UNIVERSITY OF TARTU
Institute of Computer Science
Conversion Master in IT

Norbert Metsare

Third-party services and their usage on the most visited Estonian websites

Master's Thesis (15 ECTS)

Supervisor: Arnis Paršovs

Tartu 2021

Third-party services and their usage on the most visited Estonian websites

Abstract:

Websites and used business models have changed a lot during the last few decades. While in the past a website had to create, manage, and secure all its own components, now all the site must do is to focus on its very own speciality and, by using third-party services, leave all the infrastructure, visitor analysis and security management to the services that specialize on those exact topics. This allows the website operators to create quality content, save money on development and increase their income.

While using third-party services makes the process more efficient for the website operators, it makes the users of those sites more vulnerable. Every piece of possible information about the user is gathered and shared with third-party partners to analyse user's patterns, location, and possible interests, so that the website can offer specific content to specific user. If we look more into the requests that forward the user's data, we can see that all of this ends up in the databases of only a handful of big tech companies, which, in turn, creates a handful of problems – big companies having leverage over the site contents, risks related to data centralization, and web dependency on the tech giants.

The current thesis will focus on the analysis of the most visited Estonian websites and their usage of third-party services, as well as the final destinations for gathered and forwarded user's information. Based on the findings, different web business models and third-party services are explained, as well as the connections between them. Data centralization problems are discussed along with recommendations to the end-users for more safer web browsing.

Keywords:

Third-party services, web business models, data centralization

CERCS:

T120 – Systems engineering, computer technology

Kolmanda osapoole teenused ja nende kasutamine Eesti enimkülastatavatel veebilehtedel
Lühikokkuvõte:

Viimaste kümnendite jooksul on veebilehed ning veebipõhised ärimudelid palju muutunud. Kui varem pidi veebileht looma, haldama ning turvama kõiki enda teenuse komponente, siis praegu saab see keskenduda vaid enda sisu loomisele ning läbi kolmanda osapoolte teenuste jätta muu partnerite hooleks. See võimaldab veebilehel luua kvaliteetset sisu, säästa raha ning suurendada kasumit.

Kuigi veebilehtede seisukohast on kolmanda osapoolte teenuste kasutamine kasulik, siis lõppkasutaja muutub tänu nendele järjest haavatavamaks. Kättesaadavat informatsiooni lõppkasutaja kohta jagatakse enda partneritega, et analüüsida kasutusmustreid, kasutaja asukohta ning kasutaja võimalike huvisid, eesmärgiga pakkuda võimalikult personaliseeritud sisu. Kui analüüsida päringuid, läbi mille kasutaja informatsiooni jagatakse, siis näeme, et kogu informatsioon jõuab läbi väikeste ettevõtete vaid käputäie suurte korporatsioonide andmebaasidesse, mis omakorda tekitab mitmeid probleeme – võimalikkus mõjutada lehtede sisu, andmete tsentraliseeritusega seotud riskide suurenemine ning veebi suur sõltuvus tehnoloogiakorporatsioonidest.

Käesoleva magistritöö käigus analüüsitakse Eesti enimkülastatavaid veebilehti ning nii nende kolmandate osapoolte kasutamise mustreid kui ka saadetavate andmete lõppsihtkohti. Toetudes analüüsi leidudele, käsitletakse erinevaid veebi ärimudeleid, kolmandate osapoolte teenuseid ning nende omavahelisi seoseid. Ühtlasi arutletakse andmete tsentraliseerituse probleemide üle ning esitletakse võimalusi turvalisemaks interneti kasutamiseks.

Võtmesõnad:
Kolmanda osapoole teenused, veebi ärimudelid, andmete tsentraliseeritus
CERCS:
T120 – Süsteemitehnoloogia, arvutitehnoloogia

Table of Contents

Introduction

Websites and web business models have changed a lot over the past few decades to be more reliant on the third-party service. Although this symbiosis between websites and third-party services is beneficial for the websites in order to create quality content, save money on development and increase their income, it is totally opposite for the end-user visiting those websites. End-user data along with their location, web visiting history and usage patterns are in many cases shared with the third-party services to enhance the quality of personalized content and in most of the times the user is not even aware of such transactions. The user could be more cautious when browsing the web, which, of course, requires knowledge about the existing dangers, but even then, the user can't be fully protected from all kinds of data sharing schemes or technologies.

The current thesis will analyse the most visited Estonian websites with a focus on the third-party requests made when user is visiting those websites. The analysis will determine where the visitor's data ends up and if we could see signs of data centralization. Based on the findings, different business models, third-party services, and the connections between them are discussed along with practical examples from analysis and theoretical references. Possibilities to raise user awareness are proposed and recommendations are shared for safer end-user browsing experience.

The thesis consists of two main parts. The analysis of the most visited Estonian websites will be presented in the first part where the third-party request data of websites will be taken into examination. The analysed websites and third-party requests will be categorized to see which are the most used third-party services on certain types of websites. Destinations for the end-user data and possible data centralization signs are examined. Also, different types of third-party services and different business models, accompanied with the findings from the practical analysis, will be presented. Recommendations and propositions for user awareness and safer web browsing are included in this part as well. The second part of the thesis discusses risks and benefits that using third-party services offers, with the emphasis mostly on data centralization.

1. Analysis of the most visited Estonian websites

Similar works of analysing website third-party requests already exist, and a number of those works have been conducted by Timothy Libert[1], who has developed a special tool for doing the analysis. The tool that Libert has created is called webXray and it is also a core tool used in the current analysis of Estonian websites.

Although similar works have been done, there is none for the Estonian websites. And as the main goal of the current analysis is to focus on local, Estonian, user data, the existing works might not reflect the impacts on the Estonian web user – therefore it is necessary to conduct the analysis with local data. Additionally, although Estonia is often identifying itself as a digital country, the knowledge about user data sharing with third parties is minimal and the third-party requests are basically hidden from the average internet user. In order to provide recommendations for an average user about the privacy of personal data, it is necessary to analyse the websites that they most often visit and to see where their data is ending up.

This section of the thesis is where the core analysis of the websites is done and where the data is presented in more readable format with numbers and visualizations. The theory about different third-party services and business models, along with figures from the analysis, will be presented in this section as well, opening up the numbers and discussing about the threats and opportunities of the findings from our analysis.

[1] https://scholar.google.com/citations?user=52UnefMAAAAJ&hl=en

1.1. Websites analysed

The most visited Estonian website statistics data is taken from Amazon's Alexa Ranking list, extracted on 11.04.2021. The Alexa Ranking list displays 500 most visited websites by users from specified country and is commonly used in web measurement research. The sites in the top websites list are ordered by their 1-month Alexa traffic rank. The 1-month rank is calculated using a combination of average daily visitors and pageviews over the past month. The site with the highest combination of visitors and pageviews is ranked #1.[2] From the list of the most visited Estonian 500 websites, only the ones with Estonian domains (ending with *.ee*) are extracted and analysed.

The websites were also divided into 7 different categories to see if there are differences in the results based on category. The 7 categories are following:

- Business – websites that are representing a certain business or selling a service (including online banking sites)
- E-commerce – websites for selling or mediating products online
- E-mail – online e-mail clients
- Entertainment – websites with entertaining content (online games, gambling, meeting portals)
- News – websites with general or specific news content
- Public – government websites, educational institution websites, public databases, websites related COVID-19 information and vaccination
- Search – search engines

[2] https://support.alexa.com/hc/en-us/articles/200449744-how-are-alexa-s-traffic-rankings-determined-

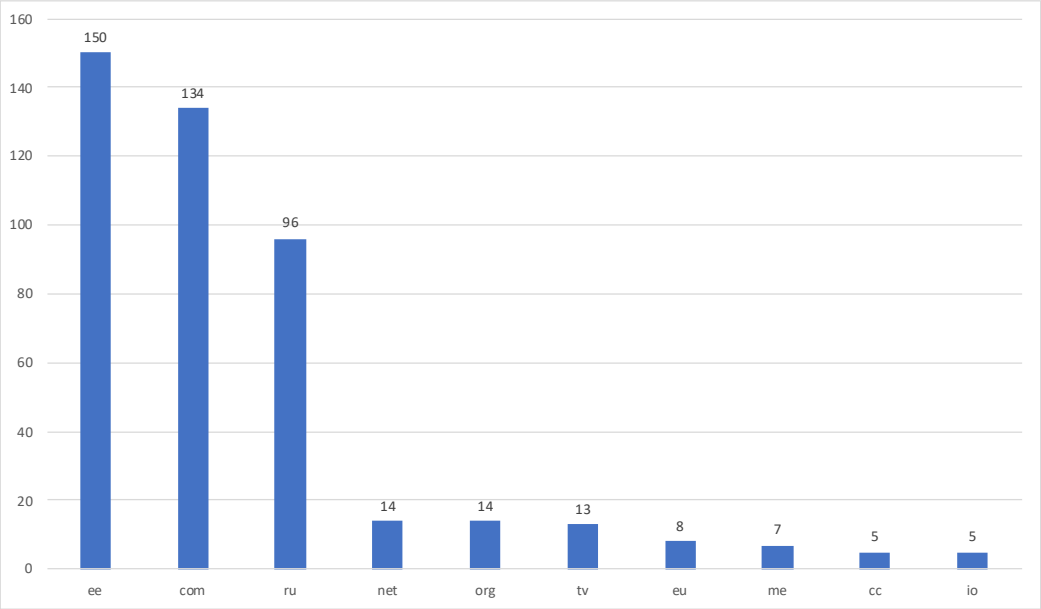


Figure 1. Top 10 domains out of 500 most visited Estonian websites based on Alexa Ranking

As seen from Figure 1, the majority of the 500 most visited websites by Estonian users are registered with Estonian domain – 150 of sites. The second place of visits is represented by general domain *.com* and third place by Russian domain *.ru*. Although Estonia has a large concentration of Russian speaking inhabitants within its population, the websites with Russian domains are left out of the scope of the analysis, because most of these websites are located in Russia. So, in order to focus on websites with Estonian origin, the scope is limited to Estonian domains only. Therefore, 150 websites ending with domain *.ee* will be chosen for the current analysis. Table 1 below lists those websites and shows which category each site presents. The sites in the table are ordered by their popularity.

Table 1. List of 150 top websites with Estonian domains

#	Website	Category
1	Delfi.ee	News
2	Postimees.ee	News
3	Google.ee	Search
4	Swedbank.ee	Business
5	Err.ee	News
6	Auto24.ee	E-comm.
7	Seb.ee	Business
8	Harjuelu.ee	News
9	Opiq.ee	Public
10	Ope.ee	Public
11	Neti.ee	Search
12	Ut.ee	Public
13	Ohtuleht.ee	News
14	Eki.ee	Public
15	Digilugu.ee	Public
16	Kv.ee	E-comm.
17	Online.ee	Business
18	Ilmateenistus.ee	News
19	Telia.ee	Business
20	Mnt.ee	Public
21	Zone.ee	Business
22	Moodle.edu.ee	Public
23	Olybet.ee	Entertain.
24	Avariilised-autod.ee	E-comm.
25	Elu24.ee	News
26	Seti.ee	E-comm.
27	Tootukassa.ee	Public

28	Lhv.ee	Business
29	Ikea.ee	E-comm.
30	Tallinn.ee	News
31	Inforegister.ee	Public
32	Okidoki.ee	E-comm.
33	Mke.ee	News
34	Bauhof.ee	E-comm.
35	Ria.ee	Public
36	Romu.ee	E-comm.
37	Juristaitab.ee	Public
38	Tv3.ee	Entertain.
39	Euronics.ee	E-comm.
40	Piletilevi.ee	E-comm.
41	Rik.ee	Public
42	On24.ee	E-comm.
43	Uueduudised.ee	News
44	Maaamet.ee	Public
45	Vvs.ee	E-comm.
46	Weby.ee	E-comm.
47	Kaubamaja.ee	E-comm.
48	Synlab.ee	Business
49	Kaup24.ee	E-comm.
50	Plusmerk.ee	E-comm.
51	Riigiteataja.ee	News
52	Rug.ee	E-comm.
53	Terviseamet.ee	Public
54	Politsei.ee	Public
55	Teatmik.ee	Public
56	Kallikalli.ee	Entertain.
57	Atyk.ee	Entertain.
58	Elisa.ee	Business
59	K-rauta.ee	E-comm.
60	Kuldnebors.ee	E-comm.
61	Purecosmetics.ee	E-comm.
62	Stena.ee	E-comm.
63	E-krediidiinfo.ee	Public
64	Dpd.ee	Business
65	Hansapost.ee	E-comm.
66	Riigihanked.riik.ee	Public
67	Soov.ee	E-comm.
68	Energia.ee	Business
69	Merit.ee	Business
70	Volley.ee	News
71	Innove.ee	Public
72	Vaktsineeri.ee	Public
73	Eestiloto.ee	Entertain.
74	Ilm.ee	News
75	Tele2.ee	Business
76	Hinnavaatlus.ee	E-comm.
77	Jahipaun.ee	E-comm.
78	Kalale.ee	News
79	Ristmik.ee	E-comm.
80	Kriis.ee	Public
81	Onninen.ee	E-comm.
82	Bauhaus.ee	E-comm.
83	Dormitorium.ee	Public
84	Maxima.ee	Business
85	Netfit.ee	Business
86	City24.ee	E-comm.
87	Barbora.ee	E-comm.
88	Ehituseabc.ee	E-comm.
89	Luminor.ee	Business
90	Arhitektuurikool.ee	Public
91	Osta.ee	E-comm.
92	Omniva.ee	Public
93	Jalgpall.ee	News
94	Lkf.ee	Public
95	I-smith.ee	Business
96	Sputnik-meedia.ee	News
97	Soccernet.ee	News
98	Mail.ee	E-mail
99	Kutsehariduskeskus.ee	Public
100	Tahvel.edu.ee	Public
101	Kalkulaator.ee	Public
102	Partnerkaart.ee	Business
103	Haigekassa.ee	Public
104	Selver.ee	E-comm.
105	Koroonatestimine.ee	Public
106	Voodi.ee	Entertain.
107	Nelli.ee	News
108	Iha.ee	Entertain.

109	Rc-est.ee	E-comm.
110	Cvkeskus.ee	Business
111	Medicum.ee	Public
112	Most.ee	Public
113	Projektid.edu.ee	Public
114	Dv.ee	News
115	Meblik.ee	E-comm.
116	Ehituskool.ee	Public
117	Tartumaraton.ee	Business
118	Envir.ee	Public
119	Eesti.ee	Public
120	Diil.ee	Business
121	Aliot.ee	Business
122	Petcity.ee	E-comm.
123	Thebodyshop.ee	E-comm.
124	Ilm24.ee	News
125	Klick.ee	E-comm.
126	Taltech.ee	Public
127	Polaver.ee	Business
128	Fin.ee	Public
129	Tln.edu.ee	Public

130	Avon.ee	E-comm.
131	Prismamarket.ee	E-comm.
132	Hiiumaa.ee	News
133	Geenius.ee	News
134	Kovtp.ee	Public
135	Estravel.ee	Business
136	Rahvaraamat.ee	E-comm.
137	Kokkama.ee	Entertain.
138	Membershop.ee	E-comm.
139	Digar.ee	Public
140	Bondora.ee	E-comm.
141	Cleankitchen.ee	Business
142	Vatteater.ee	Business
143	Aripaev.ee	News
144	Gamekeskus.ee	Public
145	Uuskasutus.ee	E-comm.
146	Molodoi.ee	News
147	Targaltinternetis.ee	Public
148	Kontaktibaas.ee	Business
149	1a.ee	E-comm.
150	Blummin.ee	E-comm.

Estonia is often identifying itself as a digital country, making many public services available online, and it is well supported also by the visitor data in Figure 2, making the public category one of the most represented among the most visited websites.

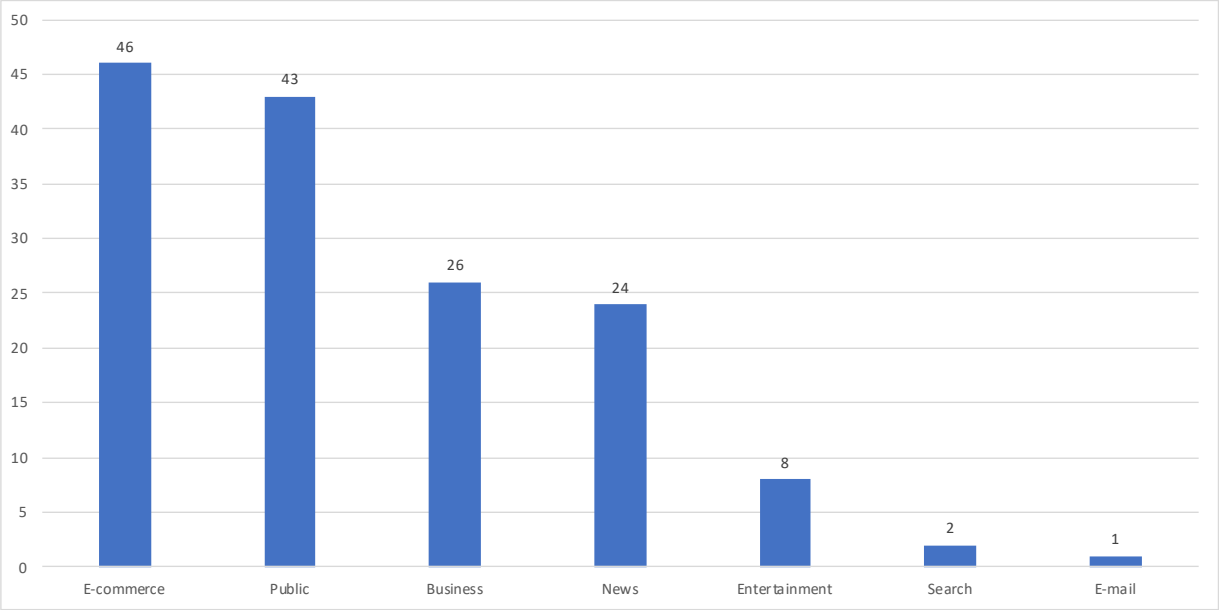


Figure 2. Categories of analysed websites

The popularity of the websites from public category is also affected by the fact that during the COVID-19 pandemic people are looking for information about the virus and vaccination, and most of the schools are functioning from the distance. E-commerce popularity is definitely affected by the pandemic and lockdowns as well.

1.2. Tools and methods

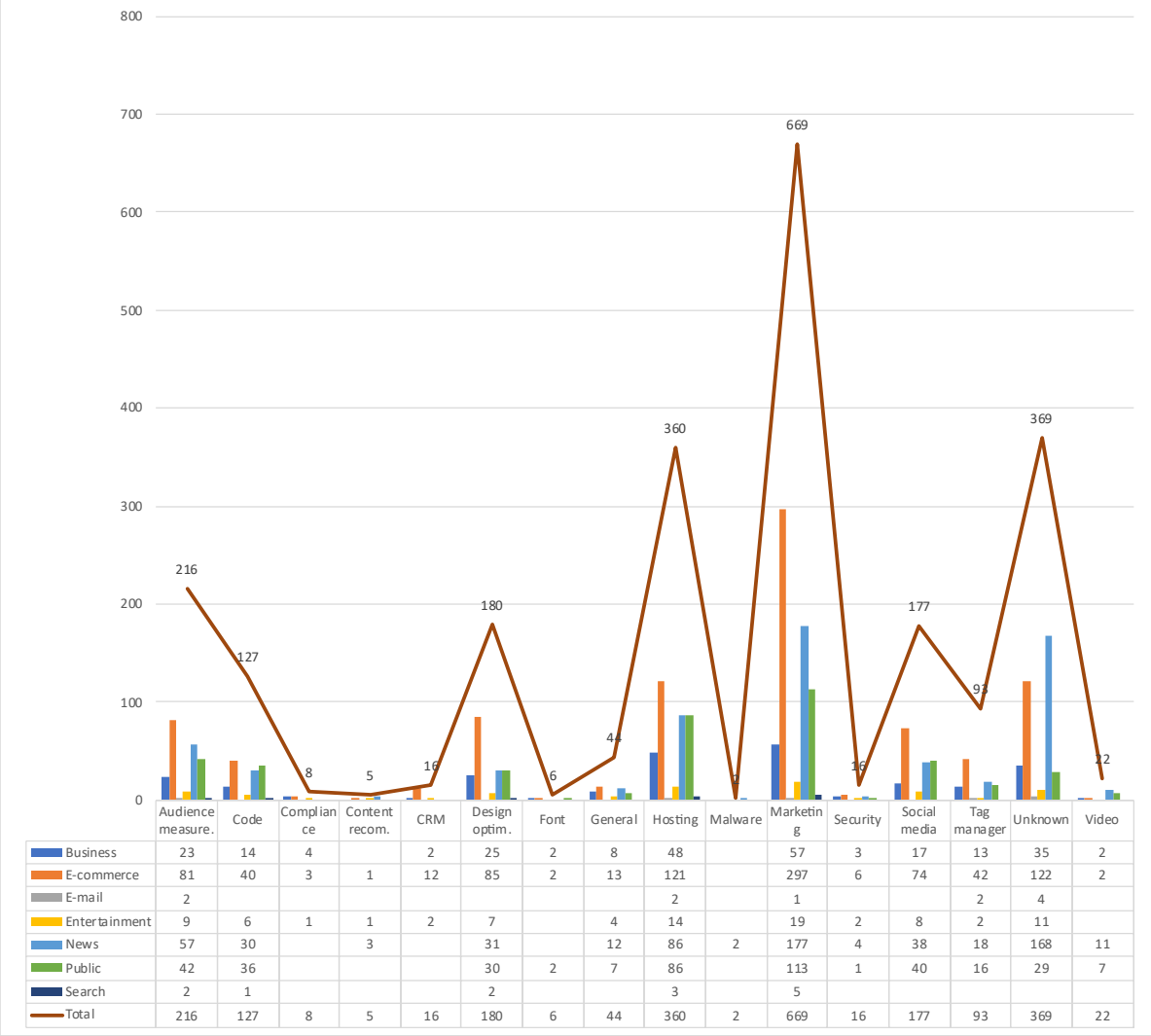
As already mentioned, the core tool for the current research is webXray, created by Timothy Libert, which will be used to analyse the most visited Estonian websites with Estonian domains (ending with *.ee*) from the previously mentioned list of 150 websites. The tool is *source available* and licenced for non-commercial use only – personal, academic, or non-profit use – and is made for analysing webpage traffic and content, extracting legal policies, and identifying the companies which collect user data.[3] Processing the list of the most visited Estonian websites with webXray results in generating following reports for further analysis:

- List of third-party requests (all third-party requests made on all analysed websites)
- Third-party request domains, domain owners and their origin country (domains of all the third-party requests, as well as the owner companies of those domains and their origin countries)
- Categorization of third-party requests by usage (all third-party requests categorized by usage type)
- Third-party requests for each analysed first-party website (list of third-party requests initiated by visiting each website)

It is important to note that the third-party request analysis is performed only on the landing pages of analysed websites. After loading a landing page of a website from a given list, webXray waits 45 seconds before loading the next page to be sure that all the third-party elements were loaded and identified. Additionally, same pages were scanned multiple times to gain more precise data. As a result of this, webXray did a total of 21 594 requests and identified 11 292 third-party requests among them. That means that a considerable amount of 52,29% of all the requests were third-party requests and carried some sort of user data via requests from first-party website to a third-party service. This is a good starting point to demonstrate the dependence of the websites on third-party services.

[3] https://webxray.org/

1.3. Types of third-party services



	Audience measure.	Code	Compliance	Content recom.	CRM	Design optim.	Font	General	Hosting	Malware	Marketing	Security	Social media	Tag manager	Unknown	Video
Business	23	14	4		2	25	2	8	48		57	3	17	13	35	2
E-commerce	81	40	3	1	12	85	2	13	121		297	6	74	42	122	2
E-mail	2								2		1			2	4	
Entertainment	9	6	1	1	2	7		4	14		19	2	8	2	11	
News	57	30		3		31		12	86	2	177	4	38	18	168	11
Public	42	36				30	2	7	86		113	1	40	16	29	7
Search	2	1				2			3		5					
Total	216	127	8	5	16	180	6	44	360	2	669	16	177	93	369	22

Figure 3. Third-party requests by website types

For each analysed website webXray provides us with data about requests made and their classification, seen in Figure 3 where the data is categorized by third-party request type and matched with website type. In total, 1713 third-party requests were made from our 150 analysed websites. Each request is categorized to one or several categories by webXray. Additionally, if there was no information about the request category in webXray database, the requests that occurred more than 3 times in total of 1713 third-party requests, were also looked up and supplemented with correct category. If there was no match in webXray category database for the request and the request was made 3 or less times from all analysed pages (total of 1713 requests) the request was categorized as unknown.

The categories of third-party requests are following:

- Audience measurement
- Code
- Content recommendation
- Customer relationship management
- Design optimization
- Font
- General
- Hosting
- Malware
- Marketing
- Security
- Social media
- Tag manager
- Unknown
- Video

As emphasized by Libert, the creator of webXray, it is important to note that the categorization of third-party requests by webXray is done from the perspective of the first party as the third-party may have different objectives. For example, while a site may utilize Google Analytics to gain insights into the site traffic, Google may use that data for marketing purposes. [1] Different categories presented above, and their purposes will be presented more in depth later in this section, but the numbers in Figure 3 by themselves present us with valuable data. While the unknown category with its 369 third-party requests holds the second place, these will be left out of analysis as these requests might belong to any other category. The most significant third-party request amounts come from marketing, hosting, and audience measurement categories (with 669, 360 and 216 total requests respectively). From the websites, E-commerce and news category websites are the ones that stand out especially when considering the amounts of different total third-party requests made – both categories dominating over other types of websites with 901 total requests for the e-commerce sites and 638 total requests for the news sites. Two websites, *mke.ee* and *sputnik-meedia.ee,* were infected with *yadro.ru* redirect adware, which is used to push certain unwanted ads to a website for ad revenue increase.

Running the analysis on our 150 websites, webXray identified 390 different third-party domains to which, during the loading process of the website, a third-party request was initiated. The number itself is not very big and makes a rough average of 2.6 requests per page, but what is more interesting and more remarkable is the concentration of those 390 requests on all analysed webpages.

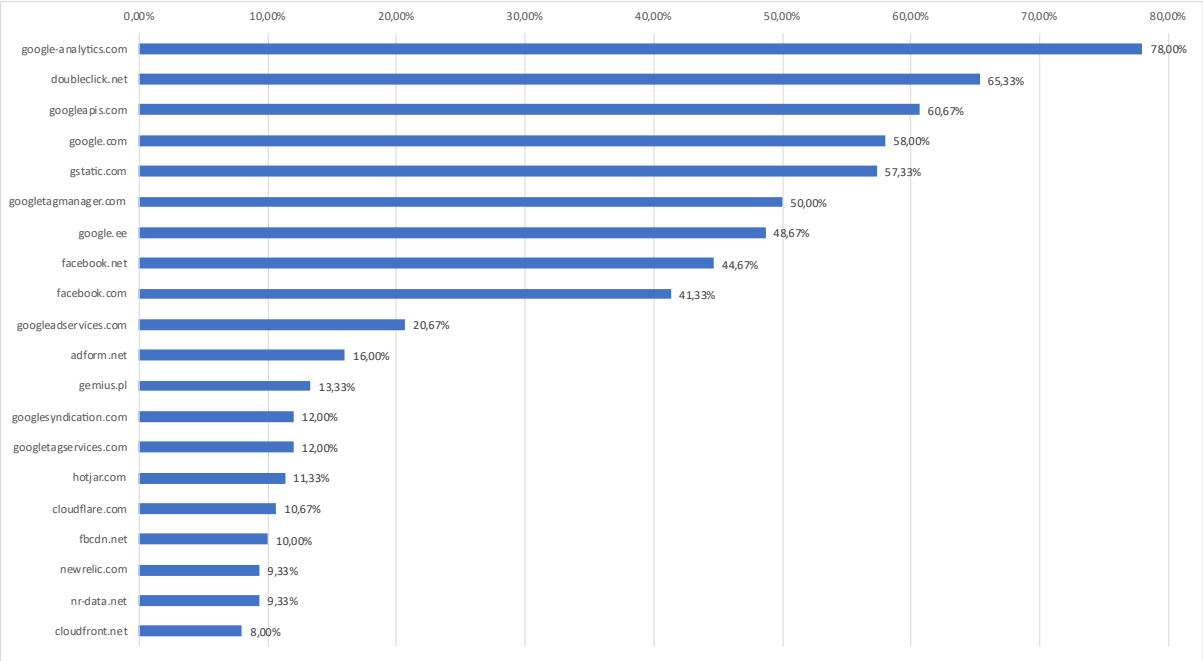


Figure 4. Top 20 third-party domains

Different third-party services and their usages will be explained later in the thesis, but as seen from Figure 4, 78% – that is 117 sites out of 150 – use google analytics third-party service, which is one of the audience measurement third-party services provided by Alphabet, parent company of Google. And that is not the only number that shows how much of user data ends up in Alphabet's servers and in the territory of the United States overall. The charts below, in Figure 5, demonstrate that the data of 10 domains out of top 20 most used ones end up in the Alphabet's servers and that the data of 17 domains out of top 20 end up in the territory of the US. From the overall results 114 domains, out of the total 390, have US origin.

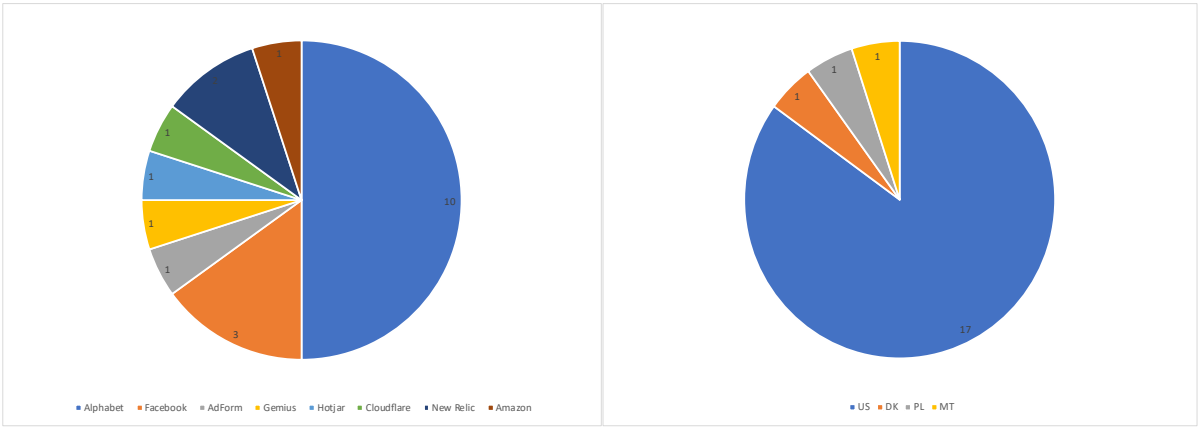


Figure 5. Parent companies (left) and origin countries (right) of top 20 third-party domains

The quantities of the data sent via the third-party requests are immense and the fact that so much of it is legally under the jurisdiction of one country shows how much of a user data can be potentially gathered and analysed by such governments if they only need to.

Figure 6 shows in which companies the user data, that is sent via our 1713 analysed third-party requests, mainly ends up.

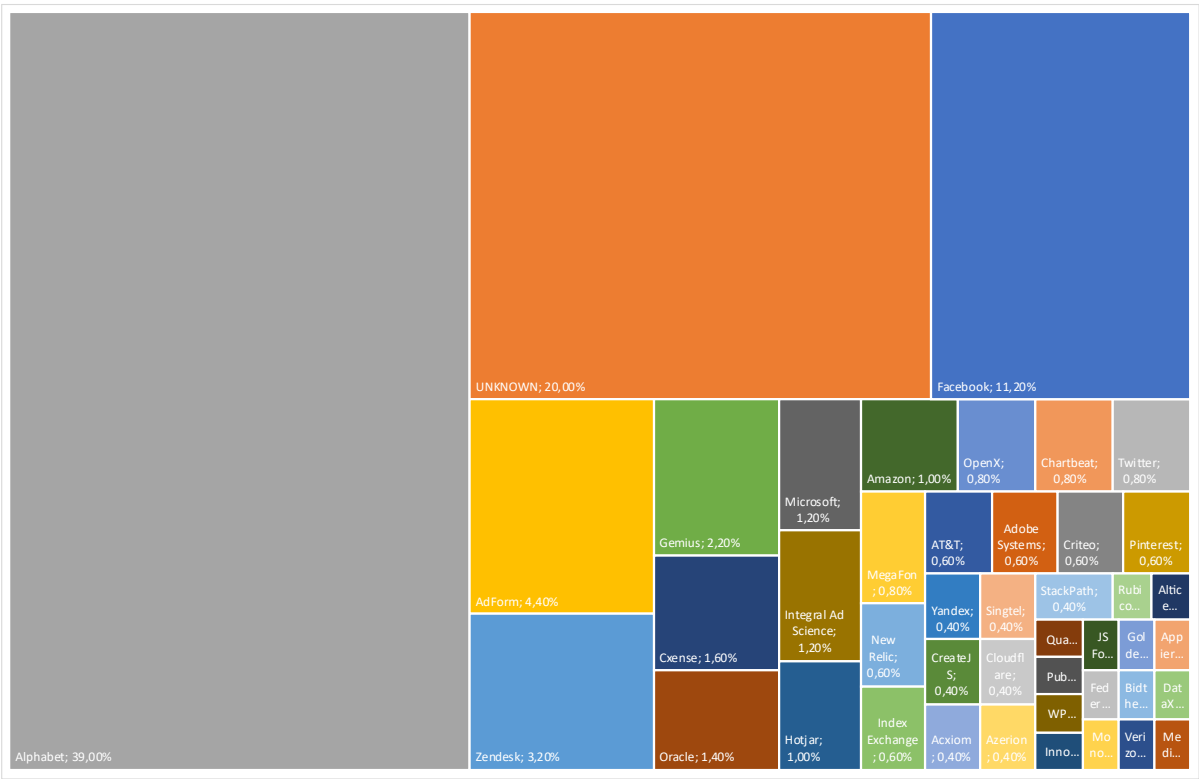


Figure 6. Companies receiving the user data via analysed third-party requests

Figure 6 clearly demonstrates the data centralization existence by indicating that 39% of all the requests end up in Alphabet servers. The second place, 20%, is taken by the companies with unknown origin, that represent smaller third-party providers possibly with local origin or

offering some other kind of niche third-party services – *kissmetrics.com, bepolite.eu, onesignal.com, livechat.com, frosmo.com,* just to name a few. Putting it into comparison with Alphabet, makes no doubt that data- and even so-called internet centralization exists and is largely dominated by only one company. Facebook with its 11.20% is firmly holding third place in our request destination analysis. Later in this thesis we will discuss more about the threats of data centralization, but it is important to note that while the number of third-party requests per site is high, it would be much safer for the end-user if the highest percentage of data centralization would be dominated by the unknown portion as it would mean that the user data is evenly distributed between many different providers and would not cause the core issues of the data centralization – ease of data manipulation, website content manipulation and surveillance.

In the following subsections different types of third-party services that have occurred from our analysis of 150 most visited Estonian websites, their functioning principles and main uses are covered. The webXray request categorization is taken as a basis for choosing the third-party services and where appropriate, possible threats to the end-users are brought out as well.

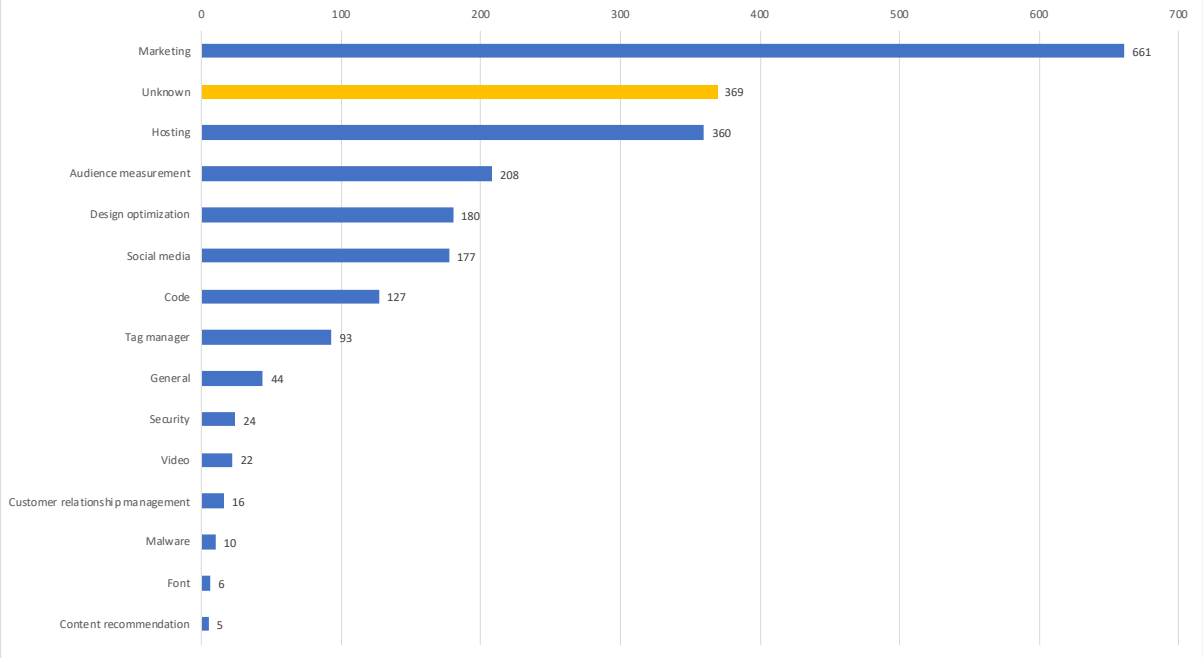


Figure 7. Third-party requests by usage types

A total of 1713 third-party requests were initiated from visiting the landing pages of the 150 analysed websites. Figure 7 presents us different third-party request categories and the amount of the respective requests done. The total amount of all the requests in Figure 7 is higher than the total number of requests done because one request might fall into several different

categories – for example a third-party request done to *facebook.com* falls into marketing and social media categories. It is also important to note that the basis for the categorization is taken from the data that webXray provides while analysing the websites and if the webXray database did not have information about some requests, we supplemented the database with appropriate categories for the requests that occurred more than 3 times. The requests that occurred 3 or less times, in total of 1713 requests, were marked as unknown and the requests in this category could be falling under other mentioned categories.

1.3.1. Marketing

Third-party service providers who focus on marketing are basically the intermediaries between the consumers, websites, and the advertisers. The marketing third-party services rely on cookies that are held in user's browser and are constantly accumulating information about user location, website visits, interests, use patterns, statuses, and everything else that can be collected from cross-site requests. When a user visits a website that is using a marketing third-party service, the identifier from the cookie is sent to the third-party service provider, which in turn matches the identifier of an already existing entry in third-party database. The response from the third-party provider contains information about the ads that should interest the user based on the previously collected information. This process is also known as behavioural targeting. The website that the user visits – the first-party website – will display targeted ads at the section that is specifically meant to display them, and the user will not see any random advertisement, but rather something of their interest. In contrast to legacy media, the web facilitates monitoring the actions of specific users, allowing advertisers to target messages based on inferences gleaned from "tracking" users as they browse the web, a process known as "online behavioural advertising" (OBA). The technological systems facilitating OBA are highly centralized, allowing a handful of companies to monitor the web browsing behaviours of billions of people and broker the flow of advertising revenue to millions of sites. [1]

When talking about marketing third-party services, the requests that forward information of a cookie to a third-party website are one of the biggest threats in user data privacy and dominate as a main resource of centralized data centres, where that data is then stored, analysed by complicated algorithms, and used to customize the content that the user is seeing. As seen from the findings of the analysis, marketing third-party services are the most used services among all the 150 sites analysed. The connections between the first party websites and third-party

services often occur without user interaction and may expose users to persistent tracking carried out by cookies, browser fingerprints, and other identifiers. [1]

Table 2 shows top 10 most used third-party marketing domains from our analysis of 150 most visited Estonian websites.

Table 2. Domains of top 10 marketing third-party services

#	**Domain**	**Occurrences**
1.	doubleclick.net	98
2.	google.com	87
3.	google.ee	73
4.	facebook.net	67
5.	facebook.com	62
6.	googleadservices.com	31
7.	adform.net	24
8.	atdmt.com	12
9.	2mdn.net	10
10.	casalemedia.com	9

From Table 2 it can be seen that the servers of Google and Facebook are the ones where most of the user data ends up in – 5 of the top 10 domains (*doubleclick.net, google.com, google.ee, googleadservices.com, 2mdn.nt*) belong to Google and 3 belong to Facebook (*facebook.net, facebook.com, atdmt.com*). 113 websites from the total of 150 used marketing third-party services.

Marketing third-party services are mostly provided by large companies that own a lot of user data already and are greatly centralized – that is where the benefit for the first-party website comes from and the amount of data that the third-party provider owns is acting as a proof of being effective with providing best personalized content. Using such services is beneficiary for the first party websites but is one of the biggest threats for the end-user.

1.3.2. Audience measurement

Audience measurement (also known as web analytics) third-party services allow first-party websites to see and track user behaviour on their website – number of visits, user actions, most

visited sections, and pages. The reasons why a website might use an audience measurement third-party service are different and might carry various purposes. Plausible.io, a vastly growing Google Analytics competitor that focuses on user privacy by creating separate data sets available only to the client website, brings out three main benefits of using a web analytics tool – to see if the website is growing, to see what people are doing when they visit the website and to see if the visitors are converting on the goals. [2] It is also important to note that a company who offers several different third-party services, can use data gathered by one service to enhance the other service. For example, Google uses data gathered through AdSense, Google Analytics and YouTube to deliver their services, maintain and improve them, develop new services, measure the effectiveness of advertising, protect against fraud and abuse, and personalize content and ads the user sees on Google and on its partners' sites and apps.[4]

Table 3 shows top 10 most used third-party audience measurement domains of our analysis. The Google Analytics service stands out above the others by far. From the total of 150 websites 119 used audience measurement third-party services.

Table 3. Domains of top 10 audience measurement third-party services

#	**Domain**	**Occurrences**
1.	google-analytics.com	117
2.	gemius.pl	20
3.	hotjar.com	17
4.	atdmt.com	12
5.	everesttech.net	8
6.	quantserve.com	8
7.	hotjar.io	7
8.	pbstck.com	4
9.	chartbeat.com	4
10.	chartbeat.net	4

In case of user data privacy matters, it depends a lot on what kind of third-party service the website uses. Opposing to the advertising third-party services, the audience measurement services can be provided very effectively by smaller and less known providers, because the service does not depend on the previously gathered data, but rather on good user experience

[4] https://policies.google.com/technologies/partner-sites?hl=en-US

and performance. For the end-users it is certainly more secure if their usage data doesn't end up in yet another tech giant server and is analysed anonymously like some smaller providers prefer to do it. Despite that, according to the W3Techs Web Technology Surveys, Google Analytics is installed on more than 55% websites on the web and is therefore the most used analytics tool in the world.[5]

1.3.3. Compliance

Compliance tools allow sites to manage their privacy policies and consent notifications in order to comply with data protection laws. [1] These third-party services don't usually gather user's personal data, but rather just remember previously selected settings by the user. Compliance services are one of the few that the user actually sees and can interact with, when used for asking user's consent for different cookies if visiting a website. The usage of such services is beneficial for the end-user, and they make following different regulations much more easier for the first-party website. In many cases though, it seems that the website is using those services just to comply with the law and not to protect the user, because in the background other third-party services that gather user information can still be found.

From the websites analysed only one compliance domain, which occurred in 8 requests, was identified – *cookiebot.com*. All 8 requests were done from 8 different websites - *1a.ee, diil.ee, ikea.ee, k-rauta.ee, maxima.ee, olybet.ee, tele2.ee, telia.ee.*

1.3.4. Design optimization

Design optimization services allow websites to experiment with different designs and help to analyse the impact of different designs on the users. It is often referred to as A/B testing and nowadays the first-party website doesn't have to create different designs and gather data themselves but can easily use third-party service to do all that. The basic concept about design optimization is to measure usage and conversion rate data to determine if one version of website design acts more effectively than the other version. The first-party website usually declares an original design version and then one or several variations of the webpage design. The users are then randomly directed to the original or a variation page and as a result the usage data of both versions is compared.

[5] https://w3techs.com/technologies/history_overview/traffic_analysis/all

Perhaps one of the most popular examples is the design optimization, or A/B testing, of Barack Obamas election website, where different background pictures and button variations were tested, and which, in the end, had remarkable results in the growth of supporters and ultimately winning the elections. A right combination of the background picture and button improved the signup rate by 40.6% and over the course of campaign that 40.6% translated to 2.8 million more email subscribers, 288 000 more volunteers and additional $57 million in donations. [3]

Design optimization third-party services require first-party website only to define the original and variation designs and everything else can be assigned to the third-party – redirecting users, data collection, data analysis, recommendations. That, of course, results in sending user activity data to the third-party and can be used to determine user preferences, online activity and sometimes might even give a hint about user's physical properties (race, gender, age, etc.). The previously mentioned election website testing is a great example of the power of such tools and the temptation of using such tools is understandable.

Top 10 design optimization domains from our analysis are included in Table 4.

Table 4. Domains of top 10 design optimization third-party services

#	**Domain**	**Occurrences**
1.	gstatic.com	86
2.	googlesyndication.com	18
3.	hotjar.com	17
4.	newrelic.com	14
5.	nr-data.net	14
6.	cloudfront.net	12
7.	hotjar.io	7
8.	searchnode.io	6
9.	googleoptimize.com	2
10.	pingdom.net	2

Table 4 presents us domination of Google also in the field of design optimization with following domains – *gstatic.com, googlesyndication.com* and *googleoptimize.com.* 98 websites out of 150 used design optimization third-party services.

1.3.5. Hosting

As the name says, the hosting third-party services, also referred as cloud hosting services, provide hosting for different content that the websites consist of. Before the hosting services became popular, the first-party website had to manage and host their own servers, securing those servers and dealing with different scalability issues. Nowadays much of the website's content is hosted by using a third-party service, which means that some of the code, fonts, images, and videos are stored on a third-party server and whenever a user visits a website, it is actually served from the third-party server. Only if the first-party website is not serving content that is already available in third-party server, the role of the first-party website is to update the content on third-party servers to make changes on the website available to the end-users, but everything from scalability, security and performance is taken care by the third-party. A good example of using a hosting service is a website that has embedded YouTube videos to its site or a website that is using certain fonts from Google services.

The hosting services usually don't gather end-user data directly and are rather used to store and analyse the first-party website data, making the content management much more easier for them. This, however, doesn't mean that hosting services do not present any threats to the users. As seen from analysis results, hosting third-party requests are made 360 times during the analysis. From the total of 150 websites 123 are using hosting services, being therefore one of the most used third-party service. The fact that so many of the websites are using hosting services and that those services are relatively centralized, presents the common risks of data centralization – content manipulation possibilities and easier unauthorized access to data due to centralization.

Following 10 hosting domains, seen in Table 5, were the most popular among our analysed websites.

Table 5. Domains of top 10 hosting third-party services

#	**Domain**	**Occurrences**
1.	googleapis.com	91
2.	gstatic.com	86
3.	googletagmanager.com	75
4.	googletagservices.com	18

5.	cloudflare.com	16
6.	fbcdn.net	15
7.	ytimg.com	9
8.	youtube.com	9
9.	ggpht.com	9
10.	createjs.com	8

Table 5 shows Google's domination once again with domains like *googleapis.com, gstatic.com, googletagmanager.com, googletagservices.com* and *ggpht.com*. From 150 analysed websites 123 used hosting third-party services.

1.3.6. Security

Security services exist to help site operators cope with threats such as distributed denial of service (DDoS) attacks and to prevent criminals using automated means to commit ad fraud and scrape content. [1] These services are used to protect the first-party website and not the end-user. In many cases, as these services analyse user behaviour in order to identify bots, they also need to gather some data about the user. The positive thing about security service providers is that they are relatively decentralized and do not rely so much on user data, but rather on user behaviour analysis.

Among all our analysed requests only one security-related domain, *cloudflare.*com, could be identified, which occurred on 16 cases on 16 different websites – *bondora.ee, hiiumaa.ee, jahipaun.ee, kalkulaator.ee, kaup24.ee, merit.ee, molodoi.ee, netfit.ee, okidoki.ee, olybet.ee, osta.ee, partnerkaart.ee, soccernet.ee, stena.ee, tallinn.ee, tv3.ee*.

1.3.7. Social media

Social media third-party services are mainly provided by social media enterprises and facilitate directing of user to the respective social media page from the first-party website. These services also allow displaying social media feeds and other content on the websites. As they are closely related to the social media companies, of which many are also providing advertising third-party services, they gather user data and use that data with previously collected data to identify users and therefore help first-party websites to personalize the content even more. Also, the fact that many of the social media users are using automatic login, the social media site can instantly

identify the user that clicked the social media link on the first-party website, allowing it to enhance other third-party services provided by the same third-party provider and ultimately even personalize user's social media feed.

Table 6 shows top 10 most used social media domains found from our analysed websites.

Table 6. Domains of top 10 social media third-party services

#	**Domain**	**Occurrences**
1.	facebook.net	67
2.	facebook.com	62
3.	fbcdn.net	15
4.	addthis.com	11
5.	ggpht.com	9
6.	twitter.com	4
7.	linkedin.com	4
8.	pinterest.com	2
9.	sharethis.com	1
10.	instagram.com	1

Although Google is represented with *ggpht.com* domain, the social media third-party services are mostly provided by Facebook as seen in Table 6 (*facebook.net, facebook.com, fbcdn.net* and *instagram.com* domains). From the total of 150 websites 79 used social media third-party services.

1.3.8. Tag management

Tags are snippets of code that are included in the first-party website source code, mostly to enable the usage of already previously mentioned advertising and audience measurement services. These snippets of code are the ones that provide third-party service with necessary information and usually they are included in several different places in the source code. Tag management services allow first-party websites to manage those tags by giving control in adding, removing, and organizing different tags as well as when those tags are triggered. An example of a tag would be a Facebook tracking pixel that has been implemented to the first-party website. Usually, websites have dozens of those tags. The tag manager allows to add or remove the Facebook tracking pixel and it allows to create triggers so that the Facebook pixel

should be activated when user clicks on a certain button for example. These services do not rely on user data, but as they make managing different advertising and audience measurement tags so much easier, they encourage the usage of more tags that gather user data.

Table 7 shows that the tag management, based on our 150 Estonian website analysis, is totally dominated by Google.

Table 7. Domains of tag management third-party services

#	**Domain**	**Occurrences**
1.	googletagmanager.com	75
2.	googletagservices.com	18

From total of 1713 third-party requests 93 were related to tag management as seen from Table 7 and from total of 150 websites 86 sites used tag management services.

1.4. Data collection methods by third-party services

There are different kind of technologies that the third-party services use to function effectively and perform at their best. Many third-party services combine different ones to get the best results and may therefore act as an even bigger risk for the end-user, who is tracked, identified, and made unique by those features. These technologies that enable the identification and tracking of a user are called online tracking technologies.

The following figures, Figure 8 and Figure 9, represent an example of data that the browser request sends to the third-party server (request headers – Figure 8) and what it receives from the third-party server (response headers – Figure 9). This is an example of Google's third-party service that is used by *delfi.ee*. As it can be seen, many different features are used to transfer data with simple request, like user agent and HTTP referer, but also cookies are sent and set from the third-party server.

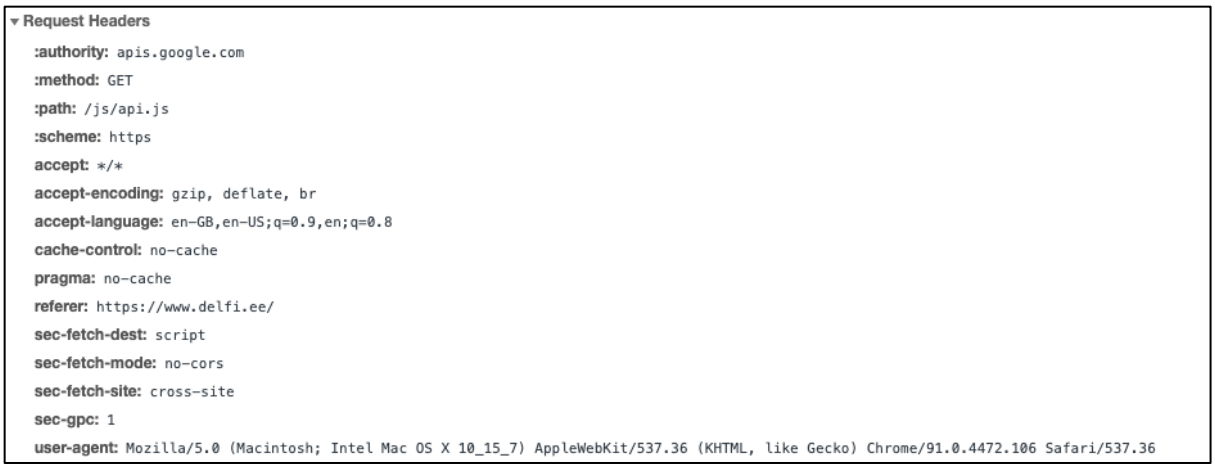


Figure 8. Request headers to google.com when visiting delfi.ee website

Figure 9. Response headers from google.com, provided to request in Figure 8

1.4.1. Online tracking technologies

IP address

Perhaps the most basic way to identify rough location of a user is using user's IP, or Internet Protocol, address. Although it cannot pinpoint the detailed user location, the information it provides to the websites and third-party services that the websites use – user's country, region, city, time zone, service provider – will still contribute a lot to the big picture of identifying the user, especially when that information is combined with other available information. The user's IP address is something that is always available to a website that the user is visiting (therefore available also to all third-party services that the website uses) and can be masked, to some extent, only by experienced and aware user (using VPN, proxy servers, Tor routing, etc.).

The IP address is allocated to a user when user starts the connection to the internet and in most cases, it is allocated dynamically from a specified pool of IP addresses that the ISP, or Internet Service Provider, owns. Therefore, as in most of the cases it is not specifically tied to one device or user only, it is only a rough estimate of a user's location. Several methods are used, such as IP geolocation databases. Accuracy of client-independent geolocation is lower, typically at city level. [4] On the other hand, a lot of research has been done trying to make IP address pinpointing more accurate, using different metrics related to web browsing – latency, number of hops, etc. [5] [6] [7] The accuracy could of course be used by giant tech companies, collecting the user data, but could be also used for tracking down criminals for example.

HTTP referring

HTTP referring is very common to advertising and affiliate marketing business models, because, as the name indicates, they refer to a website what was the website from which the user just came from. This information is usually passed by the user's browser in the HTTP request header, a part of a third-party request that is always included, and it is received and used by the website that the user is currently browsing as well as the third-party services that the website uses. For affiliate marketing business model this is the main way to know which was the website that the user used to reach the destination website and thus the source website can be established, and commission paid. Otherwise, the HTTP referring info might be used for promotional or statistical purposes.

Cookies

__Secure-3PSIDCC	AJi4QfGvI6ijiBHgeD1lK5Qc6zo-pxr3Yug71VrhanwS0RcdaPJHMlOydTLAn-McbHsc9b97HQ	.google.com
SIDCC	AJi4QfFil-pLw90xy7WWSdnJ1fW3PQtTFLR7l4yStg6_W91-4Wr-pvSXlk3fABUGt5TYY1LoLA	.google.com
NID	217=5664suzBjhyPd0uglei60_rS4vBxPQ--RYD8UPvbk-FMdzKHQPlqjCMrqAX4Eu-URlERIM…	.google.com
__Secure-3PSID	-we2suksReln2SuPoNk6k5p7HJ7E_aFv2blqEw-SYk5sFP9yJGbkxYTlQzx6o9hOoiZSxg.	.google.com
SID	-we2suksReln2SuPoNk6k5p7HJ7E_aFv2blqEw-SYk5sFP9yYDWP_ElQzA82xG_XnoSiDw.	.google.com
SAPISID	kupVJdT356CxJMvl/Arif6l0snwumarzO5	.google.com
SEARCH_SAMESITE	CgQI-pEB	.google.com
CONSENT	PENDING+687	.google.com
APISID	ACutJNX7t0ylPuwf/AzDP8MpsD8EleTEFJ	.google.com
HSID	AyWOQTU8bxxCH2R2Q	.google.com
SSID	ABikCwuQa5zfeTa80	.google.com
OGP	-19022622:	.google.com
1P_JAR	2021-6-20-10	.google.com
OGPC	19022622-1:19022591-1:	.google.com
__Secure-3PAPISID	kupVJdT356CxJMvl/Arif6l0snwumarzO5	.google.com

Figure 10. Google.com cookies set by visiting a news website that uses Google's third-party services

A "cookie," is the piece of information that the server and client pass back and forth. The amount of information is usually small, and its content is at the discretion of the server. In

general, simply examining a cookie's value will not reveal what the cookie is for or what the value represents. [8] The cookies are a little text files that are stored in the browser and will remain there until they are deleted. While the previously described IP address and HTTP referring are called stateless, meaning that the information in them always varies and is processed on one-time request basis only, the cookies are called stateful, because they are saved to the browser, constantly helping to accumulate user data about browsing. With each request the data from the cookie is being forwarded to the third-party who initially set the cookie to browser and who is now processing it for more accurate personalization of the content. It is also important to note that besides third-party cookies, some of the web services might use functional cookies, which are strictly necessary to use the service – they usually hold user login and session information and do not forward the information to a third-party.

Tracking pixels

Tracking pixels are very small invisible pictures (usually 1x1 pixels) that are placed on the websites or even in e-mails. When a website is loaded, also the invisible picture is loaded, and the loading of the picture will initiate a request which can forward information to a third-party that the website has been loaded or to the e-mail sender that the e-mail has been read by the recipient. Along with the information about the picture being loaded, additional information could be extracted from the initiated request – user-agent string, user's IP address and time of the request.

User agents

User agents are simple lines of text, that are included in browsers and contain information about user's operating system and used browser. Today, user agents generally identify themselves to servers by sending a User-Agent HTTP request header field along with each request. Ideally, this header would give servers the ability to perform content negotiation, sending down exactly those bits that best represent the requested resource in a given user agent, optimizing both bandwidth and user experience. In practice, however, this header's value exposes far more information about the user's device than seems appropriate as a default, on the one hand, and intentionally obscures the true user agent in order to bypass misguided server-side heuristics, on the other. For example, a recent version of Chrome on iOS identifies itself as:

User-Agent:	Mozilla/5.0 (iPhone; CPU iPhone OS 12_0 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) CriOS/69.0.3497.105 Mobile/15E148 Safari/605.1

While a recent version of Edge identifies itself as:

User-Agent:	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/68.0.2704.79 Safari/537.36 Edge/18.014

There is quite a bit of information packed into those strings (along with a fair number of lies). Version numbers, platform details, model information, etc. are all broadcast along with every request, and form the basis for fingerprinting schemes of all sorts. [9] The "lies" here refers to the possibility of the user-agent altering by individual vendors to overcome certain bugs or errors.

Browser fingerprinting

Browser fingerprinting checks certain properties of the browser and the computer it is run on and tries to calculate a unique identifier from the gathered information. [10] It is a highly accurate way to identify an online user by extracting information and analysing this information to create a unique user profile. In order to cross-browser fingerprint a user, a website operator has to choose some browser-independent features as a basis of identification. These are likely to include a set of, but are not limited to, the following browser- and system-dependent properties:

- Networking information. Since HTTP requests are sent via TCP/IP, the server always sees the IP address (and hostname), and the TCP port number. The location of the client can also be inferred from the IP address in most cases.
- Application layer information. The user agent string is a standard HTTP header and is sent with every request. It contains the type and version of the browser; the name and version of the operating system; the type and version of the layout engine (e.g. Gecko for Firefox); and the names and versions of certain extensions. It must be noted that some browsers (e.g. Opera) are extremely verbose about their version, so even minute patches change their UASes. Finally, the HTTP request usually contains a language preference code (e.g. 'en-us'), too.
- Information gained by querying the browser. JavaScript programs have access to the list of fonts, plugins (along with their version numbers), screen resolution, and the time

zone. Additionally, some vulnerabilities may allow access to browser history [11] or to other client-side databases that are otherwise inaccessible for the visited website. [10]

There are many websites that demonstrate the uniqueness of the user by creating user's browser fingerprint. One of those is amiunique.org[6], which will show all the data that can be gathered about the users from their browsers and is thus available to any website that the users visit. The browser fingerprint created by amiunique.org during the current work identified more than 1400 different data points about the user, starting with user agent data, installed fonts and operating system, and ending with installed plugins, screen resolution and computer battery status.

Supercookies

Not to be fooled by the name, the supercookies are technologically nothing like cookies. The supercookies are today one of the biggest challenges in data protection, user privacy, profiling, and data collection. Whereas the previously described cookies were stateful and physically existing in the user's browser, meaning that they could be deleted or blocked, the supercookies are stateless and will do the user profiling and data collection on the fly. Combining different stateless and stateful technologies, a website can identify, using a third-party service, a unique web-user and match it with already existing entry in the third-party database. Therefore, the data is not collected into a text file (into a cookie) in a user's browser but is sent straight into a third-party database. This combination of stateless technologies that can identify the user on the go is given a name of 'supercookie' and does not refer to a one specific technology, making it more challenging to eliminate and to ignore. It has also been noted that supercookies are able to recreate normal cookies when the user has deleted them and will contain all the previous information after recreation. [12]

1.4.2. User awareness

With the increased use of technological devices, activities that realize data collection increase, reaching all segments of society. As such, it becomes necessary to better understand this process which often does not occur in a perceptible way to the user who has low awareness about when, how, and where it occurs. Since data relating to such actions may reveal

[6] https://amiunique.org/fp

individuals' personal information, threats to privacy emerge. [13] To internet users around the world the data collection is inevitable, and it is also a factor that keeps the internet running and growing. Although different regulations have been accepted, the best way to ensure that the user data is protected and used only for purposes in favour of the user is to raise user knowledge so that the user itself acknowledges the possible dangers and would know, on high-level at least, how different internet technologies work. The user might not be aware, and even not interested, about how cookies are stored in the browser and how they help to collect information, but they should at least know what it means to accept or reject a cookie policy whenever a pop-up appears on a visited page. As mentioned, there are a lot of different technologies that collect user data in the background without a user knowing, but a good start to gain and raise awareness is to at least be curious about the things that a user can see and interact with – cookie policy pop-ups, privacy policies, personalized ads of the items previously browsed, personalized news about the topics that interest the user, and so on.

The best ways to raise user awareness about the data privacy and data collection is by demonstrating the data flows that the users themselves generate and by analysing the data that the users themselves provide – that is also one of the goals of the current thesis. The analysis of 150 most visited Estonian websites by the Estonian users should give a good overview of data flows and services that those websites use in order to offer services that the end-users consume. Although a thesis is not a best medium to present the ideas and possible threats to the everyday internet users, it hopefully raises the awareness of those who are involved in the process of creating, assisting, and evaluating.

1.4.3. Tracking prevention measures

A user seeking to avoid being followed around the Web must pass three tests, each trickier than the previous one. First, find appropriate settings that allow sites to use cookies for necessary user interface features, but prevent other less welcome kinds of tracking. Second, learn about all the kinds of supercookies, perhaps including some quite obscure types, and find ways to disable them. And third, fingerprinting. [14]

There are several ways in which aware users can lower the risk of their data being sent to and used by unwanted parties. Most common measures are described in this section. It is important to emphasize the awareness, as this serves as a basis to protect a user – only if the user knows

the possible ways of data collection and the threats that it poses, there is a chance for the user to protect itself.

VPN

VPN, which stands for Virtual Private Network, is a way to cover user's IP address and to avoid at least some sort of supercookies. VPN creates a connection between a user's computer and another computer, which could be located anywhere in the world. While the VPN connection is active, the network requests that the user does through its browser will have the other computer's IP address and will therefore prevent direct tracking of the user's own device by IP. This also prevents some supercookie algorithms, as IP address is one of the key components in many of those.

The pros of using a VPN connection are that it is easy to use and that the users might be already somewhat familiar with using VPN as many organizations use VPN connections to allow working from distance. The cons are that the services are usually not free and that the computer in the other end of the connection might be tracked as well, which would be able to identify the IP address of the computer where the connection was made from and the requests that the user makes. The key to find a reliable and trustworthy VPN provider is to choose one carefully, and ironically Google is a good way to determine that.

Browser extensions

Browser extensions might help the user block different kind of data that would be transferred to third-party services. There are browser extensions that allow user to edit or turn off HTTP referrers, extensions that automatically disable all non-functional cookies on a website or extensions that prevent sending analytics data to the third-party services. All these kinds of extensions serve the purpose of data privacy and installing some of those will get the user one step closer to sending less data to unwanted parties.

Cookie policies

Without knowing the purpose of cookie policy pop-ups and the concept of cookies, the user is vulnerable to saving unwanted and unnecessary cookies to the browser. As mentioned before, cookie policies are one of the few things that the users can see and interact with, however this has not always been so – the Cookie Consent was enforced through GDPR in May 2018 to all

websites that have visitors from United Kingdom or European Union and while Cookie Consent is not required in the United States, there is still a federal law that places strict restrictions on the use of cookies. So, if before there was no restrictions to using cookies on a website, it is now required to ask user's permission about using cookies. There are different kind of cookies, that might be saved to the browser when a user is visiting a website. The functional cookies are all that the user actually needs to consume the service, but often there are different analytics and advertising cookies included, from where it would be recommended to opt out. The cookie pop-ups will usually introduce two options to users: *Accept all cookies* or *Customize cookies* (both options could be rephrased differently on different sites). If the user clicks *Accept all cookies*, which is encouraged by the websites, making the button big and colourful as seen in Figure 11, he will accept saving all kinds of different cookies to the browser, allowing the transfer of user's data to the third parties.

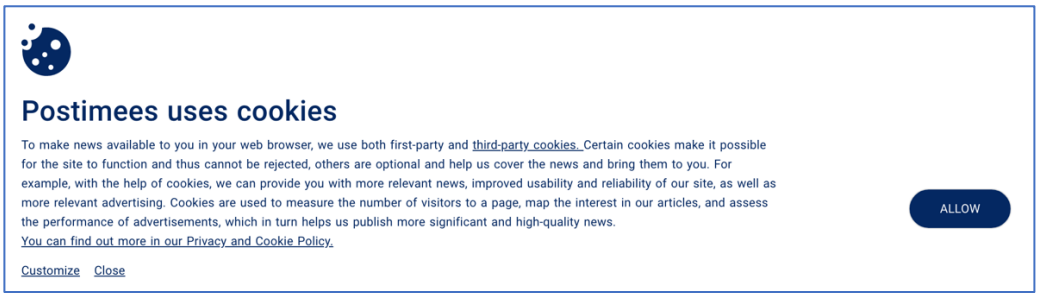


Figure 11. Postimees.ee cookie consent pop-up

Most of the unaware users do press this button to get the pop-up out of the way fast and start consuming the content. The recommended option, however, would be to click *Customize cookies*, in order to opt out from all cookies that are not functional.

Private browsing

Private browsing is a web-browsing mode that is offered by most of the popular browsers. This mode does not hide user's IP address from the website it visits and is mostly used to hide internet usage traces locally, meaning that it won't retain temporary session data [15], including the deletion of cookies from the browser after each session. This means that every time a user visits a website while in private browsing mode, the website won't recognize the user by the data previously stored in the cookies and rather acts as a user is visiting the website for the first time. After closing the browsing session, all the cookies will be deleted and if needed, the user can start a new session, without webpage knowing that the same person is visiting it, which prevents personalized content and ads. It is still important to emphasize that other tracking

mechanisms, like supercookies and browser fingerprinting, can still be used by the websites and third-party services browsing in private browsing mode to identify the user.

Browsers

Accessing the internet nowadays has become nearly inevitable, and web browsers remain the most popular tool to do so. The increased amount of web browser users and their aspiration to achieve paramount personal privacy has pushed developers to devise different ways to fulfil the users' need for anonymity and seclusion. [15] While the developers of Google's Chrome browser tend to talk a lot about privacy needs and updates, it still serves one big corporation that mainly depends on the user data. By using Chrome, not to mention logged in mode in Chrome, there is no doubt that the user's internet usage data is collected more often and more thoroughly.

One thing to keep in mind while choosing an internet browser, would be the ownership of the product. If a browser is not-for-profit or open source, it would be already a step towards user privacy. Moreover, lately there have been browsers developed specifically meant to support user privacy by focusing on ad blocking, fingerprinting prevention, cookie control, etc. – one of them, for example, is the Brave browser[7].

Tor browser is also something worth mentioning, as the connection from the browser to the website is initiated through several other computers in the Tor network. That enables user to connect to a website through another computer and the tracking technologies cannot be applied to the user directly. Although Tor browser seems like a best way to prevent tracking, many sites prohibit access through a Tor network computer, which does not make it very popular among average end-users.

1.5. Business models and their reliance on third-party services

In the early days of the web, content was designed and hosted by a single person, group, or organization. This is no longer so as webpages are increasingly composed of content from myriad unrelated "third-party" websites in the business of advertising, analytics, social networking, and more. Third-party services have tremendous value: they support free content and facilitate web innovation. But third-party services come at a privacy cost: researchers, civil

[7] https://brave.com/

society organizations, and policymakers have increasingly called attention to how third parties can track a user's browsing activities across websites. [16]

Whenever a business enterprise is established, it either explicitly or implicitly employs a particular business model that describes the design or architecture of the value creation, delivery, and capture mechanisms it employs. The essence of a business model is in defining the manner by which the enterprise delivers value to customers, entices customers to pay for value, and converts those payments to profit. [17]

Since the internet has become available for the public use, the business models of traditional businesses, who have moved their operations partly or fully to internet, have changed a lot. The internet has also enabled the emerging of new models that the traditional businesses have adapted. Even the term "business model" was considered a buzzword a few years ago and has only recently been acknowledged fully in the academic literature. [18] Besides traditional businesses, new models have created new type of businesses and the growth and innovation is constantly expanding and evolving. Most of those new internet-enabled business models are today depending on third-party services that allow the original services to focus on their original idea or content. Using third-party services, the business can either make its main online operations more efficient or even create additional flows of revenues. Some traditional models, like advertising, subscriptions, and commerce, have been around for a long time, but the internet has broadened each of those areas a lot and created many new opportunities for many companies. As mentioned before, while the traditional models have changed, also new models have emerged.

Hereby, a brief explanation of different business models is presented along with data from the practical analysis part. Seeing how different models function, allows us to analyse how third-party services fit into the models in the first place and how big role third-party services play in them.

First, we categorized our 150 most visited Estonian websites by their business models. Each website was assigned to a category based on the business model that is used to generate the main part of the website's revenue. As a business model of a website is not publicly available information and is rather considered a business secret, the only way to categorize the websites

is by evaluating the revenue sources visually by visiting a website directly – the general functionality of the website, number of ads, number of links referring other websites, free-trials and partly available content are good indicators to take into consideration when doing the categorization.

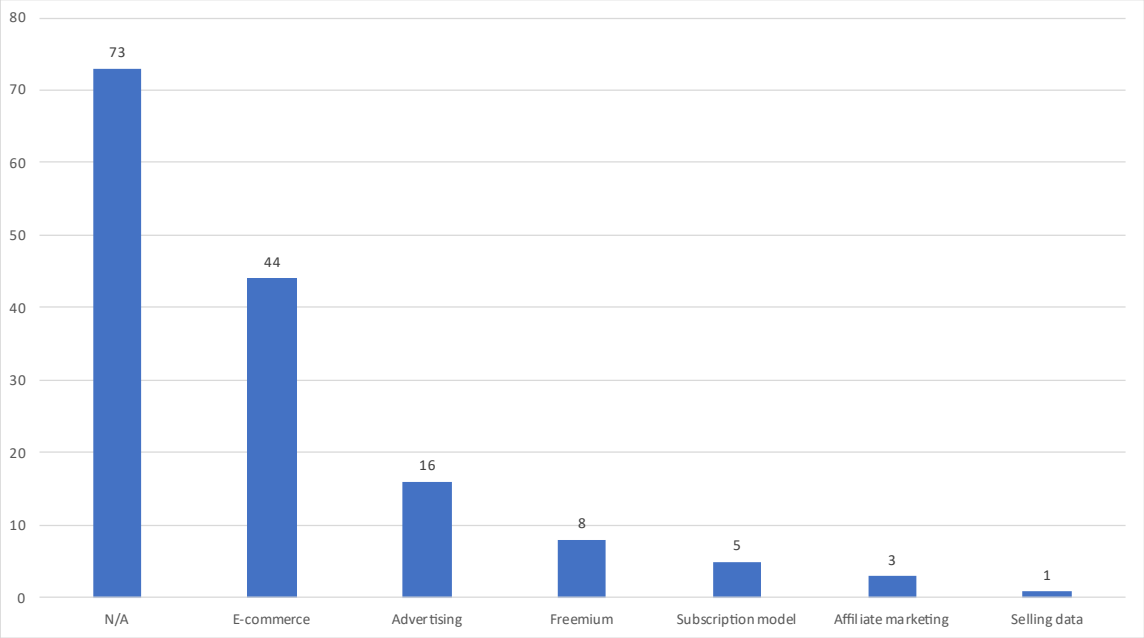


Figure 12. Business models of analysed websites

Figure 12 shows the categorization of our 150 websites by business models. Nearly half of the analysed websites, 73 with category N/A (not applicable), do not generate their revenue directly from website users, meaning that they are either public or governmental institution websites that are financed by the government and additional revenue models are not required or not even allowed, or they just represent traditional businesses that generate their revenue outside the web environment. Although these websites do not generate their revenue directly from website users, it does not mean that they do not use third-party services or that they do not send user's data to the third-party providers.

Based on the categorization above, business models are introduced and explained in following subsections, as well as the relations of such models with third-party services. For each business model, we present as an example one website from our 150 analysed websites to demonstrate the third-party services that websites using these revenue models might use. It is important to note that a website might use several different revenue models at the same time to increase its revenue, but for each model most relevant type of website is chosen, based on the main revenue source.

1.5.1. Advertising

Advertising as a business model might be one with the biggest changes and impacts. It has gone from product-based approach to customer-based approach and from impersonal advertising to more personalized advertising during its evolution.

The greatest change came with online advertising era in the beginnings of 1990s. That is when advertising started to solve customer problems and became customer-based rather than being product-based. As the services and websites were also able to gather various data points about the users, the advertising became more personalized. The personalization and customer-based approach has been perfected throughout the decades and we are now facing personalized ads whenever we use the web. That is where the third-party services come into play and make it possible for the websites to recognize user patterns, locations, interests, and various other factors, to offer personalized ads, products, and content. The amount of data points, that for example Google or Facebook collects per person, is huge and is the core of success for those tech giants. An experiment by Daily Mail[8] reporters revealed that a user's internet history over the 12 months stored by Google was the equivalent of 569,555 pages of A4 paper. [19] By sharing the insights from the collected data with their customers (different websites and services) through third-party services, they make it possible for the websites to grow revenue, number of visitors or purchases with very little effort.

We present *soov.ee* as an example of a website that uses advertising as its main revenue stream. In *soov.ee* website sellers and buyers can make postings for free and can contact each other directly without paying commission to the website. The site has high visitor count and uses business models that benefit from it. Table 8 below shows third-party request domains for the requests that were made when visiting the landing page of *soov.ee*.

Table 8. Third-party request domains of soov.ee

#	Domain	Uses
1.	gemius.pl	audience measurement
2.	google-analytics.com	audience measurement
3.	ampproject.org	code, hosting
4.	createjs.com	code, hosting

[8] https://www.dailymail.co.uk/home/index.html

5.	googleapis.com	code, hosting
6.	gstatic.com	hosting
7.	googletagmanager.com	hosting, tag manager
8.	doubleclick.net	marketing
9.	google.com	marketing
10.	google.ee	marketing
11.	googlesyndication.com	marketing
12.	facebook.com	marketing, social media
13.	facebook.net	marketing, social media
14.	googletagservices.com	tag manager

As we can see, marketing third-party services are highly represented on *soov.ee* website as expected for a site that uses advertising as its main revenue model. As many different types of websites might use advertising as one of their revenue sources, for example news sites, the example website falls into the category of making most of its revenue from the advertisements. This category is most of the times represented by sites that offer free services or data to the user and therefore have high traffic.

1.5.2. Freemium

Freemium is a rather new revenue model concept and has emerged along with popularity of internet and web applications in particular. Web applications are kind of websites that are designed for interaction with end-users and usually require logging in as opposed to traditional websites that contain and serve static content. Freemium model makes the core functionalities of a web service available for free, but there are additional features that can be used only with premium account, which could be acquired by paying the premium fee. The free part of the service is mostly used to grow customer base, through word of mouth, referral networks or organic search marketing, who are then offered premium priced value-added services or an enhanced version of the service. [20]

The freemium model, third-party services and advertising are in many cases closely related, because many services use advertising in their freemium model to make at least some revenue from the free users. Those ads are, of course, referring to user's interests and needs that are extracted by the third-party services offering the ads. So, in this case everyone wins, except the

end-user – the first-party webpage is offering ads for the users with free account, also gaining revenue from them, and will gain even more revenue when user switches to premium and the third-party service benefits from showing the advertisement.

Additionally, many sites that use freemium model, use audience measurement third-party services in order to find out what interests the end-user the most on their website and can use that information in order to leverage user to use their services by approaching the user personally with certain content.

Freemium model is mainly used by web applications. Table 9 shows an example of the third-party requests on a web application, *merit.ee*, landing page.

Table 9. Third-party request domains of merit.ee

#	Domain	Uses
1.	intercom.io	audience measurement, marketing
2.	intercomcdn.com	audience measurement, marketing
3.	fontawesome.com	fonts
4.	cloudflare.com	general, hosting, security
5.	gstatic.com	hosting
6.	google.com	marketing

It is important to note that analysing a web application landing page, like done in the current case, may not reflect the full list of third-party services used in the application and more accurate results would be revealed when logged in as a registered user.

1.5.3. E-commerce

After visiting websites that use e-commerce as their main revenue model, it becomes clear that besides their main revenue model there are many signs of other business models and therefore different kind of additional third-party services are used. Advertising, audience measurement, social media and tag management services are very commonly used to offer suitable ads, measure visitor count and visitor behaviour, and to offer recommendations to the user. While the first-party websites often use third-party services to display ads on their sites for generating revenue, they also are the sources for the ads themselves, so that the user can see their items on a totally different website.

In order to demonstrate the diversity of the third-party services on e-commerce sites, we'll extract the different third-party requests that have been done through one of the analysed e-commerce websites – *kaup24.ee*.

Table 10. Third-party request domains of kaup24.ee

No	Domain	Uses
1.	google-analytics.com	audience measurement
2.	tvsquared.com	audience measurement
3.	gstatic.com	code, design optimization
4.	searchnode.io	code, design optimization
5.	googleapis.com	code, hosting
6.	newrelic.com	design optimization
7.	nr-data.net	design optimization
8.	cloudflare.com	general, hosting, security
9.	googletagmanager.com	hosting, tag manager
10.	creativecdn.com	marketing
11.	doubleclick.net	marketing
12.	google.com	marketing
13.	google.ee	marketing
14.	googleadservices.com	marketing
15.	owox.com	marketing
16.	pigugroup.eu	marketing
17.	teads.tv	marketing
18.	facebook.com	marketing, social media
19.	facebook.net	marketing, social media

As seen from Table 10, many different types of third-party services are represented on an e-commerce site, which indicates that in many cases selling products might not even be the biggest revenue source and considerable part of the revenue might be related to marketing and different advertisements on the website.

1.5.4. Affiliate marketing

Affiliate marketing has become a major strategic consideration for all companies participating in e-commerce. The concept has certain technological complexities that have been made simpler through the development of several network companies that facilitate the tracking and settlement of payments between various companies on the internet. [21] Affiliate marketing is probably something that most of the e-commerce sites are trying to use to increase their sales and to gain new customers. That is possible through third-party services that e-commerce sites use to display their items and different sales on someone else's website – the affiliate website. Affiliate marketing is classified as a type of online advertising where merchants share percentage of sales revenue generated by each customer, who arrived at the company's website via a content provider. Content provider, also referred to as affiliate, usually places an online ad (for example a banner or text link) at its web. When visitors click at the ad, they are redirected to merchant's website and affiliation is tracked by a cookie stored on visitors' computers. [22]

The third-party service will be creating a link to a website, displaying items to the user that the user might be interested in. Whenever the user clicks on the link, the third-party request will carry different kind of information about the user to identify what was the website where the user clicked the link, what time the user clicked it, and even what websites user has visited previously. After the user clicked the link and maybe even bought the advertised item, the e-commerce site gets their revenue from direct sales and the affiliate website will get their commission for leading the user to the sale. The third-party services make it therefore very easy, being the middleman between the e-commerce site and the affiliate website, keeping track of the clicks and commissions.

The best example from our list of 150 websites is *hinnavaatlus.ee*. It creates content in form of news, blogs, forums, and is probably making most of its revenue from affiliate marketing, because the main idea of the site is to compare the price of products on different e-commerce sites and lead user to the most suitable one. Third-party requests of this website can be seen in Table 11 below.

Table 11. Third-party request domains of hinnavaatlus.ee

No	Domain	**Uses**
1.	google-analytics.com	audience measurement

2.	pbstck.com	audience measurement
3.	quantserve.com	audience measurement
4.	hotjar.com	audience measurement, design optimization
5.	everesttech.net	audience measurement, marketing
6.	jquery.com	code
7.	googleapis.com	code, hosting
8.	taboola.com	content recommendation, marketing
9.	cloudfront.net	design optimization
10.	googlesyndication.com	design optimization
11.	gstatic.com	design optimization, hosting
12.	fastly.net	hosting
13.	googletagservices.com	hosting, tag manager
14.	1rx.io	marketing
15.	adform.net	marketing
16.	adnxs.com	marketing
17.	bepolite.eu	marketing
18.	bidswitch.net	marketing
19.	bidtheatre.com	marketing
20.	casalemedia.com	marketing
21.	criteo.com	marketing
22.	deepintent.com	marketing
23.	dotomi.com	marketing
24.	doubleclick.net	marketing
25.	google.com	marketing
26.	google.ee	marketing
27.	gumgum.com	marketing
28.	loopme.me	marketing
29.	mediarithmics.com	marketing
30.	onaudience.com	marketing
31.	pubmatic.com	marketing
32.	rlcdn.com	marketing
33.	rubiconproject.com	marketing

34.	sitescout.com	marketing
35.	smartadserver.com	marketing
36.	tapad.com	marketing
37.	tribalfusion.com	marketing
38.	turn.com	marketing
39.	yahoo.com	marketing
40.	zeotap.com	marketing
41.	360yield.com	unknown
42.	4dex.io	unknown
43.	a-mo.net	unknown
44.	acuityplatform.com	unknown
45.	ad4m.at	unknown
46.	adgrx.com	unknown
47.	adition.com	unknown
48.	adleadevent.com	unknown
49.	adnetmedia.ee	unknown
50.	adsrvr.org	unknown
51.	advertising.com	unknown
52.	bidr.io	unknown
53.	bnmla.com	unknown
54.	clarium.io	unknown
55.	contextweb.com	unknown
56.	cpx.to	unknown
57.	creative-serving.com	unknown
58.	crwdcntrl.net	unknown
59.	de17a.com	unknown
60.	erne.co	unknown
61.	exelator.com	unknown
62.	hinnavaatlus.ee	unknown
63.	id5-sync.com	unknown
64.	indexww.com	unknown
65.	ipredictive.com	unknown

66.	leadplace.fr	unknown
67.	mathtag.com	unknown
68.	onetag-sys.com	unknown
69.	playground.xyz	unknown
70.	quantcount.com	unknown
71.	sascdn.com	unknown
72.	sharedid.org	unknown
73.	simpli.fi	unknown
74.	splicky.com	unknown
75.	stackadapt.com	unknown
76.	themoneytizer.com	unknown
77.	themoneytizer.net	unknown
78.	tmyzer.com	unknown
79.	unrulymedia.com	unknown
80.	w55c.net	unknown

Table 11 presents us 80 different third-party requests that were made from *hinnavaatlus.ee* website. Marketing third-party services are the most used ones along with audience measurement and design optimization services, demonstrating a large variety of third-party services that a website can use.

Generally, affiliate marketing is mainly used by websites that create content – news websites, blogs, and forums.

1.5.5. Subscription model

While freemium model offers the core product or content of the website for free, and the user must pay for premium features only, the subscription model is more restrictive and mostly offer their core features and content for a regular fee. The subscription as a revenue model has been around already for a long time outside the web – newspapers, magazines, and gym memberships. With businesses moving to the web, many have quite directly adapted the same model online, where possible, and the most outstanding of them are online news sites.

Nowadays, when many of the news companies have moved their businesses to the web, they mainly use subscription model and advertisements to maximise their revenues, but due to high website traffic, other revenue streams are also used that mostly rely on third-party services. When during the Internet boom, a vast number of websites attracted users by offering them with large amounts of free information ranging from news, business data to sports statistics, the once well-sold business model of offering free content to secure advertisement revenues yielded rather disappointing results for most of the e-service providers. Increasingly, advertising revenues alone are insufficient to meet the bottom-line needs of a company for survival. [23] [24] [25] Therefore, the subscription model has become vital for such websites.

Table 12 below shows the examples of different third-party requests made from a news site *ohtuleht.ee*, whose main business model is subscription.

Table 12. Third-party request domains of ohtuleht.ee

No	Domain	Uses
1.	chartbeat.com	audience measurement
2.	chartbeat.net	audience measurement
3.	gemius.pl	audience measurement
4.	google-analytics.com	audience measurement
5.	googleapis.com	code hosting
6.	googleusercontent.com	content recommendation
7.	googlesyndication.com	design optimization
8.	gstatic.com	design optimization, hosting
9.	googletagservices.com	hosting, tag manager
10.	2mdn.net	marketing
11.	cintnetworks.com	marketing
12.	cxense.com	marketing
13.	doubleclick.net	marketing
14.	google.com	marketing
15.	google.ee	marketing
16.	facebook.net	marketing, social media
17.	pinterest.com	social media
18.	cdn-apple.com	unknown

19.	jwpcdn.com	unknown
20.	jwplatform.com	unknown
21.	jwplayer.com	unknown
22.	ocdn.ee	unknown
23.	onesignal.com	unknown

While there are no specific third-party requests defining that a website is using a subscription model, Table 12 indicates additional ways of generating revenue on a news website with marketing services. The subscription model usage could be only identified by visiting the website directly.

1.5.6. Selling data

There is a saying: "If you are not paying for the product, you are not the customer, you are the product being sold." [26] With so many data points available about users and their patterns, many companies gather that data and make selling insights of the data their main way to create revenue. It should be noted that the data itself (in most cases) is not sold, but only the results from analysing the data. This model is the core of social media and search engines, who are also the largest data traders in the world. Our analysis of 150 most visited Estonian websites clearly showed the dominance of companies like Alphabet and Facebook in the field of data gathering and that these companies were the main destinations for most of the data that was sent from first-party websites via third-party requests.

Selling data as a business model serves on a different side than the other previously mentioned models. Although social media sites and search engines have their own websites with high traffic, these companies also provide third-party services to other websites to share insights from the data that they have gathered about the user. That enables websites to personalize their content and to do more calculated decisions.

As we didn't have any social media site in our analysed websites list, in Table 13 below we will present the third-party requests of the only search engine that appeared in our list – *google.ee*.

Table 13. Third-party request domains of google.ee

No	Domain	Uses

1.	gstatic.com	hosting
2.	google.com	marketing

It is important to note that although data from the user's search engine use is saved to search provider servers, these companies rather rely on the data gathered by the third-party services that they provide to other websites. Also, as it can be seen from Table 13, not many third-party service requests were done, because the data ends up in the first party – data trader (in this case Alphabet's) – own servers.

2. Risks of using third-party services

This section will discuss risks that using third-party services might present, with a focus on one of the biggest issues that the third-party data collection is causing – data centralization. Different aspects, positive and negative, will be brought out, as well as how these will impact the individual end-user. The information in this section is based on the notes taken by the author when studying different topics for the current thesis.

Data centralization, or sometimes even called internet centralization, is caused by the combination of all the previously discussed business models, third-party technologies, and third-party services. The internet nowadays is so dependent on those services that there is hard to find a website which does not use any third-party service, causing a continuous and growing data flow into the servers of big internet corporations, as well as to the servers of the smaller companies, who might be still selling end-user data to some of the data brokerage companies with centralized data storages. This has caused the centralization of data which comes with many risks for the end-user, but with many benefits to the web businesses.

Although this section will focus mainly on the risks of using third-party services, there are still some benefits that could be brought out. Data brokering and state surveillance through data, which will be explained briefly in the following sections, are not transparent, and one might consider those activities shady or tending to exploit the end-user. Nevertheless, there are some aspects of those activities which could end up being beneficial for the users and for the society in general. Data centralization makes it easier to allow government agencies and law enforcements to gather user data, which is then analysed by their internal tools to prevent and solve crimes, discover money laundering cases and to identify individuals who might be a threat to others. One might argue if those benefits compensate the lack of intellectual privacy, but it is important to also acknowledge the opportunities provided.

2.1. Single point of failure

The single point of failure might be considered as a risk for the functioning of web businesses. When so many websites and businesses rely on third-party services and on the information of centralized data centres, the downtime of such centres, which could be either caused by cyber-attacks, physical- or programmatic malfunctions or even natural disasters, might result in the downtime of all the websites using those services. Therefore, that kind of malfunctions not only affect the end-user, but also the websites and services relying on centralized data and servers. The occurrences are getting more often, and the malfunctions of the hosting third-party services are the ones that cause most trouble resulting in website unavailability and service disruption. One very recent example of Akamai Technologies, content delivery and security third-party service, outage perfectly expresses concerns and problems for both, end-users and the web-services. The outage caused major banks and airlines to be affected in the middle of the day with websites being inaccessible, failing authentication and even having technical problems on the Hong Kong's stock exchange. [27]

2.2. Single point of access

As for the single point of failure, the core for the single point of access is the same – too much data in the administration of one single company. From one side this will generate another issue, called data brokerage, which will be explained in the next subsection, but from the other side it presents a good way to get hold of a lot of data from one single source.

The servers where all that centralized data is being held on, are certainly secured, backed-up, and distributed, but it is still in the administration of one company and can therefore breached through an organized cyber-attack to a one and single company. Moreover, the probability that the cyber-attacks could get to all the data that a company owns is very low, but even one server or database could hold immense amounts of data. The examples of such data breaches come to public very often, demonstrating the risk of holding so much data in one location: CAM4 2020 data breach with 10,88 billion records, Yahoo 2017 data breach with 3 billion accounts or Aadhar 2018 data breach with the biometric data of 1,1 billion people.[9] These are one of the biggest data breaches in the past 4 years, but much smaller breaches, and still significantly big, come to public almost every week.

[9] https://www.upguard.com/blog/biggest-data-breaches

Another issue to be brought out here is the state surveillance and the data that is requested by the state institutions from the companies "legally", opposed to illegal cyber-attacks. The best example of this is the PRISM project of the NSA, under which the National Security Agency has obtained direct access to the systems of Google, Facebook, Apple, and other US internet giants, according to a top-secret document obtained by the Guardian[10]. It allows officials to collect material including search history, the content of emails, file transfers and live chats, the document says. [28] This goes for the big tech companies, that we've not heard of selling the data before, but there is no doubt that such institutions use every source available for accumulating data – data brokers for example.

2.3. Data brokering

Big companies like Google or Facebook, have not been selling user data directly, as far as we know of, but they rather sell their insights, based on the analysis of collected data, through their own third-party services – the data accumulated from user's search or browsing history is analysed to personalize ads and content through third-party services. Therefore, they package the data into their own products, selling it that way.

The data brokers, from the other hand, are the companies that gather user data from various sources and sell it directly to the highest bidder – be it state institution, advertising companies or even criminals. As defined by Rieke *et al.*, a data broker is a company or business unit that earns its primary revenue by supplying data or inferences about people gathered mainly from sources other than the data subjects themselves. The data brokers do not usually reveal the direct sources of the data, but mostly it is acquired from three general categories: [29]

- Publicly available data
- Non-public data obtained through private contract
- Online tracking data

While the publicly available data is gathered from the sources where the user is responsible for entering the data himself, knowing it is publicly available to everyone, the non-public data can be obtained from various invisible sources invisible to the user – other data brokers, third-party service providers, and even first-party websites might be having a contract with a data broker directly. There are a lot of data brokers with different sizes, but the top data brokers do not fall

[10] https://www.theguardian.com/world

much behind the tech giants when speaking of the amounts of processed and obtained data. Based on the Federal Trade Commission report, data brokers collect and store a vast amount of data on almost every US household and commercial transaction. Of the nine data brokers, one data broker's database has information on 1.4 billion consumer transactions and over 700 billion aggregated data elements; another data broker's database covers one trillion dollars in consumer transactions; and yet another data broker adds three billion new records each month to its databases. Most importantly, data brokers hold a vast array of information on individual consumers. For example, one of the nine data brokers has 3000 data segments for nearly every US consumer. [30]

2.4. Service and content manipulation

While this is not, luckily, a very common practice, there have been incidents indicating that the data ownership can be used against the websites using third-party services and is therefore worth mentioning as one of the risks of data centralization. If a company owns a lot of data, it has good leverage on the first-party websites that use and rely on the data holders' third-party services. An example is provided by Hill, who was told by Google that a publisher's search results might suffer if the publisher didn't include certain Google provided buttons to their articles. The real number of such incidents is unknown, partly because the traces could be easily concealed by the data owners themselves as brought out by Hill, whose initial article on the issue was removed from the search engine all along: "But the most disturbing part of the experience was what came next: Somehow, very quickly, search results stopped showing the original story at all. As I recall it – and although it has been six years, this episode was seared into my memory – a cached version remained shortly after the post was unpublished, but it was soon scrubbed from Google search results." [31]

2.5. Service providers as competitors to the first-party websites

Third-party services make the life easier for the first-party websites using them. As already mentioned several times, many businesses even depend on those services, building whole revenue streams around them – advertising or affiliate marketing for example. It is also important to understand that many of the benefits offered to the first-party websites by third-party services come from the data collected from the users – centralized data that could be analysed and used to offer better content to the end-user. As the data is often very centralized and the companies that own the data and also offer third-party services, like Google or

Facebook, might use the collected data also in their own services. If Facebook displays ads, personalized offers, or news articles in user feed, it uses data gathered about the user via the third-party services they offer to the websites. Moreover, Facebook is now a competitor to the website to whom it also offers third-party services because the news and offers are provided by Facebook itself and not by the original first-party website, whose revenue streams might be built on visiting user flow, ads displayed and clicks made. The first-party website is still dependent on third-party services and can't stop using them, but at the same time the site is losing some part of the revenue because of them. This whole situation is creating an infinite loop of influence and does not leave much room for the first-party websites to play with.

2.6. Threat to intellectual privacy

Intellectual privacy is the ability, whether protected by law or social circumstances, to develop ideas and beliefs away from the unwanted gaze or interference of others. Surveillance or interference can warp the integrity of our freedom of thought and can skew the way we think, with clear repercussions for the content of our subsequent speech or writing. The ability to freely make up our minds and to develop new ideas thus depends upon a substantial measure of intellectual privacy. [32]

The fact that the end-user data is gathered and centralized makes it easier to identify a person by its web-profile, which brings the users to two major problems:

1. The chance of being persecuted in real life for the things expressed on the internet, influencing the importance of the free speech.
2. The lack of anonymity and privacy, along with the fear of persecution, represses the expression of new ideas and thoughts that would need a discussion in order to take off or thrive.

Both of those points are very well summarized by Richards: "Intellectual privacy is vital to a robust culture of free expression, as it safeguards the integrity of our intellectual activities by shielding them from the unwanted gaze or interference of others. If we want to have something interesting to say in public, we need to pay attention to the freedom to develop new ideas in private, either alone or with trusted confidants. Free speech thus depends upon a meaningful level of intellectual privacy, one that is threatened by the widespread distribution of electronic records of our intellectual activities." [32]

Conclusions

In this work 150 most visited Estonian websites were analysed for third-party requests using a tool called webXray. All third-party requests initiated from these 150 websites were categorized along with the websites themselves and the results were compared to determine the amounts of third-party requests made from each type of website. Also, a business model for each website was manually determined and analysed in the context of third-party services that are used by the website of the particular business model.

After analysing the initial landing pages of the websites, in total 1713 third-party requests were observed. We found out that marketing (mainly dominating on e-commerce and news sites), hosting (dominating on e-commerce, news, and public sites), and audience measurement (dominating on e-commerce, news, and public sites) third-party requests were the most popular ones. E-commerce sites and news sites are the biggest consumers of third-party services – 901 total requests from e-commerce sites and 669 total requests from news sites. End-user data was found to be forwarded to only a handful of corporations – 39% of all requests ended up in Alphabet's (Google's) servers and 11.2% in Facebook servers – and dominantly inside the borders of one country – the United States.

To consider that only 150 websites were analysed, 1713 requests that all carry some sort of end-user data to relatively centralized data centres is a frightening fact and should concern every visitor of those websites. As found out, the most popular sites, like e-commerce and news sites, are the ones that hold the biggest threat to end-user data privacy and there is not much an end-user could do to prevent this. The websites and their business models are deeply reliant upon third-party services – simplifying a website's infrastructure management, content creation and security, while also creating extra revenue streams – making the usage of those services inevitable. While the data centralization is also a threat to the websites, the benefits from using third-party services are greater for the website owners. As there is little that the end-user could do about data collection, more emphasis should be put on to what the websites could do about it. Usage of more decentralized services, supporting of local services, investing more into the security and privacy of the end-users are just a few principles that the websites should follow.

Sending end-user data, mostly without average user even knowing, via requests to third parties is a result of the evolution of internet which has made data the most valuable asset to own. It is no coincidence that the companies with the biggest valuation are the companies with the most data and as the virtual world is progressively merging with the reality, these are the companies that not only control the online behaviour of a user, but also their reality.

References

[1] T. Libert and R. Binns, “Good News for People Who Love Bad News: Centralization, Privacy, and Transparency on US News Sites,” in *11th ACM conference on Web Science*, Boston, MA, USA, 2019.

[2] M. Saric, “Blog: How we use web analytics to measure our startup's progress and make better decisions,” 12 August 2020. [Online]. Available: https://plausible.io/blog/analytics-metrics-definitions.

[3] D. Siroker and P. Koomen, A / B Testing: The Most Powerful Way to Turn Click Into Customers., New Jersey: John Wiley & Sons, Inc., 2013.

[4] D. Komosny, M. Voznak and S. U. Rehman, “Location Accuracy of Commercial IP Address Geolocation Databases,” *Journal of Information Technology and Control,* vol. 46, no. 3, pp. 333-344, 2017.

[5] P. Hillmann, L. Stiemert, G. D. Rodosek and O. Rose, “Modelling of IP geolocation by use of latency measurements,” in *11th International Conference on Network and Service Management (CNSM)*, Barcelona, 2015.

[6] B. Eriksson, P. Barford, J. Sommers and R. Nowak, “A Learning-Based Approach for IP Geolocation,” in *PAM 2010: Passive and Active Measurement*, Zurich, 2010.

[7] E. Katz-Bassett, J. P. John and A. Krishnamu, “Towards IP geolocation using delay and topology measurements,” in *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement (IMC '06)*, Rio de Janeriro, 2006.

[8] K. M. David, “HTTP Cookies: Standards, privacy, and politics,” *ACM Transactions on Internet Technology,* vol. 3, no. 2, pp. 151-198, November 2001.

[9] M. Taylor, Y. Weiss and M. West, “Draft Community Group Report, 10 June 2021,” 10 June 2021. [Online]. Available: https://wicg.github.io/ua-client-hints/.

[10] K. Boda, A. M. Földes, G. G. Gulyás and S. Imre, “User Tracking on the Web via Cross-Browser Fingerprinting,” in *Information Security Technology for Applications. NordSec 2011. Lecture Notes in Computer Science.*, 2011.

[11] G. Wondracek, T. Holz, E. Kirda and C. Kruegel, “A Practical Attack to De-anonymize Social Network Users,” in *2010 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 2010.

[12] A. Soltani, “Flash Cookies and Privacy II,” 11 August 2011. [Online]. Available: https://ashkansoltani.org/2011/08/11/respawn-redux-flash-cookies/.

[13] E. P. Alfonso and R. C. G. Sant'Ana, “Privacy awareness issues in user data collection by digital libraries,” *IFLA Journal,* vol. 44, no. 3, pp. 170-182, 2018.

[14] P. Eckersley, “How Unique Is Your Web Browser?,” in *Privacy Enhancing Technologies. PETS 2010. Lecture Notes in Computer Science, vol 6205.*, Berlin, 2010.

[15] A. R. Mahlous and H. Mahlous, “Private Browsing Forensic Analysis: A Case Study of Privacy Preservation in the Brave Browser,” *International Journal of Intelligent Engineering and Systems,* vol. 13, no. 6, pp. 294-306, 2020.

[16] J. R. Mayer and J. C. Mitchell, “Third-Party Web Tracking: Policy and Technology,” in *2012 IEEE Symposium on Security and Privacy*, San Francisco, 2012.

[17] D. J. Teece, “Business Models, Business Strategy and Innovation,” *Long Range Planning,* vol. 43, no. 2-3, pp. 172-194, April-June 2010.

[18] E. V. Reime, “Exploring the Freemium Business Model,” Oslo, 2011.

[19] B. Ellery, J. Bucks and J. Hurfurt, “Worried Facebook has far too much data about you? Google has enough to make a 7ft 9in pile of paper every TWO WEEKS (which they then sell to the highest bidder!),” *The Mail on Sunday,* 21 April 2018.

[20] F. Wilson, “My Favorite Business Model,” 23 March 2006. [Online]. Available: https://avc.com/2006/03/my_favorite_bus/.

[21] D. L. Duffy, “Affiliate marketing and its impact on e-commerce,” *Journal of Consumer Marketing,* pp. 161-163, 2005.

[22] J. Gallaugher, P. Auger and A. Barnir, “Revenue Streams and Digital Content Providers: An Empirical Investigation,” *Information & Management,* pp. 473-485, August 2001.

[23] D. Addison, “Free Web access business model is unsustainable in the long term,” *Marketing,* pp. 9-10, August 2001.

[24] R. Dewan, M. Freiner and J. Zhang, “Management and valuation of advertisement-supported web sites,” *Journal of Management Information Systems,* vol. 19, no. 3, pp. 87-98, 2003.

[25] E. Turban, D. King, M. Lee and D. Viehland, Electronic Commerce 2004: A Managerial Perspective, 3rd Ed., New Jersey: Prentice Hall, 2002.

[26] R. Serra and C. F. Schoolman, Directors, *Television Delivers People.* [Film]. 1973.

[27] J. Diaz, “Airlines, Banks And Other Companies Across The World Hit In The Latest Web Outage,” *NPR,* pp. https://www.npr.org/2021/06/17/1007496797/airlines-banks-and-other-companies-across-the-world-hit-in-latest-web-outage?t=1627401769805, 17 June 2021.

[28] G. Greenwald and E. MacAskill, “NSA Prism program taps in to user data of Apple, Google and others,” *The Guardian,* 2013.

[29] A. Rieke, H. Yu, D. Robinson and J. Von Hoboken, “Data brokers in an open society,” Open Society Foundation, London, 2016.

[30] FTC, “Data Brokers: A Call for Transparency and Accountability,” Federal Trade Commission, 2014.

[31] K. Hill, “Yes, Google Uses Its Power to Quash Ideas It Doesn’t Like—I Know Because It Happened to Me,” 31 8 2017. [Online]. Available: https://gizmodo.com/yes-google-uses-its-power-to-quash-ideas-it-doesn-t-li-1798646437. [Accessed June 2021].

[32] N. M. Richards, “Intellectual Privacy,” *Legal Studies Research Paper Series,* p. 387, August 2008.

Appendix

List of figures

List of tables

Non-exclusive licence to reproduce thesis and make thesis public

I, Norbert Metsare,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,
   Third-party services and their usage on the most visited Estonian websites,
   supervised by Arnis Paršovs.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Norbert Metsare
04/08/2021