Tracking Cookies in the European Union, an Empirical Analysis of the Current Situation.

Master Thesis

by Elsa Rebeca Turcios Rodríguez [4597818]

MSc. Management of Technology

Delft University of Technology Faculty of Technology, Policy and Management Section Organization & Governance

Delft, August 2018

TU Delft

Tracking cookies in the European Union, an Empirical Analysis of the Current Situation.

Master thesis submitted to Delft University of Technology
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in Management of Technology

Faculty of Technology, Policy and Management

by Elsa Rebeca Turcios Rodríguez
Student number: 4597818
To be defended in public on [08 28 2018]

Graduation committee

Chairperson: Prof.dr. M.J.G. (Michel) van Eeten, Section Organization & Governance (TBM)
First Supervisor: Dr. Ir. H. (Hadi) Asghari, Section Organization & Governance (TBM)
Second Supervisor: Dr. S.T.H. (Servaas) Storm, Economics of Technology and Innovation (TBM)
External Supervisor: (Rob) van Eijk, Ph.D Candidate Dual PhD Centre, Leiden University.

Table of Contents

Table of Figures

Acknowledgement

First of all, I would like to express my gratitude to my mother, Evelyn, to whom I owe everything I am. Second, thanks to Bayardo who unconditionally loved and supported me in the day to day upsides and downs of this journey. Third, thanks to my sister, Linda, who makes me want to be a better person, so she can have an example to defeat. Last but not least, thanks to my grandpa Miguel, Jenny, Lucy, Victoria L, Victoria M, and the rest of my family who supported my mother and me while I was here. Thank you all for your endless love and encouragement.

Secondly, I would like to thanks to my daily supervisor, Hadi, who guided me through this process and taught me that research requires reflection. Hadi, thanks for all the patient, insight, time, and discussions to improve the quality of this work. Thanks also for suggesting such as an interesting topic in the graduation portal, and the support on the data collection. Also, I have to express my gratitude to Michel and Servass whom with their vast experience always asked questions that encouraged me to think critically about my research. Last but not list, thanks to Rob, my external supervisor, who since the first day was willing to share his knowledge, ideas, and experience to enrich the content of this thesis.

Finally, I would like to thank all my friends from the Netherlands and Nicaragua who have supported me through messages, calls, chats, singing, and dinners. All members of the "Mandatory Group" and "ISC family". Also, special thanks to Rahul, Tanya, Akhil, whom since the very first quarter when we studied technology dynamics offered me their friendship, as Akhil well explained once 'it was meant to be'. Thank you all to share knowledge, tips, open your home and hearts, and for all the support during the thesis life and during the program.

Elsa Turcios, August 2018, Delft

This page was intentionally left in blank.

Executive Summary

The internet and World Wide Web (also known as the Web) allows its users to access vast amounts of information and resources at the comfort of their homes. Together with its growth have grown several services and common practices. One such practice is the integration of third-party companies, besides the websites' owners, into websites who are capable of tracing and collecting users' "online" data. The tracking is performed by placing "cookies" in the users' devices or other mechanisms while websites are browsed. Some of these (third-party) companies are analytics, front-end services, social integration, hosting platforms, market research, content provider, and companies involved in online behavioral advertisement. Due to this third-party integration, every time a website is accessed, the users are exposed to the risk of their "online" data being tracked for different purposes; tracking represents a threat to users' privacy as they can be used for malicious purposes. For example, the threats can be in the form of identity thief, price discrimination, and even government surveillance.

In the European Union (EU), "privacy" is deemed to be a fundamental right; hence, different legal instruments are in place to safeguard it. One such instrument relevant to tracking mechanisms is the E-Privacy Directive. The aim of the E-Privacy Directive is to increase and harmonize privacy protection across member states. Similar to any other socio-technical problem, privacy is also a complex problem that involves different actors, incentives, and evolution of the Web as technology. Hence, due to this complexity, the E-Privacy Directive required amendments through the years in an attempt to re-solve the privacy problem. In the near future, the E-Privacy Directive will become the "E-Privacy Regulation". However, there are different opinions and debate about what the E-Privacy Regulation should entail, for the variety in interpretation and application—or transpositions—of the existing Directive by the member states.

The main objective of the thesis, therefore, was to streamline the opinions and bring empirical evidence to the debate in an effort to yield recommendations for the "Regulation" understanding the legal and market forces surrounding the Directive. Hence, we sought to answer the main research question(s)—**What legal and market forces factors can better explain the presence of tracking cookies across the European Union, and how the E-Privacy Regulation reform can reduce tracking?**

To achieve the stated objective and answer our main question(s), we used Open Web Privacy Measurement, a framework developed by Princeton University, to simulate users visiting the homepage of websites, and we collected data about tracking cookies and cookies notices. We used the cookies as a proxy to measure tracking since it is argued to be one of the most commonly used tracking mechanisms in the existing literature, which was also confirmed with our results. Also, we used the Top Level Domains (TLD) of the websites as a proxy for their location and the transposition of the E-Privacy Directive they follow. We looked up the top 100 country-specific websites for 15 EU countries, and 5 control countries (Australia, Canada, Japan, Switzerland, and The United States), along with 200 top global websites with TLD .com and .org . In addition, we made a cross crawl from the 15 EU countries and the control countries to simulate users' locations for a total crawl of 35,325 websites. We counted 642,362 tracking cookies, 206,787 third party domains, and 217,183 Java Script calls to third-party companies in all websites. In addition, we collected our independent variables form secondary data, and we used the categories of websites as a proxy to business models' incentives to use tracking.

To start our research and to understand what legal and market forces could explain tracking presence, first, we examined if tracking existed. We found that tracking cookies were present in 81% of the websites in the selected countries not mattering users' location. Besides, top trackers such as Google and Facebook were present in member states, and trackers had a long tail. Consistent with Englehardt and Narayanan (2016) this implies that few companies can be encountered on daily basis by users, and they might be easy to regulate.

Next, we determined which laws and norms websites follow. Websites could follow the rules and norms from where users were located or where their companies were located. In addition, we needed to understand if the laws they decided to follow led to differences in tracking and cookies notices. Based on or our results, users' location did not play a role, so websites do not follow the laws where users are located, except for websites .COM. Larger companies that use Top Level Domain .COM use geolocation to adapt their websites to the rules of users' location. This result might indicate that larger companies have the capacity to interpret different local laws, and they have international strategies that allow them to adapt their websites. In contrast, we found that websites which have a defined target market, and which are based on a certain location, such as websites with Top Level Domains .NL, .DE, do not use geolocation, and follow local rules and norms of the target market they decided to serve. In both cases, our findings indicate, that following the local laws might be related to the managerial decision to serve specific markets, which implies following its norms.

In addition, we assessed if websites following local laws and norms of the market they serve leads to differences in tracking and cookies notices. Based on our findings, we determined that harmonization on tracking and privacy protection on member states is not achieved yet. Websites based in different member states present high and significant variability on tracking and cookies notices. The lowest presence of trackers was found for websites based in The Netherlands which have 32.6% less likelihood to have trackers, while the highest presence of trackers was found for websites based in UK which have 3.09 time higher relative risk of having a tracker. Moreover, we found that there are differences in cookies notices. Websites based in France has the highest banner presence with an incident rate ratio of 6.34, while the websites based in Romania have 56.3% less likelihood of having a banner. Moreover, as an additional observation was that on average the control countries have more tracking and less cookies notices than European Union countries.

These differences in tracking and cookies notices that we observed could be related to the different transpositions of the E-Privacy Directive in member states. In the European Union, countries have flexibility in implementing directives into national laws, this led to subtle, yet important, differences in the transposition of the provisions of the E-Privacy Directive. For example, in some countries consent needs to be explicitly provided by the users to install cookies in their devices, while in other countries consent might be implied. Also, some countries decided to emit guidance on how to implement the E-Privacy Directive, while others did not. However, the businesses' incentives to use tracking and market forces in a country could drive these differences too. Hence, finally, we examined the impact of the local provisions of the E-Privacy Directive and market forces on tracking presence across member states.

We determined that the different local transpositions of the E-Privacy Directive have led to differences in tracking. We have two different outcomes of the local provisions affecting the tracking mechanisms. We observed that the provision of (users' giving their) consent alone does not have an effect, but when controlled for the business models' or the website owners' incentives to use tracking, we have observed that websites located in countries that transposed explicit consent significantly decreases the likelihood of using trackers (by 15%). This might be explained because to gain consent,

websites need to provide information to users on what they will do with their data and the purposes of tracking. Hence, consent reduces the information asymmetry and Principal-Agent problem between websites and users. Also, we found that websites located in countries that impose fines significantly decrease the likelihood of using tracking by 36%, more than we observe without controlling for the businesses' incentives to use tracking which was 32%. A possible explanation for this result is that fines may act as a punishment for the companies that do not adhere to the norms thereby reducing the businesses' incentives to track. On the other hand, for websites located in countries that promulgated Guidance via their Data Protection Authorities, strikingly, a significant increase in tracking, of 30%, was observed, but when we controlled for business models' incentives to use tracking the magnitude of the effect was reduced to 12%. This result was not expected, and this might be possible explained due to the divergence in the Guidance from the Directive. Perhaps, the transposition of Directive into Guidance might have watered-down its rigidity only to enable businesses to exploit the context. Finally, websites located in countries in which they are expected to provide more information to users, regarding the purpose of data collection, reduce the use of tracking by 38.2% (When controlling per businesses' incentives), so this might suggest that reducing information asymmetry between users and websites help to discourage websites to exert tracking. These findings have a twofold implication. First, businesses' incentives are important to understand tracking. Second, there is an opportunity to reduce tracking harmonizing the provisions. Countries that transposed consent has less tracking than countries that did not, as well as countries that impose fines. This might suggest that requiring consent, imposing fines, and providing information to users are an alternative to discourage tracking in member states. In addition, since guidance has a contrary effect, this might suggest that how to provide guidance on how to comply with the future regulation might need to be revised.

While testing the market forces which could encourage or discourage tracking, three groups of companies' websites with different business models' incentives that led to different levels of tracking were identified. The foremost, the companies whose revenue streams are highly dependent on advertisement exert more tracking. Usually, these websites build an audience and give free content to them, and their revenue streams are highly dependent on monetizing their audience, in this group, we found news. Next, those companies whose revenue streams are slightly dependent on advertisement exert less tracking compared the first group. These type of business models have other sources of income besides advertisement, but they still might use ads to have additional income and/or promote their brands. In this group, we observed technology and computing, businesses, careers, hobbies and interest, home and garden, Science, education, and food and drink. Lastly, companies whose revenue streams are independent on advertisement exert the least tracking among the groups. Here we observed businesses that main aim is to provide information, respect the anonymity of users, and they are very specific. In this group, we found government, illegal content, non-standard content which include adult websites, and health and fitness.

We determined that the different business models' incentives to use tracking were even more powerful predictors than the local laws and norms the websites followed, and a combination of the local laws and norms and businesses' incentives to use tracking are even more powerful predictors of tracking presence than the local provisions (transpositions) of the E-Privacy Directive. However, we observed that the differences in transpositions do have an effect in explaining tracking presence. Therefore, these results suggest that businesses' incentives need to be better understood to avoid the so-called tragedy of the commons, where individuals acting on their own interest deplete and affect a dearer resource, in this case, privacy. In addition, these different incentives can lead to a market

failure since businesses, especially the ones whose revenue streams are highly dependent on advertisement, acting in their self-interest can lead to an Agent-Principal problem.

In sum, answering our main research question, the legal and market forces factors that can explain the presence of tracking cookies across the European Union are the variability in business models' incentives to use tracking and the lack of harmonization of the transposition of the E-Privacy Directive by less.

The future E-Privacy Regulation reform can reduce tracking and improve privacy protection by:

- Harmonizing the local provisions of the E-Privacy Directive across member states, especially on consent and guidance's provisions.
- Better Understanding the business incentives to avoid market failures accompanied by credible enforcement capacity while strengthening fines and requiring websites to provide information to users.

These two recommendations, based on our results, were compared with the draft of the E-Privacy Regulation, in the end. It is noted that the draft already covers the first implication. In contrast, the second implication, regarding understanding the "parsimonious" factor of businesses' incentives can be considered by policy-makers to commence an important debate and re-solve the privacy problem.

This page was intentionally left in blank

Chapter 1

1. Introduction[1]

Nowadays, the internet has facilitated people's lives. Every day we visit different websites to do banking transactions, to buy products or to spend our leisure time. Also, we are familiar with search engines that allow us to look for any information we require. In addition, we have created a digital self, we share online data through social media accounts, blogs, and web profiles. However, with the use of the internet and its convenience, threats to our privacy have increased.

Different tracking mechanisms have arisen to track our online traces and collect our online data[2] (N. van Eijk, Helberger, Kool, van der Plas, & van der Sloot, 2012). One of them is canvas fingerprinting. This tracking mechanism creates a unique and hidden image through our web browser to track our online behavior (G. Acar et al., 2014). Another mechanism to track us is the battery status of our computer (Olejnik, Englehardt, & Narayanan, 2017). Not only with certain hardware characteristics, but also with certain software characteristics it is possible to identify our devices. Also, social media networks can track users and non-users of them through the web using social widgets and like buttons ('Cookies and other (illegal) recipes to track internet-users', 2018). In addition, there are embedded pixels through our emails that can leak our personal data (Englehardt, Han, & Narayanan, 2018). Although there are many more tracking mechanisms, the literature expresses that the most commonly used across the web are cookies (Fruchter, Miao, Stevenson, & Balebako, 2015; Narayanan & Reisman, 2017a; N. van Eijk et al., 2012), and one of the industries that is highly dependent and profit from them is online behavioral advertisement (Smit, Van Noort, & Voorveld, 2014). This industry shows targeted ads through websites related to the preferences they infer from our online behavior.

Cookies are small files that are placed on our computer through our web browser with a unique identifier. There are different types of cookies. However, third party cookies, cookies that third-party companies place on our computer through websites we are visiting, are mainly the ones that can track our online behavior. For example, you visit www.a.com, and if this website is affiliated with a third-party company to deliver targeted ads, this third-party company will install a cookie on your computer. Every time you visit different websites that install third party cookies, and these websites are affiliated with the same third-party company, the third-party company will be collecting data about your online behavior.

Figure 1 presents how third-party cookies allow third parties companies to collect and track your online behavior[3].

1 This is a slightly modified version of the Research Proposal.
2 In this research, we will assume that "Online" data contain the user's real-life behavioral traits and information of users since it might be the case that some users use fake profiles or non-real data while using the web.
3 The icons of this figure and subsequent figures of this master thesis were taken from Flaticon ('Flaticon, the largest database of free vector icons', 2018)

Figure 1: How do third-party cookies allow third party companies to collect online behavior?

In 2016, the digital advertising industry generated 41.9 billion euros revenue in Europe (IHS Markit, 2017) ,but targeted ads need to track our online behavior in order to work. Hence, we observe a tension between privacy, which in the European Union is a fundamental right ('EUR-Lex - 12012P/TXT - EN - EUR-Lex', 2012), technology, and economic growth. Therefore, there is a call to develop regulations and policies that ensure the protection of privacy, balancing economic growth, and keeping up with the rapid evolvement of these tracking technologies.

Different policies have been developed to safeguard privacy. However, the E-Privacy Directive (E-PD) is the regional legal instrument of the European Union that more specifically address this issue. Although the directive covers a variety of topics, one of its aims is to safeguard for the processing of personal data of individuals and the storing or retrieving of information in their devices ('EUR-Lex - 32002L0058 - EN - EUR-Lex', 2002). Besides, the Directive wants to ensure that users give consent before any installation or retrieval of information takes place in users' devices, and there are some provisions related to tracking('EUR-Lex - 32002L0058 - EN - EUR-Lex', 2002). The final aim of the E-Privacy Directive is to increase the level of privacy protection of users and harmonize privacy protection on member states.

Nevertheless, regulations or policies that try to solve social problems such as privacy are what is known in policy science "wicked problem". A wicked problem is a problem difficult to solve due to its evolving and complex nature (Rittel & Webber, 1973a). Another characteristic of this type of policies according to Ritter & Webber (1973a) is that they cannot have a definitive solution, but they are "re-solved again and again". In addition, institutions (values and norms, laws and regulations, and all type of organizations) play a role in regulatory policies (Berg, Spithoven, & Groenewegen, 2009).

Due to this complexity, there is an ongoing debate about if this law is enough or not to prevent tracking mechanisms. The European Union citizens are concerned regarding not having control of their data or their data being misused, and they ask for more privacy protection ('Data Protection Eurobarometer Factsheet', 2015)('ePrivacy', 2016). The online behavioral advertisement industry states that more strict regulation can cause economic damage and kill the web since advertisements

sponsor its content ('Proposed ePrivacy Regulation', 2017). Policy makers seem to agree that a revision of this policy is necessary. The E-privacy Directive has been in place for more than two decades with modifications in 2002 and 2009 ('EUR-Lex - 32002L0058 - EN - EUR-Lex', 2002), and it is a reality that tracking is increasing and there are different pervasive and extraneous ways in which this can be done (G. Acar et al., 2014; Englehardt et al., 2018; Olejnik et al., 2017). Hence, we have arrived at a crucial moment in which the E-Privacy Directive is under discussion to become the E-Privacy Regulation.

The complexity of privacy policies, the different actors involved, and the variety of incentives pose a challenge for policy-makers. Besides, the E-Privacy Directive has been implemented in different ways in member states, and there is scarce literature about which approaches of the transposition of the law accomplished reduction or encourage tracking mechanisms. In addition, online behavioral advertisement stake revenues, so there is a lack of understanding of how the market forces and businesses' incentives to make a profit influence tracking. Therefore, there is a debate of what elements of the E-Privacy Directive are worthy to keep or what to consider for the future E-Privacy Regulation, so policy makers can safeguard the citizens' right to privacy, reduce tracking, and promote economic growth.

1.1 Problem Statement

Privacy policy is a complex issue, and although the E-Privacy Directive pays attention to tracking mechanisms, there is a rapid evolvement of technology and tracking mechanisms on the web. There is little research about how the legal framework and territorial scope surrounding this policy impacts tracking in the European Union given that member states transpose the E-Privacy Directive in different ways. Also, there is a lack of understanding of which approaches are beneficial to re-solve the privacy problem and to protect users' privacy against tracking. In addition, privacy has an economic component, and businesses which use tracking are profiting from it, so there is also a gap of knowledge on understanding to what extent the market forces and businesses' incentives to make a profit play a role encouraging or discouraging tracking. Also, it is not clear what elements of the E-Privacy Directive the future E-Privacy Regulation (E-PR) should keep. The E-Privacy Regulation was planned to enter in rigor this year on May 25th, 2018 at the same time of the General Data Protection Regulation(GDPR). However, it seems that some actors expect that the ongoing debate of what the E-PR should consider will continue until 2019 ('The new EU ePrivacy Regulation: what you need to know', 2016). Since the literature available does not include empirical data to demonstrate which of these arguments are right, and the regulation is still in discussion, this gives an opportunity to fill a literature gap to understand through an empirical analysis which aspects of the legal framework and the market forces factors encourage or discourage tracking.

1.2 Research Objective and Question

The objective of this research project will be to **"Bring empirical evidence to this debate"** and empirically and quantitative **test what legal and market forces factors encourage or discourage tracking cookies presence in the European Union**, so we can shed some light on how the E-Privacy Regulation can reduce tracking. Hence our main knowledge gap is:

What legal and market forces factors can explain the presence of tracking cookies across the European Union, and how the E-Privacy Regulation reform can reduce tracking?

To answer this main research questions, the following sub-research questions need to be answered:

SUBQ1: What is tracking? How pervasive are they in European Union countries? and What are the type of tracking in use?

SUBQ2: Which laws do websites follow? Are there differences in tracking and cookies notices related to the laws they follow?

SUBQ3: What local provisions of the E-Privacy Directive and market forces factors, if any, encourage or discourage tracking presence across member states?

SUBQ4: What are the implication of the findings to policy makers?

1.3 The Scope of the Study

The literature states that cookies are the tracking mechanism more common on the web. Hence to narrow the scope of this study they will be used as a proxy to understand tracking. In addition, the E-Privacy Directive covers a variety of topics such as traffic data, spam and cookies (Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws (Text with EEA relevance), 2009). Hence, this thesis will focus on the provisions related to tracking and article 5(3).

Article 5(3) amended states:

'Member States shall ensure that the ***storing of information****, or* ***the gaining of access to information already stored****, in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or* ***user has given his or her consent****, having been* ***provided with clear and comprehensive information*** *in accordance with Directive 95/46/EC, inter alia about the purposes of the processing.'*

('EUR-Lex - 32002L0058 - EN - EUR-Lex', 2002)

Moreover, there might be different industries that use tracking mechanism. However, as it was expressed previously the industry that benefits the most from cookies is the online behavioral advertisement (Smit et al., 2014). Hence, we will analyze tracking cookies in the context of this industry controlling for the different business models' incentives of websites.

1.4 Research Methodology/ Approach

This thesis project opted for an empirical research with emphasis in quantitative analysis. The research has two parts. First, a thorough literature review of the provisions related to tracking of the E-Privacy Directive. This will allow codifying the independent variables of the legal framework related to tracking.

Second, the collection of data about the dependent variable will be made using OpenWPM – Open web privacy measurement framework. This framework is part of the web transparency and accountability project of Princeton University, and it allows data collection for privacy studies on a large scale (*OpenWPM*, 2014/2017). The framework uses a script that runs a simulation of visiting websites using Mozilla Firefox and it stores data related to tracking using an SQLite database.

With the results of the analysis of the data collected judgment will be formulated to answer the research sub-questions and test a conceptual model that will be explained in more details in Chapter 2 – Literature Review.

1.5 Scientific, Practical, and Managerial Relevance

1.5.1 Practical Relevance

Hopefully, the data analysis resulting from this master thesis can contribute to the discussion of what are the most relevant elements of the legal framework and market forces that can be considered in the reform of the E-Privacy Regulation.

1.5.2 Theoretical Relevance

One direct contribution of this study will be to privacy governance. Privacy is a complex phenomenon that involves a wide variety of actors, and this complexity has to be understood to derive policies that govern privacy (Bennett & Raab, 2013).

Also, this study will have as a general framework institutional economics. Tracking is a problem that involves two institutions Law and Market Forces, and we want to understand how these institutions drive or discourage tracking.

In addition, a contribution to the field of economics of privacy which focuses on the flow and use of individuals' personal information by firms will be made (Acquisti, Taylor, & Wagman, 2016). Economics of privacy addresses the economic incentives of firms in the use of personal information. Hence, this thesis will address how the economic incentives of the businesses, which are part of the market forces, can encourage or discourage tracking.

Finally, indirectly this thesis project will contribute to the literature related to Web Privacy Measurement. Web Privacy Measurement allows through the observation of websites and services "to detect, characterize, and quantify privacy-impacting behaviors" (Englehardt, Eubank, Zimmerman, Reisman, & Narayanan, 2014).

1.5.3 Managerial Relevance

Companies must comply with the E-Privacy Directive and the future E-Privacy Regulation or any future regulation related to privacy. Also, companies should respect the trust that users deposit in their companies every time they visit their website. However, it is a challenge for companies to adopt privacy policies that respect users' privacy. Hence, understanding what elements are encouraging or discouraging tracking might shed some light on how to become more responsible in their websites' privacy policies. Also, this study can lead to understanding the impact of embedding third party companies on their websites.

1.6 Structure of the Report

The deliverables of this research project are 5 chapters. The first chapter was a modified version of the research proposal.

The second chapter will consist of a literature review related to the privacy concept, the law in the context of the European Union, and tracking mechanisms. This chapter will also include a review of the literature on institutional economic, web privacy measurement, and privacy governance. Finally, this chapter will finish arriving to the problem statement, and the main research question proposed that is not addressed by the literature yet, and a conceptual model of the tracking phenomena.

Chapter 3 will present the methodology and the steps followed to accomplish the research objective and answer the research questions.

Chapter 4 will be devoted to present the results of the data analysis and hypothesis testing to answer the research sub-questions.

Chapter 5 will wrap up with the conclusions and reflection on the results as well as the limitations of the research, and the recommendations for future research.

This page was intentionally left in blank

Chapter 2

2. Literature Review

In this chapter, the relevant concepts and the theoretical foundations of tracking and privacy will be introduced. The first section of this chapter, section 2.1 will introduce briefly how the web works, section 2.2 is devoted to an introduction to the different tracking mechanisms on the web, section 2.3 address the incentives of businesses to track, section 2.4 will introduce the concept of tracking, section 2.5 describe the business models of companies that are interested in tracking, section 2.6 explain some of the risks that tracking represents to users, section 2.7 will address the future trends related to tracking, the challenges to solve this problem will be addressed in section 2.8, section 2.9 addresses the mechanisms to prevent tracking, and section 2.10 narrow the context of the thesis to cookies and the E-Privacy Directive. We finalize the chapter with the conceptual framework that depicts how we think that tracking phenomena occur.

Privacy comes from the Latin "Privus" meaning "single" (Hirshleifer, 1979). It is a concept traced back to 1888 used for first time by Judge Cooley, and described as the "the right to let be alone" (Warren & Brandeis, 1890). This concept was born in a context where digital technology did not exist, so this definition was limited to a physical space. However, as it was mentioned in the introduction chapter, with the evolvement of technology, we have created a digital self. Hence, privacy is not limited anymore to our physical surrounding. Discussions have been going on for around forty years on the concept of privacy, and it has been a contested concept (Bennett & Raab, 2013). Still, nowadays, there is not a clear definition of privacy in the digital age (Damen, Köhler, & Woodard, 2017), but it is expected that the same rights that you have in the physical context are respected when you use digital technology (Damen et al., 2017).

We will consider privacy in the context of the European Union where privacy is a fundamental right. Since 1953 under the European convention of human rights in its article 8 the respect to private life was established ('European Convention on Human Rights - Official texts, Convention and Protocols', 2010; Warbrick, 1989). The article expresses that family life, correspondence, and home are part of individual's private right and they should be respected unless national security, public safety, or protection of the rights and freedoms of others are at stake. Besides, the right to privacy is stated in the Universal Declaration of Human Rights (UDHR) in its article 12 (Assembly, 1948). Hence, we will take for granted that European citizens have the right to privacy.

Through the years privacy has been disrupted due to technology. For example, with the invention of the camera, individuals were photographed without their permission, and there were cases of legal complaints against newspapers that published pictures without individuals' consent (Warren & Brandeis, 1890). Nowadays, technologies such as big data where a huge amount of data can be collected and analyzed pose a threat to privacy. Hence, we observed that there is always a tension between privacy and the technological progress. In addition, we have witnessed how the complexity and evolution of technology have been an instrument to violate privacy. One of the well-known cases is the National Security Agency which exerted mass surveillance on citizens (Mazzetti & Schmidt, 2013). Also, another example is how Google read your emails in exchange for a free email account (Gibbs, 2014).

Due to the tension between privacy and technology, international, regional, and national legal instruments have been developed to protect this fundamental right. However, we need to keep in mind that some regulations safeguard for the right to privacy, while other safeguards for data processing or

collection of personal data of individuals. According to Zuiderveen Borgesius (2014) we need to distinguish between the fundamental right to privacy and the policies that safeguard the fairness and transparency of data collection which is often seen as different things. Zuiderveen presented as an example, that if someone stalks you through your window, he is violating your privacy, but according to the laws he is not collecting personal data, so there is no violation to the data collection laws. In this case, we will be seen the E-privacy Directive and the future E-Privacy Regulation as policies that safeguard or the fairness and transparency of data collection of individuals in the European Union. Also, it is important to highlight that tracking on the web is a mechanism to collect data of individuals which might or not violate the right to privacy.

Besides the E-Privacy Directive, there are well-known international guidelines to protect personal information. These guidelines were developed in 1980 by the Organization for Economic Co-operation and Development (OECD) (Organisation for Economic Co-operation and Development., 2003a). They stated that data collection should be limited, relevant, open, protected, and who collect the data should be accountable. In addition, the European Union developed the Data Protection Directive in 1995 as the first regional instrument to safeguard for data protection. During the development of this master thesis, this Directive was substituted by the General Data Protection Regulation (GDPR).

Although privacy is a fundamental right, and these different legal instruments are in place to safeguard for fairness and transparency of data collection, there are different tracking mechanisms that have emerged on the web that pose a threat to privacy and can collect personal data. Also, it is important to note that there might be different technologies that pose a threat to this right. For example, Internet of Things, Big Data, or Smart Cities. However, the focus on this thesis will be on "the web" as one of the most widely used technologies across the globe. Also, just in the European Union, 71% of the individuals use the internet every day ('Internet access and use statistics - households and individuals - Statistics Explained', 2016). Hence, there is a significant amount of people exposed to the conflict between the need for the use of this technology and protecting their right to privacy and data collection.

To understand "the web" as a technology. First, we will start with a brief description of how it works.

2.1 The World Wide Web (The web)

The Hypertext Transfer Protocol (HTTP) is the foundation protocol of the web. Hence, we need to understand at a high level how HTTP works to understand how the web works.

The Hypertext Transfer Protocol (HTTP) is the protocol that allows the communication between a user's web browser and the server where the website he wants to visit is hosted. When a user requests to visit a website, he writes a Uniform Resource Locator (URL), e.g. www.site.com. After, a HTTP request to the server where the website is hosted is made, and the server sends back all the files related to the website through a HTTP response. Files related to the website can be images, text, and sounds. Finally, the website appears on the user's computer screen.

Figure 2 depicts how HTTP works.

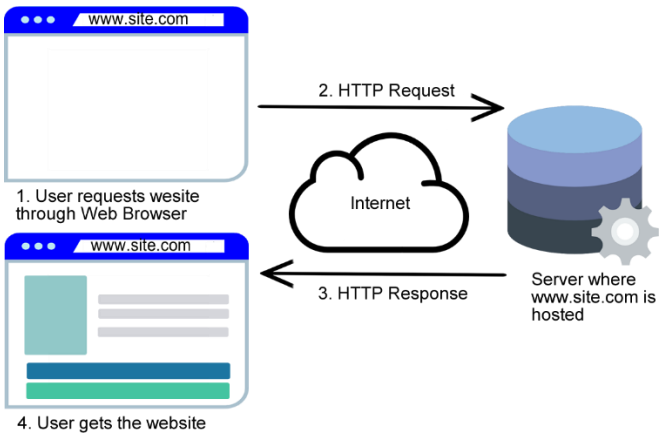


Figure 2: How does HTTP works?

At the beginning of the web, HTTP was a stateless protocol. This means that every time a user called a website each request he made was independent of the next one, and the websites could not remember his actions. For example, if the user visited www.site.com, and he selected as preferred language English, every time he visited this website he needed to select this preference. Also, websites could not remember users' passwords and emails accounts, so every time they visited a website that required them they had to type them. In addition, the statelessness of HTTP hindered some functionalities of websites. For example, E-commerce could not remember more than one product in the shopping cart, and it was not possible to know which customer has to pay what (Weber, 2000).

2.2 Tracking Mechanisms on the Web

Due to the limitation of HTTP some mechanism to track users' actions through websites were born. These mechanisms facilitated web browsing and achieve some new functionalities of websites such as online shopping. Bujlow, Carela-Español, Sole-Pareta, & Barlet-Ros (2015a, 2017) classifies the tracking mechanisms in five main categories *session tracking, cache tracking, storage base tracking, fingerprinting, and other types of tracking*. This classification will be used to make a brief description of some of these tracking mechanisms and how they work. For the interested reader, we recommend to consult Bujlow, Carela-Español, Sole-Pareta, & Barlet-Ros (2015a, 2017) to get more information on other tracking methods in each category.

2.2.1 Session Tracking

A session can be considered as the period a user enters a website and the time he leaves it. There are tracking mechanisms to ensure that the website knows the user's actions during this period, and he can be identified.

Authentication. Users type a username and password to log in on a website, so the website knows the user's preferences and actions during the session (Bujlow, Carela-Español, Solé-Pareta, & Barlet-Ros, 2015b). This is a widely known mechanism, even for users. Users log in to use social media, email accounts, or to buy products online. Hence, users need to be aware that with this unique username they are identifying themselves while using that website. This mechanism has legitimate purposes, for example, if a user enters his email account, he wants to read only his email, so this mechanism is necessary to give the information the user requires. However, the website also can keep

track of how many times he logs in, time spends on the website and the pages he visits within the website. Also, this type of mechanisms is used in online shopping to determine what a user orders, and what he has to pay.

Figure 3 depicts a basic log in of two users that allow the website to recognize each one, and what each user has bought.

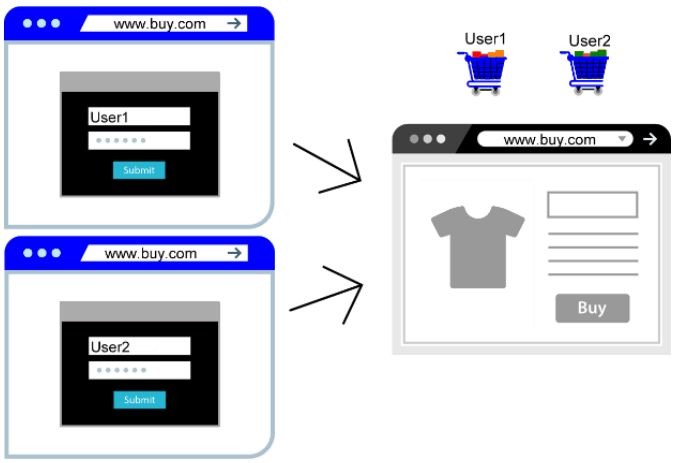


Figure 3: Authentication as a tracking mechanism

Uniform Resource Locator (URL) rewriting. Another form of session tracking is rewriting the URL the user is visiting (Facca & Lanzi, 2005). The user types into the address bar of his web browser the URL, e.g. www.site.com, and when he enters the website a unique session id number is assigned to him. This type of tracking is sometimes unnoticed by users, but in the address bar the number assigned to the session can be observed.

Figure 4 depicts how the URL is rewritten to keep track of the user session.

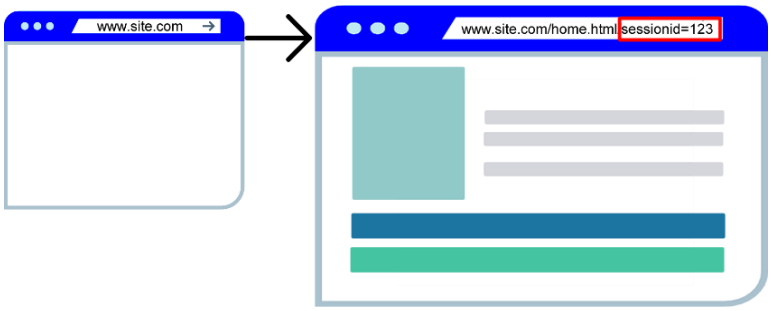


Figure 4: Tracking rewriting the Uniform Resource Locator (URL)

Identifiers stored in hidden fields. Websites can use hidden pieces of code that are not visible to users at all (‘<input type="hidden">’, 2018). This hidden fields can also assign a unique identification number to the user session and track his actions in the website without being noticed. Figure 5 tries to depict how these pieces of code are embedded in the websites.

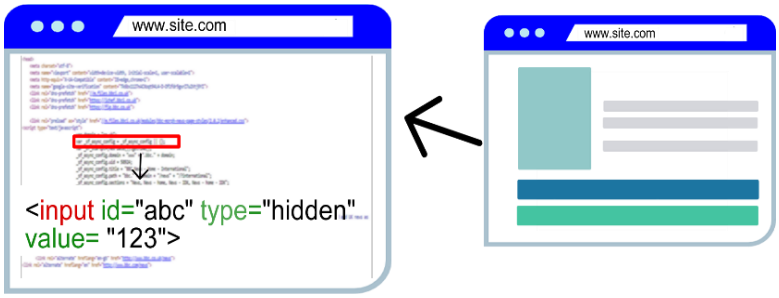


Figure 5: Hidden fields in the code of websites to exert tracking

Besides these session tracking mechanisms, there are more advanced methods such as cache base tracking. These methods are completely transparent to users. Hence, users that do not have any technical skills or even the ones who have might be not aware of them.

2.2.2 Cache Base Tracking

This type of tracking is capable of not only recognizing the user within his session in the website, but also they are capable of tracing the previous websites visited by the user (Felten & Schneider, 2000). Only caching and Domain Naming System will be briefly described.

Caching. When a user visits a website, certain documents, images, or sounds of that website are stored in the web browser cache of users. The web browser cache is a web browser temporary memory to store internet content the user has visited. Hence, the next time the user wants to visit the same website these files will be in the cache, and the user can access the website in a short time since it will not be necessary to retrieve them from the server where the websites are hosted. Also, this is an efficient way to not overload requests to the server where the website is hosted. Nevertheless, by determining the time that it takes to the web browser to access the files of a website, the website can determine if that documents, images, or sounds are or not in the cache of the user. In this way, it is possible to know if the user has visited that website before. In addition, there is the possibility to invisible embed in website images or files from other websites to determine if the user has visited other websites as well (Felten & Schneider, 2000).

Figure 6 depicts how caching involving a different website takes place.

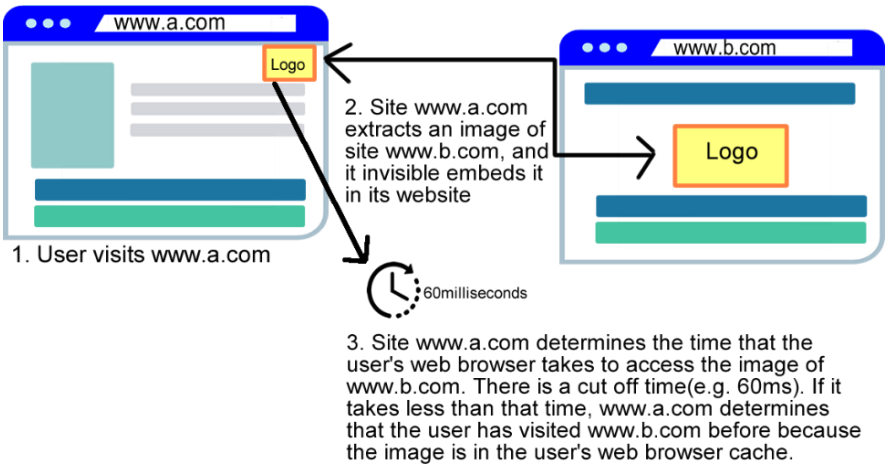


Figure 6: Caching as a tracking mechanism

Domain Naming System (DNS) Caching. When a user wants to visit a website, he types a URL, e.g. www.site.com, in the address bar of his web browser. However, to get the website, it is necessary to have the Internet Protocol address of it, e.g. 192.162.2.1. Since it is easy for users to remember the URLs than IP addresses, the Domain Naming System helps to translate the URLs the users type into their address bar to IP addresses. This process is called DNS lookup, and it involves different servers and takes some time and network resources. As a solution to save time to access the website when a user requests it, there is a Domain Naming System cache to store IP Address of recently requested websites. In this way, if the user requests a website that has been visited recently, it does not have to go all the process of finding the IP address, but retrieving it from the cache. Nevertheless, a website can make use of the DNS cache to determine if a user has visited a website before.

In addition, this mechanism can be used to determine if a user has visited another website before. For example, if the user is visiting www.a.com, this website can trigger a DNS lookup to another page, e.g. www.b.com, and measure the time it takes to get that website IP Address, so it is possible to determine whether or not the IP Address is in the cache of the user, so if the user has visited that website previously (Felten & Schneider, 2000).

2.2.3 Fingerprinting

Besides cache tracking, another type of sophisticated method to track users across the web is fingerprinting. The hardware and operation system utilized by users can be a unique fingerprint to identify them.

Canvas fingerprinting. Users' web browsers have unique characteristics such as the fonts, plugins, version, and many other parameters. In a website, it is possible to have areas designated to render graphs, and this is called canvas. The users' web browser unique characteristics make possible that canvas renders images in a particular way, detecting even subtle differences. According to (G. Acar et al., 2014), two steps are involved in this process. First, without the user knowing a script draws a text in a selected font and background color. Second, a method called ToDataURL from the canvas Application Programming Interface (API) convert this text into a Base64 representation, and then this

is used as a unique fingerprinting identifier. This fingerprinting is being completely transparent to the user.

Figure 7 tries to represent the two steps of how canvas fingerprinting happens on a website.

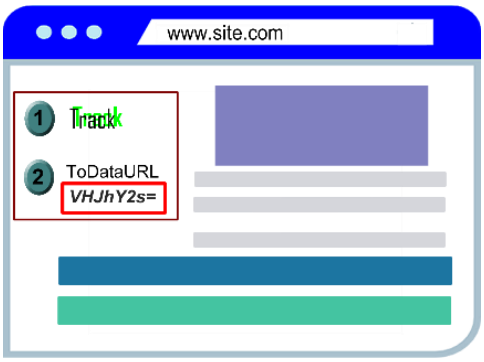


Figure 7: Canvas Fingerprinting as a tracking mechanism

Network Fingerprinting. The Internet Protocol (IP) address of users' computers can be used to track users' location, and it is unique. Hence, users' IP addresses are usually not revealed. To protect the IP address the use of an HTTP proxy server is common, a server that is used as an intermediary when a user requests a website, to assign a different IP address to the user. However, it has been demonstrated that the real IP address of the user can be obtained through flash files, called Small Web Format (SWF), embedded in websites bypassing this proxy (Nikiforakis et al., 2013). Then the Internet Protocol address can be used to track users' real location. Figure 8 shows how network fingerprinting works at a high level.

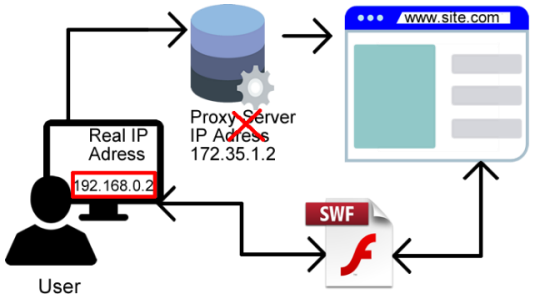


Figure 8: Network Fingerprinting as a tracking mechanism

Device and operating system fingerprinting. Any other characteristics of the users' computers such as operating system version, screen resolution, battery status, drivers installed on a computer, audio capabilities, among others provide a unique set of characteristics of each user's device and operating system (M. G. C. Acar, 2017; Bujlow et al., 2015b; Englehardt & Narayanan, 2016a). Hence, all of them or a combination of them allow to having track users' online behavior.

2.2.4 Storage Base Tracking

All the tracking mechanisms previously described do not require to store any file in users' devices. However, there are even more advanced mechanisms that have the capacity to install files on users' devices, and one of them is cookies.

Cookies. Cookies are small files that are set on users' hard drive through his web browser. Cookies are sent to the users' device through a set cookie header in the HTTP response sent from the server where the website is hosted to users' web browser. The next time the user visits the same website, the cookie is sent to the server with the HTTP request he makes to that server to obtain a website. Cookies assign a unique identifier to the user, so that when he comes back to the website, they recognize user's preferences. Figure 9 depicts how cookies are set in users' web browsers.

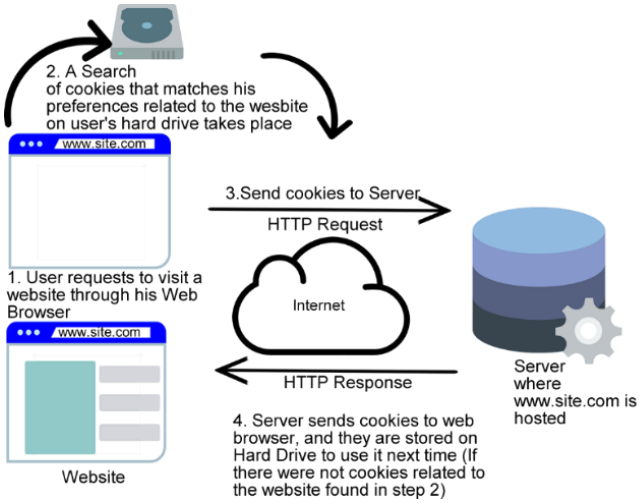


Figure 9: How do cookies work?

2.2.5 Other Types of Tracking

There are more tracking mechanisms that are not discussed. For example, emails can contain embedded pixels that can leak personal data (Englehardt et al., 2018). Also, some apps have access to the GPS of cellphones that can collect this data. For example, it was recently demonstrated that companies such as Google can collect users' location, even in airplane mode or when they are not connected to the internet ("'It Knows When I Got Out of the Car!"', 2018).

The tracking mechanisms discussed in this section hopefully gives an overview of how users are exposed every day to tracking on the web. The next section will discuss why these tracking mechanisms are used.

2.3 Why Are We Tracked?

There are legitimate purposes to use some of these tracking mechanisms. As it was expressed in section 2.2.1, session tracking is necessary when users visit online banking websites, E-commerce, or any websites that need to keep track of each user session to deliver a service or charge them correctly. In addition, IP tracking mechanisms can be used to detect fraudulent transactions, e.g. to detect someone who wants to access a bank account from a different country where the owner of the account lives. However, there are companies that can detect fraudulent transactions and also can use the same data for marketing purposes (G. Acar et al., 2014). Hence, even though there are legitimate purposes to use tracking, users are exposed to the use of this sensitive data for other purposes.

In addition, there are purely commercial purposes of tracking. One of them is price discrimination (Bujlow et al., 2017). Price discrimination has existed and it is economically desirable since transactions with different prices make companies and users better off (Odlyzko, 2003). If the companies learn about the wiliness to pay of users thanks to tracking, the company might establish the price of a product or service accordingly. For example, if a user wants to buy a flight ticket and he is willing to pay 100 Euros, and another user is willing to pay for the same ticket 200 Euros, the airline can set a price of 100 Euros and 200 Euros for each customer. Airlines are better off since both customers will be satisfied and buy the product or service, and the company can maximize its profit.

Moreover, there are also websites such as E-Bay, Amazon that uses tracking to deliver product recommendations based on the search the users do on their websites. Also, they learn from the behavior of regular users to give customized recommendations to new users who visit their websites.

Also, tracking can be used to deliver personalized services. For example, Netflix, a tv show entertainment company, needs to collect data about the tv shows or series users' watch to recommend content according to users' preferences ('How Netflix Knows Exactly What You Want to Watch', 2016) .

Another user of tracking is to improve websites. Tracking can be used to understand what the most visited pages on websites are and generate statistics and web analytics to improve them. This allows to understand better the behavior of users within a website, so websites can offer better interfaces and improve their return on investment or engaging customers to buy a product or service.

There might be other purposes to use tracking mechanisms, even national security might be one. Also, there are different incentives among different actors to learn about users' behavior. Some companies might argue that tracking allows to provide better services, help to innovate, and even prevent crime. However, nowadays one of the most contested uses of this tracking mechanisms is the online behavioral advertisement. Websites can use tracking mechanisms, especially cookies as it will become clear later, to collect data about users' online behavior, so they know what previous website an individual has visited and matched his "inferred preferences" with products or services.

2.4 But What Is Tracking?

So far, we have been discussing the different mechanisms to exert tracking across the web, and we have discussed in the previous section some reasons why there are incentives to use these tracking mechanisms. However, we did not provide yet a formal definition of tracking. In this research, tracking will be defined as proposed by Wefers Bettink, Van Eijk, & Wagner (2012):

"Web tracking is an act by a party, or host, or service, of reading or writing Unique Identifiers (UIDs) that are connected directly or indirectly to an end-user, computer, or device while the end-user is interacting with various services of the web, in order to collect, combine, or analyze data about the end-user for charitable, philanthropic, or commercial purposes".

Tracking can be exerted by websites that users are visiting, and this is considered first party tracking. Also, websites might have relationships with other companies, third party companies, that can collect, combine, or analyze data about the end-user for charitable, philanthropic, or commercial purposes, and this is considered third party tracking. However, a first party can be a third party in another context. For example, if a user has a Facebook account and he visits Facebook, Facebook can exert first party tracking. However, if the user visits www.a.com and www.a.com has a Facebook like button and the user is logged into his Facebook account, then Facebook is a third party. Hence, this concept was chosen because it covers any form of tracking and it contemplates that tracking includes unique identifiers even if the first or third party do not know the name of the users.
Now that we are clear about the definition, in the next section we will describe which the different type of third party companies that are interested in tracking users.

2.5 For Whom Are We Tracked?

From the definition of tracking that we will use, it is clear that there might be different third parties, or host, or services that are interested in tracking end-user online behavior, and their business models depend on it. Mayer & Mitchell, (2012a) proposed six different third-party business models that are being used and embedded by websites.

Advertisement companies. These are companies that sell advertisements directly to advertisers or agencies. One example of this type of company is Google Adwords which sells directly to advertisers or agencies privilege position in its search engine. In addition, another type of companies within this category are the called advertisement networks. These are companies that through online platforms connect advertisers and websites which want to publish their advertisement. Finally, another type of companies in this category are the ad exchanges. These are online platforms that in real time bid the advertisement spaces on websites to advertisers. All these companies benefit from tracking mechanisms by creating profiles that allow targeting end-users according to their demographics or inferred users' preferences, for example.

Analytics. These companies track returning visitors, unique visitor, or what pages of a website are more visited, buttons that have more clicks, the location of the users who visit the websites, among others. These types of companies are popular because they offer to the websites the possibility to improve their search engine optimization, which is the position in the organic search of search engines such as Google. Also, they can help improving the popularity of the website, and they can help websites learning about the users' behavior on their websites. However, since many websites use the same analytic, e.g. Google analytics, these third-party companies have the possibility of aggregating data from different websites and tracking users across websites.

Social Integration. These are usually the social networks companies that allow users to sign up for a website with their social networks account. Also, users can have the possibility to add comments, give like to the websites directly on social networks, read tweets or retweet information of a website

(Mayer & Mitchell, 2012a). Once users make use of social networks buttons or log in to a website using their social media accounts, these companies have the possibility to learn about the users' online behavior. Even more alarming some social networks such as Facebook can track non users of its social network thanks to like buttons or social integration ('Cookies and other (illegal) recipes to track internet-users', 2018).

Content Providers. These are companies that offer content that can be retrieved by websites for free such as weather conditions or videos (Mayer & Mitchell, 2012a). One example of this type of third party companies is Weather Stickers® ('Weather Stickers® | Free Weather Sticker | Weather Underground', 2018), which offers to add a free widget to websites to provide users with weather conditions. However, this type of widgets in some way want to monetize this piece of code they share. Hence, it is possible that they use these widgets to transform them into ads (Mayer & Mitchell, 2012a). Hence, this can lead to the use of tracking mechanisms by advertisement companies that target the users across the websites which have used the widget.

Front-end Services. These are third party companies that offer code ready to use to integrate into websites to have Application Programming Interfaces (APIs) on websites that were not developed by the website a user is visiting (Mayer & Mitchell, 2012a).

Hosting Platforms. These are companies helping websites to distribute and create their content. For example, Content Delivery Networks have copies of the content of the websites distributed in different locations of the network to minimize the time response a user has to wait to access a website and to have a more efficient network (Krishnamurthy & Wills, 2009a). In addition, there are platforms such as Joomla, Wix, Wordpress that help to create web content for free, but they can gain access to users' data.

Market Research. To the six categories proposed by Mayer & Mitchell (2012a), we can add market research. Market research companies are firms dedicated to research, analyze, and give insight about consumers attitudes, demographics, and target markets ('The AMA Gold Report 2017 Top 50 Market Research Firms', 2017) to improve companies position in the market, launch a product, etc. This is a broader group of companies that includes more than analytics, and that is also interested in tracking.

2.6 To What Are We Exposed With Tracking?

Business models that use tracking mechanisms represent a risk for end-users. With tracking mechanisms companies are growing the amount of data they collect about users. Even if companies were not collecting personal data, the data they collect can be aggregated or combined with other data such as social networks profile or even offline data that can help to reveal a user's identity (Narayanan & Reisman, 2017a; Krishnamurthy & Wills, 2009b). Users are exposed to the misuse of this aggregated data by companies that are collecting it or data breaches. Around the world, 4,923,037 data records breached every day (Gemalto, 2018), so if these records are stolen for malicious intent, users are exposed to damages such as identity thief.

In addition, there is some evidence that a person can be charged 4 times higher according to their budget inferences (Mikians, Gyarmati, Erramilli, & Laoutaris, 2012). Hence, once companies know more about users' purchase intentions, they can take advantage of that data.

Moreover, users' personal location can be exposed. For example, Google maps provide users with the most convenient routes to go to a certain place, but they are learning about where they have been. Another case of how the physical location of users might be exposed is Strava. This is an app that monitors the physical activity of cyclists, runners or any other sport that involve motion through GPS. The app allows sharing the effort and distance achieve, and it provides the location of the users while doing sports, so that family or friends could learn about their location in emergency cases. However, the app shows a heat map of the location that users shared that allows the app to aggregate data of all the users. With the heat maps, some areas, even secret military locations, were exposed (Tufekci, 2018).

Also, users can live in a "bubble". It is possible to show news or information limited to users' profiles or to show to users only what other people want to show to them. For example, in the United States a set of advertisements were targeted to certain segments of the population to support Trump campaign in the elections in the United States in 2016 ('Here are some of the Russian Facebook ads meant to divide the US and promote Trump', 2017). Certain ads were shown only to certain groups to prime the ideas of those groups. Hence, exist the risk that tracking can contribute to undermining democracy.

In addition, Zuiderveen Borgesius (2014) states that some tracking mechanisms, especially involved in online behavioral advertisement can affect the development of a person's identity. If people are being observed, they can change their behavior or limit themselves to not do a certain online search. For example, Zuiderveen Borgesius (2014) expresses that if people have doubts about their sexual orientation, they might not be comfortable looking for information about it, so this can hinder the development of their sexual identity.

The risks mentioned in this section are not exhaustive, there are much more risks which users might face when using the web such as gender and race discrimination. Besides, new technologies such as the internet of things can lead to more ways of exposure. Not only privacy is at stake, but also users' real identity that can lead to physical harm. Also, with the vast amount of data collected and aggregate it there might be future implications still unknown.

2.7 Future Trends

Tracking technologies have kept evolving, and there are different possibilities of how they can be used in the future.

First at all, tracking mechanisms in a combination of social media profiles can be used to create advertisements that use similar faces of users' friends appealing to click them (Samat, Acquisti, Gross, & Pe'er, 2013).

In addition, there is already some evidence that tracking mechanisms, especially cookies, can represent risks for payments through blockchain (Goldfeder, Kalodner, Reisman, & Narayanan, 2017).

Moreover, it is expected that 20.4 billion devices will be connected to the Internet of Things by 2020 ('Gartner Says 8.4 Billion Connected', 2017). Hence, more and more devices using sensors and collecting data can be used to reveal the real location of a person. For example, smart thermometers can collect information about the temperature of a house or automatize the use of the heater or air

conditioner, but also they can be used as an instrument to track whether there are people or not at home.

Another possibility is the use of tracking in city management. Cities are adopting artificial intelligence to predict traffic jams before they occur, monitor accidents, and crime. For example, in the Netherlands, cities such as Eindhoven and Utrecht are monitoring people in the streets to prevent crime and improve security (Naafs, 2018). However, these systems need to collect data from different sources, and also track users real location in order to work ('Alibaba Cloud Launches Malaysia City Brain to Enhance City Management', 2018)

From the last example, we can foresee that tracking mechanisms can be combined with any other future technologies. Also, since technology evolves rapidly, the combination of tracking mechanisms with them might not be even detected at the beginning, as it was the case of fingerprinting.

Data collection and profiling users thanks to their online behavior represent the "new gold" to make a profit. However, we hope it is clear to the reader that tracking mechanisms and data collection about users can undermine privacy and expose users' data to different risks. Hence, it is necessary to keep an eye on the future technologies and combination with tracking mechanisms to regulate them. However, this is not an easy task, and the next section will describe some of the reasons why.

2.8 Challenges to Tackle Tracking Mechanisms on the Web

Although tracking mechanisms pose risks for end users, and in the near future potentially more technologies can be combined with them, tackling them pose a challenge. In this section, a brief discussion of these challenges will be discussed.

Privacy a wicked problem. The first challenge to control tracking mechanisms on the web is that it is a wicked problem. A wicked problem is a problem difficult to solve due to the evolving nature and complexity (Rittel & Webber, 1973b). There are different policies created to protect privacy that will be addressed in more details in section 2.9 of this chapter. However, policies that respond to privacy are considered a wicked problem.

One characteristic of wicked problems is that they can be considered as a consequence of different problems. One might argue that the policies in place are not enough to protect privacy as a fundamental right and ensure the transparency of data collection. Other people might argue that the cause of the problem is the rapid evolvement of technology, and that it is hard for laws to keep track of the advancements in technologies and new possibilities of tracking. Another people might say that there is a lack of technologies that protect or detect tracking mechanism. Hence, determining the cause of the problem is a problem in itself.

Another characteristic of a wicked problem is that the solutions propose might generate consequences for a period of time. For example, if more strict regulations are proposed companies might need to adapt their websites to comply, and this generates costs for them. Also, regulators need to implement and enforce the laws. Also, users might need to be able to adapt to any changes on the web. Also, if the solution proposed is to include privacy by design, then finding the right balance might be a challenge for companies that want to offer products or service or slow down their innovation process. Hence, any solution proposed will generate aftereffects.

In addition, wicked problems do not have a stopping rule. Even though more regulations or new technology development to detect tracking were in place, new technologies will emerge, the market will change, the users will change. Hence, regulations, as well as preventive measures, that work today might not work in the coming years.

Economics of Privacy. Another factor to consider is that privacy has an economic component. The collection of personal data through tracking mechanisms by businesses is generating income for them. Hence, this also can pose a difficulty to solve this problem because employment and revenues of companies are at stake. Economics of privacy focuses on the flow and use of personal information of an individual by firms (Acquisti et al., 2016). Also, economic of privacy has studied the impact of privacy in price discrimination (Odlyzko, 2003; Acquisti, 2008), which means that the information that the businesses collect to make a profit can be used to charge more to users due to their online behavior. In addition, it is possible that businesses might influence the demand for certain products or services (Hagiu & Jullien, 2011). For example, if companies match certain users with certain products or services this can affect or punish certain products or services that do not have the same benefit. All this economic component lead to misaligned incentives of the actors that interact with this problem.

Actors Incentives. There is a variety of incentives surrounding tracking. As it was expressed these tracking mechanisms are used for businesses to generate revenues, so their incentive is to use tracking to increase their profitability. In addition, some users might enjoy having services tailor-made for them thanks to tracking. On the other hand, other users might care more about privacy, and the risks that tracking poses that were explained in section 2.6. Moreover, governments need to safeguard privacy as a fundamental right, but also, they need to allow the market to make a profit and incentivize economic growth. Therefore, different actors have different interests and even conflicting interests around tracking mechanisms.

Information Asymmetry. Another factor why tracking mechanisms are difficult to tackle is because there is information asymmetry present between the regulators who develop policies to protect privacy and businesses. Information asymmetry is when one party has more information about a transaction than the other (Akerlof, 1970). Although there are regulations in place to prevent tracking on the web, the governments are not having enough information about how companies are complying with them. Besides, the websites that constitute the web have more information of the tracking mechanisms being used to track than the users. However, websites are also facing information asymmetry when they integrate third parties on their websites. For example, they might use Google analytics to understand better their websites and improve it, but they might not be completely aware of what data is Google collecting from users visiting their websites.

Market of lemons. The information asymmetry that exists between websites and users lead to a market of lemon situation. There might be websites that use tracking mechanisms and others that do not. The websites are aware if they are using tracking mechanisms or not. However, the users might not be aware of the use of tracking. The user does not know if he is visiting 'a lemon', a website that uses tracking; or a website that does not use tracking. This can lead to users to think that all websites are using tracking, so the user might not want to visit websites at all if they are concern about privacy, which can result in slowing down activities such as e-commerce or other transactions on the web. However, to fix this problem some websites have banners/cookies notices and policies to notify the use of tracking mechanisms. In addition, regulations in some countries in the European Union demand the use of these notifications to users. Nevertheless, it is difficult still to determine which websites are exerting tracking and which ones are not. This is still difficult because reading these notices and policies of websites lead to transactions cost for users.

Transaction Cost. Although websites on the web exerting tracking might publish in their policies and cookies notices how they use these tracking mechanisms, there is a transaction cost reading these policies by users. Usually, when visiting a website users do not take the time to read these notifications and users click in them without being aware of what they are doing (Liu, 2014). At least in The United States, a study suggested that reading privacy policies will take at least 40 minutes every day and a total of 244 hours per year (McDonald & Cranor, 2008). Hence, it is difficult to think that a person that is in a hurry or wants immediately to access to certain content in a website will take that amount of time to read the policy of the website. Besides, this implies not only reading the policy of the website the user is visiting, but also the policies of another third party that might be integrated into the website the user is visiting.

Zuiderveen Borgesius (2014) stated that also it is necessary to consider other costs such as lock-in because even if the user does not agree with the conditions of the website, there might not be other choices from where to get the content or service the user needs. Besides, Zuiderveen Borgesius (2014) stated that there are costs of clicks, waiting time in the website, and cost of checking if the websites comply or not with what they are promising in their privacy policies or cookies notices.

Externalities. Another challenge involved in this problem are externalities. When transactions in the market impact third parties that are not involved in them this is known as an externality. They are called positive if the transaction impacts the external third parties in a positive way. In the case of tracking as a positive externality, we can consider innovation. Companies can create new business models, new services, and they can reduce the search cost of users knowing their preferences. On the other hand, if the cost of the transaction imposes costs on the external third party they are called negative externalities. For example, we can consider that people who do not agree with tracking mechanisms can be affected for people who do agree. For example, if there are people who agree to be monitored on their driving behavior to get a discount on their car insurance, they can impose a cost on people who do not agree. The insurance company can learn from the monitored users about the different causes of accidents, so they might develop new policies to refund claims. Hence, this policy can affect all their clients, not only the ones who shared the data and agreed on being tracked.

Institutional Economics. Besides online tracking being a wicked problem, encountering misaligned incentives of different actors, facing information asymmetry, the possibility of a market of lemons, externalities, and transaction cost, Institutional Economic theory states that individual's tastes are not given, but shaped by society and institutions (Hodgson, 2000). The word institution in institutional economics refers to formal and informal institutions that rule the game of society (North, 1991). Formal institutions are written laws and rules and all type of organizations, and informal institutions are social norms, customs, culture or spontaneous rules that will be respected for the self-interest of actors (Berg et al., 2009). Hence, tracking not only has to consider the economic components of privacy, but also the institutions that surround it.

Tracking is a problem that involves two institutions Law and Market Forces, and these two institutions might drive or discourage tracking. Businesses need to compete, generate revenues, and they have created rules that represent their self-interest. At every moment in time businesses need to make a profit. Hence, businesses might play a role trying to influence other institutions, such as the law, to continue businesses' operations in a way that they still can make a profit. Also, users that are part of the market forces as customers might also have different opinions about tracking and concerns. On the other hand, the law as a formal institution needs to be flexible enough to let the businesses still make a profit and protect users' privacy. Hence, there is an interplay between these two informal and formal institutions involved in tracking.

Principal-Agent Problem. Institutional economics also addresses the principal-agent problem which also explains why tracking is difficult to solve. The Principal – Agent problem states that there is a basic relationship between an agent who makes decisions that affect the principal, and due to information asymmetry a conflict of interest might arise among them (Jensen & Meckling, 1976). In this case, there is information asymmetry between the users, the principal, and the companies that use tracking on the web, the agent. Also, as it was expressed these companies have economic interest because they are making revenues which leads to a problem of conflict of interest between the principal and the agent.

Tragedy of the commons. In institutional economics, we can also find the theory of tragedy of the commons. This theory states that individuals acting rationally and in their interest can destroy a common source even if this destruction affects everybody (Hardin, 1968). Individual companies exerting tracking mechanisms on the web might have economic incentives to over-use tracking to make a profit. Hence, collecting the online behavior in the digital world, in which these data can be aggregated, have the potential to undermine users' common right to privacy.

We hope that after describing the challenges to tackle tracking on the web, it is clear that tracking is a complex problem, and privacy governance is not an easy task. Hence, due to the complexity of this topic, different approaches have been born to address possible market failures and empower users to protect their privacy, so we will discuss them in the next section.

2.9 How Are We Protected?

As it was expressed in the previous section, it is a daunting task to protect users against tracking mechanisms on the web. Hence, multidisciplinary approaches need to be combined to try to tackle this issue. There are different instruments already in place that are part of the solution to this problem. First, we have legal instruments that have the challenge to evolve and be flexible enough to allow economic growth. Second, the use of technologies to block and detect tracking is available for users. Finally, an emerging way of privacy protection is accountability and transparency. In this section, we will address each one of them.

Regulation and Legal Instruments. As we expressed at the beginning of this chapter, there are different legal instruments that try to protect the right to privacy and keep up with the evolvement of tracking technologies. Law serves as an instrument to shape social behavior, look for the common welfare, and to institutionalize values in society (Morgan & Yeung, 2007). In addition, according to Morgan & Yeung (2007), regulations are born to fix market failures when the market can take advantage of asymmetric information to safeguard the interest of society, as it is the case of tracking.

On September 23rd, 1980 eight guidelines to govern the protection of privacy were adopted by the Organization for Economic Co-operation and Development (OECD) members to protect the right to privacy and data collection (Organisation for Economic Co-operation and Development., 2003b):

1. Limited collection: The personal data collected should be minimum and the data subject should provide consent.
2. Data Quality: the data collected should be relevant for the purpose that is intended to be used, and it should be accurate, complete and up to date.

3. Purpose specification: the data subject should be aware of the purpose of the data collection before it happens, and the use of the data needs to be limited to what it was specified.
4. Use limitation: data cannot be disclosed or used for other purposes that were not specified, except if it is requested by law authorities or under the consent of the data subject.
5. Security safeguards: the data needs to be protected against loss, unauthorized access, destruction, use, modification or disclosure.
6. Openness: transparency about development, practices and policies related to personal data should exist, and the identity and residence of who is collecting the data must be available.
7. Individual participation: individuals should have the right to know if a data controller has data about him or not for a reasonable cost within a reasonable period of time, and they should be able to correct, erase, or amend that data.
8. Accountability: who is collecting data should be accountable.

These eight principles were the first international instrument that made clear the need for privacy principles to protect personal data and data flow. These principles shaped the way, different countries, and especially the European Union ratified to see privacy as a fundamental right.

After these principles, in 1995, the Data Protection Directive (95/46EC) was born. This was the first regional instrument in the European Union to safeguard for data protection (Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995). Also, in 2007 the Lisbon Treaty was signed by the European Union member states. The treaty under the Title II, article 16, also ratify the right to protection of personal data ('Article 16', 2013). This again gives a clear distinction in how data protection is seen in Europe, and why it is more important than any economic benefit.

The Data Protection Directive (95/46/EC) will be replaced by replaced during the course of this master thesis by the General Data Protection Regulation (GDPR), which will enter into force on May 25th, 2018. The GDPR has a strict approach to consent, and its recital 32 states that consent is necessary before any data collection, and this has to be freely given, informed, unambiguous, and given by an affirmative act (Vollmer, 2017). Since 1997, The E-Privacy Directive was born to complement the Data Protection Directive (95/46/EC) and to safeguard for confidentiality of information. However, now there are discussions to change the E-Privacy Directive to a regulation to complement the GDPR. In terms of legal instruments, the European Union has taken the lead in the privacy protection, and the GDPR and the E-Privacy Regulation are one of the first regulations to quake the web as we know it today to protect users' privacy ('Rise of the data protection officer, the hottest tech ticket in town', 2018).

Technology Measures. Besides the policies and regulations in place and upcoming, there are also technologies that fight back tracking mechanisms.

Ghostery, Privacy Badger, Panopticlick are browser extensions that can be used to block tracking on the web. They usually optimize users' web browsers giving control to users to block tracking related to advertisement, analytics, and even they help anonymizing users' data. However, sometimes these tools make the user's browser unique. Hence, as it was explained in section 2.2, it might be possible that installing these tools provide to the users a unique characteristic to be identified and tracked using fingerprinting. Besides, users that use these technologies might need to be aware of how to install them and use them.

In addition, most web browsers today include a feature of Do Not Track. However, this feature does not impede the installation of tracking mechanism such as cookies. Other technology mechanisms are Privacy Enhancing Technologies such as The Onion Router (TOR) that allow anonymized web browsing. Also, privacy by design in web browsers is another option that is mentioned when talking about technology measures to protect privacy on the web. However, technologies are only part of the solution since companies using tracking might learn about how users are protecting themselves, and they can learn how to circumvent them.

More technologies might exist that prevent tracking mechanisms, but we will not go in deep on them because they are out of the scope of this thesis project. Also, despite their pros and cons of them, we do not mean that these technologies do not help to prevent tracking.

Web Privacy Measurement. Although there are legal instruments and technologies in place to protect users against tracking, policy makers need quantitative evidence to know the effectiveness of policies and how the web is operating regarding tracking mechanisms ('Executive Summary of the Ex-post REFIT evaluation of the ePrivacy Directive', 2017). Hence, measurement infrastructures allow collecting data about tracking and contribute to solving the information asymmetry between users and websites as well as websites and policy makers. Web privacy measurement is objective, reliable ,fast, and usually automated, and they allow to perform longitudinal studies (Mayer & Mitchell, 2012b). However, this is a newly born discipline that faces challenges especially because the causality that wants to be established is about tracking on the web which has very complex mechanisms (Englehardt et al., 2014).

The focus on this master thesis will be in two of the multidisciplinary approaches the regulation and legal instruments and web privacy measurement. First, it is necessary to determine if the legal instruments that are in place are accomplishing their purpose, considering the complexity of privacy. Second, web privacy measurement provides transparency. Transparency is difficult when the website incorporates third parties since might be the case that websites are not completely aware of third party practices. Hence, the Open Web Privacy Measurement (OpenWPM) will be used as a tool in this study to measure tracking on the web (in the methodology section a description of OpenWPM will be made).

2.10 Narrowing the Problem

Up to now, in this chapter, we have seen that tracking can be exerted in different manners, it is challenging to solve this problem, and there is the need of multidisciplinary approaches to protect users against tracking on the web. To focus this research, the scope must be narrowed down. Hence, through the literature review, we found that the most commonly used tracking mechanisms across the web are cookies (Fruchter et al., 2015; Narayanan & Reisman, 2017a; N. van Eijk et al., 2012), and one of the industries that is highly dependent and profit from them is online behavioral advertisement (Smit et al., 2014). This industry shows targeted ads on websites using the inferred preferences from users' online behavior. In addition, from the policies available the one that pays more attention to this type of tracking is the E-Privacy Directive. Hence, we will narrow down the scope of the analysis of this thesis to tracking cookies and the E-Privacy Directive in the context of the online behavioral advertisement industry.

2.10.1 Cookies in Depth

As already was briefly introduced in section 2.2, cookies are small files that are stored on the user's computer through their web browser, and they can remember users behavior, preferences, and track users for a long period of time('Cookies - European commission', 2016; Article 29 Data Protection Working Party, 2010;Leenes & Kosta, 2015a).

The development of cookies is attributed to Louis J. Montulli II known as Lou Montulli in the year of 1994 (Weber, 2000). The name "cookies" came from the term "Magic cookie" or "opaque identifier" (Kristol, 2001). The Magic cookie term was used in the Unix manual page to describe a "Something passed between routines or programs that enable the receiver to perform some operation; a capability ticket or opaque identifier" ('magic cookie', 2003).

Cookies are usually sent by the server where the website is hosted to users' web browser through the HTTP response. The server sends a "set-cookie" header along with the HTTP response, so the cookie is stored in the users' web browser. When a cookie is installed on the web browser of users' devices, these files contain the following parameters:

1) The name of the cookie
2) The value of the cookie
3) Expiration date
4) Path the cookie is valid for
5) Domain cookies are valid for

Figure 10 shows how a cookie looks like in a computer and some of its parameters.

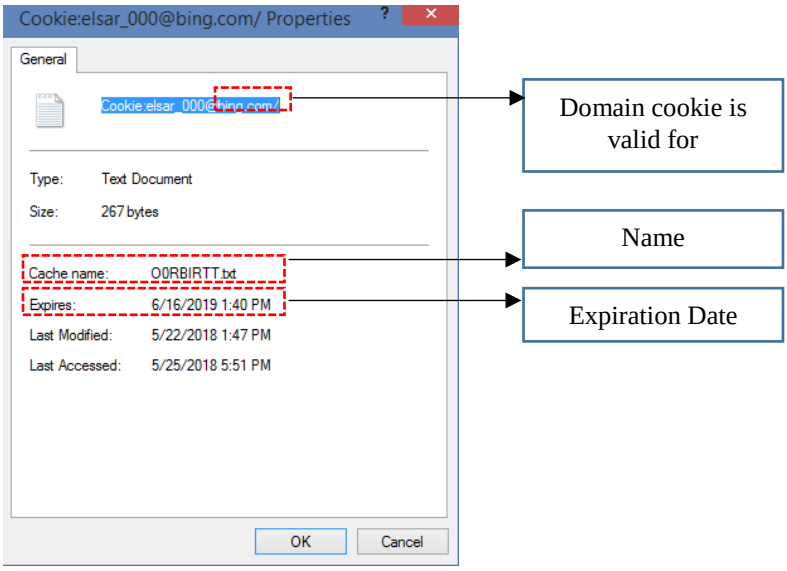


Figure 10: How does a cookie file looks like in a personal computer?

Cookies have evolved through the years, and some of the most commons types of cookies are HTTP cookies and flash cookies.

HTTP Cookies. As it was stated previously, in the beginning on the web the statelessness of HTTP hindered some activities on the web, so HTTP cookies were the first type of cookies created to offer

a solution to this problem. In 1995, the standardization of HTTP cookies took place (Kristol, 2001) with the standard RFC 2965 ('HTTP State Management Mechanism', 2000). Still, this type of cookies are the most common cookies used on websites, and they consist of a file of 4kb of data that it is stored in users' devices.

From this type of cookies, two types of cookies emerged persistent cookies and session cookies. Session cookies are cookies that only last the time a user visits a website, while persistent cookies are cookies that have an expiration date of more than the session time a user is visiting a website. In this thesis, we called persistent cookies to cookies that last more than thirty days on users' devices.

Flash Cookies. After HTTP cookies also flash cookies were created. This type of cookies are also known as "Local Shared Objects", they can store 100kb of data on users' devices, and different browsers can access them, so this means that they are more persistent and sophisticated (M. D. Ayenson, Wambach, Soltani, Good, & Hoofnagle, 2011).

Besides HTTP and flash cookies, the cookies are classified into first party and third-party cookies.

First Party Cookies. These are the type of cookies that are dropped on users' devices by the websites the users are visiting, and we will name the company that drops a first party cookie a first party. Figure 11 shows how first party cookies work.

Figure 11: First Party Cookies

Third Party Cookies. These cookies are cookies that are set by a third party company when users visit a website. Meaning a company different than the website a user is visiting is the one which drops or read the cookie. Most of the third party companies involved as third parties in websites are the ones mentioned in section 2.5 advertisers, analytics, front-end services, social integration, content providers, hosting platforms, and market research, and our focus will be on cookies related to advertisements. Figure 12 shows how third party cookies work.

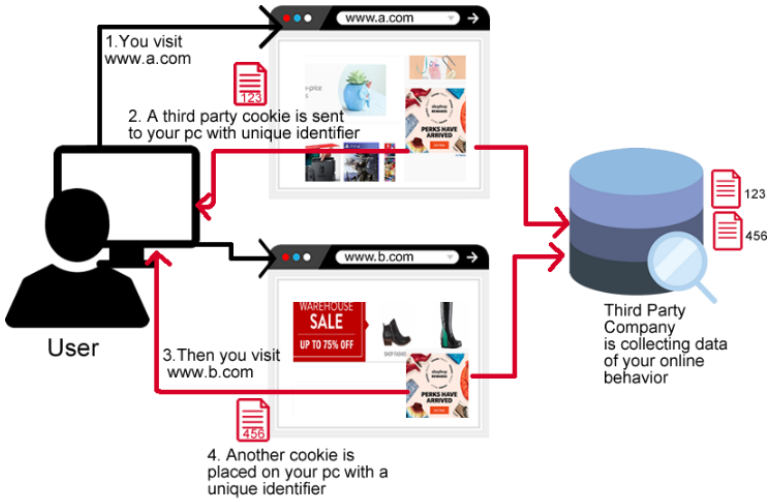


Figure 12: Third Party Cookies

Another important concept that is related to third party cookies is third party domains. Third party domains are basically the third parties' URLs which are different from the one that the user is visiting. For example, a user might be visiting www.a.com and the cookies that are placed on users' device are from www.b.com. Thus, www.b.com is called 'third party domain', which simple put is a domain that is obtaining users' data, but it is a domain that the user is not visiting and belong to a third party company.

In addition, *Java Script Calls* are also a concept related to third-party cookies. Java Script is a language commonly used in website development. Hence, this language is being used to drop or read third party cookies in users' devices. A Java Script Calls is a call to a third party domain. For example, a user might be visiting www.a.com, and www.a.com calls through the Java Script code of its website to www.b.com. This code is transparent to users, and it is not possible to observe it or delete from the web browser settings.

These concepts are relevant to understand for the elaboration of the metrics for the data analysis that we will present in the methodology section. Hence, a summary of the definitions is presented in Figure 13.

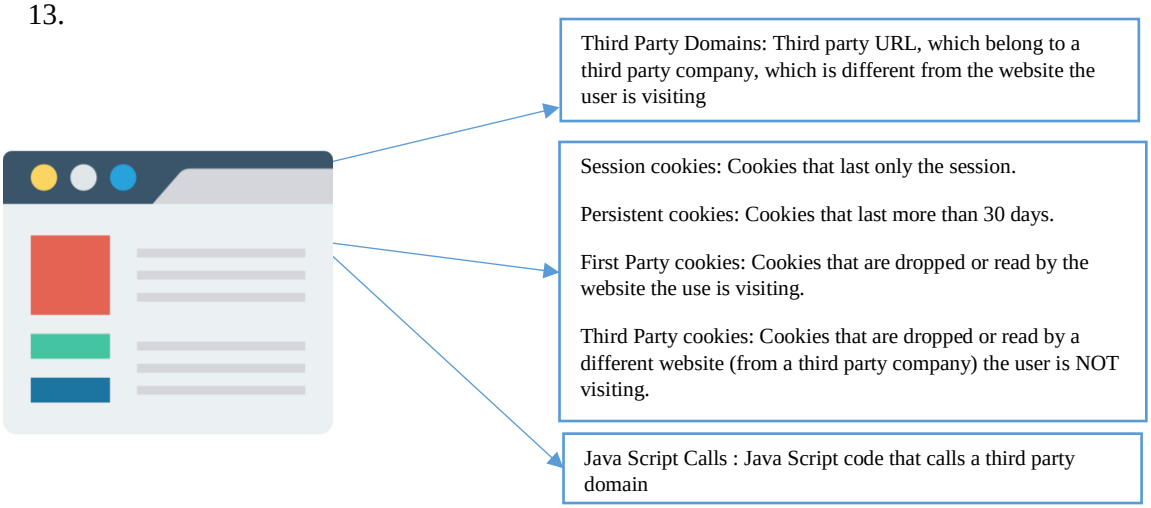


Figure 13: Principal concepts related to cookies

Finally, we need to clarify some of the misconceptions about cookies. Cookies are not a virus, or they cannot contain executable files. Also, they are not spyware. In addition, they do not collect personal data by themselves. Cookies are a text file that it is stored in users' hard drive. Usually, cookies are retrieved by the server of the website a user is visiting to obtain information about users' preferences.

2.10.2 The E-Privacy Directive in Depth

So far, we have been mentioning the E-Privacy Directive as the regional instrument that protects individuals against tracking. The directive pays special attention to cookies and tracking. In, 1997 the Directive 97/66/EC was born to complement and harmonize the privacy protection in member states focused on the telecommunication sector ('EUR-Lex - 31997L0066 - EN', 1998). Later in 2002, the directive 97/66/EC was amended by the Directive 2002/58/EC because the Directive 97/66/EC did not address the internet as a new technology, so internet technologies were introduced (Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), 2002). Finally, in 2009, the Directive 2009/136/EC amended the directive 2002/58EC having a more strict approach to the notification of data breaches, requiring consent from users to store information in their equipment, and reinforcement the protection of users against unsolicited information (Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws (Text with EEA relevance), 2009). Now, the E-Privacy Directive will become the E-Privacy Regulation, and there is already a draft of a proposal to change this regulation. The E-privacy Directive was a lex specialis of the Data Protection Directive 95/46/EC that will become the General Data Protection Regulation(GDPR) on May 25th, 2018. Lex specialis means that the E-Privacy Directive pays attention to specific subject matters and will prevail over GDPR that applies to general matter. Hence, the E-privacy Directive overrides the Data Protection Directive in specific subjects among them cookies.

From the complexity of privacy policies, it is clear that these modifications and amendments of the E-Privacy Directive have been trying to re-solve the privacy problem. However, we need to remind the reader that this is not an easy task. Besides, since the E-Privacy Directive only offered a minimum set of rules, each member state was free to implement it in their national laws either providing less or more privacy protection to their citizens (Van Eijk, Helberger, Kool, van der Plas, & van der Sloot, 2012)

Due to the freedom each national legislator had, when the E-Privacy Directive was implemented by the different member states, this implementation was fragmented (Deloitte, 2017a; European Commission et al., 2015). Fragmentation in this research will be defined as the ambiguities related to how to gain consent from users to place tracking cookies in their computers, information required to give to users about tracking cookies when they visit a website, and the varying degree of enforcement capacity, fines, and guidance that each member state applied. This fragmentation leads to finding in the literature that the E-PD did not accomplish the expected harmonization of privacy protection in

the European Union ('Briefing EU Legislation in Progress', 2017). Hence, this fragmentation has led to different interpretations and ways to tackle tracking cookies in member states.

One example of this varying degree of implementation is The Netherlands. In the Netherlands explicit consent is expected, so users have to click or to modify preferences regarding tracking cookies when they visit a website (Koninkrijksrelaties, 2016). However, websites in response started using cookie walls that did not allow users to enter the website until they modify preferences about tracking cookies ('Frequently asked questions about Dutch cookie act', 2016). The implementation of this form of consent could cost to lose customers since the first few seconds that a user enters a website are crucial to decide to stay or leave (Bonnardel, Piolat, & Le Bigot, 2011), and although in this way users' privacy is protected because tracking cookies are not placed immediately in the users' devices, the users are annoyed with the fact that they first need to perform these actions in order to access the content of the website (Leenes & Kosta, 2015b).

Another example of a different approach to the implementation of the E-Privacy Directive is France. Although they implemented consent, they request that users need to be provided with information about the use of tracking cookies, and users need to have mechanisms to reject them (Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés - Article 32, 2011). Hence, this was a more flexible approach to consent. In this case, websites respond with the implementation of banners to comply ('France | IAB Europe', 2018). However, users tend to click yes without reading notices or not being aware of how their personal data is being collected (Liu, 2014).

Each approach has advantages and disadvantages. However, up to know and to our knowledge it is not clear *how the different local transpositions of the E-Privacy Directive impact tracking cookies presence in the European Union or which approaches were better in terms of reducing tracking cookies on the web.* Besides, *the territorial scope or which law the websites should implement was not clear* leaving uncertainty about how to comply. Companies could decide to follow the law where their users were located or the rules where the company was located.

In addition, the Data Protection Authorities are institutions that were formed to safeguard for the implementation of the E-Privacy in the different member states. However, they face different challenges, and they even count on different budgets to achieve their enforcement tasks (Custers, Dechesne, Sears, Tani, & van der Hof, 2017). This has also caused some differences among the member states in terms of their enforcement capacity. Although some people might disagree that there are different implementations of the E-Privacy Directive, as a directive, each country have flexibility to implement it in different manners.

2.10.3 Online Behavioral Advertising

As it was stated before, one of the industries that nowadays profits from cookies is online behavioral advertising. Online behavioral advertisement will be defined in this research using an adapted definition of Boerman, Kruikemeier, and Zuiderveen Borgesius (2017) as a type of online advertisement that "monitors people's online behavior when they visit websites using their computers or digital devices to show them individually targeted advertisement based on the preferences inferred by their online behavior". For example, if users visit sports clothes' websites, advertisements related to this "inferred preference" have to be shown to them.

How does behavioral advertisement work? The Online behavioral advertisement is "one of the most complex computational systems in the planet" since many servers can be involved in serving one single targeted advertisement in one single website (The office for creative research, 2013). Also, there is some evidence that companies involved in tracking online share among them the users' unique

identifiers which make the ecosystem more complex (Falahrastegar, Haddadi, Uhlig, & Mortier, 2016).Hence, here, we will limit ourselves to give a brief explanation of the four major actors involved in online behavioral advertising and how they interact.

The first actor is the Publishers. They are the websites that display targeted advertisements. Targeted advertisement will be used as the definition of the ads shown to users using online behavioral advertising.

Second, the advertisers. They are companies that want to announce their products or services through publishers.

Third, the online behavioral advertisement industry. This industry is complex, and it involves different actors. Among the principal actors, we have the Advertisement Networks, Demand Side Platforms, Supply Side Platforms, Ad Exchange, and Digital Advertisement Agencies. A brief definition of these five actors will be made based on an adaptation of Yuan, Abidin, Sloan, & Wang (2012). First, we will start defining the Advertisement Networks. They are third-party companies in charge to connect publishers with advertisers, and they count on affiliate publishers to serve targeted advertisements. Second, the Demand Side Platforms. They are platforms in charge of aggregating the online behavioral advertisement's demand. Third, the Supply Side Platforms. They are the platforms in charge of aggregating the online behavioral advertisement's supply. Fourth, the Ad exchange. This is a marketplace platform that meets the aggregated supply and demand of online behavioral advertisement, here real-time bidding takes place to select which advertiser will appear in the publisher based on the inferred preferences of the user base on his online behavior.

The fourth and final main actor is users or the target audience. People like you who visit different websites every day, from whom online behavior is collected, and who are targeted with targeted advertisements.

Figure 14 depicts the main actors involved in online behavioral advertisement.

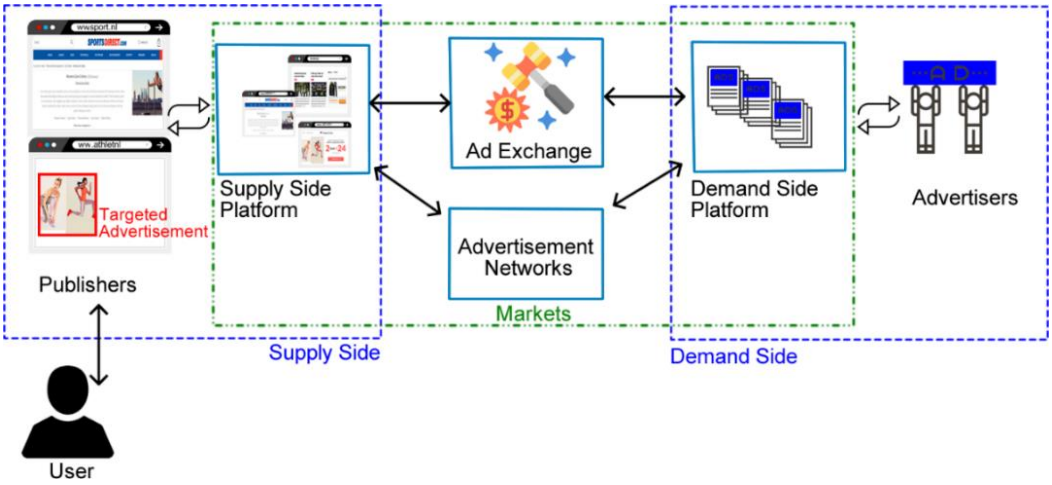


Figure 14:How does Online Behavioral Advertisement work?

(Yuan et al., 2012)

Economics Incentives of using Online Behavioral Advertisement. As we expressed in section 2.8, there is an economic component of privacy. The major actors in the online behavioral advertisement industry, previously described, the advertisers, and the expenditures in online behavioral advertisement might drive economics incentives, especially on generating revenues streams and employment, from these actors. According to IHS Markit, the likelihood of a person clicking a targeted advertisement is 5.3 higher than a normal online advertisement. By "normal online

advertisement", we mean advertisements that do not use online behavior to appear on the websites. E.g. fixed advertisements on the websites. Hence, targeted advertisement can potentially create more sales transactions for advertisers' companies motivating businesses to pay for them. In addition, online behavioral advertising is generating employment in research and development, algorithm developers, and creative advertising (IHS Markit, 2017), so there is economic growth in the European Union based on targeted ads.

On the other hand, there are arguments against this point, there are privacy advocates such as Jason Kint that states that this industry do not make more than 10% of their revenue based on this type of advertisement ('Behavioral Advertising Might Not Be As Crucial As You Think', 2014). Hence, this industry might not be generating as much revenue from this type of advertisement to justify tracking on the web.

Hence, a second problem we identify is that *we do not know how the market forces and its economic incentives can also influence tracking in the European Union, and the interplay of these the businesses' incentives with the law*.

Online Behavioral Advertisement versus the E-privacy Directive. In addition, Online Behavioral Advertising industry claims that cookies do not collect personal data. They state that cookies are used to give a better experience to users and provide relevant content (IAB Europe, 2015). However, according to the Working Party 29 (WP29), cookies fall into the category of personal data. The WP29 was formed by a group of representatives of the Data Protection Authorities of member states. They have emitted opinions and recommendations on what is tracking, how to gain consent from users, what the word consent means, among other topics. WP29 states that it is not necessary to collect personal data or to have the name of a person to identify him. For example, with the use of identifiers, IP address, and the cooperation of the internet provider is possible to identify a person (Clifford, 2014). In addition, the combination of off-line data with identifiers can lead to identify a person. Hence, although Online Behavioral Advertising industry says that they do not collect personal data, it might be possible to identify users.

2.11 Objective and Research Questions

In this crucial moment in which the E-Privacy Directive will become the E-Privacy Regulation and this ongoing debate with different sides, opinions, and approaches to protect users' privacy, there is a lack of literature about the problems encountered in the scope selected. Hence, the research objective of this master thesis will be to **"Bring empirical evidence to this debate"** and empirically and quantitative test the differences in the implementation of the E-PD, territorial scope, and determine what legal and market forces factors can explain tracking cookie presence in the EU, and how the E-Privacy Regulation can reduce this tracking mechanism.

Hence our main knowledge gap is:

What legal and market forces factors can explain the presence of tracking cookies across the European Union, and how the E-Privacy Regulation reform can reduce tracking?

To answer this main research questions, the following sub-research questions need to be answered:

SUBQ1: What is tracking? How pervasive are they in European Union countries? and What are the type of tracking in use?

SUBQ2: Which laws do websites follow? Are there differences in tracking and cookies notices related to the laws they follow?

SUBQ3: What local provisions of the E-Privacy Directive and market forces factors, if any, encourage or discourage tracking presence across member states?

SUBQ4: What are the implication of the findings to policy makers?

2.12 The Conceptual Model

After the literature review presented in this chapter, and now that the research questions of this thesis project were presented, a conceptual model where we depict our beliefs about how tracking cookies phenomena occur will be depicted.

According to (Sekaran & Bougie, 2016), the conceptual model is a graphical representation of how the concepts or variables under study are related, and a theory that explains the relationship among concepts or variables needs to explain the relationships.

We determined that there are two major institutions surrounding online tracking, the law and market forces. The legal framework is an abstract concept, so this has to be operationalized to make it measurable (Sekaran & Bougie, 2016). Hence through literature review, the main provisions of the E-Privacy Directive implemented in member states will be used as independent variables. Also, market forces will be translated into businesses and companies' websites characteristics and users or target audience' characteristics to make it measurable.

2.12.1 The Legal Framework

Six provisions of the law were determined to be relevant for the legal framework that can be measurable, so we want to determine the impact of them in tracking cookies presence. A brief description of them follows.

Consent: According to Cookie collective (2014) if a country has opt in mechanism or explicit consent, the use of cookies is blocked when the users first arrive in a website, so cookies are only set until there is an interaction between the user and the website, which is taken as consent, so usually cookies are set on the second page of a website. On the other hand, if the country relies on implied consent this means that websites set cookies on the first arrival of the user giving to them the possibility to opt out or change the cookie preferences later. There are countries which apply explicit consent and other countries applied implicit consent or flexibles forms of gaining consent.

Guidance: Data Protection Authorities acquire different tasks to ensure the implementation of the E-Privacy, one of them was to advise actors in the implementation of the law. Hence, some Data Protection Authorities did emit guidance on how to implement the law while others did not.

Fines: Fines schemes are a way to reduce the opportunistic behavior of businesses against the users. Dealing with fines will represent a cost for companies, and fines also can affect companies' reputation. Some member states developed fines schemes and others did not, so this is another provision which we want to study. It is worthy to clarify that the fines schemes which we are considering are not directly for tracking mechanisms, but for data protection.

Enforcement capacity: To ensure the application of the law enforcement capacity is necessary, and there are differences in member states can lead to different levels of enforcement capacity. Hence, this also will be included in the model.

Information required: The law required to provide information to users to gain informed consent from them. Some member states ask to provide more information than others to the users. A way to comply with this requirement has been the use of banners or notices.

2.12.2 Market Forces

Category of websites: We see categories of the websites as a proxy to understand the business model of companies, as well as their incentives to use or not tracking. To the best of our knowledge, Englehardt & Narayanan (2016a)already demonstrated in the analysis of the top 1 million websites that websites in the News business are more likely to have more tracking cookies. Also, Trevisan, Traverso, Metwalley, & Mellia (2017) did a study in some European Union countries where this also holds. Hence, we will study if the business models' incentives encourage or discourage tracking.

Location: This refers to the target market or country that companies decide to serve. This variable is not only part of the market forces, but also it is part of the laws the websites decide to apply.

Users or target audience: Bellman, Johnson, Kobrin, & Lohse (2004); Cecere, Le Guel, & Soulié (2015); Milberg, Burke, Smith, & Kallman (1995); Milberg, Smith, & Burke (2000), studied the relationship between culture and privacy concerns, and O'Neil(2001) pointed out a relationship between privacy concerns an people's level of education. Therefore, we will consider privacy concerns and education as control variables of the market forces that can influence tracking.

Besides, other elements as control variables will be added to the model. However, they are not included in it. These "other elements" such as the Gross Domestic Product per capital and rule of law.

Finally, our dependent variable will be called tracking cookies presence and notices. These variables will have different metrics that will be explained in more details in the Research Methods chapter.

Figure 15 depicts the conceptual model that represents the relationship between the law and market forces elements.

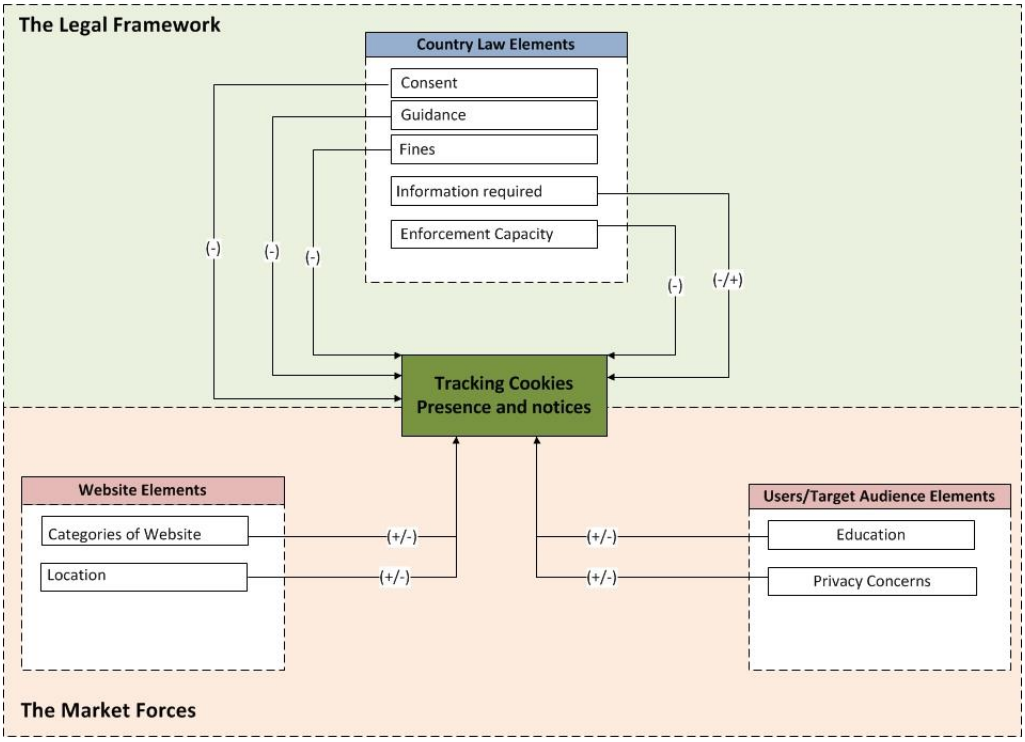


Figure 15: Conceptual Model of the tracking cookies phenomena

2.13 Chapter Summary

Nowadays, a technology that is widely used is the web, and with its progress, different tracking mechanisms have arisen that are producing tension with privacy. Some of these mechanisms are session tracking, cache tracking, fingerprinting, and storage based. The use of these mechanisms imposes risks to users such as mass surveillance, identity thief, price discrimination, among others. Nevertheless, there are different third-party companies which business models dependent on tracking, and have different incentives to use them, and one that is the most contested is the advertisement industry. In addition, privacy is a wicked problem, there are misaligned incentives among actors, and it has an economic component making this a problem difficult to solve. Hence, multidisciplinary approaches such as regulations, technologies, and transparency are being used as a mean to tackle this issue. One of the legal instruments that safeguard privacy and pays attention to tracking is the E-privacy Directive, and one of the most used tracking mechanisms are cookies, which is storage based. Hence, in this chapter, the scope of this study was narrowed down to study tracking cookies and the E-Privacy Directive on the provisions related to tracking in the context of the online behavioral advertisement industry.

This page was intentionally left in blank

Chapter 3

3. Research Methods

To test the conceptual model proposed in the previous chapter logical and systematic steps must be followed. Hence, in this section, the explanation of these steps and the decisions made will be explained. In addition, how the independent and dependent variables were collected and operationalized as proposed in the conceptual model will be detailed.

3.1 Country Selection and Control

From the conceptual model and literature review, we hope is clear that we want to test the impact of the legal framework and market forces on tracking cookies presence in the European Union. So far, we have been talking about to study tracking cookies in the European Union. However, the collection of the independent variables as well as including all the EU countries in the analysis is hard due to the time constraints of this master thesis. Hence, the first decision that was made is that not all EU countries will be studied under scope selected.

To reduce the list of countries we used two criteria broadband connections and population. We selected countries which broadband connections were higher than 2,500,000 as well as countries with a population higher than 8,000,000 inhabitants. This decision was made based on two reasons. First at all, policy makers must consider the population a policy will impact. We consider that large populations pose a bigger challenge since more people will be affected. Besides, more dynamics among different groups of actors might impact the outcome of the policy. Secondly, tracking is closely related to the use of the internet. Countries with more fix/wireless connections have more potential internet users, so there are most people exposed to the use of tracking cookies mechanisms on the web. Besides, countries with more population and broadband connection might be more interesting to be a target market for online behavioral advertisement industry.

After using those criteria, the final list of countries to study consisted of Austria, Hungary, Czech Republic, Portugal, Greece, Sweden, Belgium, Romania, Poland, Netherlands, Spain, Italy, United Kingdom, France, Germany. In total, 15 European Union countries. Table 1 depicts all EU countries, and it highlights the countries that were selected.

Table 1:Country Selection based on broadband connection and Population

Country	CC	Population_2017 (Mill)	Broadband_subs (fix/wireless)_2017
Iceland	IS	338,349	130,131
Malta	MT	460,297	171,293
Luxembourg	LU	590,667	208,500
Cyprus	CY	854,802	278,483
Estonia	EE	1,315,635	382,466
Latvia	LV	1,950,116	525,065
Slovenia	SI	2,065,895	591,749
Lithuania	LT	2,847,904	857,761
Croatia	HR	4,154,213	1,043,795
Slovak Republic	SK	5,435,343	1,363,674
Ireland	IE	4,784,383	1,378,994
Bulgaria	BG	7,101,859	1,640,921
Finland	FI	5,503,297	1,707,000
Norway	NO	5,258,317	2,140,340
Denmark	DK	5,748,769	2,460,031
Austria	AT	8,772,865	2,514,600
Hungary	HU	9,797,561	2,875,362
Czech Republic	CZ	10,578,820	3,148,731
Portugal	PT	10,309,573	3,464,636
Greece	GR	10,768,193	3,686,911
Sweden	SE	9,995,153	3,715,974
Belgium	BE	11,351,727	4,319,929
Romania	RO	19,644,350	4,448,950
Poland	PL	37,972,964	7,047,821
Netherlands	NL	17,081,507	7,184,200
Spain	ES	46,528,024	14,167,814
Italy	IT	60,589,445	16,137,606
United Kingdom	UK	65,808,573	25,334,955
France	FR	66,989,083	28,055,000
Germany	DE	82,521,653	32,529,616

Another decision made was to include 5 control countries that do not have cookies related laws to have a benchmark to compare. The countries selected were Switzerland, Australia, Japan, Canada, and The United States. The United States was an interesting control country because it has an agreement with the European Union to protect personal data when it is exchanged between them (‘Privacy Shield Program Overview | Privacy Shield’, 2016). Canada and Australia were chosen because they only have developed principles to protect their citizens’ privacy(Office of the Australian

Information Commissioner, 2014; Office of the Privacy Commissioner of Canada, 2015). Hence, their approach is softer in terms of privacy protection. On the other hand, Japan and Switzerland have data protection laws, but not specific rules about tracking(Personal Information Protection Commission Japan, 2017; The Federal Council, 1992). Also, all these countries have relied on principles, guidelines, and code of conduct. Hence, the main reason to select them as control was that they have totally different approaches to protect users' privacy than the European Union, so we wanted to compare these different approaches versus the European Union which have the E-Privacy Directive to safeguard privacy.

Table 2 depicts the population and broadband connections of the control countries.

Table 2: Control countries

Country	CC	Population_2017	Broadband_subs (fix/wireless)_2017
Switzerland	CH	8,419,550	3,831,800
Australia	AU	24,127,000	7,525,000
Canada	CA	36,286,000	13,514,552
Japan	JP	126,994,000	39,004,612
United States	US	323,127,000	108,678,000

3.2 Collection of Independent Variables

Once the countries to study were decided. The next step was to accomplish the collection of the independent variables.

3.2.1 The Legal Framework

A thorough literature review of the E-Privacy Directive paying attention to the tracking provisions of the selected member states was made. We could only find two legal firms, DLA Piper and Field Fisher, that provided the classification of the transposition of the provisions of the E-Privacy Directive in member states. The reports of two law firms, DLA Piper and Field Fisher (DLA Piper, 2014; Field fisher, 2015) disagree on the classification of countries that required consent and do not. However, we decided to use DLA Piper report for two reasons. First, DLA Piper offers an analysis of the type of consent required in order to install cookies in the users' devices, and the report offers a summary of whether or not guidance from the Data Protection was emitted. On the other hand, Field Fisher only offers information about consent. Hence, DLA Piper report had more information related to the variables we wanted to test. Second, DLA Piper is globally recognized, and has presence in 40 countries of the world, and in 10 of our selected countries ('About Us | DLA Piper Global Law Firm', 2018). In contrast, Field Fisher only has presence in 7 of the countries under analysis ('About us - Fieldfisher', 2018), and less worldwide presence. Hence, we considered that DLA Piper has more local information of the application on the law in more of the selected countries.

Table 3 presents the summary of DLA Piper's classification of consent and guidance.

Table 3: Classification of consent and guidance provisions (DLA Piper, 2014)

	Country	CC	Opt in required	Guidance
1	Austria	AT	Yes	No
2	Belgium	BE	No	Yes
3	Czech Republic	CZ	No	No
4	France	FR	Yes	Yes
5	Germany	DE	No	No
6	Greece	GR	Yes	No
7	Hungary	HU	No	No
8	Italy	IT	Yes	Yes
9	Netherlands	NL	Yes	Yes
10	Poland	PL	Yes	No
11	Portugal	PT	Yes	No
12	Romania	RO	Yes	No
13	Spain	ES	Yes	Yes
14	Sweden	SE	Yes	No
15	United Kingdom	UK	Yes	Yes

In addition, more literature review was followed to code the other variables of the model. Neither DLA Piper nor Field Fisher offered information regarding information required to provide to users, fines, and enforcement which were variables of the legal framework that we wanted to test. Hence, we had to come up with a classification of the countries making use of different sources.

The first additional resource to classify the information required to provide to users was the Interactive Advertisement Bureau (IAB), the leading association of online advertisement industry in Europe. This association, offers the "E-Privacy Directive Implementation Center" ('Europe's Cookie Laws', 2018).The implementation center is a website with information regarding how businesses need to comply with the E-Privacy Directive in each member state. Besides, the laws of each country were read to determine the degree of information necessary to provide to users.

The variable fine was operationalized using the transposition of the law of each country. It is worthy to mention again that the fines we looked at were not specifically related to tracking cookies, but to data protection. Since the law is extensive we used three keywords: Fines, Penalties, and Sanctions to determine if there were fines or not in that country. We are aware that there might be other keywords that could have been used in the national laws to establish the fines. However, we accepted this as a limitation, and we consider this as a rough classification of the fines schemes.

The next step was to determine the enforcement capacity of the Data Protection Authorities. To operationalize this variable, we used the budget of Data Protection Authorities from 2011. This information was found in a report elaborated by the International Association of Privacy Professionals (International Association of Privacy Professionals (IAPP), 2011). Thereafter, this variable was normalized with the gross domestic product of each country. However, we considered that this is a rough measure of enforcement since it is not possible to determine where the budget has been spent, and it is not up to date. Hence, we decided to test this variable as control.

It is important to mention that the transposition of the law in some countries was found in English, but in other countries was found in a different language. These countries were Belgium, France, Germany, Poland, Portugal, Romania, Spain, and The Netherlands. Hence, Google translate was used to read them, and look for the keywords related to fines. Hence, we are aware that this can create some criticisms for this classification. Besides, we are aware that the variable created through our own literature review might have different interpretations. Hence, we understand this is a limitation of the coding of these variables. Table 4 has a summary of our own classification for fines, information required to provide to users, and the budget of data protection authorities.

Table 4: Classification of countries using different sources and literature review

Country	CC	Fines	Info_users	Normalized Budget DPA 2011 (Mill)
Austria	AT	Yes	Yes/ purpose of processing, personal data identified, for how long the data will be stored.	11.6
Belgium	BE	Yes	Yes/ purpose of processing, rights	9.7
Czech Republic	CZ	No	Yes/Purpose of storing the data	23.2
France	FR	No	Yes/Purpose of any action to access, Representative, information already stored in its electronic communications terminal equipment, means available to oppose it.	2.83
Germany	DE	Yes	Yes/Nature, scope and purpose of the collection and use of personal data and the processing of his data in countries outside EU, information available anytime	2.52
Greece	GR	Yes	Yes/Purpose of data processing, processor's identity and the identity of his/her representative, if any, the recipients or the categories of recipients of such data, the existence of a right to access.	15.1
Hungary	HU	No	Yes/Purpose of data procesing	9.7
Italy	IT	No	Yes/ Purpose of the procesing	19.2
Netherlands	NL	Yes	Yes/ Purpose for which information is used, unambiguously informed about these cookies (purpose, type of cookies, etc)	9
Poland	PL	No	Yes/ Purpose of storage, conditions for storing	7.3
Portugal	PT	Yes	Yes/ Purpose of the processing, Identity of the controller and of his representative, if any;, recipients or categories of the recipients of data.	22.1

Romania	RO	Yes	Yes/Purpose of processing the information, if the provider allows third parties to store or access information the general purpose the processing of this information by third parties, and how the subscriber or user can use the settings Internet browsing application or other similar technologies to delete stored information or to refuse third parties access to this information	0.7
Spain	ES	Yes	Yes/Purpose of the processing	19
Sweden	SE	Yes	Yes/Purpose of the processing	4.6
United Kingdom	UK	No	Yes/Purpose of the storage or access of information	26.5

3.2.2 Market Forces

In addition, collection of the independent variables related to the market forces was made.

Categories of websites. To categorize websites, we used www.webshrinker.com. Web Shrinker is an Application Programming Interface (API) that allows passing the URL of a website or IP address, and it returns the categorization of them. Also, Web Shrinker uses International Advertisement Bureau taxonomy to categorize websites which is appropriate for this study. However, one limitation to consider is that although machine learning is used to perform the categorizations of websites, they might not be always as accurate as a human classification.

Users/Target Audience. For the target audience characteristics as part of the market forces as control variables, we collected data about their privacy concerns and education. First, the education index was collected from the United Nations education Index (United Nations, 2016). Second, to measure privacy concern we used as a proxy the percentage of the individuals' concern about the misuse of their personal data that was published in the Eurobarometer in 2009, and that was summarized by Cecere at all (2015).

3.3 Collection of the Dependent Variables

Sampling Frame. To collect the dependent variable a physical or digital dataset that represents the population need to be used, and this is called a sampling frame(Sekaran & Bougie, 2016). However, we could only find datasets with the most popular websites, so a redefinition of the population was made, and the sample we used was representative of the most popular websites. We compared five datasets Alexa, Similar Web, Cisco Umbrella, Statvoo, and Majestic Million to make the decision of which one to use. Although we compared Alexa and Similar web, unfortunately, they are not free, so we had to discard them as an option.

The other three free options we compared were Statvoo, Cisco Umbrella, and Majestic Million. Cisco Umbrella provides the top 1 Million websites globally. However, this dataset has a drawback for the purpose of this study. They create a rank that includes all popular URLs not only based on HTTP

requests, so this means that websites that are not necessarily visited by users can appear in the ranking. In addition, Statvoo offers also websites categorized and ranked. However, Statvoo is not clear on how they rank the websites. Hence, if the metric they use is not clear, it was not convenient to use this dataset. Finally, Majestic Million also offers the top 1 Million websites. The metric they use to rank the websites consists of the citations that other websites make to a website. In this way, the popularity and trustworthiness of the website ranked are ensured. The only drawback of this dataset is that it might be the possibility that spammy links might increase the ranking of certain websites. However, Majestic clarifies that tries to check for that.

Table 5 presents the datasets and how their ranks are calculated as well as the benefits and constraints of them.

Table 5: Options to extract websites to crawl from countries selected

Majestic Million		**Cisco Umbrella**		**Alexa overlap with Cisco Umbrella**		**Alexa**		**Similar web**		**Statvoo**	
Based on number of citations from OTHER websites (Links).		Based on open DNS. Ranking is based on internet activity not only from browsers (port 80)		Alexa List posted by Censys 2015: HTTPs://scans.io/series/alexa-dl-top1mil matched with Cisco Umbrella		Rank is based on browser behavior of people. Rank is built upon traffic data provided by users in Alexa's global data panel		No clear how the top websites are created. They stated they use global ISP data, direct measurement from apps, and public data sources to create datasets		Stavoo separates websites in different groups or categories, and it created its own rating	
Pros	**Cons**	**Pros**	**Cons**	**Pros**	**Cons**	**Pros**	**Cons**	**Pros**	**Cons**	**Pros**	**Cons**
Free, data is about websites that users visit, used for SEO (Search Engine Optimization) and marketing purposes in the marketing industry, updated	Spammy links can flaw the index	Free, updated	It is not based on HTTP request, but DNS lookup. Hence, it creates a list with URLs that are not websites visited by users	Alexa limit to websites used by users	We can miss new websites; data Alexa is outdated	Websites are categorized, updated	No free	Websites are categorized, updated	No free, no price in website, only shows 50 websites free	Free, websites are categorized	They do not express which criteria are using to create its own metric.

After analyzing the different options, Majestic Million database was chosen to collect the sample of websites from the 15 countries selected. It is free, and it seems to correct for the possible flaws of the rank. In addition, this rank is used for the online marketing industry as well as to improve the search engine optimization of websites. Also, we found that recent studies such as Castro at al (2017) used this list, and this top list is more stable than Alexa and Cisco Umbrella (Scheitle, Jelten, Hohlfeld, Ciprian, & Carle, 2018).

Sampling. To collect the sample a proportionate stratified sampling was selected. A stratified sample means that the dataset will be divided into strata or groups, and after the same number of samples will be collected per group. The Majestic Million dataset was separated into strata per top level domain. A Top Level Domain (TLD) is the last part of the URL of a website. E.g. .com, .nl, .es, .us. Thereafter, the top 100 websites were selected. One limitation we had to face is that not all TLDs had a top 100 websites, some had 99, 98 or even 80, so we had to adjust using the maximum top websites we found in the list. The decision to use the top websites was made based on the literature review of studies that used the same methodology (M. Ayenson, Wambach, Soltani, Good, & Hoofnagle, 2011; Soltani, Canty, Mayo, Thomas, & Hoofnagle, 2010; McDonald & Cranor, 2011) to analyze tracking and tracking cookies. Although a random sample could have been used, the top 100 websites represent the most visit websites in those countries and more people are exposed to tracking when visiting them.

In addition, these websites might represent large companies that can interpret the law. Besides, in the context of online behavioral advertisement these websites might be preferred to show ads since they have more popularity. Finally, the top 200 of websites .com and .org were added to the crawl as a comparison group of the websites of each country.

OpenWPM to crawl websites. To collect data on the dependent variables a measurement platform is necessary. Hence, in this study, the Open Web Privacy Measurement (OpenWPM), an open source web privacy measurement which allows simulating users visiting websites, and recording the response metadata, cookies, and behavior of scripts (Englehardt & Narayanan, 2016a), was used.

OpenWPM is a sum of the efforts of Princeton Web Transparency & Accountability Project to automate privacy measurements (WebTAP Princeton University, 2018). It allows "to detect, characterize, and quantify privacy-impacting behaviors"(Englehardt & Narayanan, 2016, 'Introduction'). This project was born in 2013 when a group of researchers from Princeton University detected around 20 studies which automated web browsers to study privacy or security, and they found out that they face the same challenges and problems (Narayanan & Reisman, 2017b). Hence, they wanted to create a framework that allows a realistic simulation of a user visiting a website using Firefox web browser, collect stateful measurements, collect network requests between the website and web browser through a proxy, simulate certain users' profiles, collect all the information that is stored in the hard drive of an user, and with stability to avoid crashes while collecting the data. This tool facilitates researchers' tasks since the only part that has to be done is the analysis once the data is collected. In addition, being an open source framework, modifications are possible to make more sophisticated analysis.

OpenWPM has a browser manager which is in charge of running each browser process and execute the orders of a task manager (*OpenWPM*, 2014/2018) . The task manager contains all the configuration settings and monitors browser managers (Englehardt & Narayanan, 2016; Englehardt & Narayanan, 2016). Then the Browser managers pass the instructions to Selenium. Selenium allows the automation of the browsers ('Selenium - Web Browser Automation', 2018). Finally, the data collected is passed to a data aggregator, in this case, SQLite (Englehardt & Narayanan, 2016a), and it is ready for data analysis. Figure 16 shows how OpenWPM works.

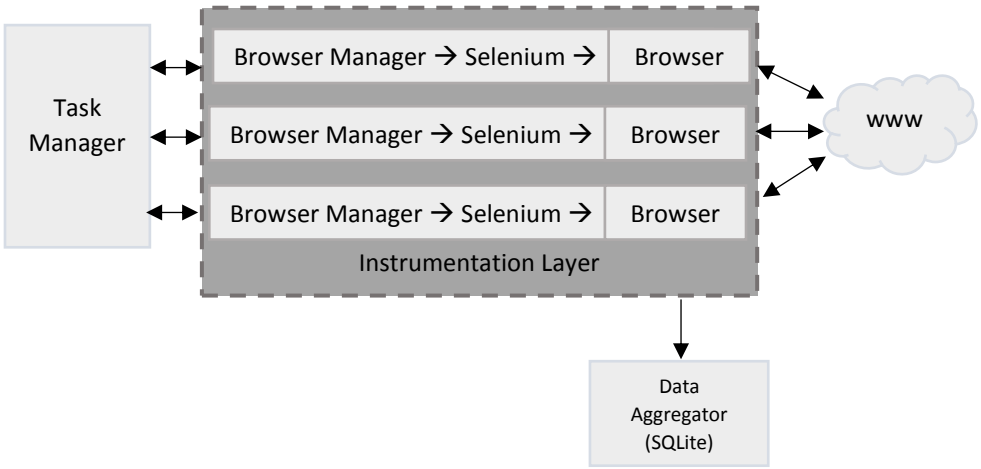


Figure 16: How does OpenWPM work?
Adapted from (Englehardt & Narayanan, 2016a)

In the data collection of this project, two additions were made to the Open Web Privacy Measurement framework. First, since the websites crawled can contain malicious files, virus, the infrastructure where the crawled data was stored could be compromised. Hence, to avoid this, dockers were used.

Dockers separate the infrastructure from applications. Secondly, a Virtual Private Network (VPN) was used to simulate a cross visit of the websites from different locations. The data collection of the dependent variable was made by Professor Hadi Asghari of the faculty of Technology Policy and Management of TU Delft as well as the modifications to the Open Web Transparency Framework.

Banners/Notices. To accomplish the collection of this dependent variable, 'I do not care about cookies' global rules list was used (Kladnik, 2018). I do not care about cookies is an add on for web browsers that remove the notices, so users do not have to deal with the banners. The limitations of using this list is that only the banners that were on the global list were collected.

3.4 Metrics of the Dependent Variables

Different studies have used different metrics to define tracking. Hence, to understand how to measure tracking cookies in this master thesis a review of them was made to determine which metrics to use.

The first column of the table, Type of crawl, contains information on the type of crawl the authors did to obtain the cookies from websites. Some of them did a homepage crawl which means that they only collected the cookies that were set on the user device when they visit the first page of the website. Other studies did a deep crawling which means the authors studied cookies that were placed when they visit different pages from a single website.

The second column, Filter, is what they defined as tracking. As it was explained in the literature review, there are different types of cookies, and some of them are considered session cookies. Hence, to only study the cookies that persist in the users' devices certain criteria have to be met. For example, the majority of studies consider only cookies that have an expiration date of more than one month once they are installed in the users' devices.

The third, the column, Comments, expresses some details about the study. Finally, the column source is the reference to the study.

Table 6 contains a summary of the literature review of papers that have measured tracking cookies on the web, and the columns previously explained.

Table 6: Literature Review of papers that have studied tracking cookies on the web

Type of crawl	Filter	Metric	Comments	Source
No found	No found	Counting third parties cookies	There is no difference between using entropy and counting third party cookies. Asghari, van Eijk, Englehardt, Narayanan, and Winter (2016) found there is a Spearman correlation of 0.957 between entropy and counting third party cookies	(R. van Eijk, 2017)
No Found	No Found	Counting cookies	Third and First party cookies were studied	(Altaweel, Good, & Hoofnagle, 2015)
Deep crawling	More than three months lifespan, unique value across different browser instances, high entropy	No mentioned	Focus on HTTP cookies since they are in every HTTP request and to send the identifier to trackers they need to be included in HTTP cookies or query in the request	(Englehardt et al., 2015)
No Found	Expiry length of more than one month.	No mentioned	Evercookies, cookie syncing, and respawning were studied	(G. Acar et al., 2014)
Home Page Crawl	Expiry length of more than one month.	No mentioned	This study was about mobile tracking. They did not clear the state	(Eubank, Melara, Perez-Botero, & Narayanan, 2013)
No Found	No Found	Logarithmic expiration time first party cookies vs. third party cookies	They found that the expiration time of first party was larger than third party. First party also had at least one cookie of more than one month expiration time, so they use log scale.	(Eubank et al., 2013)
Homepage crawl	No Found	Counting	Top 100 websites Quantcast and 500 random websites. They found only 2 companies respawning on the top 100, and non in the random sample.	(McDonald & Cranor, 2011)
10 arbitrary clicks in the domain (Deep crawling) and clean state between sessions	No Found	Counting	Top 100 websites Quantcast was used for the analysis	(M. Ayenson et al., 2011)
No Found	Flash cookies and HTTP cookies were studied. They cleared all cookies except flash and clean the state of the browser, but they kept flash cookies	Frequency of cookies	Respawning was found. Flash cookies respawned HTTP cookies. Top 100 domains Quantcast was used for the analysis	(Soltani et al., 2010)
No Found	No Found	They count the presence of third party domains	From 2005 to 2008, they study the presence of third party domains. They use the root domain approach to discover third party	(Krishnamurthy & Wills, 2009a)

After this literature review to analyze tracking cookies these decisions were made to be implemented in this master thesis:

Type of crawl:

- Home page crawl.

Filter:

- Only cookies that have an expiration date of more than one month
- Cookies that have a different domain than the website crawled
- No session cookies

Metrics:

- Count of unique third party cookies names present in a website
- Count of unique third party domains that have presence in a website and are persistent.
- Java Script Calls to third party domains.
- Average of unique third party domains per unique website.

Counting might seem a naïve metric. However, from a legal perspective, it makes sense to count how many cookies are install in the users' devices. The article 5(3) of the E-Privacy Directive expresses:

'Member States shall ensure that the ***storing of information****, or* ***the gaining of access to information already stored****, in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or* ***user has given his or her consent****, having been* ***provided with clear and comprehensive information*** *in accordance with Directive 95/46/EC, inter alia about the purposes of the processing.'*

('EUR-Lex - 32002L0058 - EN - EUR-Lex', 2002)

Therefore, each cookie can be considered as information stored in the users' device.

An important part of the regulation is to provide information to users. Hence, besides the tracking metrics that were used in other studies, we added banners presence to check for cookies notices.

To understand better the metrics, simple examples of how the metrics counted the dependent variables will be presented.

Metric 1: Count of Unique Cookies Names. This metric counted only third party unique cookies names that appear in a website. In the example in Table 7, the metric counted 2 unique cookies names for website www.a.com and www.b.com. Note that Cookie_Life represented the life span of a cookie in users' device.

Table 7: Count of unique third party cookies names

Site_url	Name	Cookie_Life
www.a.com	A	10
www.a.com	B	45
www.b.com	B	50
www.b.com	B	50
www.b.com	C	100
www.b.com	C	50

Metric 2: Count of Unique Third Party Domains(TPD). This metric only counted the Unique Third Party Domains that were persistent in a websites, which means third party companies that drop cookies that had a life span higher than 30 days. In the example of Table 8, website www.a.com and www.b.com had each 1 Third Party Domain present.

Table 8: Count of unique third party domains which are persistent

Site_url	TPD	Cookie_Life
www.a.com	a	10
www.a.com	b	45
www.b.com	b	50
www.b.com	b	50
www.b.com	b	100
www.b.com	b	50
www.c.com	c	20

Metric 3: Count of Unique Third Party Java Script Calls. A website might have multiple Java Script calls, so in this metric only unique calls to Third Party Domains were counted. Hence, in the example in Table 9, websites www.a.com and www.b.com had 1 Java Script Call.

Table 9: Count of unique third party Java Script Calls

Site_url	JScall
www.a.com	a
www.a.com	b
www.b.com	b
www.b.com	b
www.b.com	b
www.c.com	c

Metric 4: Banners. Banner was simple a binomial variable called 'Banner' that got a value of 0 if there was not banner in the website, and value of 1 if the website had banner(s). Table 10 shows that when the text of a banner is detected, the variable banners gets 1 else 0.

Table 10: Banner Presence

Site_url	Text	Banner
www.a.com		0
www.a.com	xxxxxx	1
www.b.com		0
www.b.com		0
www.b.com	xxxxx	1
www.c.com		0

3.5 Data Preparation

The data collected from OpenWPM had to be cleaned and prepared to arrive at the proposed metrics of the dependent variables presented. Also, the dependent variables were merged with the independent variables. Python was used as the tool to perform the data preparation. Python was chosen because it counts on a library, Pandas, which offers the possibility to work with data structures in a simple way optimizing the time of data processing ('Python Data Analysis Library — pandas: Python Data Analysis Library', 2018). As it will be clear in the next section the dataset had a vast amount of data points and Pandas offers the possibility to deal with the data preparation in a simple manner. Besides, Python has an intuitive syntax, it is free of semi colon, and it is open source.

Cleaning Process. We remind the reader that from the Majestic Million dataset (21/04/2018) the top 100 websites of the countries were selected. We did a cross crawl from the 15 EU selected countries and 3 control countries. The two control countries that were not crawled were Japan and Australia due to a failure. Hence, a website was simulated to the visited from 18 different users' locations. However, when doing the crawling we needed to handle some errors of sites that did not have a response. When a website visited did not have a response the HTTP status code usually get values of 400's. Also, websites where the server did not response gets HTTP status code of 500's. Hence, in line with Englehardt & Narayanan (2016a) we only used the successful responses. Besides, OpenWPM uses a flag of 0 and -1 for websites that fail to crawl. Hence, we also did not use those websites. Finally, we checked for duplicated data and delete it. Figure 17 has a diagram of the cleaning process we followed.

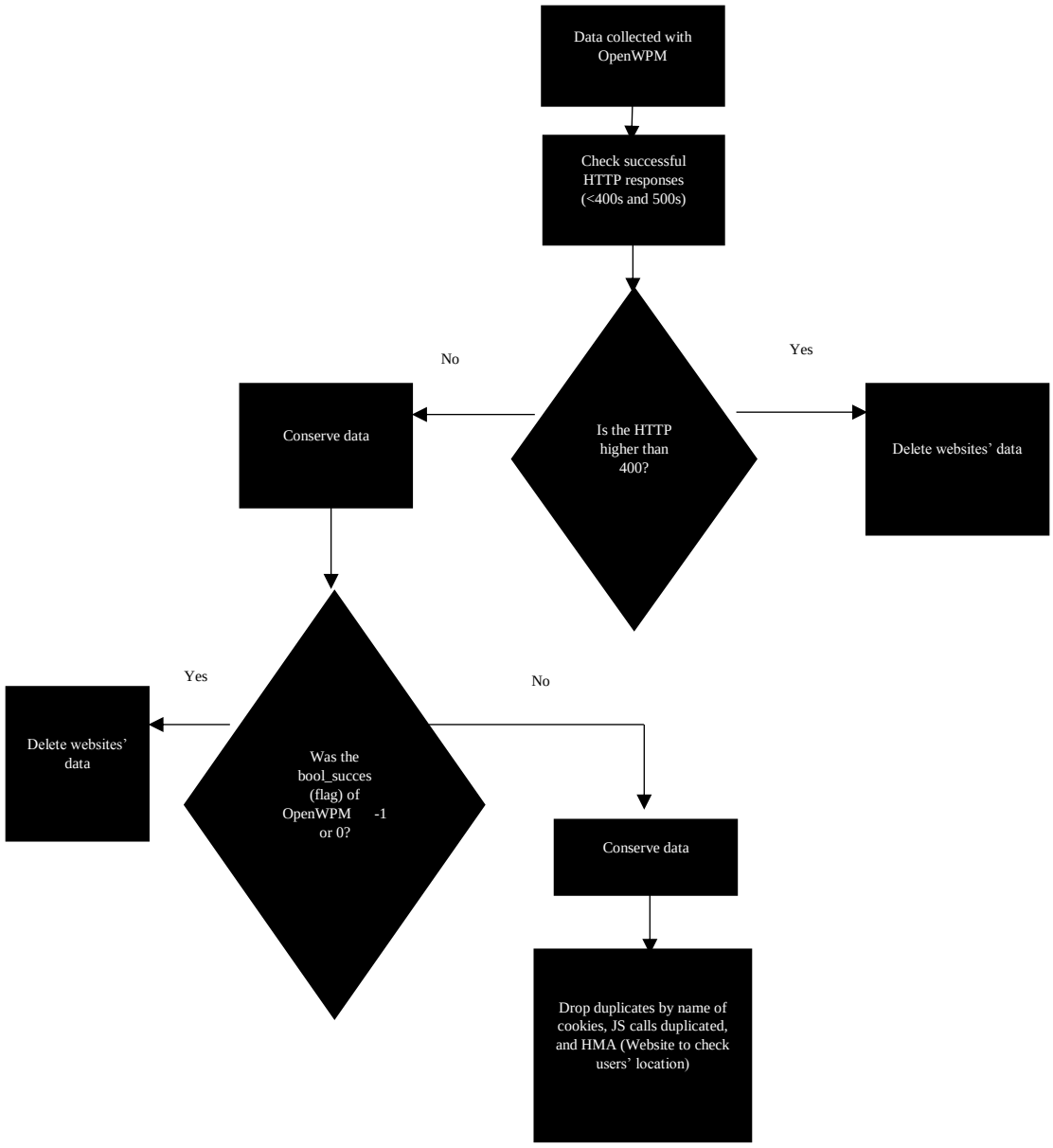


Figure 17: Cleaning Process of Data

After applying the cleaning process, the number of websites was reduced. Table 11 presents a summary of the final data. The total amount of websites in the measurement were 35,325. For the rest of the master thesis, we referred to this dataset as the measurement dataset. Since the websites were repeated per vantage points due to the cross crawl, it is important to mention that the **unique number of websites were 2010** after the cleaning process.

Table 11: Crawled data summarized after cleaning process

VP	Starting Number of Websites	After Cleaning process
AT	2271	1966
BE	2271	1970
CA	2271	19857
CH	2271	1935
CZ	2271	1967
DE	2271	1963
ES	2271	1977
FR	2271	1969
GR	2271	1962
HU	2271	1948
IT	2271	1989
NL	2271	1966
PL	2271	1953
PT	2271	1971
RO	2271	1960
SE	2271	1962
UK	2271	1959
US	2271	1951
Total Websites	48878	35325

Aggregating the dependent and Independent Variables. The independent variables related to the legal framework were converted to dummy variables leading to independent categorical variables, and they were merged with the dependent variables collected.

Three similar data frames were obtained for Unique counted cookie names, unique counted third party domains, and Unique Third Party Java Script calls. The dataset contained 5 variables. The first variable **visit_id** which was a unique number assigned to identify the website. The second variable varied as **name** for third party cookies names, **host** for third party domains, **jhost** for Java Script Calls. The third variable was the Vantage Point (**VP**) which represented users' location. Finally, **site_url** which was self-explanatory, and the Top Level Domain (**TLD**) of the website which represented the location of the website and law websites follow (More details about TLD in section 3.6). Figure 18 shows the data frame for unique counted cookie names.

visit_id	name	VP	site_url	TLD
<int>	<int>	<chr>	<chr>	<chr>
2	2	AT	univie.ac.at	at
2	2	BE	univie.ac.at	at
2	2	CA	univie.ac.at	at
2	2	CH	univie.ac.at	at
2	2	CZ	univie.ac.at	at
2	2	DE	univie.ac.at	at

Figure 18: Head of the Unique Counted Cookies Names data frame

In addition, a data frame for banners/notices was created, and we decided to call it banners for the rest of the master thesis. Banners contained 11 variables. A brief description of them will follow. **VP** represented the users' location, the variable **h** represented the height of the banner, **m** represented the css selector, **site_url** was self-explanatory, **text** contained the text of the banner or 0 if there was not banner, **visit_id** was the unique identification for websites, **w** contained the width of the banner, **x** and **y** represented the position of the banner, **banner** was the binary outcome 0 if the website has a banner and 1 if the websites did not have banner, and the **TLD** which represented the location and law websites follow (More details about TLD in section 3.6). Figure 19 shows the head of the banners data frame.

VP	h	m	site_url	text	visit_id	w	x	y	banner	TLD
<chr>	<dbl>	<dbl>	<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<int>	<chr>
US	0	0	http://univie.ac.at	0	2	0	0	0	0	at
NL	0	0	http://univie.ac.at	0	2	0	0	0	0	at
DE	0	0	http://univie.ac.at	0	2	0	0	0	0	at
BE	0	0	http://univie.ac.at	0	2	0	0	0	0	at
CA	0	0	http://univie.ac.at	0	2	0	0	0	0	at
CH	0	0	http://univie.ac.at	0	2	0	0	0	0	at

Figure 19: Head of banners data frame

After having the data frames, a code was developed to identify the cookie life either short than 30 days or longer than 30 days, and if the website had only First Party domains or Third Party domains, as well as, Unique Java Script Calls. Thereafter, we were able of counting each variable as proposed in the metric.

Table 12 depicts a summary of the independent and dependent variables that were merged and from where they were obtained, this table do not include the variables that will be used as control.

Table 12: Independent and Dependent Variables

Type of variable	Conceptual Framework	Variable	Description	Source
Independent variables	Legal Framework	Guidance	Whether the country has guidance or not emitted	(DLA Piper, 2014)
		Consent	Whether or not consent is required before placing cookies	(DLA Piper, 2014)
		Fines	Whether or not the country has defined fines	Multiple sources.
		Info_users	The level of information required to provide to users in banner	Multiple sources.
	Market Forces	Websites_categories	Category of the website as a proxy of their business models	Web Shrinker
		Location	The law the website follow and market they serve	Top Level Domain (TLD). More Details in section 3.6.2.1
Dependent Variable	Third Party Domains	TPD	Count of unique third party domains in each website crawled	OpenWPM
		TPD_unique	Average of unique TPD in each website crawled	OpenWPM
	Cookies	Name	Count of unique third party cookies names in each website crawled	OpenWPM
	Java Script Calls	JSCalls	Count of unique Third Party Java Script calls in each website crawled	OpenWPM
	Banners	Banner	Banner presence in website	I do not care about cookies

3.6 Statistical Instruments and Methods

In the data analysis, we needed to follow different methodological approaches to answer each research sub-question. Hence, in this section, we will explain these methods and statistical instruments.

Sub-question 1: What is tracking? how pervasive are they in EU countries? And what are the type of tracking in use? To answer this research question only descriptive statistics were used. The measurement dataset was used to determine the pervasiveness of tracking, using cookies as a proxy,

in the selected countries and the presence of third party trackers. Also, we used the categories of the websites as a proxy of the business models of the websites.

Sub-question 2: Which law the websites follow? Are there differences related to the law they follow? To answer this question, we used the metrics related to tracking and cookies notices explained in section 3.4. First, we inspected the shape of their distribution. Second, we tested the degree of association of the tracking metrics with a Spearman rank correlation test. A correlation measures the direction and strength of the relationship of two variables, and the value of the coefficient range between 0 and 1, values close to 1 indicate perfect correlation and close to 0 no relationship between the variables (Hair, Black, Babin, Anderson, & Tatham, 1998). Third, we checked if websites follow the law where users were located or where the companies were located. To accomplish the step, we needed to determine how we can find out which was the location of the websites, so we came up with the use of the Top Level Domains as a solution.

Determining websites location. Companies today due to globalization can be located anywhere and have different branches, so it was a bit hard to associate the websites with the physical location of the company. Hence, we used the Top Level Domains(TLD) of the websites as a proxy of the target market or country the companies decided to serve, so we interpreted it as where the websites were based. We decided to use TLDs as a proxy for websites' location because companies choose for themselves which TLD assign to their websites. This is a managerial decision to determine which is their target market. Also, Search Engine Optimization guidelines and even search engines such as Google suggest to companies to use local TLDs when they want to target a specific area(Google, 2018; Moz, 2018). Moreover, country code Top Level Domains are usually more expensive, so companies choose local Top Level Domain being aware of the implications related to cost, and they might want to have a branding opportunity to promote their websites as local, so the target audience might prefer them. In addition, the Internet Corporation for Assigned Names and Numbers (ICANN) suggests to companies that want to target a specific market to use the called Country Code Top Level Domain (ccTLD), so they can promote that they are serving a specific territory or country('Country code top-level domain - ICANNWiki', 2017). Hence, we thought that because of the cost and the implications related to the selection of the Top Level Domain by businesses, there were good reasons to assume that they are based in that country. We are aware that this method has some limitations since there might be websites such as www.bol.com of companies based in The Netherlands that serves The Netherlands market, but they do not use '.nl' as Top Level Domain. However, we accepted as a limitation since other methods were not convincing. For example, if we use the server location of the website, this also can be located everywhere, so this would not reflect the country websites were targeting.

Once we came up with this method to determine websites' location, we proceeded to determine if websites followed the rules of users' location or their location. We used the measurement dataset (35,325 Websites), and we separated the observations of the metrics base on their Top Level Domains, meaning the location of the websites. Thereafter, we run two models. A null model and a model using the users' location as dummies (Vantage Points). Since we knew we were dealing with count data we knew that we will use a form of the Generalized Linear Model, so we could compare the Akaike Information Criterion (AIC), which is a criterion introduced by Akaike in 1973, to compare which is the best model fit. Hence, we could determine if websites belonging to a certain location were using geolocation to determine users' locations (Vantage Point (VP)) and adapt their websites to users' location rules or if they stick to the rules of their location. It is important to notice that also the TLD represented for us the local law the websites should follow, so if the Model 2 was not preferred,

meaning that users' location was not relevant, websites follow the local laws of their target market. Besides, we added websites with TLDs .COM and .ORG which for us did not belong to a specific location, but as a point of comparison. Also, we need to remind the reader that users were expected to be simulated in the 15 EU selected countries and the 5 control countries as Vantage Point. However, we did not obtain data from Japan and Australia due to a failure of the crawl, so we only simulated 18 Vantage Points.

Finally, we modeled Third Party Domains and Banners as a function of the TLD as a proxy of the local laws they follow to determine if there were differences in tracking cookies presence. The dependent variables Third Party Domain and Banners were an average per all vantage points as, and we left out the TLDs .COM and .ORG. This decision was made because we did not know neither their location nor which laws they were following.

Figure 20 shows the steps followed to answer this sub-question, and the statistical instruments used.

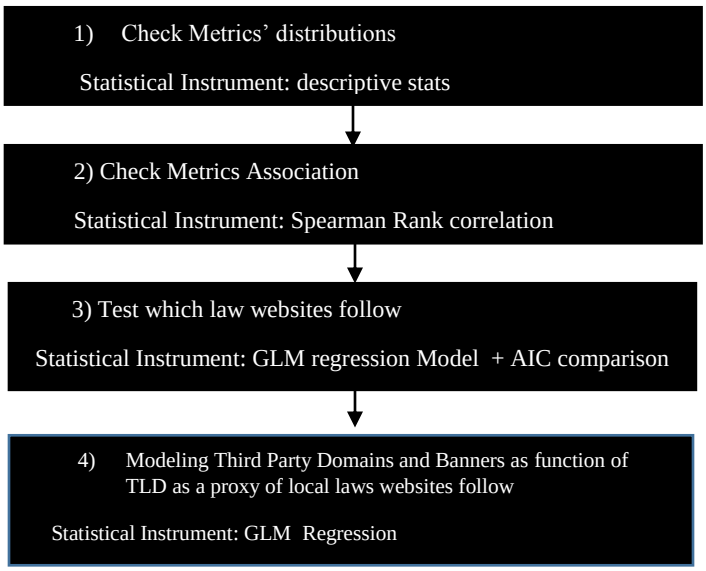


Figure 20: Statistical Instruments and Methods Sub-question 2

Sub question-3: What local provisions of the E-Privacy Directive and market forces factors, if any, encourage or discourage tracking presence across member states? To answer this question, we used the conceptual model presented in Chapter 2 to come up with the hypothesis to test. We tested the provisions of the E-Privacy Directive alone and controlling for the websites' categories, as a proxy of businesses' incentives to use tracking. We run two regressions models, and we compared their AIC values. As dependent variable, we used again the average of the counted third parties domains from all vantage points only since the focus of the question was on tracking presence, and we left out websites .Com and .Org as for the same reasons that in the previous questions.

Secondly, we modeled tracking as a function of the websites' categories as a proxy of business models' incentives to use tracking. Thereafter, we tested the effect of the law and provisions of the E-Privacy Directive versus the market forces. We added the control variables related to the target audience, and we compared a model using TLD as a proxy of the local law the websites follow with

the models of the market forces, and provisions. We used the AIC values and Log Likelihood to determine the best model fit. Hence, in the end, we could conclude which model had more predictive power, so we could determine which factors were more powerful encouraging or discouraging tracking if the local law or the market forces. Figure 21 depicts the methodology and statistical instruments used.

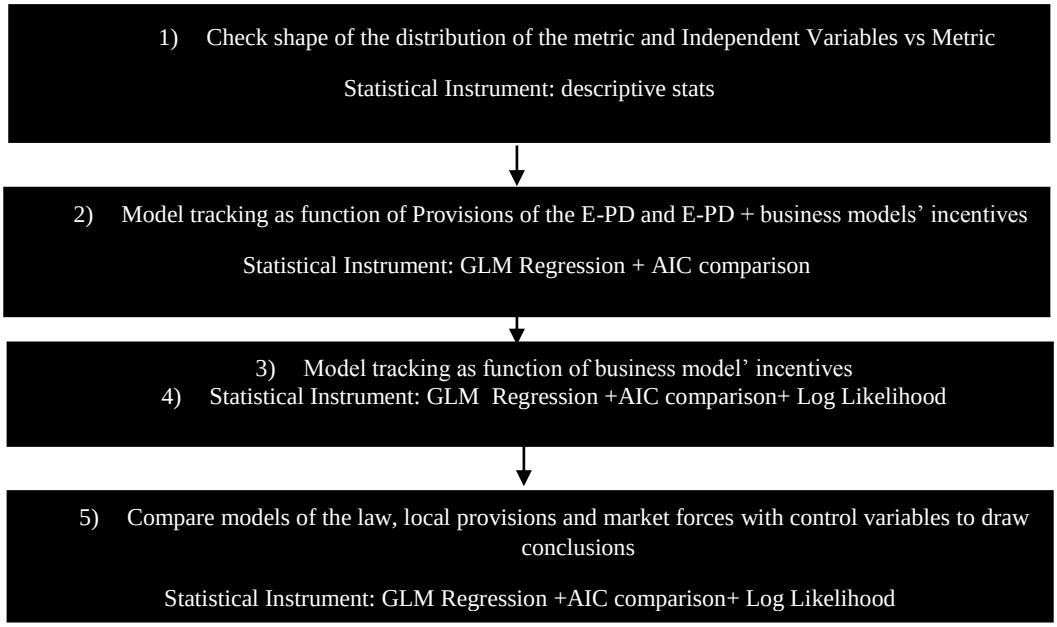


Figure 21: Statistical Instruments and Methods Sub-question 3

Sub-question 4: What are the implications of the findings for policy makers? No statistical methods were used in this question, and it was answered through reflection and interpretation of the findings.

3.7 Chapter Summary

In this chapter, we presented the choices we made and the methodologies we followed to answer our research sub-questions. We selected 15 European Union countries based on two criteria population and broadband connections, and 5 control countries that have different approaches to protect privacy. The independent variables were collected through literature review and secondary sources. The sample was selected from top 1 million websites of Majestic Million using a stratified sample, and the collection of the dependent variable was made using OpenWPM. Also, through literature review metrics to measure tracking were developed, and the data was cleaned to use only successful crawls. In addition, we described the approach to answer the research sub-questions. Also, we explained that we used TLDs as a proxy for the location of the websites and the local law the websites follow to answer sub-questions 2 and 3. Finally, we explained that comparing the predictive power of the models using AIC values and log-likelihood, we determined which models were the best fit and arrive at our conclusions and measure the impact of the market forces, law, and the provisions of the E-Privacy Directive.

This page was intentionally left in blank

Chapter 4

4. Findings

We arrived at the core of this master thesis. In this chapter, the research sub-questions will be answered using the results of the data analysis. Here, we will make use of the methods described in section 3.6 of the Research Methods chapter. The most relevant statistical outputs will be presented followed by an interpretation of the key findings. This chapter will be structured in 4 sub-sections, one section per research sub-question.

4.1 What is tracking? How pervasive are they in the European Union countries? and What are the types of tracking in use?

This question was partially theoretically answered in Chapter 2. In that chapter, we learned about the definition of tracking, as well as, the type of Third Party companies that are commonly interested on web tracking. Now with the results of the measurement data and empirical analysis, we can determine a variety of insights related to tracking using cookies as a proxy. Hence, we will complement the last two parts of this research sub question with some descriptive statistics to determine tracking' pervasiveness and the types of tracking in use.

In figure 22, we observe the count of websites that have third party cookies present, the count of websites that have first and third party cookies present, the count of websites that have zero cookies, and the websites that only have first party cookies present. The count of websites that have first and third-party cookies presence is higher than websites that do not have cookies or that only have first party cookies. Also, the count of websites that only have third party cookies presence is more than the ones without cookies or with first party cookies.

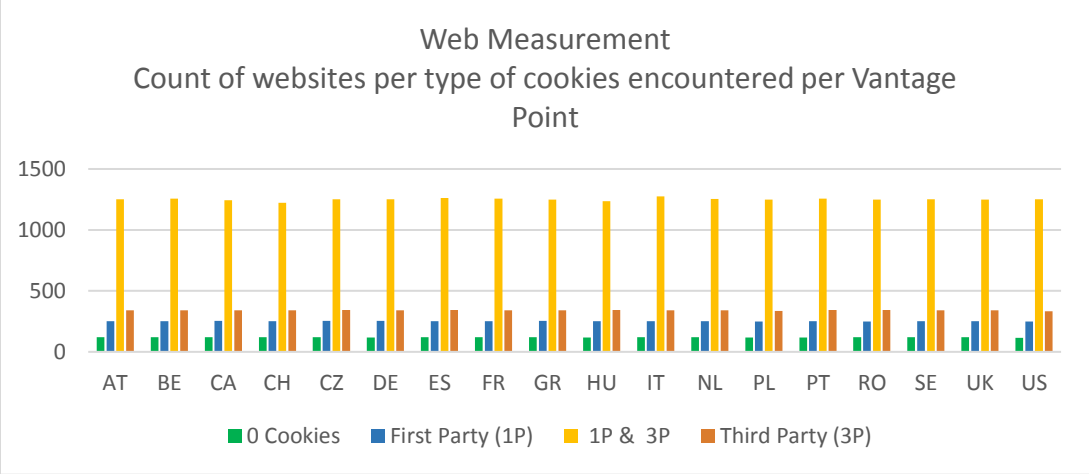


Figure 22: Web Measurement – Count of Websites per Type of cookies presence

To better understand figure 22, Table 13 offers the count of websites per vantage points, the location where the users were simulated to be, with the count of websites per type of cookie presence with a percentage of each type. This table indicates that third party cookies were present in 81% of the websites analyzed in all vantage point.

Table 13: Web Measurement - Websites per Type of cookies presence

VP	AT	BE	CA	CH	CZ	DE	ES	FR	GR	HU	IT	NL	PL	PT	RO	SE	UK	US
Web Measurement Websites	**1966**	**1970**	**1957**	**1935**	**1967**	**1963**	**1977**	**1969**	**1962**	**1948**	**1989**	**1966**	**1953**	**1971**	**1960**	**1962**	**1959**	**1951**
0 Cookies	120	120	119	119	119	118	119	119	119	117	120	120	118	118	119	119	119	115
First Party (1P)	252	252	253	251	253	253	252	252	253	251	252	252	250	252	250	251	251	250
1P & 3P	1252	1258	1245	1223	1252	1252	1263	1257	1248	1237	1275	1254	1250	1258	1248	1251	1248	1252
Third Party (3P)	342	340	340	342	343	340	343	341	342	343	342	340	335	343	343	341	341	334
% 0 Cookies	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%	6%
% 1P	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%	13%
% 3P & 1P	64%	64%	64%	63%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%	64%
% 3P	17%	17%	17%	18%	17%	17%	17%	17%	17%	18%	17%	17%	17%	17%	18%	17%	17%	17%
%Where 3P is involved	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%

Also, we found that the presence of Third Party Domains varies depending on the different categories of websites, which represent for us the companies' business models. The top three business models that have more third-party domains are news, similar to Englehardt & Narayanan (2016a), science, and art and entertainment. The business models that we found have less third-party domains presence are religion and illegal content. Figure 23 depicts the results.

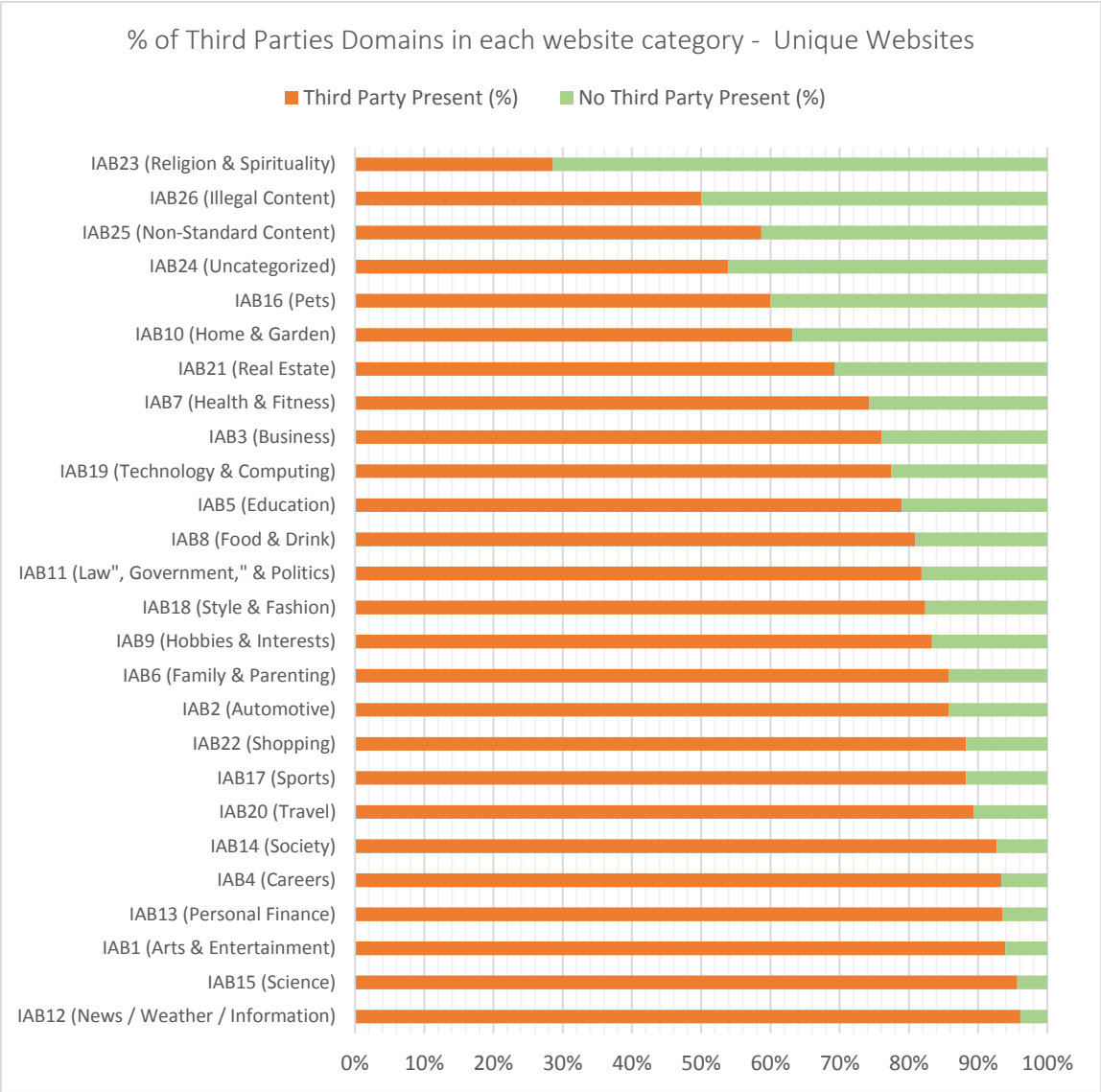


Figure 23:Websites having and not having third party domain per websites categories

To better understand figure 23, table 14 shows the count of websites that have third party domain present and the count of websites of the same category that did not have a third-party domain present. Basically, the results are the same as in figure 23, the only difference is that in the table we decided to show the count of websites per categories instead of percentages, so the reader can have an idea of how many websites were in each category.

.

Table 14: Count of websites having third party presence vs not having third party presence

	Category	Third Party Present (websites counts)	No Third Party (websites count)	Total Websites
0	IAB12 (News / Weather / Information)	368	15	383
1	IAB15 (Science)	22	1	23
2	IAB1 (Arts & Entertainment)	62	4	66
3	IAB13 (Personal Finance)	43	3	46
4	IAB4 (Careers)	14	1	15
5	IAB14 (Society)	38	3	41
6	IAB20 (Travel)	67	8	75
7	IAB17 (Sports)	45	6	51
8	IAB22 (Shopping)	45	6	51
9	IAB2 (Automotive)	12	2	14
10	IAB6 (Family & Parenting)	6	1	7
11	IAB9 (Hobbies & Interests)	45	9	54
12	IAB18 (Style & Fashion)	14	3	17
13	IAB11 (Law", Government," & Politics)	63	14	77
14	IAB8 (Food & Drink)	17	4	21
15	IAB5 (Education)	233	62	295
16	IAB19 (Technology & Computing)	320	93	413
17	IAB3 (Business)	38	12	50
18	IAB7 (Health & Fitness)	26	9	35
19	IAB21 (Real Estate)	9	4	13
20	IAB10 (Home & Garden)	12	7	19
21	IAB16 (Pets)	3	2	5
22	IAB24 (Uncategorized)	42	36	78
23	IAB25 (Non-Standard Content)	88	62	150
24	IAB26 (Illegal Content)	2	2	4
25	IAB23 (Religion & Spirituality)	2	5	7
	Total Websites	1636	374	2010

Finally, we looked at who these third-party domains were, so we could determine which trackers are commonly used. We found 3,453 Third Party Trackers in the 2010 websites with a long tail. It is important to clarify that these Third Party Domains might not be all classified as trackers for some tools such as ad blockers. However, we remind the reader that we used as a proxy Third Party Domains that last more than 30 days, so they are persistent in users' devices. Google including double click appears in 59.4% of the websites, and Rubicon Project , a California based company dedicated to automating the buying and selling of advertisement ('the Rubicon Project', 2018) appears in 44.7% of the websites. Adobe is in third place using the Third Party Domain "Demdex" which allows its audience management platform to work. Adobe Audience Management platform combines different data to create target segments, so companies can advertise effectively ('Data management platform, DMP | Adobe Audience Manager', 2018). We checked the rest of the top 20 third party domains and all of them are involved in advertisement. Figure 24 shows the top 20 third party domains we found.

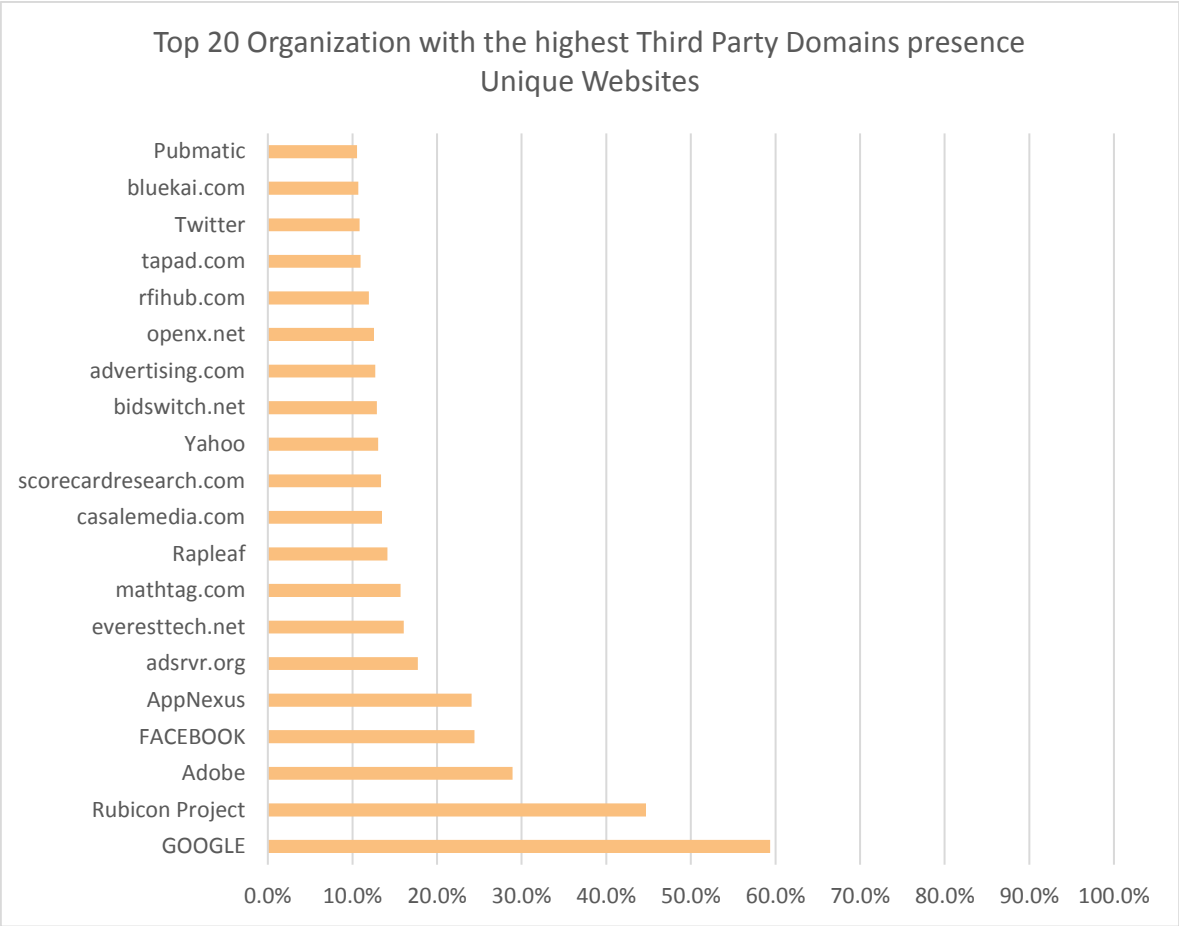


Figure 24: Top 20 Organizations with highest Third Party Domain Presence

Table 15 shows the percentages that were used in figure 24 as well as the count of how many times these Third Party Domains appeared in the 2010 websites.

Table 15: Top 20 Organizations count and presence

	Company	Count	% Presence on 2010 websites
1	GOOGLE[4]	1193	59.4%
2	Rubicon Project	898	44.7%
3	Adobe (Demdex)	581	28.9%
4	FACEBOOK	491	24.4%
5	AppNexus	484	24.1%
6	adsrvr.org	356	17.7%
7	everesttech.net	323	16.1%
8	mathtag.com	315	15.7%
9	Rapleaf	284	14.1%
10	casalemedia.com	271	13.5%

4 Google includes doubleclick.net

11	scorecardresearch.com	269	13.4%
12	Yahoo	262	13.0%
13	bidswitch.net	259	12.9%
14	advertising.com	255	12.7%
15	openx.net	252	12.5%
16	rfihub.com	240	11.9%
17	tapad.com	220	10.9%
18	Twitter	218	10.8%
19	bluekai.com	215	10.7%
20	Pubmatic	212	10.5%

4.1.1 Interpretation of the Key Findings

First at all, our results indicate that pervasiveness of third party cookies is relatively high (81%) independently of which country the users were simulated to be. These results are consistent with the literature available which has demonstrated that third party cookies are still one of the most used mechanisms to exert tracking on the web (Fruchter et al., 2015; Narayanan & Reisman, 2017a; Roesner, Kohno, & Wetherall, 2012; N. van Eijk et al., 2012). Besides, the literature indicates that tracking has been increasing over the years (Altaweel et al., 2015). For example, Krishnamurthy and Wills (2009a) demonstrated that the presence of top third parties domains exerting tracking on popular websites increased from 40% in 2005 to 70% in 2008. Also, Lener at al (2016) concluded in their longitudinal study from 1996 to 2016 that users who visit popular websites are exposed to more trackers with more complex behavior. Hence, observing this relatively high percentage (81%) is also aligned with the literature that confirms that in the recent years there is a high presence of tracking on the web. The importance of this result in terms of privacy protection is that cookies are still a preferred mechanism by websites to exert tracking, so at least this mechanism can be deleted by users from their web browsers, while with more transparent mechanisms, such as fingerprinting, it is more difficult for users to know how they are being tracked.

In addition, we observed that different companies' business models lead to different levels of the pervasiveness of tracking. Englehardt and Narayanan (2016a) proposed that websites that have sources of revenue streams different from advertisement are less likely to use third parties trackers to monetize users' visits to their websites . We agree with them since any business model needs to have a revenue stream to be successful. We observed that websites such as News and Art and Entertainment are at the top of the list using third party domains. These two categories are associated with business models that bring value to customers through free content, but they need some income to stay profitable. Historically, News has used advertisement since their beginning in 1800, and now with the use of the internet they face the challenge of not printed version circulation(Kirchhoff, 2011), so nowadays they are highly dependent on their online business models to stay profitable. Also, Art and Entertainment websites provide free content, and they also have to come up with ideas to monetize their users' visits since their principal asset might be only their audience. Hence, integrating third parties to deliver advertisement might be one of the options these businesses exploit to generate income. On the other hand, religion and illegal content and other categories on the bottom part of the list do not need advertisement as an income or are not dependent on it. More of the analysis of the differences of these business models will follow in sub-question 3, but from the descriptive statistics, we can conclude that there are variations in the pervasiveness of tracking according to companies'

business models and their revenue streams. The importance of this result is that there are certain business models that might have more incentive to exploit information asymmetry from users, as well, as they might have a conflict of interest leading to a Principal-Agent problem, and this can lead to market failures, while other businesses might have less incentives to use tracking.

Finally, we found a long tail of trackers with the most predominant third-party domains involved in advertising. The long tail we observed for trackers was also observed by Englehardt and Narayanan (2016a), and as they stated users will have the possibility to find these trackers more often. This result indicates some benefits for privacy protection. First at all, companies that are present in most of the websites are recognized, so they might have enough qualified personnel to handle users' data collected, and they might care about their reputation. Also, they might be interested in being perceived as companies which users can trust, so they can continue with their business models. In addition, users' transaction cost to check if these companies are complying with the law might be reduced, although still present. Moreover, the information asymmetry between these third party companies and regulators might be reduced since there are few of them which regulators could control for. On the other hand, the fact that few companies have a presence in most of the websites might increase their market power. Market power can lead to the deterrence of competitors and/ or abuse of their positions (Berg et al., 2009). Hence, if few companies are collecting users' online behavior, it might be important for regulators to control their incentives because they can take advantage of the information asymmetry among the users, websites, and they, and this can lead to opportunistic behavior. One might argue that there is no difference if few or many organizations are found on a daily basis by users. Users might suffer from externalities any way or price discrimination from one or multiple companies. However, we think that observing this long tail might help regulators and users to achieve better protection for the reasons previously mentioned.

After analyzing these results, we go back to the second part of our research sub-question *how pervasive is tracking in the European Union countries?, and what are the types of tracking in use?* We can answer that tracking cookies pervasiveness is 81% on the websites analyzed in the selected countries. This is in line with the literature that expresses that cookies are still one of the most used tracking mechanisms on the web. Also, the pervasiveness of the Third-Party Domains varies according to websites categories which for us is related to the business models' incentives to use tracking to make a profit. The most predominant type of tracking in use are third parties involved in the online behavioral advertising industry, and they have a long tail. The observation that the trackers in use have a long tail matters in term of privacy protection since large companies might be easy to regulate as well as might reduce some of the complexities related to information asymmetry and transaction cost for users and regulators. To finalize the answer to our question, we need to think in the generalizability of these results to the European Union. Generalizability is related to the fact if these findings might hold for others European Union countries. We consider, that due to the sampling selected these results might hold only for popular websites in the European Union. However, we will come back to this in the reflection and limitations of our work.

4.2 Which laws websites follow? Are there differences in tracking and cookies notices related to the law they follow?

As it has been stated in our literature review, websites could follow the law and norms where the companies are located or the law where users are located. To the best of our knowledge there is no literature testing which law and norms websites follow using empirical data, and for us was key to understand this to determine if there were differences in tracking across member states depending on which laws and norms websites follow. Hence, we embarked ourselves answering this research questions using the statistical instruments and methods described in section 3.6 and the metrics described in section 3.4 of the Research Methods chapter.

4.2.1 Shape of the Distributions of the Metrics

Metric 1: Unique Counted Cookies Names. We looked at 642,362 unique counted cookie names, and the dataset consisted of 35,325 observations that have 5 variables (See Research Methods section 3.5 for the description of the dataset). Figure 25 shows the head of the dataset.

35,325 obs of 5 variables

visit_id	name	VP	site_url	TLD
<int>	<int>	<chr>	<chr>	<chr>
2	2	AT	univie.ac.at	at
2	2	BE	univie.ac.at	at
2	2	CA	univie.ac.at	at
2	2	CH	univie.ac.at	at
2	2	CZ	univie.ac.at	at
2	2	DE	univie.ac.at	at

Figure 25: Print screen of dataset cookies names

The shape of the distribution of the dependent variable 'unique counted cookies names' is depicted in Figure 26.

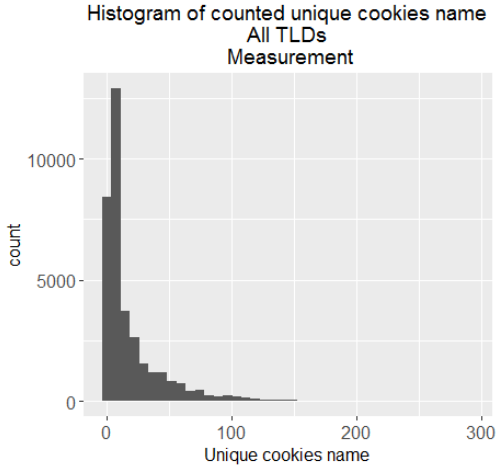


Figure 26: Histogram of counted unique cookies name

N	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD	SK
35325	0	4	8	18.18	22	290	24.91	2.74

We observe that this dependent variable is non-normal distributed, positively skew, and its shape is negative binomial. Also, we found that the mean is lower than the standard deviation (S-D/Mean> 1) suggesting over dispersion. Over dispersion means that the counts of the unique cookies names vary more widely than the mean. Also, we observe that the maximum number of unique cookies names in a website is 290. Besides, the median of unique cookie names per website is 8.

Metric 2: Unique Counted Third Party Domains. Our second dependent variable is the count of unique third party domains. We looked at 206,787, and the dataset consisted of 35,325 observations that have 5 variables (See Research Methods section 3.5 for the description of the dataset). Figure 27 shows the head of the dataset.

35,325 obs. Of 5 variables

visit_id	host	VP	site_url	TLD
<int>	<int>	<chr>	<chr>	<chr>
2136	90	US	thesun.co.uk	uk
2136	38	HU	thesun.co.uk	uk
2136	36	IT	thesun.co.uk	uk
2136	13	SE	thesun.co.uk	uk
931	89	US	sport.es	es
931	78	IT	sport.es	es

Figure 27: Print screen dataset Third Party Domains

The shape of the distribution of the dependent variable unique counted Third Party Domains is depicted in Figure 28.

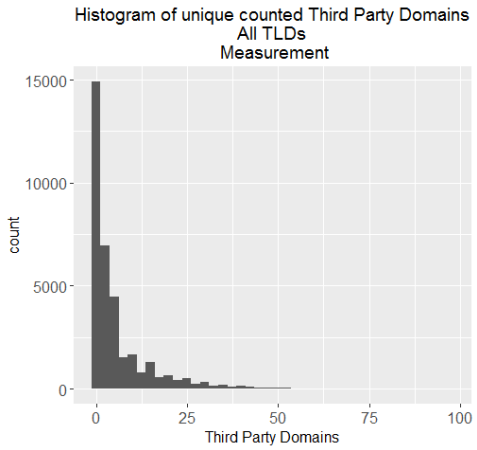


Figure 28: Histogram of Unique Counted Third Party Domains

N	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD	Sk
35325	0	1	2	5.8	7	97	9.24	2.86

We observe that this dependent variable is non-normal distributed, positively skew, and its shape is negative binomial. Also, from the descriptive statistics was found that the mean is lower than the standard deviation (SD/Mean> 1). This again suggests over dispersion. In addition, we observe that the maximum number of unique Third Party Domains present in a website is 97. Besides, the median of Unique Third Party Domains per website is 2.

Metric 3: Unique Counted Third Party Java Script Calls. Our third dependent variable is the count of unique Java Script calls to third parties. We look at 217,183 Unique Java Script Calls. The dataset consisted of 35,325 observations that have 5 variables (See Research Methods section 3.5 for the description of the dataset). Figure 29 depicts the head of the dataset.

35325 obs. Of 4 variables

visit_id	jhost	VP	site_url	TLD
<int>	<int>	<chr>	<chr>	<chr>
2136	43	US	thesun.co.uk	uk
2136	29	IT	thesun.co.uk	uk
2136	28	HU	thesun.co.uk	uk
2136	17	SE	thesun.co.uk	uk
2138	37	US	thetimes.co.uk	uk
2138	36	NL	thetimes.co.uk	uk

Figure 29: Print Screen Dataset Java Script Calls

The shape of the distribution of the dependent variable unique counted Java Script Calls is depicted in Figure 30.

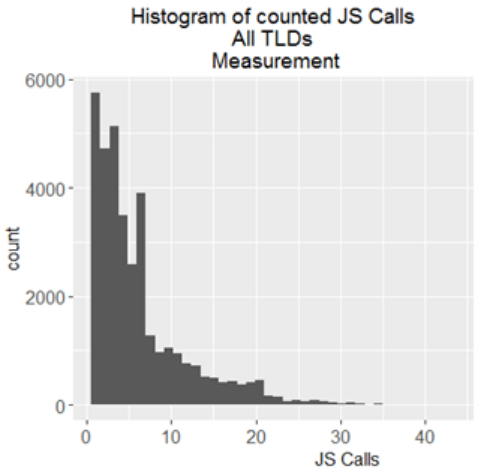


Figure 30: Java Script Calls Distribution

N	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
35325	1	2	4	6.148	8	43	5.71

We observe that this dependent variable is non-normal distributed, positively skew, and its shape is negative binomial. Also, we found that the mean is lower than the standard deviation (S-D/Mean> 1) suggesting over dispersion. Also, we observe that the maximum number of unique Java Script calls in a website is 43. Besides, the median of Unique Java Script Calls per website is 4.

Metric 4: Banners. Our final dependent variable is banners. The dataset consisted of 35,325 observations that have 11 variables (See Research Methods section 3.5 for the description of the dataset). Figure 31 depicts the head of the dataset.

35325 Obs. Of 11 variables

VP	h	m	site_url	text	visit_id	w	x	y	banner	TLD
<chr>	<dbl>	<dbl>	<chr>	<chr>	<int>	<dbl>	<dbl>	<dbl>	<int>	<chr>
US	0	0	http://univie.ac.at	0	2	0	0	0	0	at
NL	0	0	http://univie.ac.at	0	2	0	0	0	0	at
DE	0	0	http://univie.ac.at	0	2	0	0	0	0	at
BE	0	0	http://univie.ac.at	0	2	0	0	0	0	at
CA	0	0	http://univie.ac.at	0	2	0	0	0	0	at
CH	0	0	http://univie.ac.at	0	2	0	0	0	0	at

Figure 31: Print Screen Dataset Banners

Figure 32 shows the shape of the dependent variable banners.

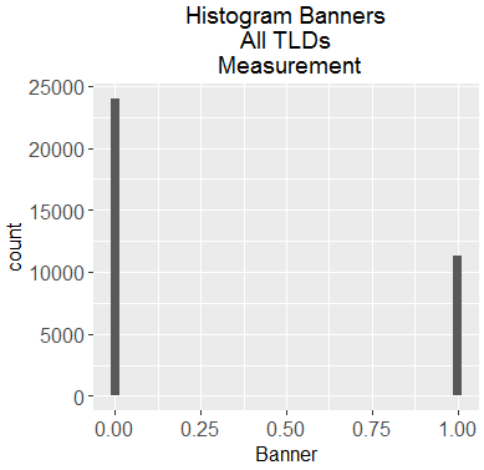


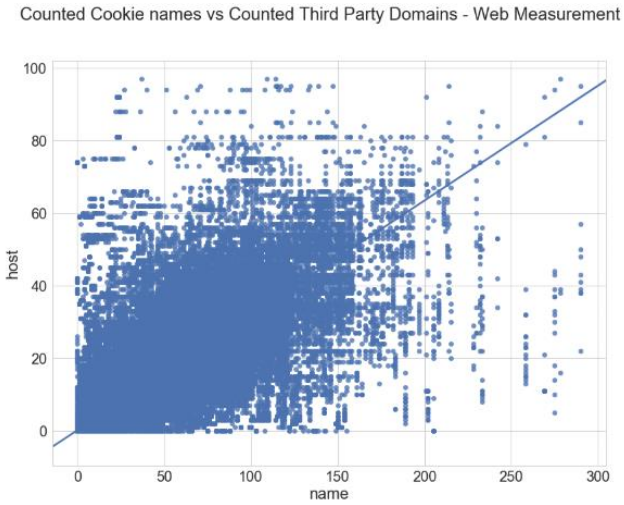
Figure 32: Shape of the distribution of variable banner

We observe that more websites do not have banner presence than websites that do have banners.

4.2.2 Assessing Metrics Association

Now that we know how the dependent variables look like, we will proceed to understand their degree of association. Since banners is a different metric, we will use this variable independently, and we will check only the association of the variables related to tracking. We will proceed with a Spearman Rank correlation analysis which suits for non-parametric variables.

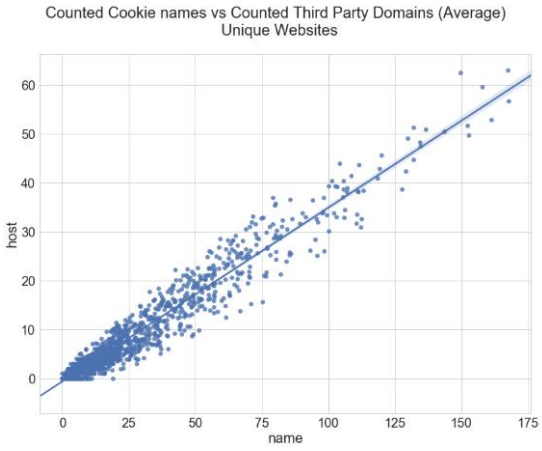
Figure 33 depicts the correlation between the dependent variable "counted unique cookies names" versus "counted unique third-party domains". We used the measurement dataset (35,000 websites), and we found a correlation of 0.88 with a significant P-value <0.05.



SpearmanrResult(correlation=0.88792705220847501, pvalue=0.0)

Figure 33: Counted Unique Cookies Names versus Counted Unique TPD – Measurement

Since the measurement dataset might have some errors due to the repetition of the measurement across all the users' location simulated, we decided to use the average of the unique counted cookies names as well as the average of unique counted third party domains to test the relationship. Figure 34 presents the results, and we can observe a clearer correlation with a Spearman rank value of 0.92 significant at a P-Value<0.05.

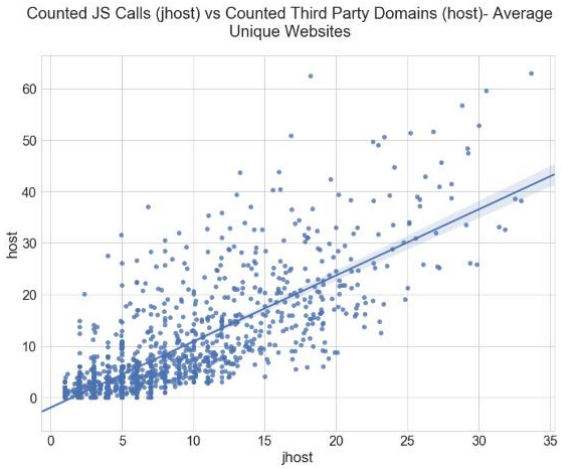


SpearmanrResult(correlation=0.91435850137104757, pvalue=0.0)

Figure 34: Counted Unique Cookie Names versus Counted Unique TPD - Average

We found a positive relationship between the number of cookies counted and the third party domains present in a website. Meaning that as the number of third party domains present in a website increases, the number of unique counted cookies names also increases. Hence as more third party companies are present in a website, more third party cookies will be found.

We continue with the Spearman rank correlation between the Java Script and Third Party Domains[5]. We use directly the average of third-party domains and unique java script calls to diminish the error of the measurement. Figure 35 depicts the results.



SpearmanrResult(correlation=0.81069119956701075, pvalue=0.0)

Figure 35:Counted Unique TPD versus Unique Counted JS Calls- Average

We observed a high a positive and significant Spearman rank correlation of 0.81 between unique counted Java Script calls and unique counted third-party domains. Meaning that as the number of java script calls increases also the unique third-party domains present in the website increases. Java script calls are one of the mechanism that websites use to allow third party company to read or drop third party cookies and to execute scripts. Hence, this association seems logical since as more calls a website does to third parties, more third parties will be present on the website.

Since we found a high and significant correlation of the three metrics, we will keep only Third Party Domains(TPD) for the rest of our analysis. The reason to use only this metric is twofold. First, since there is a high association of the metrics the results of the statistical analysis will yield to the same results. Second, Third Party Domain represents how many third-party companies learn about users' online behavior which is more relevant in terms of tracking.

4.2.3 Testing Which Law Websites Follow

After understanding the metrics, as described in the Research Methods chapter (section 3.6), we tested 2 models, Model 1 consisted of a null model, and Model 2 had the Vantage Point (VP) as dependent dummies variables, which represent users' location. Comparing the two models, we could capture if the location of the user (VP) does have an effect or if the model without the VP was preferred, so we could determine if websites located in certain countries use geolocation to adapt to users' location (VP) or not. We remind the reader that the dependent variable was grouped by TLDs which we used as a proxy of websites' location, so if the model with the VP is not preferred this means that websites stick to their localization and follow local laws. Due to the shape of the distribution of the dependent variable Third Party Domains, we run a negative binomial generalized linear models (GLM).

[5] We show here only Third Party Domains, since the results with cookies names is almost the same since cookies names and third party domains are highly correlated. See Appendix A to see Unique Counted Java Script calls versus Unique Counted Cookies Names

Model 1 = Third Party Domains ~1
Model 2= Third Party Domains ~1 + VP

Table 16 presents the results of the comparison of the 2 models, and the number of observation per TLD, which represents the location of the website.

Table 16: AIC comparison for VP vs null model- Third Party Domain as the dependent variable

	TLD	# Obs	Model 1 (AIC)	Model 2 (AIC)
1	AT	1595	7,746.091	7,778.816
2	BE	1631	8,766.113	8,792.729
3	CA	1512	8,683.857	8,706.860
4	CH	1643	9,047.080	9,075.854
5	CZ	1626	9,286.582	9,315.657
6	DE	1549	9,837.896	9,866.584
7	ES	1326	7,449.852	7,474.989
8	FR	1412	9,580.759	9,593.533
9	GR	1498	6,149.139	6,180.314
10	HU	1481	6,358.013	6,388.753
11	IT	1394	7,618.127	7,634.544
12	NL	1580	6,630.805	6,657.080
13	PL	1066	5,526.858	5,554.529
14	PT	1494	7,477.663	7,506.574
15	RO	1516	7,043.023	7,071.848
16	SE	1594	8,555.448	8,582.089
17	UK	1498	9,960.830	9,965.588
18	US	769	3,497.971	3,526.927
19	COM	3200	20,071.140	**19,979.610**
20	ORG	3388	15,213.190	15,227.020
21	JP	984	6,300.378	6,328.209
22	AU	1569	9,919.502	9,934.569

We proceed to compare the AIC values of the two Models, and we notice that only the observations with TLD '.COM' has a low AIC for Model 2, the model with vantage points. Meaning that websites with TLD '.COM' use geolocation to adapt their websites to users' location. We remind the reader than in the methodology we express that websites with TLD '.COM' do not have a specific location

for us, but they were used as a comparison. On the other hand, when the websites have a defined target market or location, such as websites which use specific TLDs like .NL, .DE, websites do not use geolocation since the low AIC was always for Model 1, and the same was observed for websites with TLD '.ORG'. Hence, they stick to the rules of their location.

Since TLD '.COM' was the only with low AIC (19,979.610) when using the location of users as dummies, we further examine the results of its GLM negative binomial regression. Table 17 shows the statistical output of Model 2. The coefficients of a GLM negative binomial regression are usually interpreted as Incident Rate Ratios or Probabilities, and they are multiplicative. We will interpret the coefficients of Model 2 as Incident Rate Ratios (IRR) or the relative risk of an event occurring, in this case, a third party being counted, with a 95% confidence interval. Once we convert the coefficients to IRR[6], we can tell how likely a Third Party Domain can be counted when a user was simulated to visit these websites from the different locations.

Table 17: Negative Binomial Regression Model TLD .COM with Incident Rate Ratios

Variable (VP)	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
VP_BE	-0.026(0.139)	0.9746571 (0.74-1.28)
VP_CA	0.495 (0.139)***	1.6408012 (1.24-2.15)
VP_CH	0.173 (0.141)	1.1890209 (0.90-1.56)
VP_CZ	-0.014(0.141)	0.9859086 (0.74-1.30)
VP_DE	0.042(0.139)	1.0432122 (0.79- 1.37)
VP_ES	0.022(0.140)	1.0219612 (0.77-1.34)
VP_FR	0.038(0.139)	1.0383505 (0.79-1.36)
VP_GR	-0.057(0.140)	0.9443579 (0.71-1.24)
VP_HU	-0.081(0.141)	0.9220945 (0.69-1.21)
VP_IT	0.197(0.138)	1.2173997 (0.92-1.59)
VP_NL	0.028(0.140)	1.0282342 (0.78-1.35)
VP_PL	-0.001 (0.139)	0.9986903 (0.75-1.31)
VP_PT	-0.005 (0.140)	0.9945339 (0.75-1.30)
VP_RO	0.042 (0.141)	1.0426297 (0.79-1.37)
VP_SE	-0.114(0.140)	0.8926841 (0.67-1.17)
VP_UK	0.055(0.139)	1.0569515 (0.80-1.38)
VP_US	0.890(0.137)***	2.4361304 (1.86-3.18)
(Intercept)	2.010(0.099)***	7.4597701 (6.17- 9.11)

*p<0.1; **p<0.05; ***p<0.01. Null deviance/residual: 3759.6/3631.6 -McFadden Pseudo R^2:0.006

6 To convert the coefficients only to Incident Rate Ratios = exp(coef)

From the results, we observe that there is high variability of incident rate ratios in the different users' locations, but only Canada, The United States are statistically significant at $p<0.01$. These results indicate that those users who visit websites with TLD '.COM' from Canada has a 64% higher likelihood of finding Third Party Domains than those in other location, all other variables being equal. In addition, users from The United States have a 2.43 times higher likelihood of finding a Third Party Domains when visiting a TLD '.COM' than those in other locations, all other variables being equal.

We followed the same procedure for our dependent variable Banners. To run two models, and we used a Logistic regression since the dependent variable as it was depicted in figure 32 is binary. We tested again a null model (Model 1) and a model using Vantage Points as dummies dependent variable (Model 2) for each TLD, which represents websites' location.

Model 1 = Banner ~1
Model 2= Banner~VP

Table 18 presents the results of the comparison of the 2 models, and the number of observation per TLDs.

Table 18: : AIC comparison for VP- Banners as dependent variable

TLD	# Obs	Model 1 (AIC)	Model 2 (AIC)
AT	1595	2,178.980	2,212.184
BE	1631	2,134.415	2,165.497
CA	1512	930.318	962.631
CH	1643	1,554.81	1587.95
CZ	1626	2,221.15	2,253.42
DE	1549	1,820.30	1,852.63
ES	1326	1,796.547	1829.933
FR	1412	1,441.34	1471.49
GR	1498	1,810.645	1,843.993
HU	1481	1,968.19	2,000.41
IT	1394	1,930.77	1,961.61
NL	1580	2,188.688	2,221.323
PL	1066	1,440.509	1,473.816
PT	1494	1,661.060	1,692.628
RO	1516	1,654.57	1,687.30
SE	1594	2,132.47	2,165.76

UK	1498	2,015.71	2,045.53
US	769	392.97	426.96
COM	3200	4,050.759	**4,048.101**
ORG	3388	2,648.750	2,681.479
JP	984	290.959	323.363
AU	1569	777.960	810.884

Again, we observed that the TLD '.COM' reveals that the preferred model with the lower AIC value is Model 2 (4,048.101). We will take Model 2 as the best, but Model 1 is not that different since Δ AIC value does not differ substantially between the two models. Guthery, Burnham and Anderson (2003) expressed that Models to be considered different should have an AIC difference of at least 10, but also they explained that AIC values can be compared with a car race and the lowest value is the one that wins since it was the first to cross the line in the race. On the other hand, we observe that for the other TLDs users' location is not relevant since the preferred model is Model 1. Meaning that the websites that have a defined target market follow their local laws.

We further examine the results of the TLD '.COM' for the dependent variable banner. The results of a logistic regression are also interpreted as Incident Rate Ratios, so once we convert the coefficients to IRR, we can tell how likely a banner is to be present when a user was simulated to visit these websites from the different locations.

Table 19 shows the results with a confidence interval of 95%

Table 19: Logistic Regression TLD .COM with Incident Rate Ratios - Banners

Variable	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
VP_BE	-0.010(0.224)	0.9899160 (0.63-1.53)
VP_CA	-1.016(0.261)***	0.3619048(0.21- 0.59)
VP_CH	-0.465(0.239)*	0.6283465(0.39- 1.00)
VP_CZ	-0.086(0.228)	0.9172414(0.58- 1.43)
VP_DE	-0.043(0.225)	0.9579832(0.61-1.48)
VP_ES	0.040(0.224)	1.0408696 (0.67- 1.61)
VP_FR	0.063(0.222)	1.0646552 (0.68- 1.64)
VP_GR	-0.119(0.226)	0.8877049(0.56- 1.38)
VP_HU	0.034(0.225)	1.0348214 (0.66-1.61)
VP_IT	0.037(0.222)	1.0378151 (0.67-1.60)
VP_NL	-0.068(0.225)	0.9341667 (0.60- 1.45)
VP_PL	0.095(0.222)	1.1000000 (0.71-1.70)
VP_PT	0.056(0.223)	1.0573913 (0.68-1.63)

VP_RO	-0.131(0.229)	0.8769231 (0.55-1.37)
VP_SE	0.080 (0.223)	1.0833333 (0.70- 1.67)
VP_UK	0.006 (0.223)	1.0058824 (0.64- 1.55)
VP_US	-0.117(0.224)	0.8896825 (0.57-1.38)
(Intercept)	-0.642(0.159)***	0.5263158 (0.38-0.71)

*p<0.1; **p<0.05; ***p<0.01. Null deviance/residual: 4048.8/4012.1- McFadden Pseudo R^2:0.009

These results indicate that the likelihood of finding a banner when visiting a website with TLD .COM decreases by 64% when a user is located in Canada (CA). Also, the likelihood of finding a banner when visiting a website with TLD .COM decreases by 38% when a user is located in Switzerland (CH), other variables being equal.

Since we determined that the location of the users did not play a role neither for third party domains nor for banners, except for websites .COM, which use geolocation to adapt to users' location local laws, this might suggest that the rest of websites, which are based on a specific location, do not adapt to users' location, and they are following local laws of their target market.

4.2.4 Testing differences in tracking and cookies notices among member states

Based on our previous results which suggest that websites follow the rules and norms of the target market they serve, we will test if there are differences on tracking related to local laws websites are following. In other words, we will test if there are differences among websites that are located and following the rules of the different member states. We remind the reader that we use the TLDs as a proxy of the location of the websites, and the local laws they should follow, so we will model tracking as a function of their location (TLD) to see if there are differences in tracking and cookies notices depending on where websites are based.

H1: There are differences in tracking and notices related to the local law websites follow	Rationale: Each country laws and market characteristics can lead to differences in tracking among member states

We will run two regressions models to test the hypothesis, one using Third Party Domains as the dependent variable and another using banner. We will use the average of Third Party Domains and banner as dependent variables, leaving out .com and .org (More details in Research Methods section 3.6). First, we will depict some descriptive statistics and then we will show the results of both regression models. Hence, as the first step, Figure 36 presents the shape of the distribution of the average of third party domains.

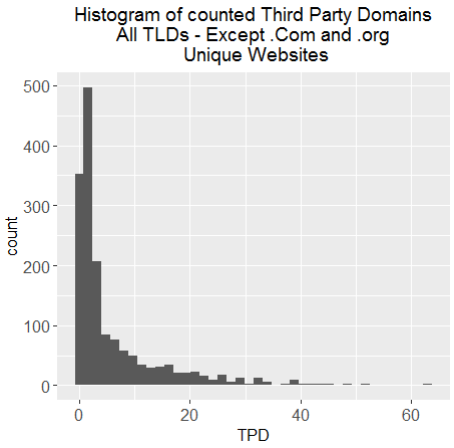


Figure 36: Histogram Counted Third Party Domains - .COM and .ORG TLDs Excluded

N	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD	SK
1634	0	1	2.06	6.11	7.54	63	9.10	2.48

We observe that the distribution of the variable is still non-normal, positively skew, and its shape is negative binomial. Also, we found that the mean is lower than the standard deviation (SD/Mean> 1) suggesting over dispersion. Over dispersion means that the counts of the unique third-party domains vary more widely than the mean. Also, we observe that the maximum number of third party domains in a website is 63. Besides, the median of unique third party domains per website is 2.

As the second step, we examine the dependent variable Third Party Domains versus TLDs, as a proxy of the local laws websites are following. Figure 37 shows the output.

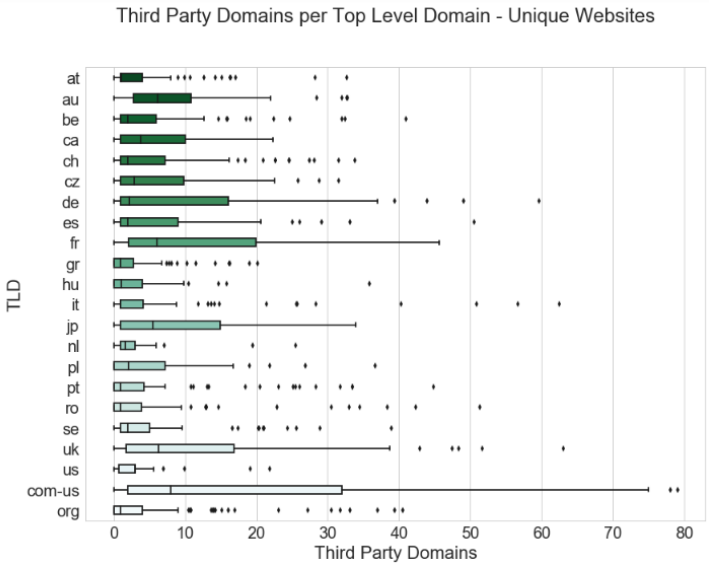


Figure 37: Third Party Domains versus TLDs as proxy of where Websites are based[7]

7 We use websites .com and .org as reference in the descriptive statistics, but they are not included in the regression models presented in this chapter. For the interested reader in Appendix B the results of a model including websites .COM visited from the US is presented.

In figure 37, we observe that there is a high variability caused by the local law that websites follow. Websites following US local laws with domain .COM has the high median, followed by UK, Australia, and France. On the other hand, websites following the local laws from Greece, Portugal , Italy and Romania has the lowest median. Table 20 shows the descriptive statistics related to the graph.

Table 20: TPD per TLD descriptive statistics

TLD	count	mean	std	min	25%	50%	75%	max
at	90.0	3.749782	5.785864	0.0	1.000000	1.083333	4.083333	32.666667
au	88.0	8.302245	7.647969	0.0	2.722222	6.166667	10.791667	32.777778
be	92.0	5.268948	7.746252	0.0	1.000000	2.000000	6.000000	40.941176
ca	85.0	6.208082	6.525450	0.0	1.000000	3.833333	10.000000	22.352941
ch	93.0	5.886378	8.110079	0.0	0.944444	2.000000	7.166667	33.833333
com-us	185.0	18.172973	20.700303	0.0	2.000000	8.000000	32.000000	79.000000
cz	92.0	6.022386	7.064562	0.0	1.000000	2.861111	9.833333	31.555556
de	88.0	9.421154	12.659703	0.0	1.000000	2.250000	16.050000	59.625000
es	75.0	6.290709	9.146509	0.0	1.000000	2.000000	9.027778	50.545455
fr	85.0	11.558065	12.074159	0.0	2.111111	6.111111	19.944444	45.625000
gr	85.0	2.742907	4.535098	0.0	0.000000	1.000000	2.833333	20.166667
hu	84.0	2.876649	4.904640	0.0	0.000000	1.058824	4.000000	35.823529
it	79.0	6.411434	12.446142	0.0	1.000000	1.000000	4.166667	62.500000
jp	55.0	8.971026	9.260377	0.0	1.000000	5.500000	14.972222	33.882353
nl	88.0	2.523479	3.572172	0.0	1.000000	1.666667	3.000000	25.500000
org	191.0	3.642743	7.092302	0.0	0.000000	1.000000	4.000000	40.461538
pl	61.0	4.874111	6.993673	0.0	0.000000	2.076923	7.166667	36.611111
pt	85.0	5.415448	9.620408	0.0	0.000000	1.000000	4.222222	44.800000
ro	87.0	5.060892	10.175332	0.0	0.000000	1.000000	3.916667	51.333333
se	89.0	4.988296	7.295771	0.0	1.000000	2.000000	5.000000	38.944444
uk	90.0	11.646613	13.960472	0.0	1.736111	6.307692	16.869048	63.000000
us	43.0	3.110275	4.406023	0.0	0.681373	3.000000	3.027778	21.823529

In addition, we will use the banner metric to analyze if there were differences in notification related to the local laws websites are following. Hence, Figure 38 depicts the shape of the distribution of the average of banners per all vantage points.

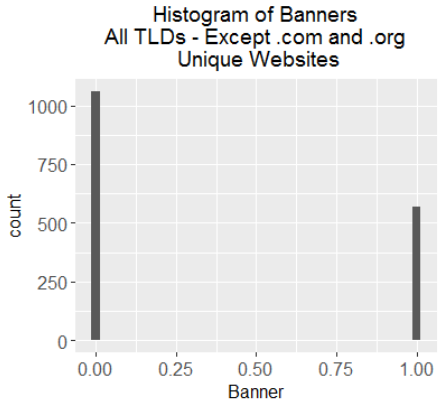


Figure 38:Unique Websites - Average Banners histogram

We observe that still more websites do not have banner presence than websites that do have banners.
Also, we inspected the dependent variable Banners versus the TLDs, as a proxy of local laws websites are following. Figure 39 depicts the output.

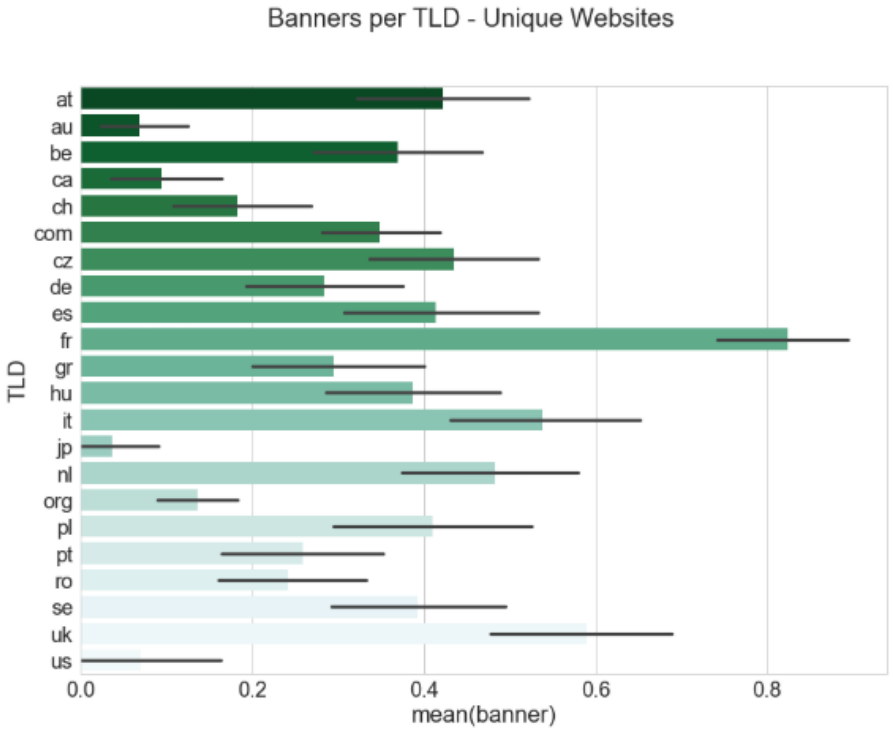


Figure 39: Banners per TLD as a proxy of where Websites are based [8]

We observe that there is a high variability in the presence of banners depending on which local law websites are following. We observe that websites that are following France local law has the higher banner presence followed by UK. On the other hand, websites that follow Japan, Australia, and The United States local laws has less banner presence.

After inspection the dependent variables and dependent variables versus the independent variable, we run two regression models. Due to the shape of the distribution of Third Party Domains for Model 1 we run a GLM negative binomial regression, and for Model 2 we run a logistic regression given that the variable banner is binary.

1) Model 1 = Third Party Domains ~ TLD
2) Model 2 = Banners ~ TLD

The next table presents the output of the results of both models:

8 We use websites .com and .org as reference in the descriptive statistics, but they are not included in the regression models.

Results Regressions TLDs as a proxy of local laws
Using Third Party Domains and Banners as dependent variables

	Third Party Domains		**Banner**	
	Negative Binomial (1)		Logistic Regression (2)	
Variable	Coefficient (Std Err)	Incident Rate Ratio(95% CI)	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
TLDau	0.795***(0.200)	2.21(1.49-3.28)	-2.301***(0.474)	0.10(0.03-0.23)
TLDbe	0.340*(0.200)	1.40(0.94-2.07)	-0.220(0.304)	0.80(0.44-1.45)
TLDca	0.504**(0.203)	1.65(1.11-2.46)	-1.951***(0.428)	0.14(0.05-0.31)
TLDch	0.451**(0.199)	1.56(1.06-2.31)	-1.184***(0343)	0.30(0.15-0.59)
TLDcz	0.474**(0.199)	1.60(1.08-2.37)	0.051(0.300)	1.05(0.58-1.89)
TLDde	0.921***(0.200)	2.51(1.69-3.71)	-0611*(0.318)	0.54(0.28-1.00)
TLDes	0.517**(0.209)	1.67(1.11-2.53)	-0.037(0.317)	0.96(0.51-1.79)
TLDfr	1.126***(0.201)	3.08(2.07-4.57)	1.854***(0.356)	6.38(3.24-13.16)
TLDgr	-0.313(0.209)	0.73(0.48-1.10)	-0.562*(0.320)	0.57(0.30-1.06)
TLDhu	-0.265(0.209)	0.76(0.50-1.15)	-0.147(0.309)	0.86(0.46-1.58)
TLDit	0.536***(0.206)	1.70(1.14-2.56)	0.466(0.311)	1.59(0.86-2.94)
TLDjp	0.872***(0.227)	2.39(1.54-3.76)	-2.963***(0.751)	0.051(0.008-0.18)
TLDnl	-0.396*(0.208)	0.68(0.44-1.01)	0.245(0.302)	1.27(0.70-2.31)
TLDpl	0.262(0.223)	1.29(0.84-2.02)	-0.051(0.337)	0.95(0.48-1.83)
TLDpt	0.368*(0.203)	1.44(0.96-2.15)	-0.738**(0.327)	0.47(0.24-0.90)
TLDro	0.300(0.203)	1.34(0.90-2.01)	-0.831**(0.329)	0.43(0.22-0.82)
TLDse	0.285(0.202)	1.33(0.89-1.97)	-0.120(0.304)	0.88(0.48-1.61)
TLDuk	1.133***(0.198)	3.10(2.10-4.58)	0.673***(0.302)	1.96(1.08-3.56)
TLDus	-0.187(0.255)	0.82(0.50-1.38)	-2.277***(0.636)	0.10(0.02-0.3100
Intercept	1.322***(0.206)	3.74(2.85-5.01)	-0.314(0.213)	0.73(0.47-1.10)
Observations	1634		1634	
Log Likelihood	-4,521.76		-921.391	
Akaike Inf. Crit.	9,083.52		1,882.78	

Note: *p<0.1; **p<0.05; ***p<0.01 / Websites .COM visited from US has a significant coefficient of 1.578(0.175)*** using the dependent variable tracking. This might be a relevant to consider since websites that use TLD .US are not that popular in The United States, while .COM are. Details of the Model where websites .COM visited from US and .ORG were included in Appendix B.

We found that there is a varying degree of tracking and notices related to the local law websites are following since we observe that the coefficients vary and most of them are significant. We observe that for Model 1 where we used Third Party Domains as the dependent variable, that websites that follow UK and France local laws have the highest coefficient, thus the highest Incident Rate Ratio, that is the likelihood of a third party domain being counted, of 3.10. and 3.08 respectively. On the other hand, the only negative and significant coefficient is for websites following The Netherlands local law. Meaning that the likelihood of counting a third party domain decrease by 32% when websites follow The Netherlands local rules.

In addition, we noticed that for banners there is also a high variability depending on the law websites apply. We observe that the coefficients vary and most of them are significant. We observe that websites that follow France and UK laws increase the likelihood of finding a banner by 538% and

96% respectively. On the other hand, websites that follow Japan and Australia local laws decrease the likelihood of finding a banner by 94.9% and 90% respectively.

Also, we observe that websites that follow UK and France local laws apart from having the incident rate ratio of Third Party Domains, they have the highest incident rate ratio of having a banner. Meaning that websites that follow UK and France local laws have the higher tracking presence and the higher use of banners.

With these results, we can conclude that there are differences in tracking related to the local laws the websites are following. Some of the local laws yield to high incident rate ratio, third party being counted, while other local laws lead to a decrease of third party domains being counted. Also, the same pattern is observed for banners.

4.2.5 Interpreting the Key Findings

Our results suggest that even larger websites follow local laws of the target market they decide to serve. When the target market of the websites is not clear, as in the case of TLDs ' .COM', websites adapt to users' location using geolocation. On the other hand, when websites have specific target market geolocation does not play a role, and websites stick to local laws of the market they serve.

To understand these results, we need to remind the reader that we are looking at the most popular websites. Hence, the TLD '.COM' might represent larger firms that have the capacity to interpret different national laws. Also, these larger firms can be located anywhere, so when these companies want to target specific markets, they need to implement international strategies to adapt their websites to the specific target audience. In fact, companies that want to capture a global market are advised to consider local laws, currencies, culture in order to succeed (Kelly, 2015), so this is not an exception for their websites. Hence, this might explain why they use geolocation to follow users' location rules. On the other hand, our results imply that websites that have a specific target market do not use geolocation to adapt to users' location. Websites might not mind where users are located because they are targeting a specific audience, so they are respecting the norms, culture, currencies, and adapting their services to the specific needs of the audience they chose to serve.

One important point to mention is that different than TLD '.COM', websites with TLD '.ORG' do not adapt to users' location. One main difference is that 'ORG' domains are used for non-commercial purposes. Hence, they might not have the capacity to adapt their websites to each target audience or they might not have commercial incentives to do it.

Also, these results indicate that there are differences on tracking and notices depending on the local laws and norms websites follow. We hypothesize that these differences might be associated to the local transpositions of the E-Privacy Directive, market forces characteristics, or general rules and norms of the target market websites serve.

Some clues to associate these differences with the local transposition of the E-Privacy Directive were observed. For example, websites that follow the Netherlands law significantly decrease the Third Party Domains by 33%, and The Netherlands is known to be one of the countries that applied the E-Privacy Directive in a more strict manner since the beginning, and opt in mechanisms were required to exert tracking (Leenes & Kosta, 2015b). On the other hand, countries such as UK and France have 3.10 and 3.08 times higher incident ratio of tracking respectively, and they also implemented consent

requirement (DLA Piper, 2014), however, UK's transposition was a more business-friendly approach, and it allowed a 'flexible consent', and even accepting implied consent as a form of consent in some cases (Oldhoff, 2013). Also, France, Belgium, and Italy required consent, but they also accept implied consent to allow websites to track users (Cofone, 2017). Hence, the sublet changes in the transpositions and guidance each country provided might be reflected in our results showing different levels of tracking.

Moreover, other signs that these differences might be related to the local transpositions of the provisions of the E-Privacy Directive were observed in terms of the cookies notices. We found that websites based in UK and France increase the likelihood of having a banner by 95% and 534% respectively. While websites that follow Portugal, Romania, Greece, and Germany local laws decrease the likelihood of having a banner. These results give us some indications that these differences might be related to the Guidance promulgated by authorities. For example, the Information Commissioners' Office use themselves a banner and suggest that businesses can copy that model if suitable for them (Information Commissioners' Office, 2012), and the National Commission for Computing and Liberties (CNIL) state in their websites how to comply with the E-Privacy Directive using banners (Commission Nationale de l'Informatique et des Libertés, 2015). In contrast, the other countries authorities did not suggest how websites should provide information to users.

Coming back to our research question *which law do websites follow? Are there differences in tracking and cookies notices related to the law they follow?* Our results suggest that websites follow the local laws and norms of the target market they decide to serve. A possible explanation is related to the commercial strategies of the companies and managerial decision to use their websites to serve specific areas. Also, companies are expected to contribute to the development of the target markets they serve and comply with its law and norms. Hence, websites .Com might want to target a broad international audience, but they still have to comply with the local laws of their target audiences, so they use geolocation to adapt to them, while companies that are local do not require to use geolocation to adapt to users' location because they websites are already following local laws of their target audience. Also, we found that the same pattern holds for banners. Moreover, there are differences in tracking and cookies notice in member states. We observed some clues that these differences might be associated to the transpositions of the E-Privacy Directive. However, the market forces or country characteristics in which websites are located might play a role too.

4.3 What local provisions of the E-Privacy Directive and market forces factors, if any, drive or discourage tracking presence in member states?

As it was stated in the previous question, the different transpositions of the E-Privacy Directive in member states might explain the variability in tracking we observed. Besides, these different transpositions have been the subject of different sides and opinions on which approaches were the best in terms of privacy protection. However, there is a lack of quantitative evidence to determine if these differences exist, and if they exist, what their impact on tracking cookies presence is. on the other hand, the market forces and business' incentives to use tracking in member states could also explain the differences we observed. Hence, answering this question, we want to provide some empirical evidence to shed some light on which approaches discouraged or encouraged tracking, and if the local provisions of the E-Privacy Directive and/or the market forces can explain the differences in tracking presence.

4.3.1 Local Provisions of the E-Privacy Directive

We will use the local provisions of the E-privacy Directive proposed in the conceptual model in Chapter 2 to test their impact on tracking[9]. Also, we decided to control for websites' categories since examining the actors' incentives have proved in other studies to be an important step to understanding why some measures work different than expected or do not work (Asghari, 2016; Moore, 2010). Hence, we will run a model using each provision alone, and a model controlling for the websites' categories, which for us represent the businesses' incentives to use tracking.

4.3.1.1 Consent

We remind the reader that we used the categorization of DLA Piper to determine which countries applied explicit consent (opt in) and which ones did not. In countries where explicit consent is expected, cookies can only be dropped or read if users give their consent.

H1: Explicit consent leads to less tracking presence.	Rationale: websites have to specify the purpose of data collection and the use of tracking, so this can lead to the reduction of information asymmetry between users and websites, so websites have less incentive to use tracking.

To start, we inspect our dependent variable 'Third Party Domains' per consent requirement. Figure 40 shows the box plot with the results.

9 We only used the provisions to test tracking presence since our question was mainly focus on tracking, but for the interested reader, we modelled each of the provisions for banners and the results are in Appendix C.

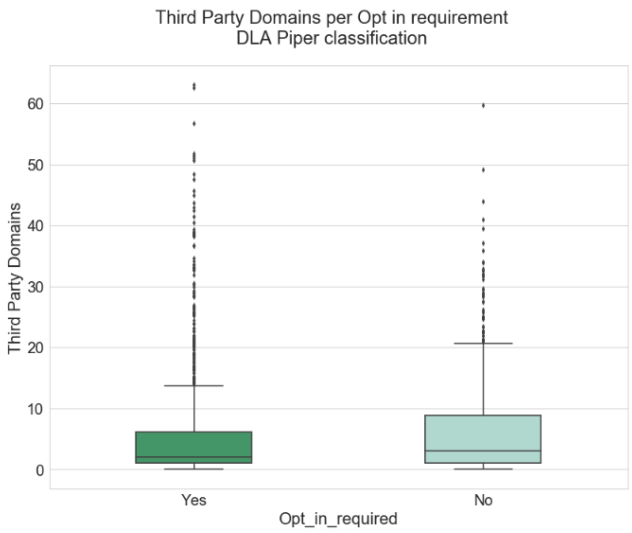


Figure 40: Third Party Domain per Opt-in requirement

We observe that websites located in countries where consent is required have less Third Party Domains than the ones that do not require consent. Table 21 depicts the descriptive statistics, and it shows that 720 websites belong to countries where consent is not required with a median of 3 Third Party Domains per website. On the other hand, 914 websites belong to countries where consent is required, and the median of Third Party Domains is 2 per website.

Table 21: Descriptive Statistics - Websites group by consent requirement

	count	mean	std	min	25%	50%	75%	max
Opt_in_required								
No	720.0	6.308845	8.27640	0.0	1.0	3.0	8.902778	59.625
Yes	914.0	5.955712	9.71602	0.0	1.0	2.0	6.111111	63.000

After inspecting the dependent variable versus the independent variable, we run two regression models.

Model 1= Third Party Domains[10] ~opt in required

Model 2= Third Party Domains ~opt in required + Websites' Categories

The next table depicts the output of the two models:

10 The shape of the dependent variable Third Party Domains that is going to be used in all the analysis of this question was depicted in figure 36.

Results Consent

	Dependent variable:	
	Third Party Domains	
	(1)	(2)
Opt-in required (consent)	-0.058 (0.069)	-0.157** (0.062)
Websites Categories	No	Yes
Constant	1.842*** (0.052)	2.269*** (0.163)
Observations	1,634	1,634
Log Likelihood	-4,603.818	-4,391.586
theta	0.562*** (0.022)	0.763*** (0.032)
Akaike Inf. Crit.	9,211.636	8,837.171
Mc Fadden Pseudo R2	0.00007	0.0461

Note: *p<0.1; **p<0.05; ***p<0.01

If we compare Model 1 versus Model 2, we observe that in Model 1, which represent the Opt-in (consent) provision alone, the coefficient is not significant, and the model that best fits the data is Model 2 with a low AIC value (8,837.171) compare to Model 1. In model 2, we observe that Opt-in requirement's coefficient (-0.157) is significant at a $p<0.05$. Hence, we proceed to convert the Opt-in requirement's coefficient controlled per websites' categories to Incident Rate Ratios (See Table 22). We find out that websites that belong to countries that transposed opt-in requirement significantly decreases the use of Third Party Domains by 15% other things being equal. Hence, we can accept this hypothesis.

Table 22: Incident Rate Ratios of Opt in (consent)

Variable	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
Opt_in_required_Yes	No significant	No significant
Opt_in_required_Yes (Controlling per businesses' incentives)	-0.157**(0.062)	0.85482320 (0.75- 0.96)

4.3.1.2 Guidance

Our next hypothesis is related to the Guidance promulgated by the Data Protection Authorities in each member state. Data Protection Authorities could emit guidance to help businesses to understand how to apply the E-Privacy Directive. However, since each country is free to transpose a Directive, some Data Protection Authorities provided guidance, while others did not. With this in mind, we wanted to test the next hypothesis.

H2: Countries which emitted guidance from the Data Protection Authorities have less tracking presence	Rationale: Data Protection Authorities will provide Guidance that does not encourage tracking. Hence, websites will not be encouraged to use Third Party Domains in their websites.

Our first step was to inspect our dependent variable 'Third Party Domains' grouped by emitted guidance. Figure 41 shows the box plot with the results. We remind the reader that this classification of which countries provided guidance and which one no was taken from DLA Piper report.

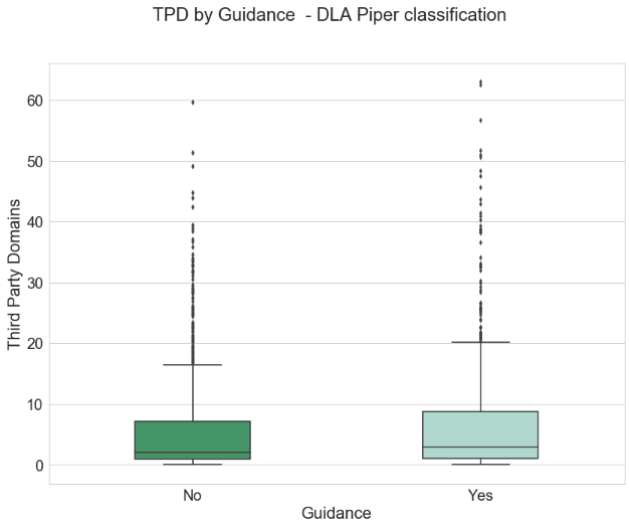


Figure 41:Third Party Domain per Guidance emitted by Data Protection Authorities

We observe that websites that belong to countries that emitted guidance by Data Protection Authorities has more Third Party Domains than the ones that did not get guidance. Table 23 shows the descriptive statics associated with Figure 41. We can observe that 1125 websites belong to countries where guidance was not emitted with a median of 2 Third Party Domains per website. On the other hand, 509 websites belong to countries where guidance was emitted, and the median of Third Party Domains is 2.94.

Table 23: Descriptive Statistics - Websites group by Guidance emitted or not in that country

Guidance	count	mean	std	min	25%	50%	75%	max
No	1125.0	5.573460	8.127165	0.0	0.944444	2.000000	7.166667	59.625
Yes	509.0	7.300091	10.886236	0.0	1.000000	2.944444	8.722222	63.000

After inspecting the dependent variable versus independent variable, we run two regression models.

Model 1= Third Party Domains ~Guidance

Model 2= Third Party Domains ~Guidance + Websites' Categories

The next table shows the output of the two regressions models:

Results Guidance

	Dependent variable:	
	Third Party Domains	
	(1)	(2)
Guidance_Yes	0.270*** (0.074)	0.121* (0.067)
Websites Categories	No	Yes
Constant	1.718*** (0.042)	2.112*** (0.161)
Observations	1,634	1,634
Log Likelihood	-4,597.345	-4,393.034
theta	0.567*** (0.022)	0.762*** (0.032)
Akaike Inf. Crit.	9,198.689	8,840.068
Mc Fadden Pseudo R2	0.0014	0.0458

Note: *p<0.1; **p<0.05; ***p<0.01

If we compare Model 1 vs Model 2 , we observed that in Model 1, which represent the guidance provision alone, the coefficient (0.270) is significant at a $p<0.01$. When we convert the coefficient to Incident Rate Ratios (See Table 24), we find out that websites in countries that emitted guidance have a 30.9% increase in tracking. However, we can observe that when we control this provision per businesses' incentives to use tracking, Model 2 has a lower AIC (8,840.068) than Model 1. Meaning that Model 2 is the best fit for the data. When we convert Guidance's coefficient of Model 2 controlled per websites' categories to Incident Rate Ratios (See Table 24), we detected that the magnitude of the coefficient decreases, and websites in countries that emitted guidance significantly increases the use of Third Party Domains by 12%, other things being equal. Hence, with these results, we cannot accept this hypothesis.

Table 24: Incident Rate Ratio Guidance

Variable	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
Guidance	0.270*** (0.074)	1.309(1.13- 1.51)
Guidance (Controlled per businesses' incentives)	0.121* (0.067)	1.128(0.98-1.28)

4.3.1.3 Fines

Some of the countries develop fines schemes related to data protection. Although not all cookies might represent personal data, as we discuss in our literature review the Working Party 29 state that it is not necessary to have a name or personal information to identify a user. Hence, reading through

the transpositions of the E-Privacy Directive as it was explained in the Research Methods chapter, we created our own classification of countries which develop fines schemes and which ones did not. Thus, we want to test the following hypothesis:

H3: Countries which developed fines schemes have less tracking presence	Rationale: the cost of dealing with fines will make businesses think twice about not notifying users correctly about the use of Third Party Domains. Since this notification requires an investment in modifying their websites, the companies will have fewer incentives to use Third Party Domains. Besides, fines in institutional economics are a mechanism to punish the companies that do not follow the law. Hence, the threat of the state might also decrease the incentive of using tracking.

First, we inspect our dependent variable 'Third Party Domains' grouped by fines schemes. Figure 42 depicts the box plot with the results.

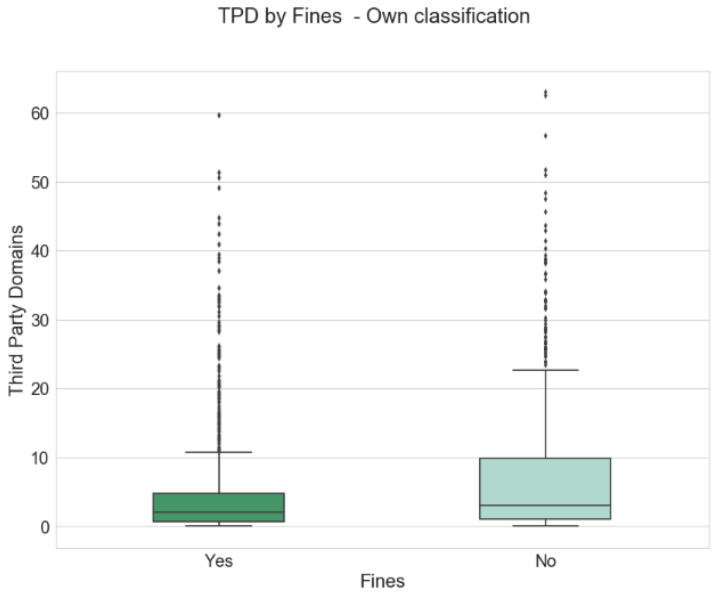


Figure 42: Third Party Domain per Fines

In figure 42, we observe that websites that belong to countries that develop fines schemes have less Third Party Domains than the ones that did not. Table 25 shows the descriptive statistics associated to figure 42, and we can observe that 812 websites belong to countries where there are no fines, and the median of Third Party Domains is 3.08 per website. On the other hand, 822 websites belong to countries where fines schemes were developed, and the median of Third Party Domains is 2 per website.

Table 25: Descriptive Statistics - Websites group by Fines

Fines	count	mean	std	min	25%	50%	75%	max
No	812.0	7.302065	9.711021	0.0	1.000000	3.084967	9.916667	63.000
Yes	822.0	4.935052	8.311820	0.0	0.638889	2.000000	4.708333	59.625

After inspecting the dependent variable versus independent variable, we run two regression models.

Model 1= Third Party Domains ~Fines

Model 2= Third Party Domains ~Fines+ Websites' Categories

The next table presents the results:

Results Fines

	Dependent variable:	
	Third Party Domains	
	(1)	(2)
Fines_Yes	-0.392*** (0.068)	-0.438*** (0.062)
Websites Categories	No	Yes
Constant	1.988*** (0.048)	2.355*** (0.159)
Observations	1,634	1,634
Log Likelihood	-4,587.985	-4,371.197
theta	0.574*** (0.022)	0.788*** (0.034)
Akaike Inf. Crit.	9,179.970	8,796.394
Mc Fadden Pseudo R2	0.0035	0.0506

Note: *p<0.1; **p<0.05; ***p<0.01

If we compare Model 1 vs Model 2 , we observed that in Model 1, which represent the fines provision alone, the coefficient (-0.392) is significant at a p<0.01. When we convert the coefficient to Incident Rate Ratios (See Table 26), we find out that websites in countries that develop fines schemes have a 33% decrease in tracking. However, when we control this provision per businesses' incentives to use tracking, Model 2 has a lower AIC (8,796.394) than Model 1. Meaning that Model 2 is the best fit for the data. When we convert fines' coefficient of Model 2 controlled per websites' categories to Incident Rate Ratios (See Table 26), we find out that the magnitude of the coefficient increases, and websites in countries that develop fines schemes significantly decreases the use of Third Party Domains by 36%, other things being equal. Hence, with these results, we accept this hypothesis.

Table 26: Incident Rate Ratio of variable Fines

Variable	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
Fines	-0.392*** (0.068)	0.675(0.591-0.772)
Fines (controlled per businesses' incentives)	-0.438*** (0.062)	0.645(0.57-0.73)

4.3.1.4 Information Requirement

To allow websites to read or drop cookies in users' devices, websites are required to provide information to users. With our own literature review, we classified countries which require that websites provide high information or low information to users. Also, 'No information' category was assigned to the control countries. Hence, we will use these categories to test the following hypothesis:

H4: Websites in countries which are required to provide more information to users have less tracking presence.	Rationale: Information provided to users diminish information asymmetry between the users and websites

First, in figure 43 we examine Third Party Domains by information requirement. We observe that the group of websites in countries which low information is required to be provided to users have higher tracking presence than the ones that are required to provide high information.

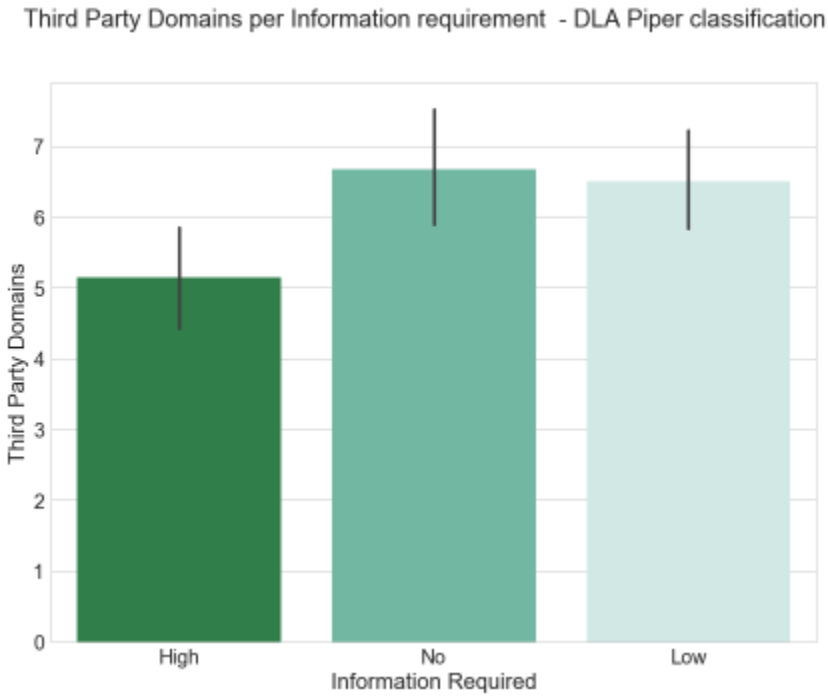


Figure 43: Third Party Domains by information requirement

Table 27 shows the descriptive statistics associated to figure 43, and we can observe that 520 websites belong to countries where high information is required, and the median of Third Party Domains is 1.6 per website. In addition, 750 websites which belong to countries where low information is required, and the median of Third Party Domains is 2 per website. On the other hand, 364 websites which belong to the control countries or where no information is required, and they have a median of 3.75 third party domains per website.

Table 27:Third Party Domains by Information Required

Info	count	mean	std	min	25%	50%	75%	max
High	520.0	5.145654	8.725622	0.0	0.0	1.666667	5.000000	51.333333
Low	750.0	6.503045	9.936136	0.0	1.0	2.000000	8.166667	63.000000
No	364.0	6.683698	7.661676	0.0	1.0	3.750000	9.888889	33.882353

After inspecting the dependent variable versus independent variable, we run two regression models.

Model 1= Third Party Domains ~Information Requirement

Model 2= Third Party Domains ~ Information Requirement + Websites' Categories

The next table presents the results:

The websites from control countries were used as reference category.

The next table depicts the output of the two regressions models:

Information Requirement

	Dependent variable:	
	Third Party Domains	
	(1)	(2)
Info_High	-0.262*** (0.095)	-0.480*** (0.086)
Info_Low	-0.027 (0.088)	-0.246*** (0.079)
Websites Categories	No	Yes
Constant	1.900*** (0.073)	2.456*** (0.169)
Observations	1,634	1,634
Log Likelihood	-4,598.889	-4,379.818
theta	0.566*** (0.022)	0.775*** (0.033)
Akaike Inf. Crit.	9,203.778	8,815.635
Note:	*p<0.1; **p<0.05; ***p<0.01	

If we compare Model 1 versus Model 2 , we observed that in Model 1, which represent the information requirements alone, the coefficient of high information (-0.262) is significant at a $p<0.01$, and the coefficient of low information requirement (-0.027) is also significant at a $p<0.01$. When we convert these coefficients to Incident Rate Ratios (See Table 28), we find out that websites in countries that required to provide high information to users decrease the use of trackers by 23.1%, while websites in countries that require to provide low information decrease the use of trackers by 2.8%. However, we can observe that when we control these two variables per businesses' incentives to use tracking, Model 2 has a lower AIC (8,815.635) than Model 1. Meaning that Model 2 is the best fit for the data. When we convert the information requirements' coefficients of Model 2 to Incident Rate Ratios (See Table 28), we find out that the magnitude of both coefficients increases and websites in countries that require to provide high information to users reduce the use of trackers by 38.2%, while websites in countries which require to provide low information decrease the use of tracking by 21.9%. Websites

in countries which are required to provide high information present less tracking, so we accept our hypothesis. However, we still see that the magnitude of tracking of websites in countries that require low information is high as well.

Table 28: Incident Rate Ratios - Low and High information required

Variable	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
Information High	-0.262*** (0.095)	0.769 (0.638- 0.926)
Information Low	-0.027 (0.088)	0.972 (0.816- 1.15)
Information High (controlled per businesses' incentives)	-0.480*** (0.086)	0.618 (0.519- 0.735)
Information Low (controlled per buinesses' incentives)	-0.246*** (0.079)	0.781 (0.666- 0.915)

We observe that the provisions by themselves are not better predictors of tracking than the model of the local laws that the websites follow. The model of the local laws that websites follow (Third Party Domains ~TLD) has an AIC value of 9,083 (This result was obtained in the previous question), but when we controlled the provisions with the businesses' incentives to use tracking, all the models were better predictors than the local laws and norms websites follow. Hence, this gives us some clues to think that the business models' incentives alone are important predictors of tracking. Hence, in the next section we will study businesses' incentive effect.

4.3.2 Market Forces Factors

In sub-question 2, we hypothesized that some of the tracking differences observed might be also related to the market forces factors and local characteristics of the target markets that websites decided to serve. In addition, analyzing the local provisions of the E-Privacy Directive, we observed that all models improved when controlling for businesses' incentives to use tracking. Hence, in this part of the analysis, we would like to test the impact of the business models' incentives on tracking. Thereafter, since in our literature review, we learned that there is an interplay between the market forces and the law, we would like to compare which institution is more powerful encouraging or discouraging tracking presence. Also, we will add the target audience characteristics proposed in the conceptual model in chapter 2 as control, as well as, other control variables related to the countries.

4.3.2.1 Business Models' Incentives

Companies have different purposes when launching a website. Some companies might want to offer information to their potential customers, some other companies might want to promote their products or services, while others might want to offer free content to monetize their audience, and many other digital strategies. Hence, each company might have different motivations to use or not use tracking mechanisms on their websites. With this in mind, we test the following hypothesis.

H5: Websites' categories, as proxy for business models' incentives to use tracking, influence the tracking presence	Rationale: Different Business Models have different incentives to use tracking on their websites.

First, we examined the dependent variable Third Party Domains per website category. Figure 44 shows the results.

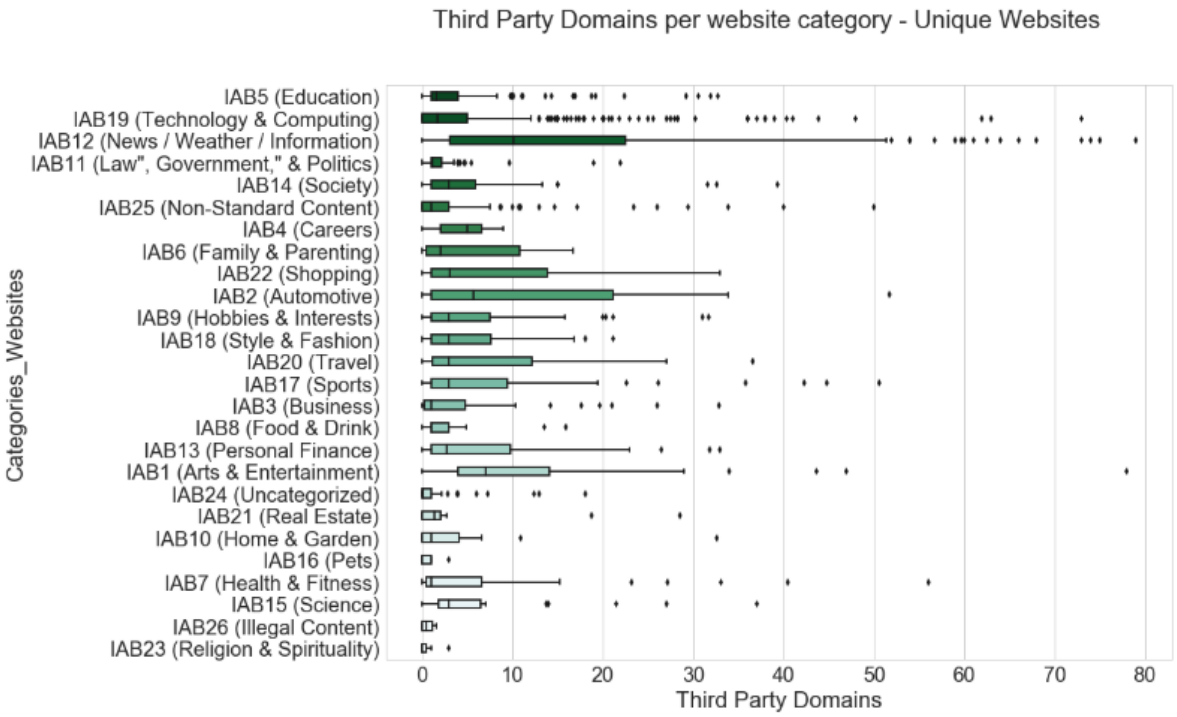


Figure 44:Third Party Domains per website categories

In figure 44, we observe that there is a high variability on Third Party Domains depending on the business models of the websites. We observe the category News/Weather/Information with the highest third party domain presence, followed by automotive. On the other hand, religious websites, government and illegal content show less tracking presence.

Table 29 presents the descriptive statistics of Third Party Domains per websites categories to better understand Figure 45. We observe that News/Weather/Information has a median of 9.16 Third Party Domains, followed by Automotive with a median of 7.44, while pets has the lowest median of 0, followed by illegal content with 0.5 Third Party Domains.

Table 29: Third Party Domains per website categories - Descriptive statistics

cat_1	count	mean	std	min	25%	50%	75%	max
IAB1 (Arts & Entertainment)	57.0	8.776293	8.226006	0.000000	3.000000	6.888889	12.000000	43.647059
IAB10 (Home & Garden)	18.0	3.767526	7.794659	0.000000	0.000000	1.000000	3.833333	32.625000
IAB11 (Law", Government," & Politics)	70.0	2.180859	3.521951	0.000000	1.000000	1.055556	2.111111	21.888889
IAB12 (News / Weather / Information)	326.0	12.975341	12.242078	0.000000	3.000000	9.166667	18.486111	62.500000
IAB13 (Personal Finance)	42.0	7.399090	9.220479	0.000000	1.000000	2.694444	9.805556	33.000000
IAB14 (Society)	24.0	5.863834	8.928749	0.000000	1.000000	3.000000	5.897876	32.666667
IAB15 (Science)	11.0	3.838384	5.987665	0.166667	1.166667	2.000000	3.027778	21.444444
IAB16 (Pets)	4.0	0.750000	1.500000	0.000000	0.000000	0.000000	0.750000	3.000000
IAB17 (Sports)	49.0	8.196651	12.469322	0.000000	1.000000	3.000000	9.888889	50.545455
IAB18 (Style & Fashion)	16.0	5.404667	6.912998	0.000000	0.777778	3.000000	4.852941	21.125000
IAB19 (Technology & Computing)	261.0	4.739438	7.751388	0.000000	1.000000	2.000000	5.625000	63.000000
IAB2 (Automotive)	13.0	13.998291	15.677490	0.000000	1.055556	7.444444	21.944444	51.700000
IAB20 (Travel)	70.0	6.865406	8.123798	0.000000	1.013889	3.000000	11.708333	36.611111
IAB21 (Real Estate)	11.0	5.030303	9.486422	0.000000	0.000000	1.000000	2.388889	28.500000
IAB22 (Shopping)	42.0	7.643480	8.604016	0.000000	1.027778	3.000000	13.708333	28.500000
IAB23 (Religion & Spirituality)	7.0	0.571429	1.133893	0.000000	0.000000	0.000000	0.500000	3.000000
IAB24 (Uncategorized)	77.0	1.377939	3.008521	0.000000	0.000000	0.294118	1.000000	18.111111
IAB25 (Non-Standard Content)	112.0	2.518207	5.366429	0.000000	0.000000	1.000000	3.000000	33.882353
IAB26 (Illegal Content)	4.0	0.638889	0.771802	0.000000	0.000000	0.500000	1.138889	1.555556
IAB3 (Business)	40.0	4.292482	7.098026	0.000000	0.000000	1.055556	4.472222	32.875000
IAB4 (Careers)	14.0	4.234432	2.712172	0.000000	2.000000	4.972222	6.375000	9.000000
IAB5 (Education)	269.0	2.948211	4.534238	0.000000	0.888889	1.058824	3.888889	32.777778
IAB6 (Family & Parenting)	8.0	4.208333	6.473826	0.000000	0.250000	1.500000	4.625000	18.750000
IAB7 (Health & Fitness)	22.0	1.121212	1.575425	0.000000	0.000000	1.000000	1.000000	6.000000
IAB8 (Food & Drink)	21.0	3.005291	4.187684	0.000000	1.000000	1.000000	3.000000	15.944444
IAB9 (Hobbies & Interests)	48.0	5.056134	6.584891	0.000000	1.000000	3.000000	6.500000	31.055556

Second, we proceed to run a GLM Negative binomial regression given the shape of the distribution of the dependent variable.

Model 1= Third Party Domains ~ Websites' categories

Table 30 depicts the result of the regression with the Incident rate ratios with 95% confidence interval.

Table 30: Negative Binomial Regression - Categories of websites

Variable	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
IAB10 (Home Garden)	-0.84564*(0.336)	0.430(6.516-12.143)
IAB11 (Law", Government, Politics)	-1.39234***(0.225)	0.248(0.159-0.385)
IAB12 (News / Weather / Information)	0.391**(0.171	1.478(1.044-2.047)
IAB13 (Personal Finance)	-0.171(0.244)	0.843(0.523-1.369)
IAB14 (Society)	-0.403(0.295)	0.668(0.380-1.216)
IAB15 (Science)	-0.827**(0.410)	0.437(0.204-1.038)
IAB16 (Pets)	-2.460***(0.829)	0.085(0.015-0.469)
IAB17 (Sports)	-0.068 (0.233)	0.933(0.591-1.480)
IAB18 (Style & Fashion)	-0.48479(0.345)	0.615(0.322-1.258)
IAB19 (Technology & Computing)	-0.61614***(0.176)	0.540(0.378-0.754)
IAB2 (Automotive)	0.46 (0.363)	1.595(0.815-3.433)
IAB20 (Travel)	-0.246(0.214)	0.782(0.511-1.189)
IAB21 (Real Estate)	-0.557(0.403)	0.573(0.271-1.346)
IAB22 (Shopping)	-0.138(0.244)	0.870(0.541-1.414)
IAB23(Religion & Spirituality)	-2.73167*** (0.680)	0.065(0.015-0.247)
IAB24 (Uncategorized	-1.851***(0.227)	0.157(0.100-0.244)
IAB25 (Non-Standard Content)	-1.249***(0.201)	0.286(0.192- 0.423)
IAB26 (Illegal Content)	-2.620***(0.863)	0.072(0.011-0.416)
IAB3 (Business)	-0.715***(0.253)	0.489(0.299-0.808)
IAB4 (Careers)	-0.729**(0.369)	0.482(0.2417-1.039)
IAB5 (Education)	-1.091***(0.177)	0.335(0.234- 0.470)
IAB6 (Family & Parenting)	-0.73499(0.533)	0.479(0.183- 1.552)
IAB7 (Health & Fitness)	-2.05764***(0.354)	0.127(0.064-0.258)
IAB8 (Food & Drink)	-1.07168***(0.322)	0.342(0.185-0.658)
IAB9 (Hobbies & Interests)	-0.55145*(0.238)	0.576(0.361-0.921)
(Intercept)	2.172*** (0.158)	8.776(6.516-12.14)

*p<0.1; **p<0.05; ***p<0.01. Null deviance/residual: 2293/1811.7 - AIC: 8841.322 Mc Fadden Ps eudo R2 = 0.0455

We observe that there is high variability of incident rate ratios, in this case, Third Party Domains being counted since most of the coefficients are significant. Highlighted in red we observe that websites which belong to the category News/Weather/Information has 47% increase in counted Third Party Domains, all other things being equal. On the other hand, websites belonging to the category Law, Government, and Politics, pets, religion and spirituality, non-standard content, and illegal content significantly decreases the likelihood of using Third Party Domains by 75.2%, 91.5%, 93.5%,71.4%, and 92.8% respectively. Hence, these results suggest that there is a varying degree of tracking presence depending on the business models of the websites. In addition, this model has an AIC value of 8,841.322 which is lower than model related to the local laws websites follow which had an AIC value of 9,083 (This result was obtained in the previous question). Meaning that the businesses' incentives to use tracking are better predictors of tracking.

Since we discovered that there is high variability of incident rate attributed to the type of websites' business models, we plot the significant coefficients of Model 2 to understand if there are any patterns associated to them. Figure 45 presents the results.

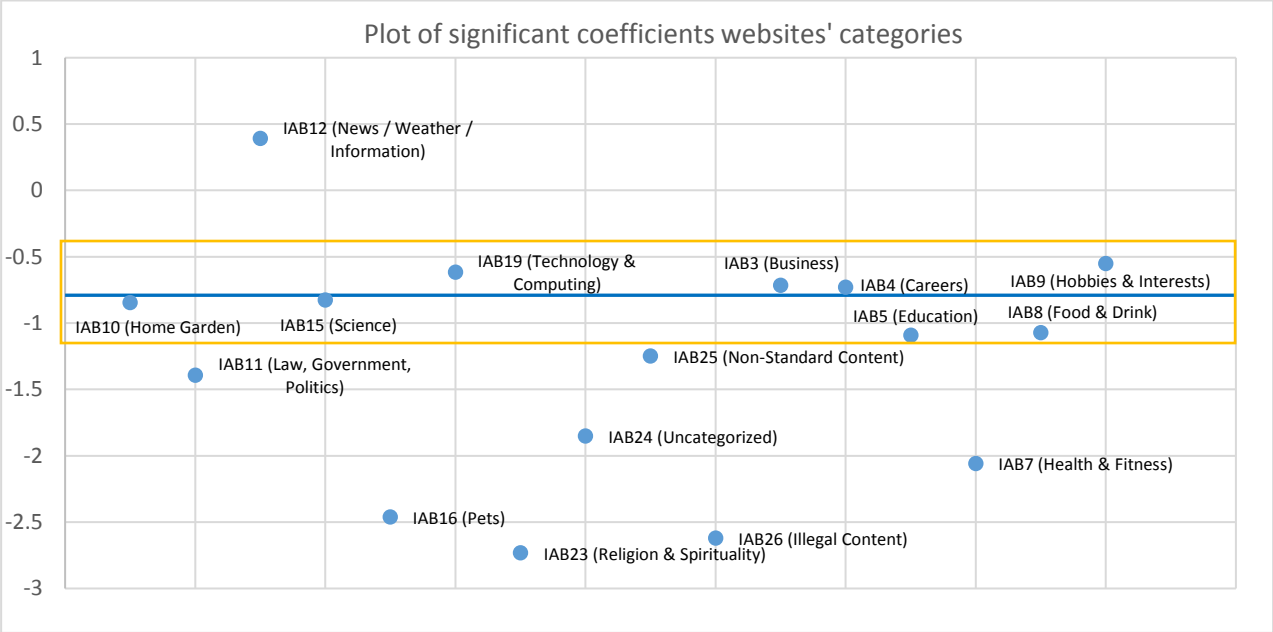


Figure 45: Coefficients of significant categories using the average of the regression coefficients as cutting line

In figure 45, the foremost, in the upper part, we observe News having the highest coefficient which means that this business models exert more tracking. Next, we observe a second group that is close to the mean, with some business models' categories which still have incentive to use tracking such as technology and computing, businesses, careers, hobbies and interest, and others business models that track by less such as home and garden, science, education, and food and drink. Finally, we observe a group with low coefficients meaning that they exert least tracking among the groups. Here we observe religion, Illegal content, health and fitness, pets, non-standard content and law, government and politics. Having in mind that these categories are a proxy for business models' incentives to use tracking, we consider that these business models can be classified into three broad categories. The first group is websites which business models' revenue streams are highly dependent on advertisement. The second group are businesses that are slightly dependent on ads to promote their

brands and/or still monetize their audience. Finally, the group in the bottom part is a group of businesses that can forgo advertisement.

Table 31 reminds the reader the number of websites that were analyzed in each websites' category.

Table 31: Number of Websites per significant categories

Websites Categories	# Websites
IAB12 (News / Weather / Information)	326
IAB5 (Education)	269
IAB19 (Technology & Computing)	261
IAB25 (Non-Standard Content)	112
IAB24 (Uncategorized)	77
IAB11 (Law, Government, Politics)	70
IAB9 (Hobbies & Interests)	48
IAB3 (Business)	40
IAB7 (Health & Fitness)	22
IAB8 (Food & Drink)	21
IAB10 (Home Garden)	18
IAB4 (Careers)	14
IAB15 (Science)	11
IAB23 (Religion & Spirituality)	7
IAB16 (Pets)	4
IAB26 (Illegal Content)	4

We observe that some of the categories have a low number of websites, while some of them have a high number of websites, so when interpreting the results of each category we found seems to be more interesting to pay attention to categories in the upper part of the table since they have a larger impact on the use or not of tracking.

4.3.3 The Law versus The Market Forces

In our literature review, we discussed that the law and the market forces are two institutions that influence tracking. Also, in our literature review, we proposed a model where all the variables that we have been partially testing influence tracking, and in which these two institutions were involved. Hence, to finalize this question we wanted to compare the models of the law websites follow, local provisions of the E-Privacy Directive and market forces to determine which institution, if the law or market forces, can encourage or discourage more tracking presence.

We will compare 10 models, and only the average of Third Party Domains will be used as the dependent variable. We remind the reader that TLD is a proxy for the local laws websites follow, and websites' categories represent businesses' incentives to use tracking.

Model 1 = Third Party Domains~ TLD

Model 2 = Third Party Domains~ Websites' Categories

Model 3 = Third Party Domains~ Websites' Categories +TLD

Model 4= Third Party Domains~ Websites' Categories + Opt in (consent)

Model 5= Third Party Domains~ Websites' Categories + Fines

Model 6= Third Party Domains~ Websites' Categories + Guidance

Model 7= Third Party Domains~ Websites' Categories + Opt in (consent) + Fines + Guidance

Model8=Third Party Domains~ Opt in (consent) + Fines + Guidance +Normalized Budged of DPA (Enforcement) + Education Index + Privacy Concerns

Model 9= Third Party Domains~ Websites' Categories + Opt in (consent) + Fines + Guidance +Normalized Budged of DPA(Enforcement) + Education Index + Privacy Concerns [11]

Model 10= Third Party Domains~ Websites' Categories + Opt in (consent) + Fines + Guidance +Normalized Budged of DPA + Education Index + Privacy Concerns + EU_Yes+ GDP per Capita 2016 + Internet Frequency Use 2016+ Rule of Law.

The next table depicts the output of the models:

[11] This model includes all the variables proposed in Chapter 2 – Literature Review. In this model we could not add TLDs, which we use as a proxy for websites' location, since there was a contingency with the provisions of the law, in other words they were not independent. In addition, Information requirement was not added because there was a contingency with consent. In addition, to increase the readability of the models the coefficients of the websites' categories were not included. The complete model can be found in Appendix D. Also, the analysis of the correlation of the Independent variables is presented in Appendix E.

Dependent variable: Third Party Domains

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Websites Categories	No	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
TLDau	0.795*** (0.200)		1.299*** (0.180)							
TLDbe	0.340* (0.200)		0.324* (0.180)							
TLDca	0.504** (0.203)		0.836*** (0.182)							
TLDch	0.451** (0.199)		0.453** (0.179)							
TLDcz	0.474** (0.199)		0.536*** (0.180)							
TLDde	0.921*** (0.200)		0.920*** (0.179)							
TLDes	0.517** (0.209)		0.589*** (0.189)							
TLDfr	1.126*** (0.201)		1.088*** (0.180)							
TLDgr	-0.313 (0.209)		-0.170 (0.193)							
TLDhu	-0.265 (0.209)		-0.125 (0.192)							
TLDit	0.536*** (0.206)		0.519*** (0.186)							
TLDjp	0.872*** (0.227)		1.033*** (0.201)							
TLDnl	-0.396* (0.208)		-0.098 (0.190)							
TLDpl	0.262 (0.223)		0.571*** (0.201)							
TLDpt	0.368* (0.203)		0.445** (0.190)							
TLDro	0.300 (0.203)		0.378** (0.185)							
TLDse	0.285 (0.202)		0.448** (0.182)							
TLDuk	1.133*** (0.198)		1.203*** (0.178)							
TLDus	-0.187 (0.255)		0.226 (0.230)							
Opt_in_required_Yes (Consent)				-0.157** (0.062)			-0.124*(0.071)	-0.043 (0.082)	-0.044 (0.074)	0.080 (0.083)
Fines_Yes					-0.438*** (0.062)		-0.396***(0.066)	-0.303*** (0.078)	-0.344*** (0.070)	-0.345*** (0.080)
Guidance_Yes						0.121* (0.067)	0.153**(0.072)	0.233*** (0.082)	0.128* (0.075)	0.175** (0.089)
Normalize_budget_DPA_2011 (Enforcement)								-0.003 (0.005)	-0.005 (0.004)	-0.009** (0.005)
Privacy Concerns								0.016*** (0.004)	0.012*** (0.004)	0.022*** (0.005)
Education_Index								3.595*** (0.901)	4.314*** (0.824)	2.701** (1.201)
EU_Yes										-0.283** (0.131)
GDP_per_capita_2016										-0.00001*** (0.00000)
Internet_Fq_Use_2016										0.008 (0.005)
Rule_law_2016										0.020*** (0.005)
Constant	1.322*** (0.143)	2.172*** (0.158)	1.543*** (0.198)	2.269*** (0.163)	2.355*** (0.159)	2.112*** (0.161)	2.337***(0.163)	-2.044** (0.878)	-2.056** (0.812)	-2.747*** (0.977)
Observations	1,634	1,634	1,634	1,634	1,634	1,634	1,634	1,634	1,634	1,634
Log Likelihood	-4,521.757	-4,394.661	-4,299.938	-4,391.586	-4,371.197	-4,393.034	-4,368.736	-4,570.333	-4,352.937	-4,336.854
theta	0.631*** (0.025)	0.760*** (0.032)	0.880*** (0.039)	0.763*** (0.032)	0.788*** (0.034)	0.762*** (0.032)	0.790***(0.034)	0.589*** (0.023)	0.810*** (0.035)	0.829*** (0.036)
Akaike Inf. Crit.	9,083.515	8,841.322	8,689.877	8,837.171	8,796.394	8,840.068	8,795.472	9,154.666	8,769.874	8,745.708
Mc Fadden Pseudo R2	0.017	0.045	0.066	0.046	0.050	0.045	0.051	0.007	0.055	0.058

Note: *p<0.1; **p<0.05; ***p<0.01

We observe that Model 2, which represent the business models' incentives to use tracking has a lower AIC value (8,841.332) than Model 1 (9,083.515) which represents the local laws that websites follow. In addition, we observe that Model 4, Model 5, and Model 6, have a lower AIC value than Model 2. Meaning that when we control the provisions with businesses' incentives, the provisions do have an effect. However, in Model 8, where the provisions are shown without controlling for businesses' incentives, we observe an AIC value of 9,154.666, so this model is not better than Model 2. Also, in Model 8, when the provisions interact with Education Index and Privacy Concerns, the consent coefficient becomes insignificant. On the other hand, in Model 9, in which the provisions are controlled with business models' incentives to use tracking, we observe a lower AIC value (8,769.874) than Model 2 (8,841.332). This means that the model of the businesses' incentives (Model 2) improved when we added the provisions of the E-Privacy Directive, so the provisions do have some effect. However, in Model 9, we also observed that the effect of consent is non-significant. A reason for consent not being significant in Model 8 and Model 9 might be that Education Index and Privacy concerns, as well as, the other local characteristics of the target market like in Model 10 might absorb its effect. In addition, in Model 10, in which we add more control variables related to the country, the control variables related to the target audience privacy concern and education are significant. As privacy concern increases tracking increase by 2.21%, and as the education of the target audience increase the incident rate ratio, or the relative risk of finding a tracker in a website, is 14.8. Also, we observe that websites that belong to EU significantly decrease the presence of third party domains by 24.5%, other things being equal. Also, enforcement capacity has a small effect in reducing tracking. As the budget of the data protection authorities increases tracking decrease by 0.9%. However, we need to keep in mind that these budgets were a bit out of date.

After comparing all the models, we observe that the proposed model in chapter 2 is not the best models. We observe that Model 9 has an AIC value of 8,769.874, and when adding more country variables as a control in Model 10 the AIC value is 8,745.708. However, from all the models we observe that Model 3 = Third Party Domains~ Websites Categories + TLD is the one that best fits the data with the lowest AIC value =8,689.877. Hence, this implies that the local laws and norms that websites follow and businesses' incentives to use tracking predict most tracking presence. Also, since we observed that when we added country related variables in Model 10, this model was not better than the parsimonious Model 3, we can think that the TLD captures most of the elements related to the location of the website. On the other hand, since we observe that Model 2 which represents the businesses' incentives to use tracking is better than Model 1, we can conclude that the businesses' incentives to use tracking explain more tracking presence. In addition, we observe that the provisions of the E-Privacy Directive do have an effect when controlling for businesses' incentives to use tracking, so we can conclude that the different transpositions of the E-Privacy Directive do have an effect explaining tracking presence by less.

4.3.4 Interpretation of the Key Findings

We found that the different local transpositions of the E-Privacy Directive lead to differences in tracking. We had two different outcomes of the provisions. First, we test the effect of the provisions alone. Second, we controlled the provisions for business models' incentives to use tracking. With the second outcome we realized that businesses' incentives have an important effect on tracking since when we did not control for them, there was bias in the parameter estimate of the provisions.

First, we observed that the provision of (users' giving their) consent alone does not have an effect, but when controlled for the business models' or the website owners' incentives to use tracking, we have observed that websites located in countries that transposed explicit consent significantly decreases the likelihood of using a tracker (by 15%). This reduction in tracking might be explained because to gain consent, websites need to provide information to users on what they will do with their data and the purposes of cookies installed in users' devices. Hence, consent reduces the information asymmetry and Principal-Agent problem between websites and users since the users might be aware of why the agent want to exert tracking, and what the agent is expected to do when the principal consent to the use of tracking. Second, we found that websites located in countries that impose fines significantly decrease the likelihood of using tracking by 36%, more than we observe without controlling for the businesses' incentives which was 32%. A possible explanation for this result is that fines may act as a punishment for the companies that do not adhere to the norms thereby reducing the businesses' incentives to track. Also, fines schemes have been used to correct "bad behaviors", so this might lead to businesses discouragement to use tracking. Besides, the threat of the state and the possibility that if businesses fail to comply correctly with the law will have to incur in fines might also reduce business incentives to use tracking. Third, for websites located in countries that promulgated Guidance via their Data Protection Authorities, strikingly, a significant increase in tracking, of 30%, was observed, but when we controlled for business models' incentives to use tracking the magnitude of the effect was reduced to 12%. This result came to surprise us since it was not expected. According to Deloitte (2017b), the guidance issued by member states vary and authorities emit guidance about different topics. Hence, we think that this result might be related to the divergence in the Guidance from the Directive in member states. Perhaps, the transposition of Directive into Guidance might have watered-down its rigidity only to enable businesses to exploit the context. Finally, websites in countries in which more information is expected to be provided to users, tracking decreased by 38.2%, and websites in countries which require to provide low information to users decreased the use of tracking by 21.9% (in both cases we controlled per businesses' incentives to use tracking), so this might suggest that the idea of requiring websites to provide information to users might be, in general, a suitable approach to reduce tracking.

The fact that countries that transposed explicit consent, impose fines, and provide information to users have a positive effect discouraging websites to use tracking should motivate policy makers to harmonize these provisions in member states. Also, our results suggest that the consent provision should be strengthened to have a more powerful effect on reducing tracking. On the other hand, the unification and clarity of guidance should be considered to avoid the unintended effect of increasing tracking. Providing clear interpretation on how to comply with the law is likely to increase the harmonization in privacy protection in member states. Besides, Guidance needs to address the challenge of not encouraging tracking or leaving opportunities open for different interpretations from businesses that can encourage the use of trackers in their websites.

Also, our findings indicate that there is a varying degree of tracking that it is associated to websites' categories, which is a proxy for the business models' incentives to use tracking. We could see that there were three groups of websites' business models which have different incentives to use tracking. First, the companies whose revenue streams are highly dependent on advertisement exert more tracking. In this category we found News. As was discussed in sub-question 1, News' business models historically have depended on advertising, but in the recent years News is struggling to stay profitable, their circulation is decreasing or even disappearing, and subscriptions too. Most of them are switching to an online presence, but their content is accessed for free (Kirchhoff, 2011). Hence, since advertisement has been part of their core business, it seems to be that the use of online behavioral

advertisement has become one of their main revenue streams. Next, those companies whose revenue streams are slightly dependent on advertisement exert less tracking compared the first group. This group is composed of businesses that have other sources of revenue streams, but they might have the incentive to use this type of ads to promote their own brands and/or promote their services. Also, they might offer some free content to a specific audience, so they still have some incentive to monetize users' visits to their websites. We observe in this category technology and computing, hobbies and interest, careers, and businesses. Also, tracking by less, but still close to the mean we have home and garden, education, food and drink, and science. The final group, companies whose revenue streams are independent on advertisement exert the least tracking among the groups. These companies are the one that can forgo online behavioral advertisement. These groups' core businesses might be to provide information, their products or services are very specific, their reputation is more important, or their revenues are associated with the anonymity of their users. In this group, we have religion, law and government, non-standard content, illegal content, pets and health and fitness. Religious and governmental websites main aim is to provide information to citizens/users about their services, and in the case of the government, reputation is important. Non-standard content, which includes adult websites, and illegal content's core businesses are based on the anonymity of their users, so they might also not be interested in using tracking since people might be less willing to visit their websites, and this can affect their revenue streams. Moreover, pets are a more specific business since pet buyers are people who have the means, want, and can afford a pet, so they might not be interested in the use of ads. Finally, to our surprise, we observed in this group health and fitness. Checking a bit further the IAB categorization, health and fitness include information related to illness and sexuality, so it might be that due other laws that have been applied to the health sector, their incentives have changed to not track users based on their clinical conditions or sexual preferences.

We think that these differences related to the business models might be explained by the intrinsic business motivation to make a profit. As proposed by Englehardt and Narayanan (2016a) the lack of other revenue streams might put pressure on some type of websites to monetize their audience, while for other businesses this might not be necessary. Besides, business models are part of a quite stable institution that through the years have been using the same revenue streams, so developing ideas or coming up with different revenues streams might impose challenges to companies, so they might use what is a common practice, which is integration of third parties, online to monetize their websites. In addition, our findings imply that business models that are highly dependent on advertisement might be the ones that can have more incentives to take advantage of information asymmetry, and that might have conflicting interest to serve the principals than the other business models.

Interestingly, we found that business models' incentives to use tracking were more powerful predictors of tracking, even more powerful than the local laws websites follow, and the local transpositions of the E-Privacy Directive. This is an important outcome in terms of policy making. This suggests the necessity to promote the right incentives for businesses, and that businesses' incentives need to be better understood. Businesses are the ones that apply the law, and from them depends on the success or failure of the future regulation and any other regulation. Hence, understanding their incentives might aid regulators to understand how to coordinate efforts to persuade businesses either through monetary or non-monetary means to reduce the use of tracking in their online business models.

Finally, some additional observations were that when adding country related control variables was observed that websites in EU countries significantly decrease the likelihood of using third party trackers. Hence, the E-Privacy Directive does have a positive effect reducing tracking in member states in comparison to our control countries Japan, Switzerland, The United States, Canada, and

Australia. This might be related to the fact that Canada, Japan, Switzerland, and Australia have relied on following the principles of the OECD guidelines and self-regulatory instruments such a code of conduct (Organisation for Economic Co-operation and Development., 2003b), and The United States does not even have 'rigth to privacy' expressed in their constitution, and the rely on self-regulation, and some sector-specific regulation (Organisation for Economic Co-operation and Development., 2003b). This outcome represents good news for the regulators given the complexity of privacy. In addition, we observed a small effect of enforcement in reducing tracking. As the Data Protection Authorities' budget Increases the use of trackers decreases by 0.9%. Even though the data used was a bit out of date, we can observe that having enough enforcement capacity to ensure companies' compliance might help to tackle this issue. Other observations were from the target audience characteristics. First, as privacy concerns increase the tracking presence increase by 2.21%. A possible explanation is that users might get concerned about privacy when the use of trackers increase. However, more study might be necessary to establish causality. Second, when the education index of a country increases, the risk of finding a tracker is 14.8 more. We need to be careful in interpreting this result since it might be possible that the education index represents other characteristics of the target audience. One possible explanation is that well-educated people have more income and access to the internet, so they might be an attractive target audience for advertisement. Nevertheless, in general, we observe that varying degree of business incentives to exert tracking go beyond the target audience's characteristics.

Going back to our research question *What local provisions of the E-Privacy Directive and market forces factors, if any, drive or discourage the prevalence of web trackers in member states?*

The provisions that discourage tracking presence are consent, fines schemes, and providing information to users, while Guidance encourages it. Also, business models' incentives as part of the market forces lead to a varying degree of tracking presence, and these incentives predict tracking presence more than the local law that websites decide to follow, and the different local transpositions of the E-Privacy Directive. This might be explained due to the fact that the companies have different core businesses and revenue streams which have been institutionalized for many years, how they decide to bring value to the target audience, and their self-interest of making a profit.

4.4 What are the implication of the findings to policy makers?

In this question through reflection and in a prescriptive manner, we determine how the main findings of this study are relevant for policy makers considering the upcoming E-Privacy Regulation.

The first point to consider is that cookies are still a pervasive mechanism and that trackers have a long tail. The fact that tracking cookies are still a widely used mechanisms by larger websites to exert tracking implies that it is relevant for the future E-Privacy Regulation to include clear definitions on cookies, how to gain access to them, and their exceptions. Besides, policy makers should pay attention to the incentives of third party companies that are at the high end of the tracking distribution. These companies are present in most of the websites, and users have more chance to find them on a regular basis. Hence, the long tail can facilitate legal actions against these companies if they do not comply with the law (Englehardt & Narayanan, 2016b)

Secondly, we found that even larger websites follow local laws of their target markets, and there is high variability of tracking depending on the local laws websites follow. The combination of these two findings implies that policy makers should strive for achieving harmonization to reduce tracking. Hence, this implies that the idea of a regulation will contribute to ensuring the same level of protection across the European Union. In addition, when analyzing the impact of the different transpositions of the E-Privacy Directive and given the different sides and opinion on which approaches were better to increase privacy protection, based on our observations, we can recommend to policy makers to pay special attention to the harmonization on consent and guidance. We observed that there is some positive effect from consent in reducing tracking when controlling for business models' incentives. Hence, this provision should be strengthened in the future E-Privacy Regulation. On the other hand, we suggest to policy makers to pay attention to how to emit guidance for the future E-Privacy regulation. The guidance promulgated by authorities can lead to non-intendent effects if it is not clear and consistent with the regulation.

Third, we learned that that business model incentives lead to a varying degree of tracking, and these incentives are even more powerful than the local transpositions of the E-Privacy Directive and local laws websites are following. This implies that policy makers should understand better business incentives to exert tracking to try to persuade businesses to try to achieve the goal of reducing tracking to increase privacy protection. The future regulation and any other future regulation will have to face businesses' incentives, so understanding them might be a workable alternative that can decrease the cost of reducing tracking, while combined with other conventional regulatory approaches. In addition, enough and credible enforcement capacity is necessary to control businesses' incentives. Companies are acting in their self-interest, and this can lead to over-use of tracking. Hence, being able to determine which companies are not complying with the law might also help to shape the incentives of other businesses and avoid their noncompliance. Collaboration among member states might be key to unify efforts and learn from each other. In addition, since we observed that fines schemesand providing information to users reduce tracking, this might imply that policy makers should strength these provisions to aid this task.

Finally, EU countries have less tracking in comparison to Japan, Canada, Australia, Switzerland, and The United States. Hence, policy makers should continue efforts to contribute to privacy protection to ensure that users' feel comfortable with their online privacy protection, so this might help to the economic growth of the digital business economy in the European Union.

This page was intentionally left in blank

Chapter 5

5. Conclusions and Discussion

The main objective of this thesis was to streamline the opinions about which approaches of the E-Privacy Directive were better in terms of privacy protection and bring empirical evidence to the privacy debate in an effort to yield recommendations to policy makers about which law elements and market forces could help to reduce tracking in the E-Privacy Regulation reform.

Up to now, the answers of our research sub-questions have partially answered our main research question and have helped us to understand how the market forces and the law can explain the tracking presence. In this chapter, we will summarize the key findings of each of the research sub-questions that once interpreted as a whole will help us to draw conclusions to answer the main research question *What legal and market factors can explain the presence of tracking cookies across the European Union, and how the E-Privacy Regulation reform can reduce tracking and improve privacy protection?*

Sub-question 1: What is tracking? How pervasive are they in the European Union countries?, and What are the types of tracking in use?

First of all, to answer which legal and market factors explain the presence of tracking, we needed to understand first if there was a tracking presence. The first part of the question was answered through the literature review, and the second part of this question was answered through descriptive statistics, and we confirmed the following:

- Tracking presence was 81% in all the websites, regardless of the location of the users.
- Trackers have a long tail, and the most predominant type of trackers in use are involved in advertisement.

Sub-question 2: Which law do the websites follow? Are there differences related to which law they follow?

Secondly, to the best of our knowledge, there was no literature that use empirical data to test which law the websites were following and understanding this was key to test if there were differences on tracking among European Union countries. our main findings in answering this question were:

- Even larger websites follow local laws. Sites .com use geolocation to adapt to users' location, while websites with specific target markets stick with the local laws of the market they serve.

- There are differences in tracking and cookie notices related to the local laws websites follow. Websites that follow The Netherland's local law had 32% less tracking, while websites that follow UK's local law had the highest use of tracking with an incident rate ratio of 3.094. These differences in tracking presence and cookies notice imply a lack of harmonization in privacy protection among member states.

Sub-question 3*: *What local provisions of the E-Privacy Directive and market forces factors, if any, drive or discourage the prevalence tracking presence in member states?

We tested the impact of the local transpositions of the E-Privacy Directive (alone and controlling per businesses' incentives to use tracking) and the role of the businesses' incentives encouraging and discouraging tracking presence. Our main findings in answering this question were:

- The local provisions Consent, Fines, and Information requirement discourage tracking's presence, while Guidance encourages tracking's presence. Table 32 summarize the results of the hypothesis when controlling for businesses' incentives.
- The market forces factors that encourage or discourage tracking's presence is business models' incentives to use tracking. The different businesses' incentives to use tracking led to a varying degree of tracking presence, and this was associated with the use of ads as their main revenue stream. Table 33 shows the websites' categories listed from exerting more to less tracking.

Table 32: Hypothesis tested in sub-question 3 regarding local provisions of the E-Privacy Directive

Hypotheses	Accept /Reject
H1: Explicit consent leads to less tracking presence.	Accepted. Consent reduces tracking by 15%.
H2: More guidance from the Data Protection Authorities leads to less tracking presence	Rejected. Guidance increases tracking by 12%
H3: Fines schemes established in the country leads to less tracking presence	Accepted. Fines reduce tracking by 36%
H4: Websites in countries which are required to provide more information to users have less tracking presence	Accepted. Information reduce tracking by 38.2%

Table 33: Summary of hypotheses tested in sub-question 3 regarding market forces

Hypotheses	Accept /Reject
H5: Websites categories, as a proxy of business models' incentives to use tracking, influences the tracking presence	Accepted. Ad-dependent business models have high use of tracking. Category: News. Slightly a- dependent Business models track by less but still has the incentive to use tracking. Categories: Hobbies and interest, Technology and computing, Business, Careers, Home and Garden, Food and Drink, Education. Businesses that forgo Ads has low tracking. Categories: Non-standard content, Government and Politics, Health and fitness, Pets, Illegal content, Religion.

Finally, we compared the models of the law that websites follow, local provisions of the E-Privacy Directive and market forces to determine which institution, if the law or the market forces, is more powerful encouraging or discouraging the tracking presence. The main finding was:

- The different business models' incentives, as part of the market forces, to use tracking were more powerful predictors of tracking than the local laws that websites follow, and the local transpositions of the E-Privacy Directive. The regression model of businesses' incentives has an AIC value of 8,841 versus the regression model of local laws 9,083.515, and the regression model of the provisions of the E-PD alone 9,154.7. Hence, the market forces explain tracking presence more than the legal framework.

When comparing the law and the market, we had some additional observations:

- Education influences tracking presence. When the education of the country increases, the relative risk of finding a tracker in a website is 14.8.
- Privacy concerns influence tracking presence. When privacy concern increases. the tracking presence increase by 2.21%
- When enforcement capacity increase, the tracking presence decrease by 0.9%.
- Websites in the European Union countries significantly decrease the use of trackers by 24.5%.

Sub question - 4: What are the implications of the findings to policy makers?

In this question, we reflected on the implications of the findings of the previous questions to policy makers. The main conclusions were:

- Policy makers should ensure harmonization, especially on consent and guidance.
- Policy maker should better understand businesses' incentives to use tracking accompanied by a credible enforcement capacity while strengthening fines and providing information to users.
- Policy makers should continue efforts to protect privacy.

After reviewing the key findings of the research sub-questions we can easily answer our main research question ***What legal and market forces factors can explain the presence of tracking cookies across the European Union, and how the E-Privacy Regulation reform can reduce tracking?***

The legal and market forces factors that can explain the presence of tracking cookies across the European Union are the variability in business models' incentives to use tracking and the lack of harmonization of the transposition of the E-Privacy Directive by less.

The future E-Privacy Regulation reform can reduce tracking by:

- Harmonizing the local provisions of the E-Privacy Directive across member states, especially on consent and guidance.
- Better Understanding the business incentives to avoid market failures accompanied by credible enforcement capacity while strengthening fines and requiring websites to provide information to users.

5.1 Discussion and Reflection

From the beginning, the purpose of this study was to shed some light on policy makers on how the E-Privacy Regulation could reduce tracking understanding what legal and market forces encourage or discourage tracking basing our analysis in the E-privacy Directive. From the legal framework, our findings emphasize the need to promote harmonization of the provisions of the E-Privacy Directive in member states, especially on consent and guidance. We showed that the different transpositions of the member states led to different outcomes encouraging or discouraging tracking, so these results suggest that there is an opportunity for the regulation to reduce tracking by the means of harmonization of the provisions to the approaches that reduce tracking presence. In addition, from the market forces, our work revealed that businesses' incentives play an important role in explaining the variability of tracking presence. Therefore, our findings indicate that businesses' incentives need to be better understood, especially the incentives to use tracking by businesses where revenue streams are highly dependent on advertisement.

To determine the practical implication of our main findings, we revised the draft the of the E-Privacy Regulation published May 4th, 2018, so we could identify if these points are addressed in the draft of the E-Privacy Regulation or if there is still room for improvement.

The first point we recommended was the *Harmonization the local provisions of the E-Privacy Directive across member states, especially on consent and guidance.* One of the main findings in this study is that larger websites follow local laws, and the different local transpositions of the E-Privacy Directive led to differences in tracking. The impact assessment of the E-Privacy Regulation mentions that the different transpositions on member states lead to challenges for businesses, and that there is a need for harmonization ('Proposal for a Regulation on Privacy and Electronic Communications', 2017). Now with our quantitative evidence, we agree that there is a lack of harmonization, and harmonization of the provisions can help to reduce tracking. Hence, the fact that the Directive will become a regulation already might help to harmonize the different approaches across member states since countries must adopt it without changes. In addition, based on our empirical analysis, we can propose especial attention to consent and guidance. Some countries transposed consent, and we observed that it led to decreasing tracking presence. However, some of these countries have flexible approaches to consent, and even implied consent is accepted as consent. Hence, if consent is strengthened, the reduction in tracking might be stronger than right now. We observed that the draft of the E-Privacy Regulation now has a clear definition of consent, assign the European Data Protection Board, a single entity, to provide guidance regarding consent, and examples of exceptions on when consent is not necessary are clear. Hence, this might prevent subtle changes in member states or flexible interpretations of consent. In contrast, Guidance is encouraging tracking, so correcting this non-intended effect is necessary. In the regulation, it is stated that Guidance will be in hands of the European Data Protection Board, so this will facilitate the interpretation of the regulation for all member states. Therefore, we think that these points are already considered in the draft.

Regarding the second point which suggests that policy makers should *Understand better business incentives to avoid market failures accompanied by credible enforcement capacity while strengthening fines*, we think that there is room for improvement. An important finding in our study that different business models lead to a high variability of tracking. Also, another observation was that the business models' incentives were a powerful predictor for tracking even more than the local law that websites follow and the local transpositions of the provisions of the E-Privacy Directive. However, the E-Privacy Regulation is still a 'one size fits all' approach. All websites must comply

with the same regulation. Nevertheless, the differences we found might suggest that a different approach should be taken depending on the business models of the websites.

In addition, although with a regulation 'businesses must comply', it is necessary to have credible enforcement capacity that allows to find out if businesses are not complying with the regulation to avoid non-compliance from other businesses. Moreover, one of the provisions that can help to shape businesses incentives is fines since they showed to reduce the use of tracking by websites. The draft of the regulation establishes now that each member states must have one or more than one supervisory authorities to enforce the regulation, authorities shall cooperate and be independent, and have the power of imposing fines. Hence, we think that this point is partially covered, but there is the opportunity to commence an important debate regarding understanding the "parsimonious" factor of businesses' incentives to try to re-solve the privacy problem.

After the analysis of our two recommendations with the draft of the E-Privacy Regulation, we reflected more on the second implication since there are still opportunities to improve. Businesses are vital for any economy, and they are part of an institution which is the market forces. Institutions shape what we do, so we observe through the years that business models and their revenues stream has been established as default. Businesses are part of a network of actors that have different interests on using tracking depending on which business models they chose to bring value to their customers. Hence, businesses have adapted the use of tracking according to their self-interest. Hence, this might explain the variability of tracking we observed in our results.

In addition, businesses were part of the set of actors that wanted to try to influence how the E-Privacy Directive was going to be transposed in member states. Businesses probably wanted to ensure to continue doing their business as they were used to do and continue generating profits. Hence, we can think that businesses' incentives influenced how the E-Privacy Directive was transposed in member states, and these subtle differences were reflected in our data analysis showing differences in tracking in member states and regarding the local transpositions of the Directive.

Apart from the economic component that tracking can represent to businesses' revenues streams, we think that business models' incentives can predict more tracking because the regulators might be in disadvantage to regulate companies. Businesses evolve faster and have more up to date personnel regarding tracking technologies, so the government has the challenge to keep up at the same level. In addition, even if the government has enough enforcement capacity, it might be impossible to check every single website. Hence, the businesses will be in a better position than the government due to information asymmetry.

In addition, the E-Privacy Directive has required amendments through the years in an attempt to re-solve the privacy problem, but our results have demonstrated that the variability of incentives of the business models still leads to different levels of tracking. Hence, more ad-dependent business models and business models that have some incentives to use tracking can lead to problems such as externalities, Principal-Agent problem, and the tragedy of the commons. This implies that a regulatory approach to controlling these business incentives in different ways might be necessary. Trying to address this issue with command and control approaches might provide to companies with certain rules that they can follow, but they might not have the right incentives to by themselves look for other solutions. Also, imposing more strict regulation on business models which use more tracking might change other businesses' incentives. For example, other businesses that would not be regulated at the same level might have incentives to start using tracking. Therefore, we think that informative instruments can facilitate transparency and accountability, and they might help in the task of policy

makers to shape businesses' incentives. Also, encouraging firms to not use tracking to a point that it is profitable for them might help to tackle this issue. Moreover, compromising the reputation of firms that do not comply with the law and imposing, fines might also aid this task.

Regarding the informative instruments, we observed that websites in countries that require to provide more information or some information to users reduced the use of trackers. Hence, simple approaches where the information asymmetry between websites and users are reduced might facilitate shaping the incentive of businesses. Openness as one of the principles proposed by the OECD guidelines might ensure fairness in the data collection from websites. In addition, some studies are aiming to promote transparency regarding the use of tracking to generate accountability of companies that are exerting it (G. Acar et al., 2014; Englehardt & Narayanan, 2016b), and transparency has helped to stop some businesses from using tracking mechanisms such as fingerprinting (Englehardt, 2016). In the moment users trust websites, to get some content or simply surf the web for leisure, users face information asymmetry, and users have transactions costs associated to check if websites are using tracking or not or even lock-in costs. Hence, informative instruments could minimize the information asymmetry, so this could align the incentives of the principal, websites, and reduce websites' information advantage. Thus, we think that this approach might be efficient and less costly for policy makers to shape businesses' incentives, especially because there is a varying degree of them.

Secondly, we have argued that businesses self-interest in making a profit might encourage the use or not of tracking mechanisms in their websites, so if profitability is associated and aligned with this intrinsic interest, an approach that help businesses to make a profit without using trackers might also help to tackle this issue. It might be possible to promote innovation or subsidies to encourage digital businesses that do not use tracking. This might create a disruption in the institutionalized business models and their way of making money, so this might discourage the use of tracking as a mean to make a profit.

Besides to these two approaches, it is important to consider reputation and fines. We found that trackers are companies that are well-known which might care about their reputation. Therefore, this might be an incentive for them to comply with the regulation. Therefore, making public when businesses exert tracking and fail to comply with the law might aid in shaping their future behavior and others' businesses behavior. Besides, our findings confirmed that fines decrease the use of tracking by websites, so imposing as conventional approach some extra costs to businesses in case they do not comply might help to shape their incentives too.

The fact that the E-Privacy Directive will become a regulation might be only part of the solution to try to harmonize the privacy protection of member states. Controlling businesses powerful and strong incentives might require credibility in the enforcement capacity since as we observed can help to reduce tracking. Up to now, a problem has been the differences in the budget that these authorities have to face (Custers et al., 2017). Hence, although now the regulation is appointing responsibilities to the enforcement authorities and the capacity to impose fine to business that do not comply with the regulation, the application of these dispositions of the regulation might be a more difficult task in reality, even if all authorities have the same budgets. For example, enforcement can be hindered by the technical capabilities of the personnel authorities can hire. Even though fines are clearly defined, if the enforcement authorities do not have enough capacity to check for compliance, the fines might not be as effective as could be.

The task of the regulation and policy makers is not easy. Besides better understanding businesses' incentives to use tracking and implementing different approaches to persuade businesses to use less

tracking, the effort to protect privacy strengthening provisions such as consent, information requirement, fines, and guidance is necessary, which was related to our first recommendation. The European Union wants to achieve the Digital Single Market, and promoting digital businesses is one of the reasons why the harmonization of privacy protection is necessary. Hence, the tasks of regulators are to try to minimize the possibility of market failures and this might need not only shaping businesses' incentives, but also regulatory approaches. Some types of businesses can produce much tracking if they profit only from it, so this can lead to the 'tragedy of the commons' and the privacy right can be undermined. If this happens, users' might distrust to use the internet, and this can lead to a market of lemons, and then the growth from a digital economy might not be as expected. As we observed from the results, at least EU countries do have less tracking that the countries that adopted different approaches, so this result should encourage regulators to continue creating policies that protect users' privacy which in turn, in the long term, can help to promote the digital economic growth.

As some additional observations, we detected that education increase tracking with a high incident rate, but we did not want to draw big conclusions on this point because education can be related to other characteristics of the country. However, a possible explanation of this result is that education might be associated with the income of the users which makes more well-educated people an attractive target market. Also, we observed that as privacy concern increase, the level of tracking increase. A possible explanation is that as tracking increase, users become more concerned, but more studies might be necessary to confirm this. However, we always observed that businesses' incentives were more powerful predictors of tracking, so their self-interest might go beyond the characteristics of the target audience, so this just confirmed once again the importance of better understanding them.

Privacy is a complex matter, and each member states have to apply the future regulation, in different context, cultures, and facing differences in institutions. Hence, there are political decisions that can affect the development of the regulation, in fact, once again businesses might be motivated to shape the outcome of the regulation. We cannot foresee how the E-Privacy Regulation will evolve, but based in our results, we perceive that the E-Privacy Directive has helped to increase privacy protection, even though there is room for improvement, especially in understanding businesses' incentives. As Shapiro at al. (1998, p. 2) expresses "Technology changes. Economics laws do not", so we think that economic analysis to understand privacy and businesses' incentives to use tracking might be a useful tool for policy makers to understand this complex and wicked problem.

Summing up this study, we sought to provide empirical evidence for the privacy debate. The evaluation of the E-Privacy Directive already pointed out the need for harmonization, and with our empirical data, we have confirmed this necessity too since there are differences in tracking in member states. In addition, our analysis helps us to expose that consent and guidance need special attention in the future E-Privacy Regulation. Moreover, our finding emphasizes the need to better understand businesses' incentives to use tracking. Regulators have the task to change the behavior of business, especially the ones that have more incentives to use advertisement, to try to achieve the goal to reduce tracking. Conventional command and control approaches such as credible enforcement capacity accompanied by fines schemes, more informative instruments, transparency, accountability, and exposing the reputation of businesses that do not comply might help to shape these incentives. Shaping the incentives alone might not be 'the solution' since privacy is a complex problem that needs multidisciplinary approaches to be tackled. However, this might be the starting point for policy-makers to commence an important debate and re-solve the privacy problem.

5.2 Limitations and Future Research

In this section, we want to discuss the limitations of our research as well as the challenges we faced. This research was divided into three main parts the first related to the literature review and collection of the independent variables, collection of the dependent variable, and the empirical analysis.

In the first phase, a difficulty that we had to face was mainly related to the collection of the independent variables through secondary data. Secondary data is data available that was not collected for the purpose of the study. We had to deal with the fact that there was scarce and not up to date data related to the legal framework and the market forces, especially because we wanted to study many countries. Hence, we had to accept that variables such as the budget of the data protection authorities and privacy concerns were a rough proxy of what we wanted to measure. Therefore, for future research, if new data is available, these results can be compared.

In addition, another main constraint that we had to accept was the codification of the law. In our analysis, to check the impact of the provisions of the E-Privacy Directive, we used dummies variables that get a value of 0 or 1 when the countries have implemented certain provisions or not. However, with this white or black classification, we lost the grey. There might be subtle differences that we could not capture. For example, if the country applied consent we have a 1, but if this consent strict or flexible we could not say it like was the case on some countries such as UK, France, Spain, and Italy. A classification that considers all these subtle differences might need a team of legal experts to define them. Hence, we accepted to use DLA Piper classification, and still, we could see the impact of the provisions missing the impact of these differences. Another limitation related to this point is that some scholars might disagree in the classification we used from DLA Piper. Although DLA Piper is a recognized firm around the world, this comes back to the fact that the legal framework is hard to code. Besides, if using a law firm classification might raise some criticisms, we are aware that our own categorization for fines, and information required to provide to users might be even more criticized. Hence, as a suggestion for future research might be to use a different approach to classify countries provisions or if in the future any other legal firm offers a different classification, this can be used to compare these results.

Moreover, regarding the data collection of the independent variables related to the market forces, the most time-consuming tasks was to find a tool or method that allows to do the categorization of the websites. Machine learning was considered, but it would represent almost a whole master thesis just to come up with an acceptable categorization of the websites. Also, we face the challenge that datasets such as Alexa that offer categorization of the websites were not free. Hence, we used Web Shrinker as the solution. However, this brings as a limitation that the results related to the categories of websites are limited to the categorizations of Web Shrinker, which is far from perfect. For example, Science category came out in sub-question 1 as the second category exerting tracking, however, according to Web Shrinker this category includes websites of Biology, Geology, Astronomy, Physics, Space, and Weather ('Websites in IAB15 Category', 2018). Hence, websites related to weather in the Science category might not be appropriated, and this categorization might drive the results we observed. Since we were aware that it will be hard to find a perfect website categorization, unless at least three humans were coding the categories and validate that websites belong to a certain type, we accepted this limitation. However, we are aware that if a different categorization is used, some new categories might arise, or the current categories can change a bit.

An additional challenge was to determine the location of the websites. Using TLDs as a proxy for the location and the local laws the websites are following was a solution. However, we took the risk that

TLDs could be associated with the dependent variables for other reasons. However, in the end, we could confirm that the TLDs capture most of the elements related to the location of the website, even more than the country related variables. Hence, this might be a useful approach for future analysis on the performance of the websites in terms of the future regulation.

In the second stage, that was mainly in charge of Prof. Asghari, we also had some lessons learned. First, finding a sampling frame or a database from where to collect the sample required its own literature review. We learned that finding the right place to collect the sample is an important decision in which the whole study will be based on, so this decision must be taken carefully. Each dataset had his pros and cons, so we needed to think which one was suitable for our purpose. Second, a limitation of the results presented is that the data collected is only based on visiting the home page of the website. We did not do any additional click on the website. Hence, the number of Third Party Domains present might vary if a click would have been done. This represents an opportunity to compare these results with new results clicking any random place in a website.

In the last stage, the data analysis, a challenging task was to come up with a metric to measure tracking cookies. This also required its own literature review. Different scholars have used different methods to measure tracking on the web. However, no all the metrics used in the papers published use the same approach. Hence, we consider that it is important to come up with meaningful and unified metrics. In our case, we used a simple approach that was related to the law, and we count the third parties. However, Krishnamurthy (2010) expressed that since blocking third-party cookies is common, it might be that some websites are using the first-party cookies to disguise third-party cookies. Hence, we might have the risk to miss some tracking cookies if they are disguised as first party. Future work in automatic classification trackers might be useful to solve this issue. For example, making use of lists of add blockers lists or any other lists available might be easy to determine which trackers belong to what type of third party companies, so it is easy to determine if they are trackers or not. Also, the elaboration of meaningful metrics to measure tracking related to the legal framework to aid policy makers in their complex task to understand websites' performance might be a worthy task

Finally, another challenge in the data analysis was that the researcher did not have experience in analyzing big datasets as well as using R and Python for data analysis. Hence, in the beginning, this seemed an overwhelming task. We came up with some reflections to share with any future researcher that wants to embark on a quantitative analysis in which a big amount of data is involved. First, it is important to follow systematic steps, and get things right from the beginning, if not, a lot of time will be spent re-processing the data. Second, it is important to start the data analysis in stages. Using a small part of the dataset to check if the outputs are the ones we need and getting used to the dataset might facilitate to scale up the data analysis. Third, learn R and Python. A common question is which one is better to start. Python is a friendly language that is powerful and easy to use to create data frames, and R is convenient to run the statistical analysis. Hence, both will be useful in the end, the good part is that when you learn one it is not hard to learn the other.

Apart from the opportunities that are open to solve the limitations of this research, there are some questions still open to continue understanding tracking. For example, the effect of enforcement capacity on tracking. The data we used to measure enforcement was a rough proxy that can be improved through surveys or interviews to the Data Protection Authorities. Also, another opportunity can be to the study business models' incentives to use tracking on their websites. Since incentives were powerful predictors of tracking might be interesting to understand why businesses use tracking to help policy makers to make decisions on how to control them. In addition, another interesting topic

might be to understand the value chain of tracking cookies. Moreover, more tracking mechanisms can be studied in future research to have a more complete overview of tracking.

In addition, a quantitative approach was taken due to the scarcity of quantitative evidence to evaluate the outcome of the E-Privacy Directive. Hence, we consider that it is important for future research to use empirical analysis, so policy makers can base their decisions in a more objective way.

Also, we are aware that we were dealing with a complex topic, so we cannot capture the whole reality in our models, and we know that the proposed best model cannot explain it all. Also, based on the McFadden Pseudo R2, the explanation of the best model is not high nor for the other models, but we consider that in policy making small differences can produce important changes especially when a fundamental right is at stake. In addition, it is important to keep in mind that the incident rate ratios that our models predict with this data might not predict what will happen in the future since actors' incentives might change when the future E-Privacy Regulation will enter in force.

Besides the challenges we faced and the possibilities of future research, as part of our limitations, we need to reflect in the internal and external validity of the results, as well as, the reliability of the measurement instrument we used.

Internal and External Validity. Web privacy measurement face problems with internal validity since it is hard to demonstrate causality. The web is so complex, and it is a changing environment that might be the case that if a new measurement following the same steps the results might differ. Besides, since the data collection tries to simulate a natural environment in which a user visits a website, it is hard to control other confounding effects such as if websites at that moment were running more advertisement campaigns or if the simulated user was selected or not to drop a cookie. There is a trade-off between internal and external validity (Sekaran & Bougie, 2016). Hence, we strived for external validity to try to generalize these results. However, not all research finding can be generalizable to all different settings, but a defined applicability does not affect the scientific value of the findings (Sekaran & Bougie, 2016). Due to the sample we used, which was the most popular websites, the results obtained might be only generalizable for the most popular websites in EU. We think that the results might differ from other websites mainly for two reasons. First, websites that are less popular might be less attractive for the advertisement industry. Second, we think that small companies might not have the capacity to interpret the law, so other websites might show a different behavior.

In addition, we need to reflect about the construct validity. Did we measure what we want to measure? We want to measure tracking on the EU, and we used as proxy cookies. We are aware that there are other tracking mechanisms out there, but as our findings confirmed, cookies are not out fashion yet. Hence, we consider that cookies are representative of what is happening with tracking, in general, on the web. However, to be cautious, we think that these results might only be generalizable to tracking exert it through cookies.

Reliability of the Measurement. OpenWPM has been used in 23 published papers already ('Studies Using openWPM', 2018). Hence, although there is a percentage of crashes when crawling the websites, these are due to the fact that the websites do not respond. Besides, since the crawl was simulated from 18 different locations or Vantage Points (VP), it is observed that the measurement is stable. Therefore, based on the validity and the reliability of this research, we consider that the scientific rigor of these results allow the generalizability of them to the most popular websites in EU and tracking exert it through cookies.

This page was intentionally left in blank

References

<input type="hidden">. (2018). Retrieved 6 February 2018, from https://developer.mozilla.org/en-US/docs/Web/HTML/Element/input/hidden

About Us | DLA Piper Global Law Firm. (2018). Retrieved 25 June 2018, from https://www.dlapiper.com/en/netherlands/aboutus/

About us - Fieldfisher. (2018). Retrieved 25 June 2018, from https://www.fieldfisher.com/about-us

Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., & Diaz, C. (2014). The Web Never Forgets: Persistent Tracking Mechanisms in the Wild (pp. 674–689). ACM Press. https://doi.org/10.1145/2660267.2660347

Acar, M. G. C. (2017). Online Tracking Technologies and Web Privacy.

Acquisti, A. (2008). Identity management, privacy, and price discrimination. *IEEE Security & Privacy*, *6*(2).

Acquisti, A., Taylor, C. R., & Wagman, L. (2016). The economics of privacy.

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 488–500.

Alibaba Cloud Launches Malaysia City Brain to Enhance City Management. (2018, January 29). Retrieved 10 July 2018, from https://www.businesswire.com/news/home/20180128005084/en/Alibaba-Cloud-Launches-Malaysia-City-Brain-Enhance

Altaweel, I., Good, N., & Hoofnagle, C. J. (2015). Web privacy census.

Article 16. (2013). Retrieved 28 February 2018, from http://www.lisbon-treaty.org/wcm/the-lisbon-treaty/treaty-on-the-functioning-of-the-european-union-and-comments/part-1-principles/title-ii-provisions-having-general-application/158-article-16.html

Article 29 Data Protection Working Party. (2010). Opinion 2/2010 on online behavioural advertising. Retrieved 20 November 2017, from http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2010/wp171_en.pdf

Asghari, H. (2016). *Cybersecurity via Intermediaries: Analyzing Security Measurements to Understand Intermediary Incentives and Inform Public Policy*. TU Delft.

Assembly, U. G. (1948). Universal declaration of human rights. *UN General Assembly*.

Ayenson, M. D., Wambach, D. J., Soltani, A., Good, N., & Hoofnagle, C. J. (2011). Flash cookies and privacy II: Now with HTML5 and ETag respawning.

Ayenson, M., Wambach, D., Soltani, A., Good, N., & Hoofnagle, C. (2011). Flash cookies and privacy II: Now with HTML5 and ETag respawning.

Behavioral Advertising Might Not Be As Crucial As You Think. (2014). Retrieved 31 January 2018, from http://adage.com/article/datadriven-marketing/behavioral-advertising-crucial/291858/

Bellman, S., Johnson, E. J., Kobrin, S. J., & Lohse, G. L. (2004). International differences in information privacy concerns: A global survey of consumers. *The Information Society*, *20*(5), 313–324.

Bennett, C. J., & Raab, C. D. (2013). *The governance of privacy: Policy instruments in global perspective*. Routledge.

Berg, A. van den., Spithoven, A. H. G. M., & Groenewegen, J. (2009). *Institutional economics : an introduction*. Basingstoke : Palgrave Macmillan,.

Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2017). Online Behavioral Advertising: A Literature Review and Research Agenda. *Journal of Advertising*, *46*(3), 363–376. https://doi.org/10.1080/00913367.2017.1339368

Bonnardel, N., Piolat, A., & Le Bigot, L. (2011). The impact of colour on Website appeal and users' cognitive processes. *Displays*, *32*(2), 69–80.

Briefing EU Legislation in Progress. (2017). Retrieved 29 January 2018, from http://www.europarl.europa.eu/RegData/etudes/BRIE/2017/608661/EPRS_BRI(2017)608661_EN.pdf

Bujlow, T., Carela-Español, V., Solé-Pareta, J., & Barlet-Ros, P. (2015a). Web tracking: Mechanisms, implications, and defenses. *ArXiv Preprint ArXiv:1507.07872*.

Bujlow, T., Carela-Español, V., Solé-Pareta, J., & Barlet-Ros, P. (2015b). Web Tracking: Mechanisms, Implications, and Defenses. *ArXiv:1507.07872 [Cs]*. Retrieved from http://arxiv.org/abs/1507.07872

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

Castro, D., Nurko, G., & McQuinn, A. (2017). Benchmarking US Government Websites.

Cecere, G., Le Guel, F., & Soulié, N. (2015). Perceived Internet privacy concerns on social networks in Europe. *Technological Forecasting and Social Change*, *96*, 277–287. https://doi.org/10.1016/j.techfore.2015.01.021

Clifford, D. (2014). EU Data Protection Law and Targeted Advertising: Consent and the Cookie Monster - Tracking the crumbs of online user behaviour. *JIPITEC*, *5*(3). Retrieved from http://www.jipitec.eu/issues/jipitec-5-3-2014/4095

Cofone, I. N. (2017). Privacy Tradeoffs in Information Technology Law. Retrieved from https://d-nb.info/1124591397/34

Commission Nationale de l'Informatique et des Libertés. (2015). Example of a cookie banner | CNIL. Retrieved 25 July 2018, from https://www.cnil.fr/fr/exemple-de-bandeau-cookie

Cookie collective. (2014). Five Models for Cookie Law Consent. Retrieved 27 September 2017, from https://www.cookielaw.org/media/105101/five-models-for-cookie-law-consent.pdf

Cookies - European commission. (2016). Retrieved 24 September 2017, from http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm

Cookies and other (illegal) recipes to track internet-users: latest episode of the Facebook saga. (2018). Retrieved 21 May 2018, from https://www.law.kuleuven.be/citip/blog/cookies-and-other-illegal-recipes-to-track-internet-users-latest-episode-of-the-facebook-saga/

Country code top-level domain - ICANNWiki. (2017). Retrieved 2 July 2018, from https://icannwiki.org/Country_code_top-level_domain

Custers, B., Dechesne, F., Sears, A. M., Tani, T., & van der Hof, S. (2017). A comparison of data protection legislation and policies across the EU. *Computer Law & Security Review*.

Damen, J., Köhler, L., & Woodard, S. (2017). The Human Right of Privacy in the Digital Age.

Data management platform, DMP | Adobe Audience Manager. (2018). Retrieved 25 June 2018, from https://www.adobe.com/la/data-analytics-cloud/audience-manager.html?origref=https%3A%2F%2Fwww.google.nl%2F

Data Protection Eurobarometer Factsheet. (2015). Retrieved 1 February 2018, from http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_eurobarometer_240615_en.pdf

Deloitte. (2017a). Evaluation and review of Directive 2002/58 on privacy and the electronic communication sector. Retrieved from https://ec.europa.eu/digital-single-market/en/news/evaluation-and-review-directive-200258-privacy-and-electronic-communication-sector

Deloitte. (2017b). Evaluation and review of Directive 2002/58 on privacy and the electronic communication sector.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Pub. L. No. 31995L0046, OJ L 281 (1995). Retrieved from http://data.europa.eu/eli/dir/1995/46/oj/eng

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic

communications sector (Directive on privacy and electronic communications), Pub. L. No. 32002L0058, OJ L 201 (2002). Retrieved from http://data.europa.eu/eli/dir/2002/58/oj/eng

Directive 2009/136/EC of the European Parliament and of the Council of 25 November 2009 amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services, Directive 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector and Regulation (EC) No 2006/2004 on cooperation between national authorities responsible for the enforcement of consumer protection laws (Text with EEA relevance), Pub. L. No. 32009L0136, OJ L 337 (2009). Retrieved from http://data.europa.eu/eli/dir/2009/136/oj/eng

DLA Piper. (2014). EU Law on Cookies. Retrieved 26 November 2017, from https://iapp.org/media/pdf/resource_center/DLA_EU_cookie_implementation_9-14.pdf

Eijk, R. van. (2017). On Browser Settings, Cookies, and (Not) Being Tracked by Digital Advertisements (Presentation Slides).

Englehardt, S. (2016). *The Web Privacy Problem is a Transparency Problem*. Presented at the FTC Privacy Conference, Washintong, D.C. Retrieved from https://senglehardt.com/presentations/2016_01_ftc_the_web_privacy.pdf

Englehardt, S., Eubank, C., Zimmerman, P., Reisman, D., & Narayanan, A. (2014). Web privacy measurement: Scientific principles, engineering platform, and new results. *Manuscript Posted at Http://Randomwalker. Info/Publications/WebPrivacyMeasurement. Pdf*.

Englehardt, S., Han, J., & Narayanan, A. (2018). I never signed up for this! Privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies*, *2018*(1), 109–126.

Englehardt, S., & Narayanan, A. (2016a). Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1388–1401). ACM.

Englehardt, S., & Narayanan, A. (2016b). Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1388–1401). ACM.

Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., & Felten, E. W. (2015). Cookies that give you away: The surveillance implications of web tracking (pp. 289–299). Presented at the Proceedings of the 24th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee.

ePrivacy: consultations show confidentiality of communications and the challenge of new technologies are key questions. (2016). Retrieved 2 February 2018, from https://ec.europa.eu/digital-single-market/en/news/eprivacy-consultations-show-confidentiality-communications-and-challenge-new-technologies-are

Eubank, C., Melara, M., Perez-Botero, D., & Narayanan, A. (2013). Shining the floodlights on mobile web tracking-a privacy survey (Vol. 2). Presented at the Proceedings of the IEEE Workshop on Web.

EUR-Lex - 12012P/TXT - EN - EUR-Lex. (2012). Retrieved 17 January 2018, from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT

EUR-Lex - 31997L0066 - EN. (1998). [text/html; charset=UNICODE-1-1-UTF-8]. Retrieved 1 March 2018, from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31997L0066:EN:HTML

EUR-Lex - 32002L0058 - EN - EUR-Lex. (2002). Retrieved 24 September 2017, from http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32002L0058

European Commission, Directorate-General for the Information Society and Media, Time.lex, & Spark. (2015). *ePrivacy directive, assessment of transposition, effectiveness and copmatibility with the proposed data protection regulation final report.* Brussels: European Commission. Retrieved from http://bookshop.europa.eu/uri?target=EUB:NOTICE:KK0415268:EN:HTML

European Convention on Human Rights - Official texts, Convention and Protocols. (2010). Retrieved 27 February 2018, from http://www.echr.coe.int/pages/home.aspx?p=basictexts

Europe's Cookie Laws: ePrivacy Directive Implementation Center | IAB Europe. (2018). Retrieved 27 September 2017, from https://www.iabeurope.eu/eucookielaws/

Executive Summary of the Ex-post REFIT evaluation of the ePrivacy Directive. (2017).

Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*, *53*(3), 225–241.

Falahrastegar, M., Haddadi, H., Uhlig, S., & Mortier, R. (2016). Tracking personal identifiers across the web (pp. 30–41). Presented at the International Conference on Passive and Active Network Measurement, Springer.

Felten, E. W., & Schneider, M. A. (2000). Timing attacks on web privacy (pp. 25–32). Presented at the Proceedings of the 7th ACM conference on Computer and communications security, ACM.

Fieldfisher. (2015). Cookie 'consent' rule: EEA implementation. Retrieved 26 November 2017, from http://www.fieldfisher.com/media/2927368/EU-Cookie-Consent-Tracking-Table-Fieldfisher-21-April-2015.pdf

Flaticon, the largest database of free vector icons. (2018). Retrieved 6 January 2018, from https://www.flaticon.com/

France | IAB Europe. (2018). Retrieved 26 November 2017, from https://www.iabeurope.eu/eucookielaws/fr/

Frequently asked questions about Dutch cookie act. (2016). Retrieved 2 February 2018, from https://www.acm.nl/sites/default/files/old_publication/publicaties/11917_veelgestelde-vragen-cookiebepaling-oktober-2016-engels-new.pdf

Fruchter, N., Miao, H., Stevenson, S., & Balebako, R. (2015). Variations in tracking in relation to geographic location. *ArXiv Preprint ArXiv:1506.04103*.

Gartner Says 8.4 Billion Connected. (2017). Retrieved 2 March 2018, from https://www.gartner.com/newsroom/id/3598917

Gemalto. (2018). Data Breach Statistics by Year, Industry, More. Retrieved 19 February 2018, from http://breachlevelindex.com

Gibbs, S. (2014, April 15). Gmail does scan all emails, new Google terms clarify. Retrieved 19 January 2018, from http://www.theguardian.com/technology/2014/apr/15/gmail-scans-all-emails-new-google-terms-clarify

Goldfeder, S., Kalodner, H., Reisman, D., & Narayanan, A. (2017). When the cookie meets the blockchain: Privacy risks of web payments via cryptocurrencies. *ArXiv:1708.04748 [Cs]*. Retrieved from http://arxiv.org/abs/1708.04748

Google. (2018). Managing multi-regional and multilingual sites - Search Console Help. Retrieved 16 July 2018, from https://support.google.com/webmasters/answer/182192?hl=en

Guthery, F. S., Burnham, K. P., & Anderson, D. R. (2003). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. *The Journal of Wildlife Management*, *67*(3), 655. https://doi.org/10.2307/3802723

Hagiu, A., & Jullien, B. (2011). Why do intermediaries divert search? *The RAND Journal of Economics*, *42*(2), 337–362.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1998). *Multivariate data analysis* (Vol. 5). Prentice hall Upper Saddle River, NJ.

Hardin, G. (1968). The Tragedy of the Commons by Garrett Hardin - The Garrett Hardin Society - Articles. Retrieved 24 January 2018, from http://www.garretthardinsociety.org/articles/art_tragedy_of_the_commons.html

Here are some of the Russian Facebook ads meant to divide the US and promote Trump. (2017, November 2). Retrieved 19 February 2018, from http://www.businessinsider.com/russian-facebook-ads-2016-election-trump-clinton-bernie-2017-11

Hirshleifer, J. (1979). Privacy: Its origin, function, and future. Retrieved 13 December 2017, from http://www.econ.ucla.edu/workingpapers/wp166.pdf

Hodgson, G. M. (2000). What is the essence of institutional economics? *Journal of Economic Issues*, *34*(2), 317–329.

How Netflix Knows Exactly What You Want to Watch. (2016). Retrieved 22 February 2018, from https://www.makeuseof.com/tag/how-netflix-knows-exactly-what-you-want-to-watch/

HTTP State Management Mechanism. (2000). Retrieved 6 January 2018, from http://www.ietf.org/rfc/rfc2965.txt

IAB Europe. (2015). Internet cookies increasing and enhancing your internet surfing experience. Retrieved 24 June 2018, from https://www.iabeurope.eu/wp-content/uploads/2015/11/Internet_Cookies_-_Increasing_and_enhancing_your_internet_surfing_experience.pdf

IHS Markit. (2017). The Economic Contribution of Digital Advertising in Europe. Retrieved 9 December 2017, from https://www.iabeurope.eu/wp-content/uploads/2017/09/DigitalAdvertisingEconomicContribution_FINAL.pdf

Information Commissioners' Office. (2012). Guidance on the rules on use of cookies and similar technologies. Retrieved 25 July 2018, from https://ico.org.uk/media/for-organisations/documents/1545/cookies_guidance.pdf

International Association of Privacy Professionals (IAPP). (2011). Data Protection Authorities 2011 Global Survey. Retrieved 8 June 2018, from https://iapp.org/media/pdf/knowledge_center/DPA11_Survey_final.pdf

Internet access and use statistics - households and individuals - Statistics Explained. (2016). Retrieved 5 February 2018, from http://ec.europa.eu/eurostat/statistics-explained/index.php/Internet_access_and_use_statistics_-_households_and_individuals

'It Knows When I Got Out of the Car!': Tucker's Special Report on How Google's Tracking You. (2018, February 8). [Text.Article]. Retrieved 12 February 2018, from

http://insider.foxnews.com/2018/02/07/google-tracking-you-tucker-carlsons-report-silicon-valley-surveillance-capitalism

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360.

Kelly, N. (2015, September 7). The Most Common Mistakes Companies Make with Global Marketing. Retrieved 18 July 2018, from https://hbr.org/2015/09/the-most-common-mistakes-companies-make-with-global-marketing

Kirchhoff, S. (2011). The U.S. Newspaper Industry in Transition. In J. Detrani (Ed.), *Journalism* (pp. 270–293). Apple Academic Press. https://doi.org/10.1201/b13161-13

Kladnik, D. (2018). I don't care about cookies. Retrieved 16 July 2018, from https://www.i-dont-care-about-cookies.eu/

Koninkrijksrelaties, M. van B. Z. en. (2016). Telecommunicatiewet [wet]. Retrieved 22 January 2018, from http://wetten.overheid.nl/BWBR0009950/2016-01-01#Hoofdstuk6a

Krishnamurthy, B. (2010). I know what you will do next summer. *ACM SIGCOMM Computer Communication Review*, *40*(5), 65–70.

Krishnamurthy, B., & Wills, C. (2009a). Privacy diffusion on the web: a longitudinal perspective (pp. 541–550). Presented at the Proceedings of the 18th international conference on World wide web, ACM.

Krishnamurthy, B., & Wills, C. E. (2009b). On the leakage of personally identifiable information via online social networks (pp. 7–12). Presented at the Proceedings of the 2nd ACM workshop on Online social networks, ACM.

Kristol, D. M. (2001). HTTP Cookies: Standards, privacy, and politics. *ACM Transactions on Internet Technology (TOIT)*, *1*(2), 151–198.

Leenes, R., & Kosta, E. (2015a). Taming the cookie monster with Dutch law – A tale of regulatory failure. *Computer Law & Security Review*, *31*(3), 317–335. https://doi.org/10.1016/j.clsr.2015.01.004

Leenes, R., & Kosta, E. (2015b). Taming the cookie monster with Dutch law – A tale of regulatory failure. *Computer Law & Security Review*, *31*(3), 317–335. https://doi.org/10.1016/j.clsr.2015.01.004

Lerner, A., Simpson, A. K., Kohno, T., & Roesner, F. (2016). Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016. In *USENIX Security Symposium*.

Liu, Y. (2014). User control of personal information concerning mobile-app: Notice and consent? *Computer Law & Security Review*, *30*(5), 521–529.

Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés - Article 32 (2011).

magic cookie. (2003). Retrieved 8 January 2018, from http://www.catb.org/~esr/jargon/html/M/magic-cookie.html

Mayer, J. R., & Mitchell, J. C. (2012a). Third-party web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on* (pp. 413–427). IEEE.

Mayer, J. R., & Mitchell, J. C. (2012b). Third-party web tracking: Policy and technology (pp. 413–427). Presented at the Security and Privacy (SP), 2012 IEEE Symposium on, IEEE.

Mazzetti, M., & Schmidt, M. S. (2013, June 9). Edward Snowden, Ex-C.I.A. Worker, Says He Disclosed U.S. Surveillance. *The New York Times*. Retrieved from https://www.nytimes.com/2013/06/10/us/former-cia-worker-says-he-leaked-surveillance-data.html

McDonald, A. M., & Cranor, L. F. (2008). The cost of reading privacy policies. *ISJLP*, *4*, 543.

McDonald, A. M., & Cranor, L. F. (2011). A survey of the use of adobe flash local shared objects to respawn http cookies. *ISJLP*, *7*, 639.

Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2012). Detecting price and search discrimination on the internet (pp. 79–84). Presented at the Proceedings of the 11th ACM Workshop on Hot Topics in Networks, acm.

Milberg, S. J., Burke, S. J., Smith, H. J., & Kallman, E. A. (1995). Values, personal information privacy, and regulatory approaches. *Communications of the ACM*, *38*(12), 65–74.

Milberg, S. J., Smith, H. J., & Burke, S. J. (2000). Information privacy: Corporate management and national regulation. *Organization Science*, *11*(1), 35–57.

Moore, T. (2010). The economics of cybersecurity: Principles and policy options. *International Journal of Critical Infrastructure Protection*, *3*(3–4), 103–117.

Morgan, B., & Yeung, K. (2007). *An introduction to law and regulation: Text and materials*. Cambridge University Press.

Moz. (2018). Moz - International SEO. Retrieved 16 July 2018, from https://moz.com/learn/seo/international-seo

Naafs, S. (2018). 'Living laboratories': the Dutch cities amassing data on oblivious residents. Retrieved 13 March 2018, from http://www.theguardian.com/cities/2018/mar/01/smart-cities-data-privacy-eindhoven-utrecht

Narayanan, A., & Reisman, D. (2017a). The Princeton Web Transparency and Accountability Project. In *Transparent Data Mining for Big and Small Data* (pp. 45–67). Springer.

Narayanan, A., & Reisman, D. (2017b). The princeton web transparency and accountability project. In *Transparent Data Mining for Big and Small Data* (pp. 45–67). Springer.

Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., & Vigna, G. (2013). Cookieless monster: Exploring the ecosystem of web-based device fingerprinting (pp. 541–555). Presented at the Security and privacy (SP), 2013 IEEE symposium on, IEEE.

North, D. C. (1991). Institutions. *Journal of Economic Perspectives*, *5*(1), 97–112.

Odlyzko, A. (2003). Privacy, economics, and price discrimination on the Internet. In *Proceedings of the 5th international conference on Electronic commerce* (pp. 355–366). ACM.

Office of the Australian Information Commissioner. (2014). APP quick reference tool - Office of the Australian Information Commissioner (OAIC). Retrieved 11 July 2018, from /agencies-and-organisations/guides/app-quick-reference-tool

Office of the Privacy Commissioner of Canada. (2015, December 24). Privacy Toolkit for Businesses. Retrieved 11 July 2018, from https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda-compliance-help/guide_org/

Oldhoff, E. (2013). *Ambiguities in the revised "cookie law" and its national implementations*. Retrieved from http://arno.uvt.nl/show.cgi?fid=128802

Olejnik, L., Englehardt, S., & Narayanan, A. (2017). Battery Status Not Included: Assessing Privacy in Web Standards.

O'Neil, D. (2001). Analysis of Internet users' level of online privacy concerns. *Social Science Computer Review, 19*(1), 17–31.

OpenWPM: A web privacy measurement framework. (2017). Python, CITP. Retrieved from https://github.com/citp/OpenWPM (Original work published 2014)

OpenWPM: A web privacy measurement framework. (2018). Python, CITP. Retrieved from https://github.com/citp/OpenWPM (Original work published 2014)

Organisation for Economic Co-operation and Development. (2003a). *Privacy online : OECD guidance on policy and practice.* Paris : OECD,. Retrieved from Table of contents http://catdir.loc.gov/catdir/toc/fy043/2004371604.html

Organisation for Economic Co-operation and Development. (2003b). *Privacy online : OECD guidance on policy and practice.* Paris : OECD,. Retrieved from Table of contents http://catdir.loc.gov/catdir/toc/fy043/2004371604.html

Personal Information Protection Commission Japan. (2017). Amended Act on the Protection of Personal Information (Tentative Translation). Retrieved 5 March 2018, from https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf

Privacy Shield Program Overview | Privacy Shield. (2016). Retrieved 13 January 2018, from https://www.privacyshield.gov/Program-Overview

Proposal for a Regulation on Privacy and Electronic Communications. (2017). Retrieved 24 July 2018, from https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications

Proposed ePrivacy Regulation: An Internet without data flows? | IAB Europe. (2017). Retrieved 30 January 2018, from https://www.iabeurope.eu/blog/proposed-eprivacy-regulation-an-internet-without-data-flows/

Python Data Analysis Library — pandas: Python Data Analysis Library. (2018). Retrieved 12 March 2018, from https://pandas.pydata.org/

Rise of the data protection officer, the hottest tech ticket in town. (2018, February 14). *Reuters*. Retrieved from https://www.reuters.com/article/us-cyber-gdpr-dpo/rise-of-the-data-protection-officer-the-hottest-tech-ticket-in-town-idUSKCN1FY1MY

Rittel, H. W., & Webber, M. M. (1973a). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169.

Rittel, H. W., & Webber, M. M. (1973b). Dilemmas in a general theory of planning. *Policy Sciences*, *4*(2), 155–169.

Roesner, F., Kohno, T., & Wetherall, D. (2012). Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (pp. 12–12). USENIX Association.

Samat, S., Acquisti, A., Gross, R., & Pe'er, E. (2013). Visceral Targeting: Using Personalized Face Composites for Implicit Targeted Marketing. Presented at the presentation at 32nd Ann. Advertising and Consumer Psychology Conf.—Consumer Psychology in a Social Media World.

Scheitle, Q., Jelten, J., Hohlfeld, O., Ciprian, L., & Carle, G. (2018). Structure and Stability of Internet Top Lists. *ArXiv Preprint ArXiv:1802.02651*.

Sekaran, U., & Bougie, R. (2016). *Research methods for business: A skill building approach*. John Wiley & Sons.

Selenium - Web Browser Automation. (2018). Retrieved 12 March 2018, from https://www.seleniumhq.org/

Shapiro, C., Carl, S., & Varian, H. R. (1998). *Information rules: a strategic guide to the network economy*. Harvard Business Press.

Smit, E. G., Van Noort, G., & Voorveld, H. A. M. (2014). Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in Europe. *Computers in Human Behavior*, *32*, 15–22. https://doi.org/10.1016/j.chb.2013.11.008

Soltani, A., Canty, S., Mayo, Q., Thomas, L., & Hoofnagle, C. J. (2010). Flash Cookies and Privacy. (Vol. 2010, pp. 158–163). Presented at the AAAI spring symposium: intelligent information privacy management.

Studies Using openWPM. (2018). Retrieved 28 January 2018, from https://webtransparency.cs.princeton.edu/webcensus/index.html#Users

The AMA Gold Report 2017 Top 50 Market Research Firms. (2017). Retrieved 14 March 2018, from https://www.ama.org/publications/MarketingNews/Pages/the-ama-gold-report-2017-top-50-market-research-firms.aspx

The Federal Council. (1992). CC 235.1 Federal Act of 19 June 1992 on Data Protection (FADP). Retrieved 5 March 2018, from https://www.admin.ch/opc/en/classified-compilation/19920153/index.html

The new EU ePrivacy Regulation: what you need to know. (2016). Retrieved 24 June 2018, from https://www.i-scoop.eu/gdpr/eu-eprivacy-regulation/

The office for creative research. (2013). Behind The Banner. Retrieved 29 January 2018, from http://o-c-r.org/behindthebanner/

the Rubicon Project. (2018). Retrieved 25 June 2018, from http://rubicon.koalition.com

Trevisan, M., Traverso, S., Metwalley, H., & Mellia, M. (2017). Uncovering the Flop of the EU Cookie Law. *ArXiv:1705.08884 [Cs]*. Retrieved from http://arxiv.org/abs/1705.08884

Tufekci, Z. (2018, January 30). Opinion | The Latest Data Privacy Debacle. *The New York Times*. Retrieved from https://www.nytimes.com/2018/01/30/opinion/strava-privacy.html

United Nations. (2016). Human Development Data (1990-2015) | Human Development Reports. Retrieved 11 July 2018, from http://hdr.undp.org/en/data

van Eijk, N., Helberger, N., Kool, L., van der Plas, A., & van der Sloot, B. (2012). Online tracking: questioning the power of informed consent. *Info*, *14*(5), 57–73. https://doi.org/10.1108/14636691211256304

Van Eijk, N., Helberger, N., Kool, L., van der Plas, A., & van der Sloot, B. (2012). Online tracking: Questioning the power of informed consent. *Info*, *14*(5), 57–73.

Vollmer, N. (2017, December 16). Table of contents EU General Data Protection Regulation (EU-GDPR) [text]. Retrieved 14 February 2018, from http://www.privacy-regulation.eu/en/

Warbrick, C. (1989). Federal Aspects of the European Convention on Human Rights. *Mich. j. Int'l L.*, *10*, 698.

Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review*, 193–220.

Weather Stickers® | Free Weather Sticker | Weather Underground. (2018). Retrieved 13 March 2018, from https://www.wunderground.com/stickers/?MR=1

Weber, T. E. (2000, March 6). Web Marketers Take Techie's Cookie Recipe And Run With It. *Tribunedigital-Chicagotribune*. Retrieved from http://articles.chicagotribune.com/2000-03-06/business/0003060043_1_web-sites-cookies-on-line

Websites in IAB15 Category. (2018). Retrieved 25 June 2018, from https://www.webshrinker.com/tech-demo/categories/iab15/

WebTAP Princeton University. (2018). OpenWPM platform. Retrieved 7 March 2018, from https://webtap.princeton.edu/research/

Wefers Bettink, W., Van Eijk, R., & Wagner, F. (2012). *Strictly Speaking: Cookies, Consent and Compliance. Europe Data Protection congress. [Presentation]. Brussels: IAPP.*

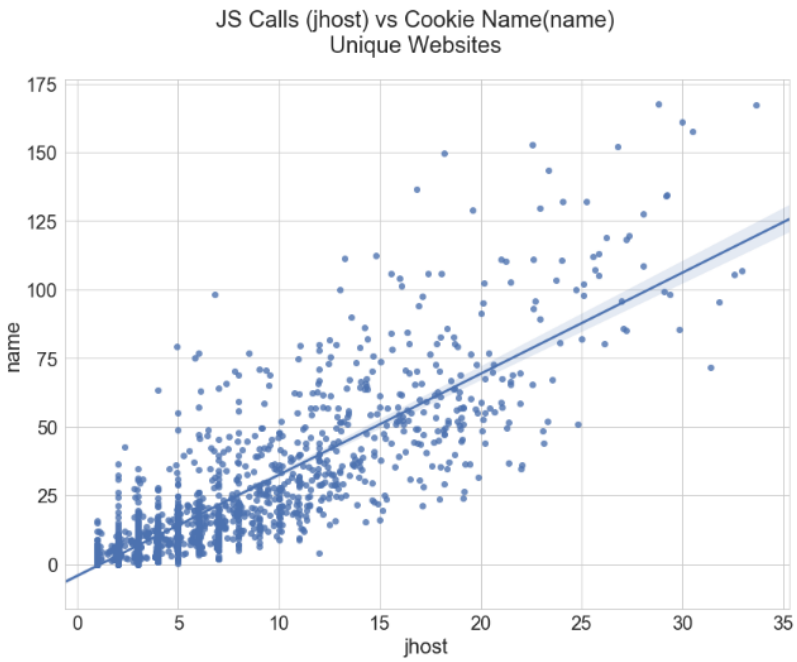
Yuan, S., Abidin, A. Z., Sloan, M., & Wang, J. (2012). Internet advertising: An interplay among advertisers, online publishers, ad exchanges and web users. *ArXiv Preprint ArXiv:1206.1754*.

Zuiderveen Borgesius, F. J. (2014). Improving privacy protection in the area of behavioural targeting. Retrieved 6 March 2018, from https://pure.uva.nl/ws/files/2141314/154442_Thesis_complete_.pdf

This page was intentionally left in blank

Appendix

Appendix A: Unique counted cookies names vs Java Script calls



SpearmanrResult(correlation=0.85396492093549514, pvalue=0.0)

Appendix B: Regressions using TLD as proxy of local laws, keeping TLD .COM with US location

Table 34: Negative Binomial Regression Model TLDs as proxy for the local law with Incident Rate Ratios keeping Websites visited from US with TLD .COM and ORG

Variable (TLD-Proxy of local law)	Coefficient (Std Err)	Incident Rate Ratio(95% CI)
TLD_AU	0.795(0.204)***	2.2140606 (1.48-3.30)
TLD_BE	0.340(0.203)*	1.4051344 (0.94- 2.09)
TLD_CA	0.504(0.206)**	1.6555846 (1.10- 2.48)
TLD_CH	0.451(0.202)**	1.5697921 (1.05-2.33)
TLD_COM-US	1.578(0.175)***	4.8464077 (3.42-1 6.79)
TLD_CZ	0.474(0.203)**	1.6060628 (1.07- 2.39)
TLD_DE	0.921(0.203)***	2.5124538 (1.68- 3.74)
TLD_ES	0.517(0.213)**	1.6776198 (1.10- 2.55)
TLD_FR	1.126(0.204)***	3.0823298 (2.06- 4.60)
TLD_GR	-0.313(0.212)	0.7314843 (0.48- 1.11)
TLD_HU	-0.265(0.212)	0.7671509 (0.50- 1.16)
TLD_IT	0.536(0.210)**	1.7098150 (1.13- 2.58)
TLD_JP	0.872(0.231)***	2.3924127 (1.52- 3.79)
TLD_NL	-0.396(0.211)*	0.6729669 (0.44- 1.01)
TLD_ORG	-0.029(0.177)	0.9714545 (0.68- 1.36)
TLD_PL	0.262(0.227)	1.2998386 (0.83- 2.04)
TLD_PT	0.368(0.207)*	1.4442033 (0.96- 2.17)
TLD_RO	0.300(0.206)	1.3496497 (0.90- 2.02)
TLD_SE	0.285(0.205)	1.3302895 (0.88- 1.99)
TLD_UK	1.133(0.201)***	3.1059438 (2.090- 4.61)
TLD_US	-0.187(0.259)	0.8294549 (0.50- 1.39)
(Intercept)	1.322(0.146)***	3.7497821 (2.84- 5.04)

Note: *p<0.1; **p<0.05; ***p<0.01. Null deviance/residual: 2601.4/ 2246.5 - McFadden Pseudo R^2:0.02

Appendix C: Impact of the E-Privacy Directive provisions in banner presence

Results banners - Consent

	Dependent variable:	
	banner	
	(1)	(2)
Opt_in_required_Yes(Consent)	0.965***	1.077***
	(0.111)	(0.116)
Websites Categories	No	Yes
Constant	-1.193***	-0.190
	(0.088)	(0.289)
Observations	1,634	1,634
Log Likelihood	-1,017.342	-950.587
Akaike Inf. Crit.	2,038.684	1,955.174

Note: *p<0.1; **p<0.05; ***p<0.01, Websites in countries that require explicit consent increases the use of banners by 192% when controlled for businesses' incentives to use tracking

Results banners - Guidance

	Dependent variable:	
	banner	
	(1)	(2)
Guidance_Yes	1.160***	1.178***
	(0.112)	(0.118)
Websites Categories	No	Yes
Constant	-1.014***	-0.030
	(0.067)	(0.286)
Observations	1,634	1,634
Log Likelihood	-1,003.003	-945.757
Akaike Inf. Crit.	2,010.006	1,945.513

Note: *p<0.1; **p<0.05; ***p<0.01 Websites in countries that emitted guidance increase the use of banners by 222% when controlled for businesses' incentives to use tracking

Results banners - Fines

	Dependent variable:	
	banner	
	(1)	(2)
Fines_Yes	-0.124	-0.011
	(0.104)	(0.109)
Websites Categories	No	Yes
Constant	-0.556***	0.469*
	(0.073)	(0.277)
Observations	1,634	1,634
Log Likelihood	-1,056.692	-996.168
Akaike Inf. Crit.	2,117.383	2,046.336

Note: *p<0.1; **p<0.05; ***p<0.01 / The provision was not significant for banners' presence.

Results Banners

	Dependent variable:	
	Banner	
	(1)	(2)
Info_Low	1.903***	1.963***
	(0.190)	(0.194)
Info_High	1.888***	2.094***
	(0.197)	(0.203)
Websites categories	No	Yes
Constant	-2.209***	-1.232***
	(0.176)	(0.333)
Observations	1,634	1,634
Log Likelihood	-981.928	-914.993
Akaike Inf. Crit.	1,969.857	1,885.986
Mc Fadden Pseudo R2	0.0713	0.1346

Note: *p<0.1; **p<0.05; ***p<0.01 / Websites in countries that require to provide to users more information have more banner presence when controlling for businesses' incentives to use tracking.

Appendix D: Comparison of all regression models including coefficients of websites categories

Model comparison

	Dependent variable:									
	host									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
TLDau	0.795*** (0.200)		1.299*** (0.180)							
TLDbe	0.340* (0.200)		0.324* (0.180)							
TLDca	0.504** (0.203)		0.836*** (0.182)							
TLDch	0.451** (0.199)		0.453** (0.179)							
TLDcz	0.474** (0.199)		0.536*** (0.180)							
TLDde	0.921*** (0.200)		0.920*** (0.179)							
TLDes	0.517** (0.209)		0.589*** (0.189)							
TLDfr	1.126*** (0.201)		1.088*** (0.180)							
TLDgr	-0.313 (0.209)		-0.170 (0.193)							
TLDhu	-0.265 (0.209)		-0.125 (0.192)							
TLDit	0.536*** (0.206)		0.519*** (0.186)							
TLDjp	0.872*** (0.227)		1.033*** (0.201)							
TLDnl	-0.396* (0.208)		-0.098 (0.190)							
TLDpl	0.262 (0.223)		0.571*** (0.201)							
TLDpt	0.368* (0.203)		0.445** (0.190)							
TLDro	0.300 (0.203)		0.378** (0.185)							
TLDse	0.285 (0.202)		0.448** (0.182)							
TLDuk	1.133*** (0.198)		1.203*** (0.178)							
TLDus	-0.187 (0.255)		0.226 (0.230)							
cat_1IAB10 (Home Garden)		-0.846** (0.336)		-0.813** (0.325)	-0.856** (0.336)	-0.821** (0.332)	-0.807** (0.336)	-0.777** (0.333)	-0.826** (0.331)	-0.824** (0.330)
cat_1IAB11 (Law", Government, Politics)		-1.392*** (0.225)		-1.546*** (0.216)	-1.379*** (0.224)	-1.425*** (0.222)	-1.384*** (0.224)	-1.400*** (0.222)	-1.441*** (0.220)	-1.470*** (0.219)
cat_1IAB12 (News / Weather / Information)		0.391** (0.171)		0.428*** (0.162)	0.387** (0.171)	0.426** (0.169)	0.403** (0.171)	0.439*** (0.169)	0.434*** (0.167)	0.465*** (0.166)
cat_1IAB13 (Personal Finance)		-0.171 (0.244)		-0.111 (0.231)	-0.152 (0.244)	-0.091 (0.241)	-0.152 (0.244)	-0.054 (0.240)	-0.090 (0.238)	-0.115 (0.236)
cat_1IAB14 (Society)		-0.403 (0.295)		-0.106 (0.280)	-0.408 (0.295)	-0.380 (0.291)	-0.375 (0.295)	-0.344 (0.291)	-0.278 (0.287)	-0.184 (0.285)
cat_1IAB15 (Science)		-0.827** (0.410)		-0.698* (0.390)	-0.828** (0.409)	-0.873** (0.408)	-0.833** (0.411)	-0.876** (0.407)	-0.969** (0.403)	-0.724* (0.398)
cat_1IAB16 (Pets)		-2.460*** (0.829)		-2.067*** (0.801)	-2.517*** (0.836)	-2.205*** (0.822)	-2.431*** (0.822)	-2.236*** (0.820)	-2.296*** (0.831)	-2.154*** (0.823)
cat_1IAB17 (Sports)		-0.068 (0.233)		0.144 (0.223)	-0.061 (0.233)	0.051 (0.230)	-0.045 (0.234)	0.080 (0.230)	0.121 (0.228)	0.143 (0.226)
cat_1IAB18 (Style Fashion)		-0.485 (0.345)		-0.481 (0.329)	-0.451 (0.344)	-0.468 (0.341)	-0.462 (0.345)	-0.408 (0.341)	-0.455 (0.338)	-0.450 (0.336)
cat_1IAB19 (Technology Computing)		-0.616*** (0.176)		-0.567*** (0.167)	-0.635*** (0.176)	-0.629*** (0.173)	-0.589*** (0.176)	-0.608*** (0.174)	-0.621*** (0.172)	-0.573*** (0.171)
cat_1IAB2 (Automotive)		0.467 (0.363)		0.271 (0.345)	0.452 (0.362)	0.434 (0.359)	0.471 (0.363)	0.432 (0.359)	0.402 (0.356)	0.427 (0.352)
cat_1IAB20 (Travel)		-0.246 (0.214)		-0.416** (0.204)	-0.245 (0.214)	-0.343 (0.212)	-0.241 (0.214)	-0.327 (0.212)	-0.381* (0.210)	-0.408** (0.208)
cat_1IAB21 (Real Estate)		-0.557 (0.403)		-0.938** (0.395)	-0.589 (0.403)	-0.705* (0.400)	-0.536 (0.404)	-0.689* (0.400)	-0.771* (0.400)	-0.872** (0.401)
cat_1IAB22 (Shopping)		-0.138 (0.244)		-0.206 (0.232)	-0.155 (0.244)	-0.139 (0.241)	-0.132 (0.244)	-0.145 (0.241)	-0.212 (0.239)	-0.186 (0.237)

cat_1IAB23 (Religion Spirituality)	-2.732*** (0.680)	-2.496*** (0.672)	-2.685*** (0.678)	-2.514*** (0.671)	-2.672*** (0.681)	-2.421*** (0.670)			-2.379*** (0.674)	-2.489***(0.677)
cat_1IAB24 (Uncategorized)	-1.851*** (0.227)	-1.581*** (0.223)	-1.858*** (0.227)	-1.659*** (0.224)	-1.807*** (0.229)	-1.623*** (0.225)			-1.578*** (0.225)	-1.539*** (0.224)
cat_1IAB25 (Non-Standard Content)	-1.249*** (0.201)	-1.091*** (0.193)	-1.270*** (0.201)	-1.234*** (0.199)	-1.218*** (0.201)	-1.211*** (0.198)			-1.189*** (0.197)	-1.189*** (0.196)
cat_1IAB26 (Illegal Content)	-2.620*** (0.863)	-2.635*** (0.854)	-2.647*** (0.864)	-2.427*** (0.845)	-2.561*** (0.863)	-2.389*** (0.847)			-2.506*** (0.861)	-2.453*** (0.858)
cat_1IAB3 (Business)	-0.715*** (0.253)	-0.748*** (0.243)	-0.727*** (0.252)	-0.759*** (0.250)	-0.705*** (0.252)	-0.751*** (0.250)			-0.815*** (0.248)	-0.783*** (0.246)
cat_1IAB4 (Careers)	-0.729** (0.369)	-0.925*** (0.351)	-0.754** (0.368)	-0.712* (0.364)	-0.703* (0.369)	-0.699* (0.364)			-0.731** (0.361)	-0.773** (0.358)
cat_1IAB5 (Education)	-1.091*** (0.177)	-1.173*** (0.170)	-1.120*** (0.177)	-1.111*** (0.174)	-1.065*** (0.177)	-1.099*** (0.175)			-1.118*** (0.173)	-1.133*** (0.172)
cat_1IAB6 (Family Parenting)	-0.735 (0.533)	-0.949* (0.526)	-0.715 (0.531)	-0.830 (0.531)	-0.745 (0.535)	-0.813 (0.531)			-0.893* (0.533)	-0.964* (0.535)
cat_1IAB7 (Health Fitness)	-2.058*** (0.354)	-1.995*** (0.350)	-2.065*** (0.355)	-2.055*** (0.354)	-2.024*** (0.354)	-2.015*** (0.354)			-2.017*** (0.352)	-2.091*** (0.356)
cat_1IAB8 (Food Drink)	-1.072*** (0.322)	-0.798*** (0.309)	-1.100*** (0.322)	-0.968*** (0.317)	-1.091*** (0.322)	-1.027*** (0.318)			-0.990*** (0.315)	-0.998*** (0.314)
cat_1IAB9 (Hobbies Interests)	-0.551** (0.238)	-0.543** (0.227)	-0.604** (0.238)	-0.481** (0.235)	-0.517** (0.238)	-0.484** (0.235)			-0.536** (0.233)	-0.510** (0.232)
Opt_in_required_Yes			-0.157** (0.062)			-0.124* (0.071)	0.125 (0.099)	-0.044 (0.074)	0.080 (0.083)	
Fines_Yes				-0.438*** (0.062)		-0.396*** (0.066)	-0.305*** (0.079)	-0.344*** (0.070)	-0.345*** (0.080)	
Guidance_Yes					0.121* (0.067)	0.153** (0.072)	0.177** (0.083)	0.128* (0.075)	0.175** (0.089)	
Info_High							-0.312*** (0.111)			
Normalize_budget_DPA_2011							-0.007 (0.005)	-0.005 (0.004)	-0.009** (0.005)	
worried_use_pd_cecere							0.022*** (0.005)	0.012*** (0.004)	0.022*** (0.005)	
HDI_edu_index							2.792*** (0.917)	4.314*** (0.824)	2.701** (1.201)	
EU_Yes									-0.283** (0.131)	
GDP_per_capita_2016									-0.00001*** (0.00000)	
Internet_Fq_Use_2016									0.008 (0.005)	
Rule_law_2016									0.020*** (0.005)	
Constant	1.322*** (0.143)	2.172*** (0.158)	1.543*** (0.198)	2.269*** (0.163)	2.355*** (0.159)	2.112*** (0.161)	2.337*** (0.163)	-1.553* (0.881)	-2.056** (0.812)	-2.747*** (0.977)
Observations	1,634	1,634	1,634	1,634	1,634	1,634	1,634	1,634	1,634	1,634
Log Likelihood	-4,521.757	-4,394.661	-4,299.938	-4,391.586	-4,371.197	-4,393.034	-4,368.736	-4,567.054	-4,352.937	-4,336.854
theta	0.631*** (0.025)	0.760*** (0.032)	0.880*** (0.039)	0.763*** (0.032)	0.788*** (0.034)	0.762*** (0.032)	0.790*** (0.034)	0.591*** (0.023)	0.810*** (0.035)	0.829*** (0.036)
Akaike Inf. Crit.	9,083.515	8,841.322	8,689.877	8,837.171	8,796.394	8,840.068	8,795.472	9,150.108	8,769.874	8,745.708

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix E: Correlation coefficients Independent variables

	PRIVACY CONCERNS	EDUCATION INDEX	GDP PER CAPITA 2016	INTERNET FQ USE 2016	RULE LAW 2016	NORMALIZED BUDGET DPA 2011	EU_YES	FINES_YES	GUIDANCE_ YES	OPT_IN_ REQUIERED _YES
PRIVACY CONCERNS	1									
EDUCATION INDEX	0.2619	1								
GDP PER CAPITA 2016	0.3871	0.6256	1							
INTERENET FQ USE 2016	0.5668	0.6419	0.5806	1						
RULE LAW 2016	-0.5215	0.5279	0.7725	0.7284	1					
NORMALIZED BUDGET DPA 2011	0.1034	-0.0851	-0.2813	0.0379	-0.1537	1				
EU_YES	0.3120	-0.5144	-0.5844	-0.3168	-0.4050	0.3203	1			
FINES_YES	0.1979	-0.3301	-0.0691	-0.2068	-0.0197	-0.1455	0.4121	1		
GUIDANCE_YES	0.1439	0.0551	0.0484	0.2074	0.0028	0.29065	0.3601	-0.0027	1	
OPT_IN_REQUIRED_YES	0.2853	-0.2292	-0.3079	-0.2729	-0.2844	0.24028	0.6031	0.34320	0.3521	1