measurements. In: Proceedings of the international workshop on traffic monitoring and analysis, Springer, Berlin, pp 111–125
9. Wu Q, Liu Q, Zhang Y, Liu P, Wen G (2016) A machine learning approach for detecting third-party trackers on the web. In: Proceedings of the European symposium on research in computer security, Springer, Berlin, pp 238–258
10. IAB (2015) Digital Ad Revenues Surge 192015. http://www.iab.com/news/digital-ad-revenues-surge-19-climbing-to-27-5-billion-in-first-half-of-2015-according-to-iab-internet-advertising-revenue-report/. Accessed 4 Sept 2017
11. Bloomberg Technology (2016) Google and Facebook Lead Digital Ad Industry to Revenue Record. http://www.bloomberg.com/news/articles/2016-04-22/google-and-facebook-lead-digital-ad-industry-to-revenue-record. Accessed 4 Sept 2017

12. European Commision (2016) Cookies. http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm. Accessed 4 Sept 2017
13. Upathilake R, Li Y, Matrawy A (2015) A classification of web browser fingerprinting techniques. In: Proceedings of the 7th international conference on new technologies, mobility and security (NTMS), IEEE, New Jersey, pp 1–5
14. Laperdrix P, Rudametkin W, Baudry B (2016) Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. In: IEEE symposium on security and privacy (SP), IEEE, New Jersey, pp 878–894
15. Borko H (1968) Information science: what is it? J Assoc Inform Sci Technol 19(1):3–5
16. Hazan E, Banfi F (2013) Leveraging big data to optimize digital marketing. http://www.mckinseyonmarketingandsales.com/leveraging-big-data-to-optimize-digital-marketing. Accessed 3 Sept 2017
17. Trusov M, Ma L, Jamal Z (2016) Crumbs of the cookie: user profiling in customer-base analysis and behavioral targeting. Mark Sci 35(3):405–426
18. Williams M, Axon L, Nurse JRC, Creese S (2016) Future scenarios and challenges for security and privacy. In: Proceedings of the IEEE 2nd international forum on research and technologies for society and industry (RTSI), IEEE, New Jersy, pp 1–6. https://doi.org/10.1109/RTSI.2016.7740625
19. Adblock (2016) Chrome Web Store: Adblock. https://chrome.google.com/webstore/detail/adblock/gighmmpiobklfepjocnamgkkbiglidom. Accessed 1 Sept 2017
20. AOL (UK) (2017) 37,000 Chrome users downloaded a fake Adblock Plus extension. https://www.engadget.com/2017/10/09/fake-adblock-plus-chrome-extension/. Accessed 11 Oct 2017
21. Nikiforakis N, Joosen W, Livshits B (2015) Privaricator: deceiving fingerprinters with little white lies. In: Proceedings of the 24th international conference on world wide web, ACM, New York, pp 820–830
22. Melicher W, Sharif M, Tan J, Bauer L, Christodorescu M, Leon PG (2016) (Do Not) Track me sometimes: users' contextual preferences for web tracking. Proc Priv Enhanc Technol 2016(2):135–154
23. Leon PG, Ur B, Wang Y, Sleeper M, Balebako R, Shay R, Bauer L, Christodorescu M, Cranor LF (2013) What matters to users?: factors that affect users' willingness to share information with online advertisers. In: Proceedings of the 9th symposium on usable privacy and security, ACM, New York, pp 7
24. Agarwal L, Shrivastava N, Jaiswal S, Panjwani S (2013) Do not embarrass: re-examining user concerns for online tracking and advertising. In: Proceedings of the 9th symposium on usable privacy and security, ACM, New York, pp 8
25. Norberg PA, Horne DR, Horne DA (2007) The privacy paradox: personal information disclosure intentions versus behaviors. J Consum Affairs 41(1):100–126
26. TRUSTe (2016) GB Consumer Privacy Index 2016. https://www.truste.com/resources/privacy-research/ncsa-consumer-privacy-index-gb/. Accessed 4 Sept 2017
27. Alexa Internet, Inc. (2016) Top websites. http://www.alexa.com/topsites/countries. Accessed 8 Sept 2017
28. Wu Q, Liu Q, Zhang Y, Wen G (2015) Trackerdetector: a system to detect third-party trackers through machine learning. Comput Netw 91:164–173. https://doi.org/10.1016/j.comnet.2015.08.012
29. Mozilla (2016) Lightbeam for Firefox. https://addons.mozilla.org/en-gb/firefox/addon/lightbeam/. Accessed 14 Jan 2017
30. Nurse JRC, Pumphrey J, Gibson-Robinson T, Goldsmith M, Creese S (2014) Inferring social relationships from technology-level device connections. In: Proceedings of the 12th annual international conference on privacy, security and trust (PST), IEEE, New Jersy, pp 40–47. https://doi.org/10.1109/PST.2014.6890922
31. Internet Live Stats (2017) Internet users by Country. http://www.internetlivestats.com/internet-users-by-country/. Accessed 4 Sept 2017
32. IHS Markit (2016) TV continues to Lead Japan's advertising market, but online advertising is coming on strong, IHS says. http://news.ihsmarkit.com/press-release/technology/tv-continues-lead-japans-advertising-market-online-advertising-coming-stron. Accessed 14 Sept 2017
33. Business Insider (2015) Microsoft is handing off yet more of its advertising sales business to ad tech company AppNexus. http://uk.businessinsider.com/microsoft-expands-appnexus-partnership-2015-12. Accessed 14 Sept 2017
34. TechCrunch (2011) Adobe buys behavioral data management platform DemDex. https://techcrunch.com/2011/01/18/adobe-buys-behavioral-data-management-platform-demdex/. Accessed 6 Sept 2017
35. Yahoo Ltd. (2017) Yahoo! Advertising. https://advertising.yahoo.com/. Accessed 3 Sept 2017
36. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):P10008
37. Rees DG (2000) Essential statistics. Chapman & Hall texts in statistical science series, 4th edn. Taylor & Francis, Florida