ARTICLE

Look Who's Tracking – An analysis of the 500 websites most-visited by Finnish web users

John Bailey
https://orcid.org/0000-0003-3040-7779

Mikael Laakso
Hanken School of Economics
https://orcid.org/0000-0003-3951-7990

Linus Nyman
Disobey Outreach
linus@disobey.fi
https://orcid.org/0000-0001-5051-1683

Though research into online tracking prevalence as a topic is not new, we still know little about who is tracking and profiling Finnish web users. This study examines tracking on the 500 websites most frequently visited by Finnish users. We also compare trackers on Finnish websites versus non-Finnish websites. We found trackers on 410 of the 500 websites, and a total of 466 unique trackers from 408 different organizations. Similar to most previous studies, Google had the greatest tracker coverage, mostly through Google Analytics and Doubleclick, reaching 75 % of the websites analyzed. The second-most prevalent tracking organization was Facebook, present on 46 % of the websites. After Google and Facebook came a number of organizations with fairly similar tracking coverage, followed by a long tail of others. There were notable differences when comparing Finnish websites to non-Finnish sites, displaying some level of geographical preference in publishers' choices of advertising platforms and analytical tools.

Keywords: cookies, tracking, remote monitoring, web pages, privacy

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Permanent identifier: https://doi.org/10.23978/inf.87841

Introduction

When the commercial Internet was taking its first steps, Steiner (1993) accidentally created an adage in his New Yorker comic strip "On the Internet, nobody knows you're a dog". What was apparently (Cavna 2013) the result of a creative doodle could now be seen as an accurate contemporary view on the new technology that was the Internet. However, over the decades that followed, tracking became more prevalent and complex (e.g. Lerner et al., 2016) and online anonymity became increasingly rare. A decade after Steiner's dog, Solove (2004) noted that the Internet's greater targeting potential and the fierce competition for the consumer's attention had given companies an unquenchable thirst for information about web users. A further decade later, data security expert Bruce Schneier (2015) described the Internet as a kind of surveillance capitalism, a concept also explored in Zuboff (2019), with advertising being the primary goal of corporate Internet surveillance. The Center for Democracy & Technology (2011:3) defines tracking as "[T]he collection and correlation of data about the Internet activities of a particular user, computer, or device, over time and across non-commonly branded websites, for any purpose other than fraud prevention or compliance with law enforcement requests." Given Steiner's adage, with all the data, machine power, and algorithms available today, "they" would most certainly know you are a dog – possibly even before you knew it yourself.

There are many reasons for tracking users. In addition to corporate interests in things like documenting our interests and surfing habits, there are also technical motivations and motivations related to useability. A simple example would be an online shop. Your computer communicates with the online shop using something called Hypertext Transfer Protocol, or HTTP. HTTP is a so-called stateless protocol. In practice, this means that the store will not remember you from previous visits to the website, or even remember you from one page to the next as you visit different pages on its website. This means that even something as seemingly simple and commonplace as adding items to a shopping cart and then buying those items would be impossible: when you clicked your way to the checkout, the website would not have any memory of you or any items you had placed in your shopping cart. From the perspective of companies, tracking users online enables things like customized advertising, tracking users across websites, and gathering data on the popularity of various kinds of content.

A 2017 literature review by Bujlow, Carela-Español, Sole-Pareta & Barlet-Ros (2017) identified 26 different tracking mechanisms. However, given that their list only covered mechanisms previously identified by academia,

the complete list of tracking mechanisms is likely to be greater still (Sanchez-Rola & Santos 2018). Studies (e.g. by Bujlow et al. 2017; Li, Hang, Faloutsos & Efstathopoulos 2015) have shown that cookies are the most common way of tracking users. A cookie is a small text file the browser receives from the requested webpage that is saved on the user's computer. An initial implementation of cookie functionality was developed for the Netscape Navigator browser in 1994 by Lou Montulli to address the issue mentioned earlier of providing stateful browsing (Kristol 2001). The protocol was then made into an IETF protocol standard in 1997 (Kristol & Montulli 1997). The IETF develops and promotes voluntary Internet standards, and the inclusion of the cookie among IETF standards aided the spread of the cookie. The cookie made it possible for the user to leave the website and then return later, without (e.g.) losing the contents of their shopping cart (Montulli US Patent 5,826,242 1998). This was a big improvement over using forgetful URL (or Uniform Resource Locator, the text in the address bar of the browser) methods, i.e. saving shopping cart information in the URL. Li et al. (2015:4) present three arguments for the prevalence and popularity of the cookie: "*Firstly, all browsers can accept and send cookies. Secondly, other non-HTTP cookies exist and can be used for tracking, but they are inefficient or will create legal issues for the entities who utilize them. Finally, even though third-party websites can track a user by their browser fingerprint (Eckersley 2010), this method incurs a much higher overhead, thus is unlikely to adopted widely*".

There are many different kinds of cookies. For the purposes of this study we differentiate mainly between first and third party cookies. First party cookies are cookies which belong to the website on which they are located. Third party cookies belong to a company or site other than the site that places the cookie. By way of example: if you visit an online shop, cookies belonging to that shop would be considered first party cookies, while cookies belonging to someone else – say, a cookie related to a social media "like" button on the product pages of that online shop – would be third party cookies. The first party cookies could, for instance, be used by the online shop to remember you and the contents of your shopping cart, while third party cookies could, for instance, be used by an external provider to track your movements on that site. Websites can sell the right to place cookies on their site, commonly done in conjunction with selling ad space. Cookies enable ad brokers, like Google AdSense, to display personalized ads based on data collected through (e.g.) the web browsing conducted on pages that have signed up for the service.

One core functionality of cookies, domain matching, was originally implemented in 1994. Domain matching requires the cookie and the website to share domains, i.e. example.com can read and write to a cookie set by example.com,

but not one set by example2.com. This capability was developed as a privacy enabler, to make sure that users were not trackable between different sites (Shah & Kesan 2009). While domain matching worked as planned, history has shown it was not enough to protect users from cross-domain tracking. If a website contained any third party elements, such as pictures, font libraries, or other content components, the third party could also write and read their own cookie. If the third party's elements were used on other websites as well, this third party could also track users over these websites. This technical design, unforeseen by Netscape, eventually enabled the dawn of third party tracking and advertising networks, through companies like DoubleClick. Even though the IETF identified the risks as early as in December 1995, the fact that Netscape had already launched support for cookies, and websites had taken that specification into use, meant that the formal technical standardisation process of cookie-use, had to largely accept the privacy concerns since Netscape had a strong first-mover advantage in defining the fundamental functionality (Shah & Kesan 2009).

Worth highlighting is the fact that even when browsing ad-free, non-commercial pages one is often tracked by third party providers. One common example is the Google Analytics tracker cookie, a cookie which has also been implemented on the Informaatiotutkimus website (among many others) for many years (`https://journal.fi/inf`, see also Eriksson-Backa 2013). With such a cookie, the website administrators get the benefit of easily gathering and analysing visitor statistics, and the cookie-owner (in this case Google) gets the benefit of also gaining access to that data. However, this kind of integration of third-party services can become regulatory risks for websites, depending on the content and context the web user is in. In 2015, a Finnish bank implemented Google Analytics in their online bank service. Räisänen (2015) used demo credentials to gain access and analyze the data being shared. One of her findings was that the demo user's account number was sent as a parameter to Google using only a simple hash to protect it. She wrote a 25-line program that could brute-force the real account number in 0.5 seconds. Overall, she noted that the information shared with Google was in breach of the Finnish bank secrecy guidelines (S-Pankki 2015).

Aim of the study

The aim of the study is to gain a better understanding of tracking on sites that are popular among Finnish web users. This is done through an exploration of the four following questions:

1. How common is tracking? On how many websites are there trackers? How many trackers per page?
2. What kinds of trackers can be identified? What purpose do they serve?
3. Who is doing the tracking? What companies or entities are behind it?
4. How does tracking differ between Finnish sites and non-Finnish sites? Does tracking behaviour differ on Finnish versus non-Finnish websites?

This paper is structured as follows. The next section covers previous research. We then discuss and describe our research methods. This is followed by a presentation of the data gathered and our analysis. We conclude with a discussion and some suggestions for future research.

Previous research

During the past decade there have been several studies related to web tracking. The topic has been approached using many different methodologies and research aims, with some key streams of research being: studies connecting trackers to organizations, studies focusing primarily on tracking growth over time, studies focusing primarily on tracking prevalence and tracker ownership at single points in time, and studies focusing on specific geographic regions of the world. In order to frame the study at hand, this section will first review the main findings from these lines of research.

One initial challenge with research on web tracking has been that, based on information that can be collected from a webpage itself, it is not always immediately apparent what organization owns each tracker. One stream of research has sought to connect trackers to organizations. Libert (2015) performed manual detective work in order to link trackers and their owners. For further mapping of domains to organizations, Englehardt & Narayanan (2016) combined the manual work by Libert (2015) and a list of known web trackers made available by Disconnect, a tracking protection application with open source components. Falahrastegar et al. (2014B) used similar reference

data available from the Firefox add-on Collusion. Thanks to such initial work, there are now publicly available lists documenting the connections between trackers and their owners.

Another area of research focus has been tracker growth over time. One of the most extensive longitudinal investigations of online tracking was presented by Lerner, Simpson, Kohno & Roesner (2016). Using the Internet Archive's Wayback Machine, they analyzed the top 500 yearly sites from 1996 to 2016 and managed to document a historical context for the growth of online tracking, both for the reach of different trackers and for the number of separate trackers per website. They note that the use of Wayback Machine as the source of data underestimated the integrity of website snapshots, as some scripts and resource calls do not function properly. Even with this handicap, their results clearly indicate both a growth of third party requests over time, as well as an ever-increasing number of third party requests per site. Krishnamurthy & Wills (2009) found that the prevalence of the 10 most prolific third party trackers grew from 40 % to 70 % between October 2005 and September 2008. A key feature in this growing coverage of the top trackers was acquisitions, resulting in a reduced number of independent trackers and an increased dominance of the top five companies: Google, Omniture, Microsoft, Yahoo, and AOL. This study highlights the connection between trackers, the organizations behind them, and the motivations for consolidation in the landscape as central actors can easily extend their reach in tracking by merging or acquiring other actors.

Other studies have focused on tracking prevalence and tracker ownership at single points in time. In an analysis of the 500 most popular websites on Alexa, Roesner et al. (2012) identified over 500 unique third party trackers. They also found that the most prevalent cross-site tracker was the Google-owned Doubleclick advertising platform, which could record user visits from almost 40% of the top 500 pages. Li et al. (2015) analyzed the Alexa top 10k sites using machine learning. They only found Google (.com and Doubleclick) on 25% of the sites. Compared to other sources, and even to earlier sources such as Roesner et al. (2012), this number seems low. However, Li et al. (2015) point out that their numbers did not take into consideration Google Analytics, because: "*[…] by contract, Google Analytics provides statistics only to the 1st party websites and the cookies set by Google Analytics are always associated with the domains of the 1st party websites and therefore are not 3rd party cookies. Furthermore, the same user who visits different websites monitored by Google Analytics will likely receive different IDs, which makes tracking him or her non-trivial.*" (Li et al. 2015:9). The current version of the Google privacy policy states the following: "*…when you visit a website that uses advertising services like AdSense, including analytics tools like Google Analytics, or*

embeds video content from YouTube, your web browser automatically sends certain information to Google. This includes the URL of the page you're visiting and your IP address. We may also set cookies on your browser or read cookies that are already there. Apps that use Google advertising services also share information with Google, such as the name of the app and a unique identifier for advertising." (Google 2019). This leaves it unclear to what degree Google themselves integrate and leverage tracking data that they collect.

Englehardt & Narayanan (2016) studied a sample of 1 million top sites provided by Alexa using OpenWPM, a web privacy measurement tool they created. The authors found a long tail of over 81,000 third party trackers that were present on at least two websites, out of which only 123 were found on more than 1 % of the sites. They note that by including subpages of the 1 million websites, instead of only the home pages, the average number of trackers would have increased from 22 to 34, indicating that their results were lower than what a real user would experience. Google took up all the spots of the top 5 for most prevalent third party trackers, and 12 spots of the top 20. Google also dominated the top organizations behind third party resources, followed by Facebook, Twitter, Amazon, and Adnexus.

Karaj, Macbeth, Benson & Pujol (2018) analyzed a dataset of over 780 million page-loads from a time period of 10 months, using data provided by over 500 000 users of the Cliqz and Ghostery browser extensions. The main advantage of their approach is the use of actual user browsing data, compared to the use of automated scripts that try to mimic browsing behavior that almost all other studies utilize. The downside, as Karaj et al. (2018) noted, is the loss of data granularity enforced by privacy constraints on the user-generated browser data. The authors found that a website loads about 10 trackers on average, and 89% of the traffic to the top 600 websites contains tracking (Karaj et al. 2018:9). Regarding owners of third party scripts, Google was present in about 80 % of the web traffic, with Facebook and Amazon being the second and third most prevalent operators. The authors noted that as the data was from a mostly German userbase, some German services, such as InfOnline, ranked in the top 10.

Research with a regional focus

Previous research with a regional focus includes Falahrastegar, Haddadi, Uhlig & Mortier, who published two separate papers in 2014 analyzing the tracking ecosystems from a regional perspective. In the first study, Falahrastegar et al. (2014A) looked at the top 500 sites of the USA, UK, Australia, China, Egypt,

Iran, and Syria. The results showed clear regional differences, but with significant cultural or language-based similarities. Google's dominance was evident, and two Google properties - DoubleClick and Google Analytics - were the only ones found in the top 20 of every country researched. In their second paper, Falahrastegar et al. (2014B) examined third party tracking among the top 500 sites in 29 countries divided into geographical regions: North America, South America, Europe, East Asia, Middle East, and Oceania. The study included Sweden and Norway, but not Finland. The study found a very clear indication of strong local trackers in all regions and countries. Overall, they found a strong presence of trackers from the USA, Russia, and Germany in most countries. In the Nordic countries, Sweden and Norway had a clearly similar tracker ecosystem. As for the actual trackers, Google-owned properties were by far the most prevalent in all regions, with Facebook, Amazon, Yahoo, and Twitter sharing the next spots. One clear exception to this rule was East-Asia, where Baidu and Sina were the top contenders, with Facebook and Twitter completely outside the top 20.

Fruchter, Miao, Stevenson & Balebako (2015) examined if and how different privacy regulations affect the amount of web tracking in four countries: the US, Japan, Germany and Australia. They were unable to identify any clear relationship between privacy regulations and tracking from their data, as for example the US has much more tracker activity than Japan, even though they have similar regulation. Their paper suggested cultural or societal reasons but stated that more research is needed to verify this.

Purra & Carlsson's (2016) research on tracking and HTTPS showed that the top 10,000 sites in Sweden and Denmark have a similar tracker hierarchy as the global top 10,000 sites, dominated by Google and followed by Facebook and Twitter. Facebook and Twitter were more prominent in the News media category of sites, but much less so in other categories.

A study by Ruohonen & Leppänen (2017) is currently the only academic research paper investigating tracking prevalence from a Finnish perspective. They measured the number of third party cookies on the top 206 Finnish sites, as identified by the TNS Metrix media measurement service, which consists of (primarily media) sites that have placed a TNS Metrix tracker on their website. Rubiconproject.com topped the list of third party cookies, with the Google-owned doubleclick.net coming in at only fourth place. This, as the authors themselves note, does not follow previous global results. One explanation for this difference may be that while studies on third party cookie prevalence commonly measure the share of sites that the third party trackers are set on, Ruohonen et al. instead measured the number of cookies that were set.

Research methods

This paper uses a research design similar to that of other research in the area (e.g. Roesner et al. 2012; Falahrastegar et al. 2014A; Falahrastegar et al. 2014B; Liber, 2015; Englhardt & Narayanan 2016; Karaj et al. 2018). The research approach involved four main methodological process stages:

1. Selecting a sample of websites.
2. Deciding how to gather and measure data.
3. Collect tracking data.
4. Identifying tracker ownership.

In selecting a sample set of websites to use as a proxy for Finnish web surfing, we chose to use Alexa.com, an Amazon-owned web service. Alexa is a popular tool in tracking research, and arguably the default choice (see e.g. Roesner et al. 2012; Casteluccia et al. 2013; Falahrastegar et al. 2014A, 2014B, 2016; Fruchter et al. 2015; Metwalley et al. 2015; Lerner et al. 2016; and Kyrölä 2018). Alexa provides tools and data for marketing and analytics purposes, with one key functionality for academic research being their worldwide tracking of popular websites, updated daily, called Top Sites. A list of these sites is available both globally as well as per country. An alternative source for a list of websites popular among Finnish users could have been the now discontinued TNS Metrix; however, they only provide data on websites that have added the TNS tracker code, which is mostly limited to Finnish media websites.

With the sample set decided upon, the following step was to decide on how to simulate user browsing and record our observations. To accomplish this step, we chose the Tracker Tracker measurement tool. Tracker Tracker is a web service based on the phantomJS scripting language, using the popular Ghostery browser extension for tracker detection, that is designed to identify and optionally block third party calls on a website. As noted by Englehardt & Narayanan (2016), all third party calls can be used for tracking and are therefore also counted as trackers in this study.

To collect tracker data, we ran five separate requests for five subsets of 100 sites between 19.8.2017 and 20.8.2017. The tool used a tracker database from March 24, 2017. The combined CSV files produced by the Tracker Tracker tool contained 89,001 lines of data, with 88,976 lines of site tracking data and 25 lines (5 separate runs for each incremental top 100 group) of CSV column headers. The tracking data from the measurement runs had high variation, supporting the decision to conduct multiple runs, the number of trackers identified in the top 400 sites varied between 5854 in run 3 and 1050 in run 2. The

number of sites on which Tracker Tracker found third party trackers varied considerably between runs. For example, 95 % of the top 400 sites had third party trackers identified during run 3, whereas only 83 % had tracker data in run 2. This explains some of the variation, but not its cause. Only 47 sites had third party scripts successfully identified in all five runs of which they were a part.

To find the link between trackers and their owners, we used open source resources used in a popular tracking detection application called Disconnect. We used the Disconnect tracking protection list to identify ownership of the tracker's domains and then to cross-reference the data from the measurement tool (GitHub 2017). Similar approaches have also been used in previous research (e.g. Englehardt & Narayanan, 2016).

The final dataset is provided as open data through Zenodo to improve reproducibility and future utility of the conducted study (Bailey et al. 2019).

Limitations

The main limitations of this study are set by the quality and correctness of the top 500 site list provided by Alexa, the tracker-owner relationship provided by Disconnect, and the tracker measurement capabilities provided by the Tracker Tracker tool. By using the third party tracking measurement tool Tracker Tracker, the authors had little influence on the environment from which the actual data gathering was conducted. Geographically attributable information, like the IP address of the server the browser was running on, which discloses the country the webiste traffic is coming from, is a typical example of an environmental variable that would most probably influence the results.

Results and analysis

Our results are discussed by research question. Our four research questions were: 1) How common is tracking? 2) What kinds of trackers can be identified? 3) Who is doing the tracking? 4) How does tracking differ between Finnish sites and non-Finnish sites?

Research question 1: How common is tracking?

On the top 500 sites we analyzed, we found 466 unique third party trackers from 408 organizations. After removing duplicate site-tracker combinations, we were left with 7253 site-tracker pairs for analysis.

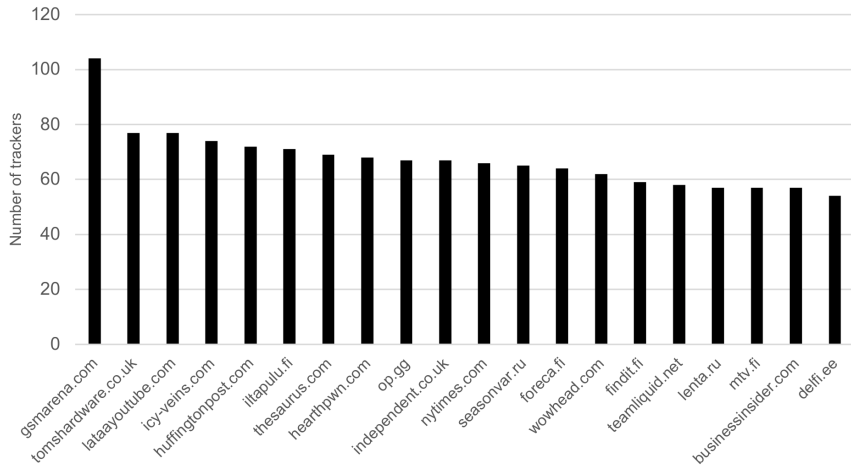

Figure 1: Top 20 pages with the most trackers

Figure 1 shows the top 20 most tracked sites, i.e. the sites with the highest number of tracker scripts. The most tracked site was gsmarena.com, where we observed 104 separate tracking scripts. Overall, news and gaming sites dominate this list, with 11 news- or weather-related sites and 5 gaming-related sites. Table 1 shows the bigger picture and the long tail of tracking. As 90 sites had no observed trackers, that leaves 245 sites with 1-9 trackers and 165 with over 10 trackers. Some sites, such as google.com or googleusercontent.com owned by Google, or t.co owned by Twitter, contain no third party trackers, as the tracking organization can utilize first party tracking instead. We found an average of 14 trackers per site, while the study by Karaj et al. (2018) found 10 trackers on average.

Number of trackers	Number of sites
>110	0
100-109	1
90-99	0
80-89	0
70-79	5
60-69	8
50-59	11
40-49	12
30-39	25
20-29	27
10-19	76
0-9	335
Total	500

Table 1: Number of third party trackers per site

Some of these 90 trackerless sites, like Wikipedia.org or governmental sites (vero.fi, suomi.fi), are credible results. Others, like the media site anna.fi, the movie rating site rottentomatoes.com, or the travel site momondo.fi, are presumably due to the limitations of Tracker Tracker. Indeed, a manual measurement using Chrome and the Ghostery extension identified third party trackers on each of these pages. As a consequence, the results presented in this paper should be seen as the minimum level of tracking imposed on Finnish web users.

Research question 2: What kinds of trackers can be identified?

As can be seen in Table 2, advertising trackers - as categorized by Ghostery (2019) - were by far the most prevalent trackers identified. Out of the 466 separate tracking scripts identified, 308 (66 %) were advertising trackers, with site_analytics and customer_interaction scripts trailing far behind. Compared to the results of Karaj et al. (2018), where the advertising category only represented 42 % of the trackers, these results are more skewed towards advertising.

Category	Number of unique trackers
advertising	308
site_analytics	79
customer_interaction	26
social_media	18
essential	16
pornvertising	10
audio_video_player	5
comments	4
Total	466

Table 2: Trackers per category for the top 500 sites

Research question 3: Who is doing the tracking?

The top 20 list of tracker prevalence is shown in Figure 2. Google dominated the list, with three out of four most prevalent trackers. Google Analytics and the Google-owned DoubleClick have a 20 % lead over the next most prevalent tracker, Facebook Connect. Apart from Google Analytics, the most prevalent tracker, the rest of the top 20 trackers were all advertising trackers. Within the site_analytics category, Google Analytics had a very dominant position with 65 % of sites utilizing the tool. The second most prevalent site_analytics tracker was TNS, with only a 15 % share. These results are similar to those of other research (Englehardt & Narayanan 2016; Macbeth 2017; Karaj et al 2018) as to the top trackers, although with higher prevalence.

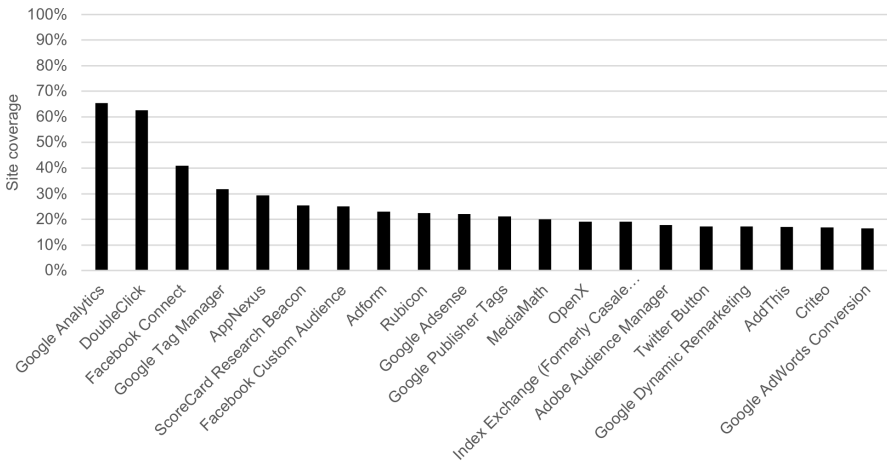

Figure 2: The most common trackers for the top 500 sites

After the top trackers, there was a sharp decline in the number of sites using each tracking script. Google Analytics and Google-owned DoubleClick were the only trackers to reach over 60% of the analyzed sites, with Facebook Connect reaching 41 % and Google Tag Manager (which focuses on tracking and analytics) reaching 32 % of the sites. No other tracker was found on over 30 % of the sites. The majority of trackers, 92 %, had a prevalence of less than 10 %, and 139 trackers (30 %) were only found on single sites. This result somewhat mirrors the long tail found by Englehardt & Narayanan (2016) – if not in numbers, then at least in shape.

Only 17 tracking organizations used more than one tracker script, the top represented by Google (19 tracker scripts), Microsoft, Facebook, and Yahoo (6 tracker scripts), and Adobe and Yandex (5 tracker scripts). The other 391 organizations were only observed to use one tracker script per organization.

Considering the site coverage results presented in Figure 2, it comes as no surprise that Google and Facebook had the widest reach. In particular Google's dominance is clear, with the ability to track Finnish users on 75 % of the top 500 pages. Facebook's reach is 46 %, while the third most prevalent tracking organization AppNexus, an online advertising platform, had a reach of 29 %. Organisations active in tracking, e.g. Google, Facebook, Twitter themselves own many sites in the top 100, which were then manually identified as a part of their reach for this analysis.

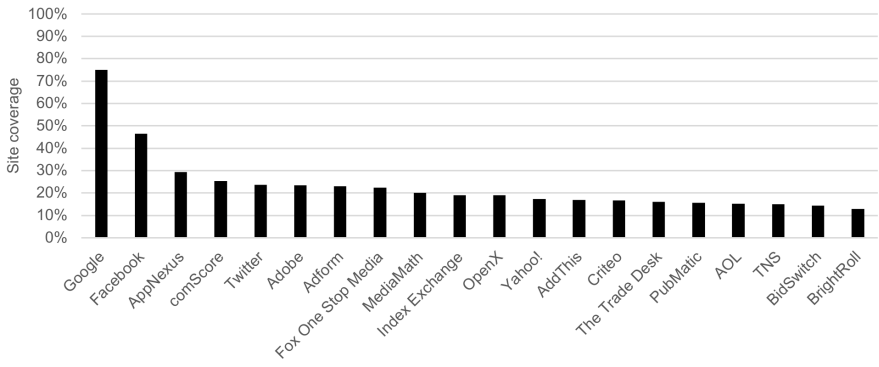

Figure 3: Reach of the top 20 tracker organizations

Research question 4: How does tracking differ between Finnish sites and non-Finnish sites?

We categorized the 500 sites into two categories: Finnish and non-Finnish. Finnish companies or Finnish sites were categorized as Finnish sites, and all other sites were categorized as non-Finnish. The 500 sites were manually labeled according to three rules classifying Finnish sites: Content is available in at least one of the three official languages (Finnish, Swedish, Sami), content originally created for the Finnish market (not only translated), and possible physical presence in Finland. Of the 500 sites, 187 were labeled Finnish and 313 were labeled non-Finnish. Of the 90 sites with no third party tracking scripts identified, 31 were Finnish and 59 were non-Finnish.

Overall, the average number of trackers per site was almost the same: 14.6 for the 187 Finnish sites and 14.4 for the 313 non-Finnish sites. This is considerably higher than the 10 trackers per site average measured by Karaj et al. (2018). The order of tracker category prevalence was the same between both groups, and the top three most used tracker categories showed little difference. Advertising was by far the largest category, with site_analytics, social_media and essential coming far behind.

Figure 4 shows the top 10 most prevalent trackers on Finnish sites. The top three trackers on Finnish sites are the same as the top three trackers overall: Google Analytics, DoubleClick, and Facebook Connect (see Figure 2). However, these three trackers were much more prevalent on Finnish sites than on non-Finnish sites. From the fourth position (Google Tag Manager) onwards, the Finnish tracker prevalence order differs from the total top 500.

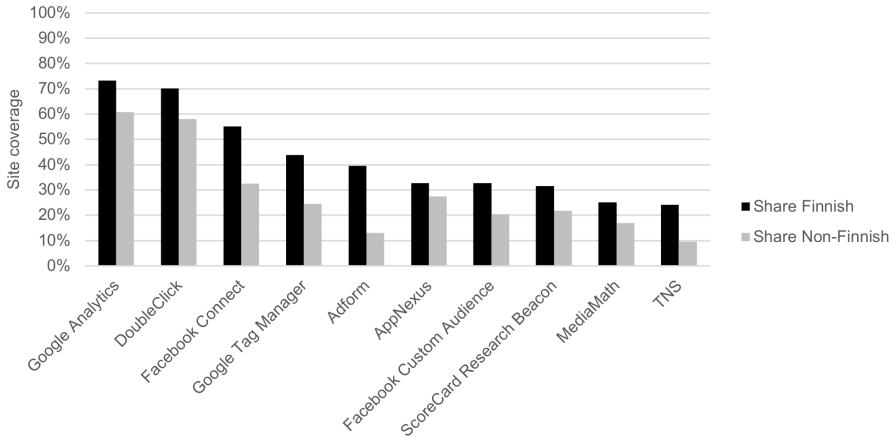

Figure 4: The top 10 most common trackers on Finnish sites

In order to illustrate the different tracker findings, Figure 5 shows the top 10 trackers that were most prevalent on Finnish sites compared to non-Finnish sites. This shows that not only did local (Nordic) trackers such as Adform, TNS, Frosmo Optimizer, and Enreach enjoy a higher reach, but also that Facebook Connect, Facebook Custom Audiences, Google Tag Manager, and Google Analytics were used more frequently on Finnish sites than on non-Finnish sites. The website tracking and analysis tools Crazy Egg and Hotjar were fourth and ninth, respectively, possibly indicating some difference in the kind of content between the Finnish and non-Finnish labeled sites, or even some preference or practices by Finnish web developers.

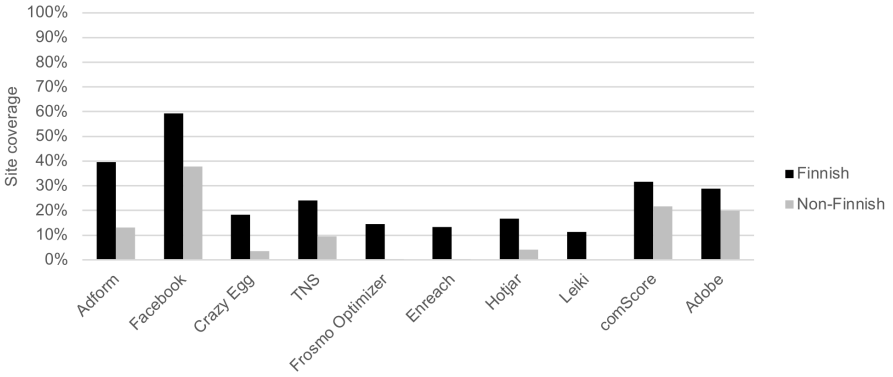

Figure 5: The top 10 trackers with the largest coverage discrepancy in favor of Finnish sites

Figure 6 presents the top 10 trackers that were most prevalent in non-Finnish sites compared to Finnish sites. The trackers represented here were clearly more prevalent on non-Finnish sites, with none of the trackers having over 10 % reach on Finnish sites and only LiveRamp having over 5 %. The Russian-focused Yandex.Metrics and the email tracking tool TopMail had no presence on Finnish pages.

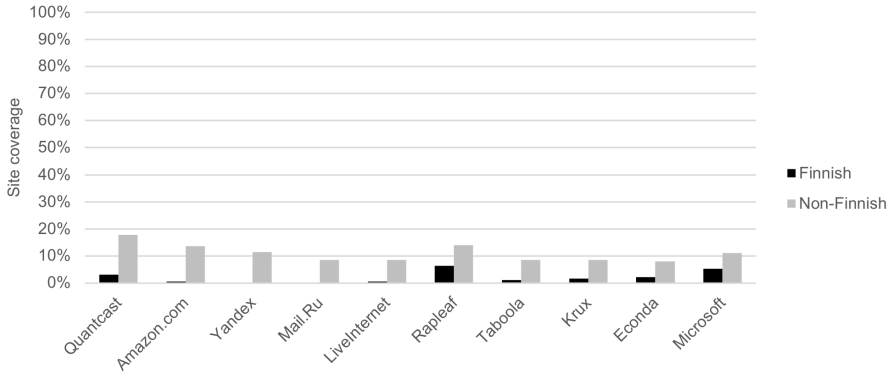

Figure 6: The top 10 trackers with the largest coverage discrepancy in favor of non-Finnish sites

Figure 7 shows the top 10 tracking organizations for Finnish sites compared with their prevalence on non-Finnish sites. As with the results from separate trackers, this figure shows a clear stronger relative presence for Facebook, Adform, and TNS on Finnish sites.

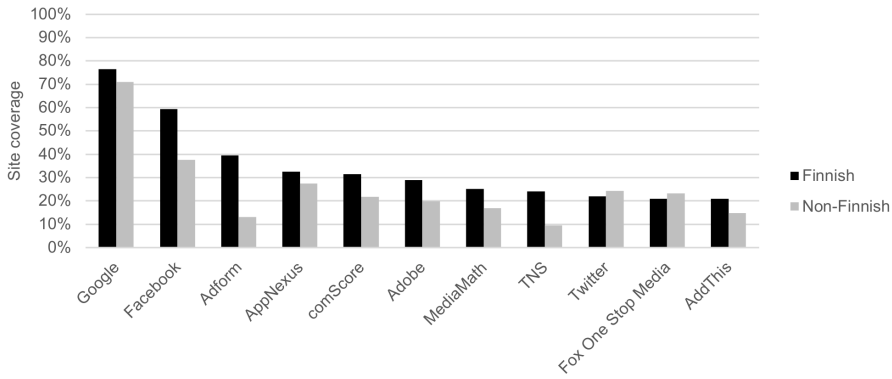

Figure 7: Reach of the top 10 tracking organizations on Finnish sites compared with their reach on non-Finnish sites

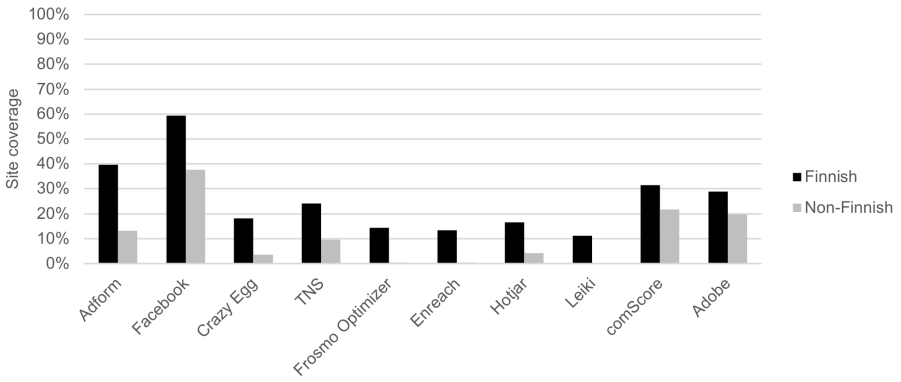

Figure 8: The top 10 tracking organisations with the largest site coverage discrepancy in favor of Finnish sites

Figure 8 illustrates the top 10 tracking organisations with the largest site coverage discrepancy in favor of Finnish sites, compared to their site coverage on non-Finnish sites. As with the results from separate trackers, this figure shows a preference for the local (Nordic) organizations such as Adform, TNS, Frosmo Optimizer, Enreach, and Leiki.

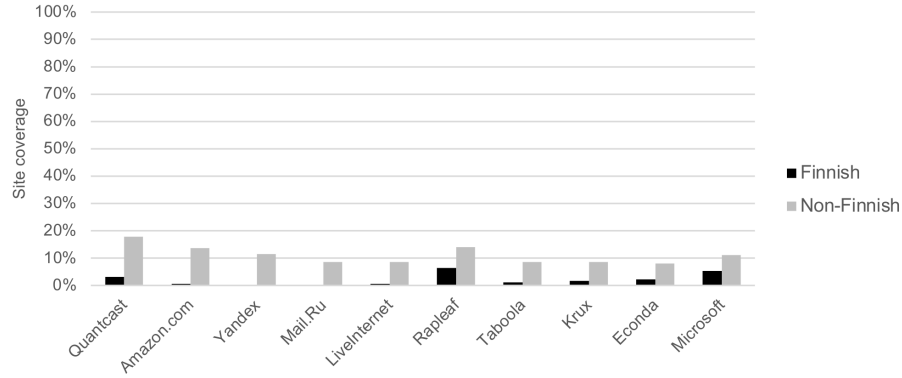

Figure 9: The top 10 tracking organisations with the largest site coverage discrepancy in favor of non-Finnish sites

Figure 9 presents the top 10 of tracking organizations that were most prevalent in non-Finnish sites compared to Finnish sites. As with the results from the separate trackers, the organizations presented here were clearly more prevalent on non-Finnish sites, with none of the organizations having over 10 % reach on Finnish sites and only Rapleaf and Microsoft having over 5 %.

Yandex and Mail.ru had no presence on Finnish pages, with all tracked sites being Russian or with Russian content. Amazon.com was only present on one site labeled Finnish, lataayoutube.com, which was an uncertain labeling choice and apparently an outlier from a tracking point of view.

Discussion

Our main findings regarding the prevalence of trackers are in line with those of Roesner et al. (2012), who found over 500 unique third party trackers on their top 500 Alexa sites. Google has a dominant presence when it comes to online tracking. Google can capture the behavior of Finnish web users on at least 75 % of the top 500 pages. This should be seen as a minimum level, as there were quite a few sites on which the automatic measurement failed to identify any trackers, even where manual sample observations proved that trackers were, indeed, present. Some have previously argued that Google Analytics is a site analytics service and that it should not be counted as tracking. But Google does provide the data from Google Analytics to the site owner for use in targeted advertising. Furthermore, the focus of this study has been on the companies behind the trackers; Google is able to track the users on sites with Google Analytics, whatever their actual data usage is. Even if one were to remove Google Analytics from the results, Google's next largest tracker, DoubleClick, itself has a reach of 63 %. It is nigh impossible to evade Google's trackers when surfing the world wide web.

Facebook was the second-largest tracker, present on 46 % of the top 500 sites which Finns frequent. This coverage is by itself considerable, yet to it one could further add the data gathered from people who use any of the Facebook-owned platforms like Facebook, Instagram, Messenger, or WhatsApp. Facebook can track Finnish users on 46 % of the sites they frequent as well as on most of the key social media platforms they use. Considering the kind of information that can be gathered when visiting websites versus when using social media, the overall understanding Facebook has of its users certainly competes with that of Google.

The dominant position of Google, with notable trackers such as Google Analytics and DoubleClick, has apparently not changed during the past 10 years. E.g. Purra & Carlsson (2016) reported that Google trackers have a coverage of over 70 % in all their domain categories, including .se (Swedish) and .dk (Danish) sites, supporting the results of this study. Facebook's runner-up status is likewise mirrored in similar studies (e.g. Karaj et al. 2018). After those two, the situation becomes more varied – there are seven other

organizations that have the capability to track Finnish users on over 20 % of the top 500 sites. These were mostly advertising platforms (e.g. AppNexus with a 29 % reach), which enable both the selling of ad space on publisher's websites and the behavioral analysis that fuels the targeting of each ad. Indeed, the advertising category is the predominant tracker category throughout the list of identified trackers, accounting for 66 % of the trackers. Common tracking organizations, e.g. AppNexus, comScore, Twitter, Adobe, Amazon, and Yahoo, are present and even have a higher-than-average coverage on Finnish sites.

The long tail of tracking is visible both from the perspective of trackers and websites. Over 90 % of the observed trackers were each found on fewer than 10 % of the sites, and 30 % of the trackers were only found on a single site. On the other hand, although there was only one site with over 100 trackers, there were 88 other sites (out of 500) that had over 20 trackers each. Every pageview on these sites could share the user's actions to a plethora of tracking organizations. To say that online tracking is common practice is to downplay the realities of the current landscape.

Even though tracking prevalence is on a similar level on Finnish and non-Finnish webpages, there is a clear geographical influence on the trackers observed. Not only are Google and Facebook's trackers much more prevalent on the Finnish sites compared to the non-Finnish ones, but there are Nordic organizations present as well. Trackers from the likes of Adform, TNS, Enreach, and Leiki are much more common on the Finnish pages. This geographical disparity is supported by previous research (e.g. Falahrastegar et al. 2014B). Organizations with roots in the Nordic countries, such as Adform, Frosmo, and Leiki, were all much more prevalent on the Finnish sites compared to the non-Finnish sites. On the other hand, the difference was even more distinct when reversing the comparison. Quantcast and Amazon were clearly more prevalent on non-Finnish sites, and the Russian Yandex and Mail.ru were not found on any Finnish sites. The russian trackers Yandex.Metrics and Mail.ru were observed on Russian sites that Finns frequent, whereas none of the Finnish sites had these trackers.

One surprising finding in the empirical data was that the website measurement tools Crazy Egg and Hotjar were much more prevalent on Finnish than on non-Finnish sites. One reason might be that as Finland is quite a small market, the preference and familiarity of a few key developers for these tools have kept the tools top-of-mind in the local organizations and therefore a preferred choice when selecting site analytics tools.

Currently online tracking is a core part of online business models (e.g. Zuboff, 2019). In the "free-to-use" model the user data is the asset being exploited, in the online commerce model it is personalization of content and

retargeting of the user, and in the advertising model it is user targeting, impression verification, and behavioral analytics loops that require the tracking of users. In a recent Finnish report on online privacy and anonymity by Sirkkunen & Haara (2017), 68 % of the respondents worried about the growing amount of online data tracking and 76 % wanted to have a better understanding of what data was collected and for what purpose. However, there is a discrepancy between what people say and what they actually do with regards to protecting their privacy. The same study showed that few users took the time to acquaint themselves with the terms of use of the services they use: Facebook 63 %, Google 40 %, Instagram 38 %, and WhatsApp 36 %. Even these numbers were seen as implausibly high by experts consulted in the report. Most respondents felt that they must give away their data in order to be able to use the services (64 %), and that the data would be collected anyway, regardless of what they do (69 %). There are many ways users can limit the extent to which they are tracked. Two of the main ways are by limiting cookie acceptance in browser settings and by installing browser extensions focused on tracker blocking.

Limitations and future research

Some technical limitations were identified, resulting in the interpretation that the findings of this study should be viewed as the minimum level of tracking prevalence, i.e. that tracking is even more prevalent than indicated here. One technical limitation was the quality of the results provided by the Tracker Tracker tool, exemplified by the high variation in the number of trackers identified and the high number of sites with no trackers identified, even after four redundant measurement runs. Further, there are many tracking mechanics that are not measured by the Tracker Tracker tool and which, although possibly less prevalent, could identify other interesting entities that are tracking users. Taking these limitations into account, the results presented in this paper should be understood as the minimum level of tracking prevalence, as any missing trackers would only increase the reach of the identified tracking organizations.

We conclude with some ideas for further research into this topic. One avenue of resaerch could focus on the GDPR (EU's General Data Protection Regulation). The data for this study was gathered before the GDPR came into effect. The GDPR requires websites to ask a user's consent to be tracked, which is something this study does not reflect. The data for this study was gathered using a virtual browser that is technically incapable of giving

cookie or tracking consent. One avenue for further research would be repeating the data gathering process using the same Tracker Tracker measurement tool and then analyzing the possible effect that the GDPR has had on how Finnish web users are tracked. Another avenue for future resaerch could use the above-mentioned whotracks.me datasets and compare the pre- and post-GDPR tracking practices of the same list of the top 500 sites frequented by Finnish web users. Finally, further research could be repeat the data gathering process utilizing the datasets made available by the researchers behind whotracks.me (owned by Cliqz, who also owns the Ghostery tracking extension used to identify trackers in the Tracker Tracker tool).

References

Bailey J., Laakso, M., & Nyman, L. (2019). Web tracking data for 500 websites popular among Finnish web users [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3543444

Bujlow, T., Carela-Español, V., Sole-Pareta, J., & Barlet-Ros, P. (2017). A survey on web tracking: Mechanisms, implications, and defenses. *Proceedings of the IEEE, 105*(8), 1476-1510. https://doi.org/10.1109/JPROC.2016.2637878

Cavna, M. (2013). 'NOBODY KNOWS YOU'RE A DOG': As iconic internet cartoon turns 20, creator Peter Steiner knows the joke rings as relevant as ever. The Washington Post. July 31, 2013. Retrieved from https://web.archive.org/web/20190906110052/https://www.washingtonpost.com/blogs/comic-riffs/post/nobody-knows-youre-a-dog-as-iconic-internet-cartoon-turns-20-creator-peter-steiner-knows-the-joke-rings-as-relevant-as-ever/2013/07/31/73372600-f98d-11e2-8e84-c56731a202fb_blog.html Accessed September 5th 2019.

Center for Democracy & Technology. (2011). *What does "do not track" mean?* (Proposal). Washington: Center for Democracy & Technology. Retrieved from https://web.archive.org/web/20190906105058/https://www.cdt.org/files/pdfs/CDT-DNT-Report.pdf Accessed September 5th 2019.

Eckersley P. (2010). How Unique Is Your Web Browser?. In: Atallah M.J., Hopper N.J. (eds) *Privacy Enhancing Technologies. PETS 2010.* Lecture Notes in Computer Science, vol 6205. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-14527-8_1

Englehardt, S., Eubank, C., Zimmerman, P., Reisman, D., & Narayanan, A. (2015). *OpenWPM: An automated platform for web privacy measurement*. Technical report, Princeton University, March 2015.

Englehardt, S., & Narayanan, A. (2016). Online Tracking: A 1-million-site Measurement and Analysis (pp. 1388–1401). Presented at the *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. New York, NY, USA: ACM. http://doi.org/10.1145/2976749.2978313

Eriksson-Backa, K. (2013). Informaatiotutkimus tänään - Informationsvetenskapen idag. *Informaatiotutkimus, 31*(4). https://journal.fi/inf/article/view/7528

Falahrastegar M., Haddadi H., Uhlig S., Mortier R. (2014A). The Rise of Panopticons: Examining Region-Specific Third-Party Web Tracking. In: Dainotti A., Mahanti A., Uhlig S. (eds) *Traffic Monitoring and Analysis. TMA 2014.* Lecture Notes in Computer Science, vol 8406. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-54999-1_9

Falahrastegar, M., Haddadi, H., Uhlig, S., & Mortier, R. (2014B). Anatomy of the third-party web tracking ecosystem. *ArXiv Preprint arXiv:1409.1066*. https://arxiv.org/abs/1409.1066v1

Falahrastegar M., Haddadi H., Uhlig S., Mortier R. (2016). Tracking Personal Identifiers Across the Web. In: Karagiannis T., Dimitropoulos X. (eds) *Passive and Active Measurement. PAM 2016.* Lecture Notes in Computer Science, vol 9631. Springer. https://doi.org/10.1007/978-3-319-30505-9_3

Fruchter, N., Miao, H., Stevenson, S., & Balebako, R. (2015). Variations in tracking in relation to geographic location. Paper presented at the *Proceedings of the 9th Workshop on Web 2.0 Security and Privacy (W2SP) 2015*. https://arxiv.org/abs/1506.04103v1

Ghostery. (2019). What are the new tracker categories? Retrieved from https://web.archive.org/web/20190203203947/https://ghostery.zendesk.com/hc/en-us/articles/115000740394-What-are-the-new-tracker-categories- Accessed February 3rd 2019

GitHub (2017). Canonical repository for the Disconnect services file. https://github.com/disconnectme/disconnect-tracking-protection/blob/master/services.json Accessed August 20th 2017.

Google (2019). How Google uses information from sites or apps that use our services. https://web.archive.org/web/20190905102725/https://policies.google.com/technologies/partner-sites?hl=en-US Accessed September 5th 2019.

Karaj, A., Macbeth, S., Berson, R., & Pujol, J. M. (2018). WhoTracks .me: Monitoring the online tracking landscape at scale. *ArXiv Preprint arXiv:1804.08959*. https://arxiv.org/abs/1804.08959

Krishnamurthy, B., & Wills, C. (2009). Privacy diffusion on the web: A longitudinal perspective. *Proceedings of the 18th International Conference on World Wide Web,* 541-550. Retrieved from https://web.archive.org/web/20190906104518/http://www2009.eprints.org/55/1/p541.pdf Accessed September 5th 2019.

Kristol, D. M. & Montulli, L. (1997). HTTP state management mechanism. Technical Report. RFC 2109 (Feb.), IETF. https://web.archive.org/web/20190906104432/https://www.ietf.org/rfc/rfc2109.txt Accessed September 6th 2019.

Kristol, D. M. (2001). HTTP Cookies: Standards, privacy, and politics. *ACM Transactions on Internet Technology, 1*(2), 151–198. https://doi.org/10.1145/502152.502153

Lerner, A., Simpson, A. K., Kohno, T., & Roesner, F. (2016). Internet Jones and the Raiders of the Lost Trackers: An Archaeological Atudy of Web Tracking from 1996 to 2016. *25th USENIX Security Symposium (USENIX Security 16).* Retrieved from https://web.archive.org/web/20190906104336/https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_lerner.pdf Accessed September 6th 2019.

Li, T., Hang, H., Faloutsos, M., & Efstathopoulos, P. (2015). Trackadvisor: Taking back browsing privacy from third-party trackers. *International Conference on Passive and Active Network Measurement*, 277-289. Retrieved from https://web.archive.org/web/20190906104258/https://www.symantec.com/content/dam/symantec/docs/research-papers/trackadvisor-taking-

back-browsing-privacy-from-third-party-trackers-en.pdf Accessed September 6th 2019.

Libert, T. (2015). Exposing the hidden web: An analysis of third-party HTTP requests on one million websites. *International Journal of Communication, 9*, 3544–3561.

Macbeth, S. (2017). *Tracking the Trackers: Analysing the Global Tracking Landscape with GhostRank*. Retrieved from https://web.archive.org/web/20190906103605/https://www.ghostery.com/wp-content/themes/ghostery/images/campaigns/tracker-study/Ghostery_Study_-_Tracking_the_Trackers.pdf Accessed September 6th 2019.

Metwalley, H., Traverso, S., Mellia, M., Miskovic, S., & Baldi, M. (2015). The online tracking horde: A view from passive measurements. Paper presented at the *International Workshop on Traffic Monitoring and Analysis*. https://doi.org/10.1007/978-3-319-17172-2_8

Montulli, L. (1998). U.S. Patent No. 5,826,242. Washington, DC: U.S. Patent and Trademark Office. Retrieved from https://web.archive.org/web/20191128073514/https://patents.google.com/patent/US5826242A/en Accessed November 28th 2019.

Purra, J., & Carlsson, N. (2016). Third-party tracking on the web: A Swedish perspective. Paper presented at the *Local Computer Networks (LCN), 2016 IEEE 41st Conference on*, 28-34. Retrieved from https://www.diva-portal.org/smash/get/diva2:1071640/FULLTEXT01.pdf Accessed May 5th 2019.

Roesner, F., Kohno, T., & Wetherall, D. (2012). Detecting and defending against third-party tracking on the web. Paper presented at the *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*. Retrieved from https://web.archive.org/web/20190906101953/https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final17.pdf Accessed September 6th 2019.

Ruohonen, J., & Leppänen, V. (2017). Whose hands are in the Finnish cookie jar? Paper presented at the *2017 European Intelligence and Security Informatics Conference (EISIC)*, 127-130. https://doi.org/10.1109/EISIC.2017.25

Räisänen, O. (2015). Trackers leaking bank account data. Retrieved from https://web.archive.org/web/20190906102954/http://www.windytan.com/2015/04/trackers-and-bank-accounts.html Accessed September 5th 2019.

Sanchez-Rola I., Santos I. (2018). Knockin' on Trackers' Door: Large-Scale Automatic Analysis of Web Tracking. In: Giuffrida C., Bardin S., Blanc G. (eds) *Detection of Intrusions and Malware, and Vulnerability Assessment. DIMVA 2018*. Lecture Notes in Computer Science, vol 10885. Springer. https://doi.org/10.1007/978-3-319-93411-2_13

Schneier, B. (2015). *Data and goliath: The hidden battles to collect your data and control your world*. New York: WW Norton & Company.

Shah, R. C., & Kesan, J. P. (2009). Recipes for cookies: How institutions shape communication technologies. *New Media & Society, 11*(3), 315-336. https://doi.org/10.1177/1461444808101614

Sirkkunen, E., & Haara, P. (2017). *Yksityisyys ja notkea valvonta: Yksityisyys ja anonymiteetti verkkoviestinnässä-projektin loppuraportti*. Tampere: Tampereen Yliopisto. http://urn.fi/URN:ISBN:978-952-03-0331-0

Solove, D. J. (2004). *The digital person: Technology and privacy in the information age*. New York: New York University Press.

S-Pankki. (2015). Google analytics -palvelun käyttö S-pankin verkkopalveluissa. Retrieved from https://web.archive.org/web/20190906101353/https://www.s-pankki.fi/fi/tiedotteet/2015/

google-analytics--palvelun-kaytto-s-pankin-verkkopalveluissa/ Accessed September 6th 2019.

Steiner, P. (1993). On the internet, nobody knows you're a dog. *The New Yorker, 69*(20), 61.

Tsai, J. Y., Egelman, S., Cranor, L., & Acquisti, A. (2011). The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research, 22*(2), 254–268. https://doi.org/10.1287/isre.1090.0260

Zuboff, S. (2019). *The Age of Surveillance Capitalism – The fight for a human future at the new frontier of power*. London: Profile Books.