Warehouse of Information:

Amazon's Data Collection Practices and their Relation to the GDPR

Dimitri Koehorst

Student ID: 10563970

Master's Thesis

Graduate School of Humanities

New Media and Digital Culture

Supervisor: Melis Bas

Second Reader: Niels van Doorn

August 31st, 2020

Abstract

In recent times, data has become increasingly central to a variety of different companies. While the use of data has become widespread, there are some companies whose entire business model revolves around the use of data. One such company is Amazon. Initially it was merely an online bookstore, but as the company grew it incorporated multiple new branches, such as Amazon Web Services, which allow the company to collect data from a variety of different sources. Companies such as Amazon use this data to optimize their services, which allows them to gain certain advantages over their competitors. However, this usage of data is bound by international regulations, one of which is the GDPR, the new data protection legislation of the European Union. By using data collected from the Amazon.com webstore as a case study, this thesis investigates the shift of companies towards a data-oriented business model, and investigates certain problems that this shift brings. This is done through the research question: *How can we conceptualize the data collection practices of Amazon in relation to the General Data Protection Regulation?*

Table of Contents

Chapter 1: Introduction

1.1) Introduction to data-oriented companies

In the twenty-first century, due to changes in digital technologies, data has become increasingly central to firms and their relations with workers, customers, and other capitalists. (Srnicek, 2017) While data usage and collection through surveillance has a longer history, there are business models that support the commoditization of data in a way that transverses the economic, political, and societal dimensions of technology. (West, 2017) Data collection is not limited to online behavior but is instead present in a rapidly growing amount of industries, ranging from insurance companies to sports apparel companies, and from smart home speakers to Bluetooth-enabled toothbrushes. (Zuboff, 2016) However, some companies dedicate their entire business model to the collection and analysis of data. They offer a wide variety of services, each of which serving as its own data collection tool.

The most prominent examples of such companies are tech giants like Google and Facebook. As Srnicek points out, while Google started originally as an online search engine, the company has since then massively expanded into many different industries. This includes the hosting of online web services, artificial intelligence research, internet access infrastructure, and many others. This is not an example of vertical integration, which occurs when companies integrate different stages of production within the same industry. (Investopedia, 2020) Rather, this is an example of what Srnicek calls rhizomatic integration. This refers to the integration of different industries in which similar tools for data extraction can be used as in pre-existing components of these companies. (Srnicek, 2017)

The sudden profitability of Google was a direct result of this rhizomatic integration. Instead of merely feeding the data back into the same system in order to optimize search results, this data was instead used to link advertisers to various websites they may want to advertise on. It became apparent that this decision to feed the analytical capacity of behavioral data into the challenge of increasing an

advertisement's relevance to users, has marked a significant shift in Google's operation of a company. This is considered by many as a historic turning point. (Zuboff, 2016)

The focus of this research is however not on Google, but on Amazon, one of their direct competitors. Amazon, similar to Google, started as a company focused on one central business practice. In the case of Amazon this business was the online sale of books. (Stone, 2013) As Amazon started expanding their business, they started integrating the usage of data for the optimization of sales. At first, Amazon expanded their online bookstore to incorporate more products as time went on, then they allowed independent merchants to sell their products through the Amazon web store. To some, this marks the shift into Amazon as a platform. (Srnicek, 2017) Then Amazon kept expanding into other industries. Most notably, with the introduction of Amazon Web Services, Amazon allowed other companies to outsource their software and hardware needs, making them in a certain sense a platform for platforms. (Srnicek, 2017)

By using such data data, Amazon and other tech giants have achieved massive commercial growth by turning this data into products and services. (Moore, 2016) This usage of data even allows these companies to hold monopolies in certain industries. For example, in 2015, Amazon had a 95 percent share of the e-book market in the UK. (Moore, 2016) While it is important to talk about this economic power and this ability to dominate particular markets, Amazon and other tech giants also have a significant amount of civic power. As Rebecca MacKinnon points out, Amazon and other tech giants are so powerful because they not only create and sell products, but also provide the digital spaces upon which citizens increasingly depend. (MacKinnon, 2012; Moore, 2016) By controlling these digital spaces they are gaining some of the civic power that has traditionally been held by large media organizations. This includes the power to communicate news and information, and the power to command public attention. (Moore, 2016)

Numerous other people have also expressed their concerns about the amount of power that Amazon and other companies gain from their data collection practices. In our contemporary capitalist society, companies are expected to have continuous growth. While data collection facilitates this growth for

companies, it also presents problems and concerns for the users. The surveillance tools that are being used to collect data are intruding into an increasing number of facets of our society, to the point that some critics are clamoring the 'death of privacy'. (Preston, 2014) Others go even further than this. In her 2016 essay 'The Secrets of Surveillance Capitalism', Shoshana Zuboff writes:

"The assault on behavioral data is so sweeping that it can no longer be circumscribed by the concept of privacy and its contests. This is a different kind of challenge now, one that threatens the existential and political canon of the modern liberal order defined by principles of self-determination that have been centuries, even millennia, in the making. I am thinking of matters that include [...] the sanctity of the individual and the ideals of social equality; [...] norms and rules of collective agreement; the functions of market democracy; the political integrity of societies; and the future of democratic sovereignty." (Zuboff, 2016, par. 3)

While the list of arising problems is too long for the scope of this research, the quote above offers insight into different aspects of our society that are being assaulted by different companies' data collection practices.

Since Amazon is such a large company that operates within a variety of industries, this thesis limits itself by focusing on two core services that Amazon uses for data collection. The first of these is the Amazon.com web store, where they offer a wide variety of products that will ship to almost anywhere in the world. The second of these is Amazon Web Services, an online platform provided by Amazon, which allows for other companies to outsource certain digital needs to Amazon. By collecting data from the web store, this thesis looks at the types of data that Amazon collects and the ways in which it might be used. Additionally, by analyzing this data, this thesis also investigates whether there is a connection between the data collected by Amazon Web Services, and the search results on the web store. This research is done through the lens of user privacy. Specifically, this thesis perceives the European General Data Protection Regulation, or GDPR, as a framework to analyze Amazon's data collection practices. The collection of data from the Amazon web store allows this research to investigate closely whether or not Amazon is compliant with the GDPR, as well as what the

implications are if Amazon complies with the GDPR or not.

This is done through the research question: *How can we conceptualize the data collection practices of Amazon in relation to the General Data Protection Regulation?*

In this introduction, this research goes into different aspects of Amazon and its business practices. It gives an overview of how Amazon started as a company, and in what ways they have since expanded. While there are a lot of different aspects of Amazon as a business, this thesis focuses specifically on the Amazon.com web store, and on Amazon Web Services. This section also explains the ways in which the web store and Amazon Web Services are related, and why this relation is important for this research. It answers the sub-question: *What are Amazon's business practices and what role does data play in these practices?*

After the introduction, the second chapter of this thesis is dedicated to literature review which problematizes Amazon's status as a platform. It draws on multiple aspects of platform studies within the field of Media Studies, including questions regarding privacy of platform-centric data collection and usage. It further explores three models of modern business practices within the field of online data collection. First, Sarah Myers West's Data Capitalism model is explained. It describes data capitalism, as "a system in which the commoditization of our data enables an asymmetric redistribution of power that is weighted toward the actors who have access and the capability to make sense of the information." (West, 2017, pp.23) Second, Nick Scrinek's Platform Capitalism model is explored since it gives a historic and economistic account of the rise of the platform as a data collection tool. According to Scrinek, this is necessitated among other reasons by the decline of profitability in the once-dominant manufacturing sector. (Scrinek, 2017) The third model that is explored is Shoshana Zuboff's Surveillance Capitalism model. Zuboff focuses on the intrusive nature of surveillance as a means of private data collection, and calls for the importance of increased regulation of the usage of data as a behavioral surplus. (Zuboff, 2016) This finally leads to an analysis of the GDPR as a means to regulate this data usage within the EU. This section answers the sub-question: *What implications arise from analyzing Amazon as a data-oriented capitalist organization?*

In the third chapter, the methodology of the research is introduced and theorized. This methodology is based upon previous work done by Tracking Exposed. They are a research group that works together with independent researchers as well as universities, whose mission statement is 'to put a spotlight on users' tracking, profiling, on the data market and on the influence of algorithms'. (Tracking Exposed, 2019) During the 2020 Digital Methods Winter School Data Sprint, the Tracking Exposed team worked with myself and other students from the University of Amsterdam, as well as some other universities, in a project where data from Amazon was being scraped and analyzed. Their methodology will be used for the collection of data from the Amazon.com web store. This section will then go over qualitative content analysis, and why it was chosen as the method of analysis for this research.

In chapter 4, a qualitative content analysis is employed on the data that was collected from Amazon in the previous chapter. It describes the data through different categories that are defined in this chapter, whereby the focus lies on finding evidence that points to whether or not Amazon is compliant with different articles of the GDPR. By looking for anomalies in the collected data, this chapter investigates whether using Amazon Web Services has any clear effect on the search results of the web store. Additionally, this chapter looks at similarities and patterns in the data, which can be seen as evidence that Amazon creates data profiles of its users without asking their permission, using a technique called device fingerprinting. This chapter seeks to answer the question: *To what extent can Amazon be shown to comply with the GDPR's data collection regulations through analysis of the online web store?*

In chapter 5, based on the findings of the previous chapter, this thesis will argue that the current limitations being put on platform holders are insufficient, and will further problematize the increasing need for rapidly changing platform regulations. The argument being made in this chapter is that the intrusive nature of capitalist data collection can not be reconciled with concepts of privacy and data protection within our current legislative system. It will draw on the example of a similar case made against rival tech giant Google to apply this analysis to the platform business model at large.

This is done to answer the question: *To what extent does compliance with GDPR influence the future of data oriented capitalist organizations?* Finally, chapter 6 summarizes the conclusions made within the different chapters of the thesis, answers the main research question, and offers relevant avenues for further research.

1.2) Introduction to Amazon

Amazon is an American based company founded by Jeff Bezos in 1994. Originally it was just an online bookstore, but it has since extended its business practices much wider. Amazon quickly expanded its business to include a wider variety of products, and shortly after established services in other countries outside of the United States, including Germany and the United Kingdom. Later, they introduced different services alongside its online web store including Amazon Web Services, or AWS. AWS can be described as a platform for platforms, a service which allows businesses to outsource the hosting of their online services to Amazon. These services initially included data storage for sellers within its platform and an API, or Application Programming Interface, that can run individually built applications. AWS expanded further eventually to include cloud computing. This means the on-demand availability of data storage and computing power for users, which is being provided by various data centers around the world. They provide cloud computing services to a variety of companies, as of today including Spotify, Reddit, NASA, Ubisoft, Netflix, and many more. (Contino, 2020)

Despite Amazon's massive international success, the company has been subject to a number of controversies. As an example, Amazon was selling illegal dog fighting magazines on its web store, for which they were later sued. (Humane Society of the United States, 2010) Additionally, Amazon garnered controversy by selling counterfeit products on its web store, which has resulted in certain companies such as Birkenstock and Nike to pull their products from the platform. (Shepard, 2018) More worryingly, there have been numerous cases of Amazon violating its workers' rights, including unsafe working conditions, opposing the formation of unions, and enforcing unreasonable performance standards to its workers. (Gruendelsberger, 2019)

What this research focuses on, however, is the way in which Amazon handles the data that it receives through consumers using their online services, both the Amazon.com web store and Amazon Web Services. An example of a way in which they gain such data is the usage of Amazon Echo and similar devices. These allow the user to use their voice to interact with the device, which is directly

connected to various Amazon services including its web store. In 2019 it was revealed that Amazon keeps recordings of any command given to its devices by the user, and stores this data indefinitely until the user requests for it to be deleted. (Ng, 2019) Additionally, transcripts of these voice recordings are stored even after the recordings themselves have been deleted. (Ng, 2019)

There are many different sources that Amazon collects consumer data from, since the company operates around the world in many different industries. This research focuses on the Amazon Web Store on Amazon.com as it operates in different countries within the European Union. This thesis argues that Amazon uses the data collected from people using the Amazon.com web store and AWS for the purpose of optimizing their services for profits. These include a personalized recommendation system, real-time price optimization, its patented anticipatory shipping model where products which are expected to be purchased are sent to distribution warehouses in advance of the sale being made in order to optimize efficiency, and more. (Wills, 2020) The following subchapter explains how the Amazon web store can be viewed as a platform, and why this distinction is important.

1.3) From pipe to platform

Amazon started out with a very linear business model wherein they sold books online through their website. Gradually it changed its business to incorporate more and more aspects of platforms. On his website Pipes to Platforms, economist Sangeet Paul Choudary created a timeline that highlights three broad properties of a platform and shows how Amazon incorporated more of these properties as time went on.

This section will first explain the three properties that Choudary describes. Then, this section will explain the steps that Amazon has taken to acquire these three properties. As a result of acquiring these properties, Amazon has moved from being a 'pipe' to being a platform.

The first of these three properties he defines is that of a magnet: a platform should draw in both producers and consumers.

The second property is that of the toolbox: a platform should provide all the necessary tools for producers and consumers to interact with each other.

The final property is that of the matchmaker: 'A platform needs to match producers and consumers, leveraging data.' (Choudary, 2015)

Choudary describes three different steps that Amazon took that resulted in them becoming a platform rather than a pipe. To describe Amazon's transformation over time, Choudary first defines what he calls a 'pipe': a linear model in which products go into one end of the pipe and come out at the other. During this time, Amazon simply concerned itself with sourcing products, managing inventory and selling the down the pipe, Amazon.com, within which they were the sole producer of value. (Choudary, 2015)

Step 1 - Introduction of user reviews

The first step to acting like a platform came with the introduction of individual user reviews. As Choudary writes, this allows users to create value in the form of reviews, thereby fulfilling the role of producer themselves. In doing so, Amazon takes on the role of being a magnet for producers and consumers alike, which is a key property of platforms according to Choudary as described earlier.

(Choudary, 2015)

Step 2 - Using data to improve services

The next step Amazon took from being a 'pipe' to being a platform is their introduction of the recommended products feature. By using data from both users and producers, Amazon functions as a matchmaker between the two, and as more users began utilizing this feature, the algorithms used to facilitate it became more and more accurate, which led to this feature becoming more prevalent as a reason for consumers to use Amazon's web store over its competitors. By introducing this feature, Amazon fulfills the role of matchmaker, which is another key property for platforms that Choudary described earlier. Choudary writes:

"Unlike pipes, platforms are intelligent. Also, platforms exhibit network effects of data. The more the number of users using a system, the more valuable the system becomes for every individual user because of the usage data it collects." (Pipes to Platforms, par. 4) He adds that this kind of network effect of data is, especially at the time, completely absent in 'traditional pipes in the offline world', i.e. brick-and-mortar stores. (Choudary, 2015)

Step 3 - Opening the marketplace

The feature that then definitely came to define Amazon as a platform is the introduction of the Amazon Marketplace, which allows external merchants to sell their products on the Amazon website. By opening their service to individual retailers and allowing them to use every aspect of their underlying infrastructure, Amazon now fulfills the three key properties of platforms that Choudary describes, since they now also have the property of the toolbox. With this final step, Amazon moved away from being a company that could be described with Choudary's pipe model, to one describable by his platform model.

Amazon then took this further by introducing new features such as the Amazon Affiliate program which allows producers to use Amazon as an advertising platform, even rewarding them with a share of the revenue through this program (Choudary, 2015). Another crucial feature of Amazon that it shares with many other platforms is the release of their API which allows developers to extend

the functionality of the platform. Choudary then explains their usage of Kindle as a platform with the introduction of the Amazon App Store, where they sell applications made by third-party developers. While Amazon sells its e-reader Kindle at a considerable loss (Zurb, 2011), it is the increased amount of data that they collect from Kindle users that more than makes up for this loss.

While these are all important ways in which the Amazon web store operates as a platform, there is another platform that Amazon operates, on a much larger scale. The next section will go into the creation of Amazon Web Services, how these services expanded over time, and how data collection became increasingly important for these services.

1.4) Amazon Web Services

The Amazon.com web store is already a massive online platform in its own right. As of 2019, Amazon is the largest e-commerce retailer by online revenue in the world. (Angelovska, 2019) However, a much bigger platform which is also owned by Amazon, and one that is in fact the main source of the company's income, is Amazon Web Services. Initially, Amazon Web Services, or AWS for short, referred to the programmable aspects of the Amazon web store itself, serving users a collection of application programming interfaces (APIs) and tools that allow them to interact with various parameters within the website. (Miller, 2016) In 2006 Amazon issued the following press statement upon the full launch of its Web Services:

> "Amazon Web Services today announced "Amazon S3(TM)," a simple storage service that offers software developers a highly scalable, reliable, and low-latency data storage infrastructure at very low costs. Amazon S3 is available today at http://aws.amazon.com/s3.
>
> Amazon S3 is storage for the Internet. It's designed to make web-scale computing easier for developers. Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites [sic]. The service aims to maximize benefits of scale and to pass those benefits on to developers." (Businesswire.com, "Amazon Web Services Launches", 2006)

Since the introduction of this service it has expanded its scale to the point that main industry leaders are using Amazon Web Services to host their online data, ranging from mere web pages to massive databases. Currently Amazon is the world industry leader with its AWS cloud-computing

services A recent report shows they account for 47% of the total market, with Microsoft's Azure as the next closest competitor at 22% market share. (Stalcup, 2019)

As mentioned, many other platforms, including Airbnb, Slack, Uber, and others, are using AWS to host their services online. As Scrinek points out in his book 'Platform Capitalism', platforms like AWS are oriented towards building and owning the basic infrastructures necessary to collect, analyse, and deploy data for other companies to use. (Scrinek, 2017) Their respective data collection practices are each being facilitated by AWS. This gives Amazon even more data in the process, which may then be further used to optimize Amazon Web Services to allow Amazon and its client companies to collect even more data.

What is important for this thesis, is the fact that AWS is the most ubiquitous cloud-computing service currently available. This gives them not only a large amount of economic power, but also civic power, as described in the introduction of this thesis. As pointed out in this chapter, their services are used even by some of their direct competitors. To investigate what kind of data Amazon stores and analyzes from all businesses that use AWS would be worth researching entirely by itself. This thesis focuses however on the data that Amazon collects from the clients of these businesses, not the businesses themselves. More precisely, one of the things this thesis seeks to investigate is the extent to which Amazon collects data from individuals who use websites or platforms that run on AWS. The way that Amazon collects this data will be explained in the upcoming section.

1.5) Usage of data

One of the types of data that Amazon tracks from users is the clickstream, the 'digital breadcrumb trail' left by users as they visit different websites and multiple pages within those websites. An example of a clickstream is given by Wang et al. in their 2017 article 'Clickstream User Behavior Models'. In this paper, Wang et al. propose the clickstream model as a solution of regulating user behavior on online services that are driven by 'users and user generated content'. (Wang et al., 2017) According to Wang et al., clickstreams are 'timestamped server-side traces of click events, generated by users during their web browsing sessions or interactions with mobile apps.' They further describe how different clickstream analytics can have different functions: their proposed system uses different systems for the detection and interpretation of human behavior. (Wang et al., 2017) There are certain advantages and disadvantages to using clickstreams for data collection. Advantages include identifying user groups that share similar clickstream activities; inferring user interests; predicting future user behaviors; and furthering the design and operation of online services. (Wang et al., 2017) Disadvantages include many of these clickstream models becoming 'black boxes' that focus on specific tasks, while offering little explanation on how users behave and why. Additionally, for these clickstreams to work properly, they either need constant supervision, or large amounts of data and fine-tuned parameters. (Wang et al., 2017)

Through requests for personal data that individuals have done to Amazon, we know that Amazon collects and uses the clickstream data of its users. In a 2018 news article, Riccardo Coluccini reports about his request to Amazon for all of his personal data. Initially, he only received very basic information that was already available through his personal account panel on Amazon. (Coluccini, 2018) However, earlier that same year, German politician Katharina Nocun published a similar story in which she requested more data from Amazon that they initially provided. After 90 days her request was processed, and among all of the data she received, the clickstream and all data it contains was included. Nocun writes that each click contains up to 50 additional details, including: the time, article number and category, the pages that were accessed before and after Amazon, whether she added

something to the shopping cart or performed a search, the web address from which she accessed Amazon, how many milliseconds her browser needed to load the page, language settings, device settings, and the country she was based in. (Nocun, 2018)

Nocun's research provided the knowledge that not only does Amazon collect this clickstream of its users, but also that it collects data outside of Amazon's own website. There are however still some limitations. First, while we have evidence from Nocun's research that Amazon registers which sites users visit before and after they visit Amazon, we do not know how this information is being processed or used. As described earlier, we know that Amazon uses data to improve its services. However, we do not know the ways in which Amazon achieves this. We also cannot know the scope of this data usage, since all of this happens within the inner workings of the website in a way that is invisible to the user.

Second, there is no evidence that suggests that Amazon also collects a similar clickstream on websites that run on AWS, or whether or not these websites all fall under this same clickstream. If this were the case, it would mean that Amazon uses the data that it collects from all of its AWS affiliates for the improvement of its own services. This fact would raise new questions about user privacy entirely, since the scope of AWS is so large that users might be completely unable to opt out of their data being collected by Amazon. Since we can not know this for sure based on current evidence, part of the analysis conducted in chapters 3 and 4 of this thesis is designed in a way that could show support to the idea that Amazon uses the data from its AWS affiliates to improve their web store.

The next chapter will explain some theoretical frameworks through which Amazon's data collection practices can be analyzed.

Chapter 2: Literature Review

This chapter explains the various theoretical backgrounds and frameworks that this research is based upon. The argument made in this chapter is that Amazon is a capitalist platform whose data collection practices are a threat to user privacy. First this chapter starts with a definition of contemporary capitalism, and will also explain why Amazon needs to be analyzed within this definition. Then, three similar frameworks for data-centric capitalist practices are analyzed and compared. They each highlight different aspects of Amazon's business practices, and together these different frameworks synthesize into a data-oriented capitalist model that explains these different business practices. Each offers its respective insights into Amazon's data collection, and based on this, this research proposes a hybridization of the three frameworks provided, since they each complement each other's shortcomings. Finally, this chapter ends with a concise analysis of the GDPR. It starts with an explanation of why compliance with the GDPR is essential for securing the privacy of the user, and in which ways GDPR differs from its predecessors. Then, this chapter highlights the articles which are important for the methodology of this research. This chapter answers the sub-question: *How has Amazon been theorized as a platform, and why is this problematization necessary to investigate its data usage?*

2.1) Capitalism

To arrive at our classification of Amazon as a data-oriented capitalist organization, we must define precisely what we mean by a capitalist organization, and what it means when such an organization or company is specifically data-oriented. Capitalism is a system that is focused on the creation of wealth 'through advancing continuously to ever higher levels of productivity and technological sophistication.' (Gilpin, 2006, pp. 3) Amazon is a prime example of this practice: it started as an online book retailer, and through Amazon Web Services it innovated their creation of wealth by offering a new service. They further innovated this by using data to reinforce their existing services, such as personalized advertising on the web store.

Amazon is not the only company that relies upon online services to continuously reinvigorate this capitalist production of wealth. Already in 2005, author Nigel Thrift wrote of the 'new economy' that relies upon information and communications technology (ICT) in order to keep growing. According to Thrift: "Nowadays the idea of the new economy has been stabilized; it consists of strong non-inflationary growth arising out of the rising influence of information and communications technology and the associated restructuring of economic activity." (Thrift, 2005, pp. 122) He argued that ICT created a new kind of market economy, which was facilitated in part by ICT's capability of rapid technological changes, facilitated by constant technological critique. (Thrift, 2005) The technological critique Thrift mentions here refers to the potential improvements to a service that arise from the analysis of data, which is a form of critique that is technological in nature. In the next three sections, this thesis will detail three different frameworks in which we can understand data collection and analysis as a capitalist practice.

2.2) Data capitalism

In her 2017 article titled ‘Data Capitalism: Redefining the Logics of Surveillance and Privacy’, Sarah Myers West offers a history of how the advent of commercial surveillance in the form of data collection, became centered around a logic of data capitalism. (West, 2017) West describes this logic of data capitalism as ‘a system in which the commoditization of our data enables an asymmetric redistribution of power that is weighted toward the actors who have access and the capability to make sense of [data].’ (West, 2017, pp. 23) West argues this through the notion that communication and information are historically a key source of power, as posited by Manuel Castells. (West, 2017; Castells, 2007) She relates this to historical efforts to quantify human behavior, such as the use of ‘political arithmetic’ in late 17th century England, an effort to seek a better understanding of everyday life. (West, 2017; Herbst, 1993) These early forms of data collection evolved into surveillance networks, for example as a means of evaluating and monitoring the credit of American businesses. (West, 2017)

While these historical examples are both cases of political and monetary value being assigned to the collection of data, their scope was inhibited by the inability of technologies at the time to retain and make sense of it. (West, 2017) While new technologies were developed over time that improved and eventually even automated this collection of data, West argues that ‘the introduction of internet commerce brought with it a new scope and scale of tracking that proved transformative for data collection practices.’ (West, 2017, pp. 25) Though initially the lack of profitability of early dotcom-companies was compensated by massive venture capital investments, the early 2000s marked a shift towards companies such as Google. These companies were able to monetize their massive amounts of data through construction of different services. Google was able to monetize data through the introduction of Adwords. Adwords was the end result of an effort of Google to finance its online business by translating the data Google collected from across the web into content-targeted advertising, branded as Google AdWords.’ (West, 2017)

West concludes her analysis of data capitalism by highlighting three narratives of technological utopianism that serve to convince customers to overcome their concerns about privacy. The first of these narratives is the value of the free and open network. The idea is that users should be willing to participate in data capitalist practices since the value that they gain from participating in a free and open network outweighs the value of their personal data. The second narrative is the potential to make customers' internet experience personal. Using data for personalized advertising benefits the companies as established earlier, but it should also help the user since they are more likely to be interested in products selected specifically for them. The final narrative regards the technocratic value placed on data and its potential to augment consumer power. (West, 2017)

The essence of the argument that West makes is the fact that the usage of data to improve a company's services, has led to the facilitation of the use of surveillance itself as a business model. The distinction here is that the data collection does not simply improve the quality of the services a company offers, but that the quality of the data collection is also improved through this process. West uses the term data capitalism to draw attention to the fact that data is both the means to improve the quality of services, and the service which is most valuable for the company itself to keep improving and growing.

While West alludes to the importance of user privacy within the framework she provides, the framework by itself is ultimately insufficient. It fails to consider the central role of platforms in this shift in scope. This shift was, according to her, merely brought forward by the introduction of internet commerce, focusing too heavily on advertising platforms like Google. Where West's framework focuses on the redistribution of power, the next section's framework instead emphasizes the problematization of surveillance itself as a business model.

2.3) Surveillance capitalism

In her 2016 article titled ‘The Secrets of Surveillance Capitalism’, Shoshana Zuboff discusses what she describes as ‘a wholly new genus of capitalism, a systemic coherent new logic of accumulation we should call surveillance capitalism.’ (Zuboff, 2016, par. 4) The first example she uses is a quote from an auto insurance industry consultant:

> “Most Americans realize that there are two groups of people who are monitored regularly as they move about the country. The first group is monitored involuntarily by a court order requiring that a tracking device be attached to their ankle. The second group includes everyone else.”
>
> (Zuboff, 2016, par. 1)

This quote was intended by this consultant as a defense of the ‘astonishingly intrusive surveillance capabilities of the allegedly benign systems that are already in use or under development.’ (Zuboff, 2016, par. 1) Zuboff uses this quote to argue that data collection practices through the use of surveillance are being integrated into a wider and wider variety of industries, and that this data is used to ‘change people’s actual behavior at scale’. Zuboff describes this as an assault on behavioral data, and claims that ‘is so sweeping that it can no longer be circumscribed by the concept of privacy and its contests’. (Zuboff, 2016, par. 3)

Zuboff describes constant surveillance of the consumer for monetary benefit as surveillance capitalism. According to Zuboff, surveillance capitalism is a ‘novel economic mutation bred from the clandestine coupling of the vast powers of the digital with the radical indifference and intrinsic narcissism of [the] financial capitalism and its neoliberal vision that have dominated commerce for at least three decades. [...] It is an unprecedented market form that roots and flourishes in lawless space’. (Zuboff, 2016, par. 4) She points to a quote of Google Chairperson Eric Schmidt’s book: ‘the online world is not truly bound by terrestrial laws… it’s the world’s largest ungoverned space’. With the quote above, Zuboff shows how the consistent lack of proper legislation is what led us to this intrusive form of surveillance that seemingly penetrates every aspect of our daily lives.

One of the most important aspects of Zuboff's work is the distinction she makes between privacy and secrecy. Rather than opposites, Zuboff argues that they are moments in a sequence; secrecy is an effect; privacy is the cause, and therefore, privacy rights are decision rights. (Zuboff, 2016) According to Zuboff then, surveillance capitalism does not completely remove these decision rights, but rather concentrates them within the surveillance regime as being maintained by these large private companies. Zuboff writes:

> "Surveillance capitalism reaches beyond the conventional institutional terrain of the private firm. It accumulates not only surveillance assets and capital, but also rights. This unilateral redistribution of rights sustains a privately administered compliance regime of rewards and punishments that is largely free from detection or sanction." (Zuboff, 2016, par. 16)

In summary, Zuboff argues for new interventions that regulate the extraction and application of user data, as well as the use of this data as free raw material, and the monetization of the results of these operations. (Zuboff, 2016) Zuboff goes so far to say that the only thing that can alter surveillance capitalism's claim to 'manifest data destiny' is nothing short of a social revolt. Through Zuboff's analysis of surveillance capitalism that she gives in this article, she indicates that it might already be too late to change its course through normal means such as increased regulations. Instead something needs to change about the operations of these companies themselves. Zuboff writes that it becomes clear that demanding privacy from surveillance capitalists or lobbying to end commercial surveillance on the internet is 'like asking Henry Ford to make each Model T by hand.' (Zuboff, 2016)

While Zuboff presents a very clear argument about the dangers of surveillance capitalism and the importance of new regulations and their enforcement, like West, she does not attribute the dangers of these companies to the platform business model. Instead this next section will introduce a framework that does precisely that.

2.4) Platform capitalism

In his 2017 book titled ‘Platform Capitalism’, Nick Srnicek provides a historical and economically-focused account of the development of capitalist practices that eventually led to the conception of what he calls platform capitalism. In the first part of his book, Srnicek defines capitalism as marked by ‘generalized market dependency that ensures a systemic imperative to reduce production costs in relation to prices for goods and services, which requires the constant optimization of labor processes and productivity through technological innovation.’ (Van Doorn, 2018, pp. 104; Srnicek, 2017, pp. 11) Next, he describes platform capitalism as a form of capitalism that emerged around the business model of the platform as ‘an efficient way to monopolise, extract, analyse, and use the increasingly large amounts of data that are being recorded’.(Van Doorn, 2018, pp. 104; Srnicek, 2017, pp. 11) This focus on data is crucial, as he argues that ‘data have come to serve a number of key capitalist functions.’ (Srnicek, 2017). According to Srnicek:

> “[Data] educate and give competitive advantage to algorithms; they enable the coordination and outsourcing of workers; they allow for the optimisation and flexibility of productive processes; they make possible the transformation of low-margin goods into high-margin services; and data analysis is in itself generative of data, in a vicious cycle. Given the significant advantages of recording and using data and the competitive pressures of capitalism, it was perhaps inevitable that this raw material would come to represent a vast new resource to be extracted from.” (Srnicek, 2017, pp. 16)

Srnicek describes platforms as ‘digital infrastructures that enable two or more groups to interact. Platforms therefore position themselves as intermediaries that bring together different users: customers, advertisers, service providers, producers, suppliers, and even physical objects.’ (Srnicek, 2017, pp. 17) As the intermediary between different groups, platforms have privileged access to the data that results from the interactions between these groups. They are far more than internet companies or tech companies, since they can operate anywhere, wherever digital interaction takes

place. (Srnicek, 2017) They also rely upon the so-called ‘network effect’: the more users a platform has, the more valuable that platform becomes to everyone else.

Additionally, he describes five different types of platforms in order to give an overview of the emerging platform landscape. The first type is advertising platforms, including Google and Facebook, which extract information from users and repurpose it to sell ad space. Facebook, for example, will recommend certain promoted pages and other advertisements based on the user’s behavior within their platform. The second type is cloud platforms, including AWS and Salesforce, which ‘own the hardware and software of digital-dependent businesses and are renting them out as needed.’ (Srnicek, 2017) The third type is industrial platforms such as General Electric and Siemens, which are similar to cloud platforms but instead created with the intent of transforming traditional manufacturing ‘into internet-connected processes that lower the costs of production and transform goods into services.’ The fourth type is product platforms such as Spotify, which generate value by turning a traditional good such as music into a subscription based service. The final type is ‘lean platforms’, which ‘attempt to reduce their ownership of assets to a minimum and profit by reducing costs as much as possible.’ (Srnicek, 2017) An example of such a lean platform is Uber. Their service functions as a platform that brings drivers and passengers together, while they do not employ any of their drivers, nor do they own any of the vehicles that the drivers use. (Srnicek, 2017)

Srnicek concludes his analysis of these different types of platforms by pointing to the example of Amazon. He points out how Amazon grew from an e-commerce company, into a logistics company, into a multi-faceted company that somehow includes most aspects of all different types of platforms listed. (Srnicek, 2017) With the exception of lean platforms, Srnicek’s different types of platforms each encompass services that Amazon continues to provide to this day. The following section gives examples of each of these.

First, the Amazon web store can be viewed as an advertising platform. Based on user data and preferences, personalized advertisements are created for each individual user while using the Amazon web store, as chapter 4 of this thesis will further clarify. Second, AWS is an example of a cloud

platform. Offering cloud computing services to other companies is the core business tenet of AWS, making it a prime example of a cloud platform. Third, the Amazon web store can also be viewed as an industrial platform. By taking control of the entire sales process from product selection through distribution, and allowing companies to make use of this functionality, they function as an industrial platform by lowering the physical costs of production and distribution. Fourth, Amazon also owns several different product platforms through which they provide traditional goods as a subscription based service. Examples of these include the Kindle e-book platform, as well as Amazon Prime Video which creates and distributes audiovisual content such as film and television series.

The next section will synthesize the three different theorizations of data-oriented capitalism given so far in this chapter, and show how we can theorize Amazon as a potential threat to privacy by using this synthesized model.

2.5) Amazon and privacy

By combining the three different frameworks provided in this chapter so far, we can examine the ways in which Amazon conducts their various business practices. For each industry that Amazon operates in, the collection and analysis of data is crucial for the services that it provides.

First, we need to consider Sarah Myers West's argument that communication and information are historically a key source of power. As shown, the earliest efforts to quantify human behavior have resulted in the creation of surveillance networks as early as the end of the nineteenth century. (West, 2017) As technology improved, new systems were developed that could make sense of data, and gradually became capable of collecting this data automatically. Considering that information is in essence data that has been given meaning through relational connections (Bellinger et al., 2004), these new technologies therefore allow companies such as Amazon to collect at a large scale what has historically been considered as a key source of power.

Additionally, Amazon makes use of the three narratives of technological utopianism that West describes in order to ease customers into agreeing to have their data collected and used. They encourage businesses and entrepreneurs to make use of the free and open networks they provide to improve their services. Meanwhile, the scope at which they collect data from these businesses for the improvement of their own services is not clear based on evidence, as described earlier when discussing the limitations of what is known about Amazon's clickstream. The second narrative, the potential to make a customer's experience personal, is perpetrated heavily by the layout of the web store in which personalized advertisements are featured prominently, as will be further shown in chapter 4. The third and final narrative regarding the technocratic value placed on data and its potential to augment consumer power is not as straightforward to attribute to a single aspect of Amazon's services. Rather, the focus on the technocratic value placed on data stems from Amazon's core philosophy of using data to make each of their services as user friendly as possible. (The Manifest, 2019)

Next, Shosanna Zuboff's argument regarding surveillance capitalism focuses around a 'systemic coherent new logic of accumulation.' (Zuboff, 2016) She argues that surveillance capitalism accumulates not only surveillance assets and capital, or in other words data, but also different rights in regard to this data. In other words, by using Amazon's different services, the user inherently forfeits their right to secrecy since secrecy is a result of privacy, and according to Zuboff privacy can not exist in the system within which data is being collected and analyzed at scale. (Zuboff, 2016) She argues that this results in a unilateral redistribution of rights in a way that is largely free from detection or sanction. (Zuboff, 2016) The crucial point here is that the only way to opt out of a user's data being collected by Amazon is to avoid the company and all of its services altogether. Since Amazon has access to such a large amount of data through Amazon Web Services, in combination with the non-transparency of what happens with this data, opting out of this data collection entirely becomes completely unfeasible. This means that users are giving up privacy rights, and therefore decision rights, at a scale that according to Zuboff would take nothing short of a social revolt to prevent.

Finally, Nick Srnicek's platform capitalism model focuses on the fact that data have become central for not just Amazon, but for all companies that operate on a platform model. He argues that the entire platform capitalism business model is centered around the usage of data in the 'key capitalist functions' that it nowadays fulfills. (Srnicek, 2017) It is clear that both the Amazon web store and AWS fulfill Srnicek's description of platforms, namely as digital infrastructures that enable two or more groups to interact; intermediaries that bring together different users. (Srnicek, 2017) Further, the previous section has shown that Amazon shows aspects of four out of five different platform types that Srnicek describes.

In conclusion, the synthesis of these three models regarding data-oriented capitalism results in the following analysis. Srnicek says that Amazon can be viewed as a platform in a variety of ways, and that platforms are the most efficient business models to collect data from. West argues that platforms such as Amazon collect data, and therefore information, which is historically seen as a key source of power, automatically and at an unprecedented scale. Zuboff posits that complying to

Amazon's collection of user data forfeits privacy rights in a way that is unavoidable through the logic of surveillance capitalism. From the combination of these three models we can conclude that these mass surveillance practices are a major threat to the secrecy and privacy of individuals, that this happens automatically with no way for the user to opt out, and that this accumulation of data results in a large amount of control regarding user privacy rights for companies such as Amazon.

Knowing that Amazon's business practices have the ability to make such a large impact on the privacy of individuals, this next section closely examines the GDPR and the ways in which it attempts to protect user privacy on an international level.

2.6) GDPR Analysis

The final section of this chapter examines the GDPR, and asks the question why GDPR compliance is important for platforms in light of user data privacy. As the previous section detailed, scholars have pointed to the dangers of large scale data collection through the use of platforms. This section explains the GDPR as an attempt to limit this data collection, and to prevent some of the dangers that the authors in the previous sections draw attention to.

According to GDPR.eu, the General Data Protection Regulation (GDPR) is the toughest privacy and security law in the world. (GDPR, 2018) Although this regulation was passed by the European Union, it imposes obligations on any organization that targets or collects data related to people who reside within the EU. The GDPR is an update to Europe's 1995 European Data Protection Directive. This older legislation was very limited in its scope, which can be mostly traced back to the unprecedented growth that the internet has seen since after the law had been implemented. Both the GDPR and its predecessor have its origins in the right to privacy as stated in the 1950 European Convention on Human Rights. This convention states: "Everyone has the right to respect for his private and family life, his home and his correspondence." (Council of Europe, 1950)

What differs the GDPR from most other privacy legislations around the world is its extremely broad and detailed definition of privacy rights that an individual has, and that must be protected. These are, in no particular order: the right to be informed when your data is being used; the right of access to any data that has been collected; the right of rectification of any data that the data subject deems false or inaccurate; the right of erasure of any such data; the right to restrict processing; the right to data portability; the right to object; as well as rights in relation to automated decision making and profiling.

In the GDPR, European legislators have made a tremendous step in the direction of data security, as its definitions are very thorough in terms of what is and is not allowed specifically for platforms. Article 6 states that, unless a data subject has provided informed consent to data processing for one or more purposes, personal data may not be processed unless there is at least one legal basis to do so.

(GDPR, 2018) These lawful purposes are: if the data subject has given consent to the processing of their personal data; to fulfill contractual obligations with a data subject; to comply with a data controller's legal obligations; to protect the vital interests of a data subject or another individual; to perform a task in the public interest or in official authority; and for the legitimate interests of a data controller or third party, unless these interests are overridden by the rights of the data subject. (Article 6, GDPR) In summary, Amazon is only allowed to process user data in the case of explicit user consent.

Article 7 explains that this consent must be a specific, freely-given, plainly-worded, and unambiguous affirmation given by the data subject, or in other words, the user must be directly and unambiguously prompted to give their consent for the processing of their data in order to make it legal. Article 25 'requires data protection to be designed into the development of business processes for products and services'. (Article 25, GDPR)

The final few articles that are important specifically to this research are those related to the rights of the data subject. Article 12 requires that, when requested, the data controller provides information to the data subject in an intelligible way. Article 15 describes how one can access this data from a company through use of the right of access, which gives people 'the right to access their personal data and information about how this personal data is being processed. Further, 'the data collector has to inform the data subject on details about the processing, such as the purposes of the processing, with whom the data is shared, and how it acquired the data'. (Article 15(3); 15(1)(a); 15(1)(c); 15(1)(g), GDPR) This both includes data provided by the data subject, and data observed from the data subject.

Article 17 states the right of erasure, or the right to request erasure of personal data related to them on any one of a number of grounds within 30 days, including noncompliance with the aforementioned Article 6. Article 20 describes the right to data portability, which means that data cannot be stored in closed databases that would make the data subject liable to a vendor lock-in, meaning the data is being held hostage in some archaic database which infringes upon the rights of the

data subject. Article 21 provides the data subject with the right to object to the processing of personal information for non-service related purposes such as marketing or sales. This also applies to algorithmically made decisions based upon the data subject's information.

In this thesis at large I am investigating whether or not Amazon complies with the GDPR in one or more ways. Going from this perspective, there are seemingly two main problems with non-compliance to the GDPR. The first problem stems from the fact that analysis of data from within the platform is being used to match consumers and producers. Namely, nontransparent usage of this data could lead to an unfair advantage both between different
producers within the Amazon platform, as well as with Amazon's competitors outside of their platform. Since data is being used to make Amazon valuable as a service, unjust acquisition or analysis of this data could lead to a monopolization of both the platform itself and the various producers that sell through the platform.

The second problem is that the usage of data in a way that is nontransparent, or worse, nonvoluntary, leads to the individual losing control over their data. This could lead to the data being shared with third parties, or it being stored somewhere indefinitely, unbeknownst to the individuals whose data has been collected. This leads to the near apocalyptic scenarios earlier described by Zuboff in her account on surveillance capitalism. Granted, the usage of data is one of the key features of Amazon that makes it not only a platform, but a desirable one at that for producers and consumers alike. However, problems arise when this usage of data can be identified for individuals that have not explicitly consented to Amazon's Terms of Use. By making an account and logging in, the individual consents Amazon with the usage of their data, but this is not the case when the user chooses not to make an account or not to log in.

In conclusion, the GDPR gives very clear outlines about which practices related to data collection are allowed, and which practices are forbidden. The most important articles within the GDPR are article 6, explaining the necessity of user consent, and article 7, which necessitates for the user to be directly and unambiguously prompted about the usage of their data. This is because, when

visiting Amazon.com for the first time, the user does not get any sort of prompt which relates to their data collection. According to article 7, this either means that no data is being collected, or that the GDPR is being violated. This means that, if the Amazon.com web store can be shown to be able to be influenced by the existence of collected user data, in a situation where the user is not logged in and therefore has not explicitly agreed to Amazon's Terms and Services, Amazon would be in direct violation of article 7 of the GDPR.

In the next chapter, this thesis will introduce the methodology that will be used to conduct a case study, that will be able to show whether or not Amazon is in compliance with article 7 of the GDPR.

Chapter 3: Methodology

This chapter gives an overview of the methodology that was used for this research. It outlines the different steps that were taken, and explains the decisions that were made. First, it gives a succinct description of Tracking Exposed, who created the amTREX application. amTREX is an application that is designed to extract data from the Amazon web store's search results page. Second, an explanation of amTREX's functionalities is given. The following section gives an explanation of qualitative content analysis, and also explains why this method of analysis was chosen. Finally this section concludes with a step-by-step explanation of the method as it has been applied in chapter 4.

3.1) Tracking Exposed

Tracking Exposed is an independent research group based in Italy. They develop applications that allow for individual researchers to scrape data from large online platforms which can then be used to analyze the behavior of these platforms among other things. Their research started in 2016 with the Facebook Tracking Exposed project, but has since extended to include different projects based around YouTube, PornHub, and Amazon.

As previous work done by the Tracking Exposed team has shown, it turns out that Amazon is not transparent on their data usage, even to individuals who have consented for their data to be processed by creating and logging into an Amazon account. In their initial report on the Amazon Tracking Exposed project, the Tracking Exposed team writes: "We knew, from past research, that Amazon.com Inc. was collecting a detailed log of personal activities, called 'clickstream'. By performing a GDPRData Subject Access Request (DSAR), we however do not get access to this information." (TrackingExposed, 2019) As explained in chapter 1, we do not have insight into the specific details of the data collected by Amazon's clickstream. While Amazon do not share the reason for this noncompliance, this already appears to be in violation of the 'Right of access' within the GDPR, which states: 'The right of access, commonly referred to as subject access, gives individuals the right to obtain a copy of their personal data as well as other supplementary

information. It helps individuals to understand how and why you are using their data, and check you are doing it lawfully.' (ICO, 2020) Within the GDPR, this is covered by article 12.

Aside from this active non-compliance, Amazon's data usage is also obscured through non-transparency, which is inherent to algorithms within platforms, especially algorithms that use machine learning or other forms of artificial intelligence. This non-transparency is derived from the fact that Amazon uses artificial intelligence to optimize its algorithms.These artificial intelligence algorithms all have an inherent downside, namely the 'black box problem'. This problem is described in a keynote at the Thirty-Third AAAI COnference on Artificial Intelligence, where researchers from the University of Pisa explain:

> Black box AI systems for automated decision making, often
> based on machine learning over (big) data, map a user's features
> into a class or a score without exposing the reasons why.
> This is problematic not only for lack of transparency, but also
> for possible biases inherited by the algorithms from human
> prejudices and collection artifacts hidden in the training data,
> which may lead to unfair or wrong decisions. (Pedreschi et al., 2019, pp. 9780)

When querying Amazon it can be very difficult to determine why results differ between different test phases as a result of this non-transparency. As a result, researchers at Tracking Exposed were left to investigate what the output is of the Amazon.com results page, and how this output can differ between certain variables such as the specific products that were shown, or the order in which they were shown. Examples of these variables can be controlled within queries by doing the queries repeatedly at different times, or using a different device or a different web browser. In order to collect the search results of different queries in a way that they can be meaningfully analyzed, the amTREX tool was created. The next section explains how the amTREX tool functions

3.2) Using the amTREX tool

The amTREX tool that has been developed by the Tracking Exposed team can be used to scrape the search results of the Amazon.com web store. The amTREX application has been written in Python, a programming language, and its code is made publicly accessible on GitHub, which is an online platform in which developers are able to share their code along with the accompanying documentation. There are different versions of the application for both Firefox and Google Chrome that are otherwise functionally identical. For this research, a modified version of Google Chrome called Brave was used. This version of Google Chrome is different because there are certain features added which gives the user greater control of the kinds of data that the browser shares. This is the same browser that was previously used during the DMI Winter School 2020, since it has a high degree of customizability in regard to its privacy settings. Additionally, the usage of the same browser in both experiments will turn out to be significant later in this research. The crucial finding that will be shown later in chapter 4, hinges on the fact that I used the Brave browser both during the previous Amazon Tracking Exposed project, as well as for the research of this thesis.

The search results that the application scrapes are then automatically placed in spreadsheets which can be accessed through the control panel. Search queries done through the amTREX tool are linked to individual users, who each receive a pseudonym when first accessing the application. The application also allows for the creation of tags that can be used to categorize different queries. This can be used to differentiate between different sessions, of different types of data being collected. Each query needs to be assigned a tag before data can be collected, and these tags can either be made private or public, which is useful for coordinating data collection between different people during a single session.

There are two components to the amTREX tool. The main visual component for users is the control panel or dashboard page. This page gives an overview of the most recent queries that have been made, and for each of these queries it automatically displays the amount of search results as well as the average price of these search results. This is done because the amTREX tool was created

specifically to investigate different types of price discrimination on Amazon. Aside from this, the dashboard page also displays the user's pseudonym, and the tag they are currently using. Finally, the dashboard page allows users to create new tags, and switch over to other previously made tags.

The other main component of the amTREX tool is the part that actually collects the data. In order to collect the data, the user first needs to create or join a tag. Next, the user enters their query in the search bar on Amazon.com. When the results page loads, the application works when the user scrolls to the bottom of the page. A prompt will be shown when the data collection starts working, and another prompt is shown when the process is finished. After it is finished, the user needs to reload the dashboard page and wait for the new query to show up. When it does, the data can now be accessed. This is done by typing a specific URL that has the following format: https://amazon.tracking.exposed/api/v2/flexibleCSV/<QUERY>/[tagName] When this URL is entered, the data will then be automatically downloaded in a .csv format.

In the data sheet that the amTREX tool creates, many different categories are consequently analyzed. These are: the pseudonym of the different amTREX users that used the same query and tag; the names of the different products, as well as their product ID and the thumbnail image that shows on the results page; the direct hyperlink to each specific product; the time and day that the query was made; the order of the search results on Amazon.com; the average price of all products within the query; and finally the original price, discount, and total price for each individual item.

The variety of these different types of data allows for different types of analysis. The original research done by Tracking Exposed with this tool mostly focused a quantitative analysis on the differences in average prices between different users, as this can be done to prove that Amazon uses different kinds of price discrimination. The type of analysis used for this research is qualitative content analysis, for which the amTREX tool lends itself particularly well as the following section will explain.

3.3) Qualitative content analysis

Qualitative content analysis is a method that systematically describes the meaning of qualitative data, which is achieved through assigning successive parts of the data to the categories of a coding frame. (Schreier, 170) This method is categorized by three key features: it reduces data, it is systematic, and it is flexible. (Schreier, 170) By using qualitative content analysis, the amount of data is reduced significantly by focusing on specific aspects of the data being analyzed. This is done by creating different categories within the data which define whether or not certain outcomes are significant. As a result, using this method allows this method to look at large amounts of data and reduce them to very specific results, which can then be compared between different queries, and analysed further using data visualization.

Qualitative content analysis is an iterative approach that relies on a few crucial steps. First the initial coding frame has to be created. This is done by creating categories in both content-driven and data-driven ways. Content-driven categories are usually predefined and based on assumptions made by the researcher upon collecting the data. (Schreier, 174) Data-driven categories are made after initial inspection of the data in order to account for unforeseen results and other anomalies. This step of the process is iterative, as the initial analysis of the data might be inconclusive, or raise the need for different categories to be used. Once the final coding frame has been established, it can be used on the entire dataset in order to categorize it. After this final categorization is done, the data can then be analyzed for further conclusions.

There are several reasons for the choice of the method of qualitative content analysis. The primary reason for this decision is the very nature of the dataset that this research collected from Amazon, and the way in which it is presented by amTREX. On one hand, the data is presented by amTREX in a very quantitative manner, meaning that most variables are expressed by a certain numerical value. Several different categories focus on numeric values such as price or discount, which can be analyzed and interpreted through various calculations as well as data visualization. On the other hand, looking at the data merely in a quantitative manner is insufficient, as there are many

variables that could exist between different Amazon results pages that can not be expressed in quantity alone. For example, one query might result in specific products being offered by certain vendors within Amazon. Additionally, the very layout of the Amazon results page is contingent, as not only different specific products might be shown, but also certain advertisements or timed offers, as well as the order in which these results are displayed. This data could be analyzed in a quantitative manner, but this would forego these subtle differences that Amazon introduces, which would not be rendered visible by quantitative analysis.

It is this very duality between quantitative and qualitative analysis that makes qualitative content analysis more suitable to this research than other qualitative methods. An example of such a qualitative method is interpretation. This method is used in order to interpret the meaning of data beyond what is initially visible. (Willig, 136) Depending on the approach, interpretation can be used to better understand different parts of the data in certain ways, such as what the author's intended or unintended meaning is, or what the underlying context is which made it possible for the author to express whatever it is they expressed. (Willig, 137) While this research and qualitative content analysis in general certainly rely on some amount of interpretation, by itself this method is insufficient as it is not well suited for data that is in some sense quantitative in nature. (Schreier, 173)

Another example of a similar method that would have been potentially suitable for this research is the use of inductive coding within grounded theory. This is a research approach in which data collection and analysis take place simultaneously, whereby each part informs the other in order to construct theories of the phenomenon under study. (Charmaz and Thornberg, 153) While this is similar to qualitative content analysis in its approach, they ultimately differ in the ways in which they iterate upon themselves. When using inductive coding, data collection and analysis need to constantly continue and iterate upon each other, while these two steps are necessarily separated within qualitative content analysis. (Schreier, 174) Due to the nature of data collection through the amTREX tool, qualitative content analysis was deemed more suitable for this research since the amTREX tool is restrictive in the ways that iterations of the categories could take place.

In the final section of this chapter, a step-by-step guide to collecting and analyzing the amTREX results is given, as it will be used in chapter 4 of this thesis.

3.4) Overview of the data collection

The goal of this method is to have two different data sheets for each individual search query so that the differences can be observed between them. The first data sheet is created in a ‘clean’ environment, in which all trackers and cookies have been deleted from the browser before the data is collected. The second data sheet is created in a ‘polluted’ environment, where various trackers and cookies have been put into place by browsing through various different websites. This is done in order to see if the data collected from AWS, influences the search results of the Amazon web store. Each query, then, is divided into three different steps.

1) The first step is the scraping of the clean data from Amazon.

2) The second step is the pollution of the browser which is achieved through browsing the internet for a period of forty-five minutes.

3) The third and final step is the scraping of the polluted data from Amazon.

Before data can be collected, a research browser must first be installed. The purpose of the research browser is to provide a fresh canvas that has been untouched by one’s browsing history. The research browser used for this thesis is Brave, as explained earlier. After installing the research browser, the amTREX tool must be downloaded and installed from the Tracking Exposed website. Finally, the amTREX tool needs to be opened in order to create a tag, which is necessary for amTREX to function as explained earlier.

After this has been done, the clean data can then be scraped from Amazon. Before doing a clean scrape, the researcher first deletes all browsing history and cookies to make sure no variables are unintentionally being introduced. As soon as the clean scrape is finished, the second phase begins where the researcher pollutes their browser by browsing through various websites. There is no specific set of websites that the researcher visits, however within this research there are three different types of browsing behavior being used, in order to introduce enough variables to see what kind of influence the data from AWS has. The first behavior only looks at various news sites; these users read some articles on one website before moving on to another, making sure to accept all cookies and

clicking on at least one advertisement if present. The second behavior only looks at various YouTube videos related to certain news stories. The final browsing behavior that this research used had no preference in the types of website it visited, and used Google Search to browse through a variety of different websites. These different forms of user behavior are not necessary, but serve as a way to introduce another variable to the research, as some browsing behaviors might prove to show more significant results than others.

This process was done for 45 minutes, which should be long enough to collect various cookies and trackers from AWS. Doing this process for a longer time is possible, and this could yield more significant results, however this process was limited to 45 minutes due to time constraints. After this, the polluted data can be scraped from Amazon. Now the query is finished and ready to be downloaded, after which it will be analyzed using the qualitative content analysis method. This process can be repeated as many times as deemed necessary for a variety of different queries in order to introduce specific variables into the results. Within this research, these variables were introduced by doing the same query with multiple devices, at the same location and at the same time. This was done in order to eliminate any server-side variables which could affect the outcome of this research in nontransparent ways.

After the data has been collected it is then ready to be analyzed. In the next chapter, the process of collecting and analyzing data for this thesis will be laid out and explained.

Chapter 4: Findings

This chapter goes over the analysis of the amTREX data that has been collected, as was described in the previous chapter. The first section will explain how the analysis was conducted, which search queries were used, which browsing behaviors were chosen, and why they were selected. Before going over the analysis, there are some technical problems that occurred during data collection which need to be addressed, since they impact the scope and limitations of this research. The next section concerns the conditions of the analysis itself. It goes over the various content-driven and data-driven categories that were used, as explained in chapter 3.3 regarding qualitative content analysis. The following sections, 4.1 and 4.2, contain the actual analysis, in which the criteria set for the various categories in the previous section will be compared to the results that follow from the amTREX data. The final section of this chapter summarizes these findings so that they can be discussed in the next chapter.

4.1) Variables

This section explains the different variables that are present when using amTREX to collect data from Amazon.com, as well as some additional variables that were introduced by this research.

In order to collect data from Amazon with the amTREX tool, a query needs to be entered into the Amazon.com search bar. After this, the researcher must scroll down to the bottom of the page in order for the amTREX tool to start working. Once the data from a page has been collected, a new query can be entered in order to repeat the process.

4.1.1) Variable 1: Queries

For this research, six different search queries were chosen. These were: ‘face mask’, ‘toilet paper’, ‘sunscreen’, ‘hand soap’, ‘smart watch’ and ‘phone case’. These queries were chosen to be general enough, so that Amazon.com’s search algorithm is not automatically biased towards specific products or brands. Additionally, the first four of these queries are all related to personal hygiene, and these were chosen in order to reflect the relevance of personal hygiene in the news due to 2020’s Covid-19 outbreak. The final two of these queries are not related to personal hygiene, so that any patterns that emerge in queries related to personal hygiene can be observed by comparing them to these two more general queries.

Another path that could have been chosen in regard to selecting search queries, is to select one single query which would be entered in periodic intervals. In doing so, certain biases that exist in Amazon.com related to time-sensitivity could be identified. The most notable of these are the hourly, weekly, and daily offers that comprise a large part of the search results page. However, as the upcoming section will explain, there were already enough difficulties in conducting the data collection only once. Doing the same processes multiple times would result in more time constraints, as well as more difficulty in organizing the data collection. Therefore it was decided to do six different queries at the time of data collection, so that certain patterns in product difference can be checked across both different participants and different queries.

4.1.2) Variable 2: Devices

As mentioned above, using different queries is a variable that this research chooses to introduce during data collection. However, another variable that is inherently present in data collection is a technique called ‘device fingerprinting’. This is a tracking technique that companies use which collects ‘identifying information about unique characteristics of the individual computers people use.’ (Nikiforakis & Acer, 2014) This technique is used primarily by advertisers, operating under the assumption ‘that each user operates his or her own hardware, identifying a device is tantamount to identifying the person behind it.’ (Nikiforakis & Acer, 2014) While there is no explicit evidence yet that Amazon makes use of such device fingerprinting techniques, it is plausible to assume that these techniques are being used due to the fact that they have been a standard practice in online data collection for years. (Nikiforakis & Acer, 2014) In addition, if Amazon were to use such device fingerprinting techniques, the amTREX tool lends itself particularly well for identifying this behavior.

If there is a device-specific bias in the Amazon web store search results that persists between the ‘clean’ and ‘polluted’ data scrapes, it has to be the result of Amazon making use of device fingerprinting techniques, because all other variables such as cookies and trackers are eliminated in the ‘clean’ scrape. If such a pattern can be shown for the different queries that will be analyzed in this chapter, it would mean that Amazon is in this case noncompliant with several different articles of the GDPR, namely article 6, which explains the necessity of user consent, and article 7, which states that the user should be directly and unambiguously prompted about the usage of their data.

4.1.3) Variable 3: Browsing behavior

The final variable of the data collection is another one that this research intentionally introduces, namely browsing behavior. Using different browsing behaviors, the participants of the research visit different websites and therefore collect different specific trackers from AWS. There were three different browsing behaviors used. The first browsing behavior limited itself to news websites, both national and international. On each news website, the participants allow the site to collect cookies by pressing 'Accept' on any given prompts related to data collection, and they would additionally click on at least one advertisement per website, if present. This is done to ensure that the maximum possible number of different cookies and trackers are collected. The second browsing behavior limited itself entirely to videos related to current news within YouTube. The participants would browse through different videos belonging to different channels, and again clicking on at least one advertisement if present. The final browsing behavior is a combination of the first two. First the participant will watch some news-related YouTube videos before moving on to news websites, holding themselves to the same standards regarding advertisements as the other two browsing behaviors. Each of these browsing behaviors is conducted by the participants for approximately 45 minutes before the ensuing data is collected.

In these browsing behaviors, a specific emphasis was made towards news-related websites and content. This was done in conjunction with the selection of specific keywords related to personal hygiene as explained earlier in this section. Due to the variance in availability of different personal health related products around the world as caused by the Covid-19 crisis, this research hypothesized that such articles would be especially contingent within the search results of Amazon.com. The browsing behavior variable was introduced in order to make the search data as contingent as possible in comparison to the 'clean' scrape. It was therefore deemed beneficial to select specific browsing behaviors that were thought to introduce as much contingency as possible.

In conclusion of this sub-section, there are three important variables related to the data collection using the amTREX tool. The first variable, the use of different queries, was introduced by

this research in order to increase the total amount of data collected in the face of time constraints. The second variable, the use of different devices, is inherent in doing the research since the different participants will be following the same steps simultaneously, which would be possible when using a single device. This variable allows this research to test the validity of its hypothesis, since the amTREX tool lends itself well to proving the usage of device fingerprinting by looking for specific patterns in unique search results as explained earlier. The third and final variable, the use of different browsing behaviors, was also introduced by this research and also serves to potentially prove this research's hypothesis. In this case, this would be achieved by showing a high degree of contingency between the 'clean' and 'polluted' data scrapes.

Before the analysis itself, the next section will briefly explain some technical difficulties that were experienced during data collection, and the limitations that these difficulties introduced.

4.2) Limitations

There were various limitations to this research that were introduced by a variety of technical and non-technical actors. First of all, in order to limit the number of variables introduced to the data collection, the process of data collection needed to happen simultaneously and at the same physical location. The reason behind this is to avoid the search results to be skewed by using different IP-addresses, or different GPS locations. However, this meant that the data collection had to be planned well in advance since not all of the participants of the case study were available at all times. As a result, it was not possible to include time and temporality as one of the defining categories of this research, since it proved impossible to plan multiple sessions in quick succession.

The second limitation was technical in nature. During the first data collection session that was planned, the amTREX data collection tool was not working properly. Originally when a new user opens the tool for the first time, they should be automatically assigned a pseudonym that is used in order to identify search results between different participants. However, such a pseudonym could not be made during this time. At the time, an improvisation allowed the participants to use an automated screenshot tool that would make an image out of each individual search result page, which would then be manually converted into spreadsheets. However, it became apparent that this process would be much too time-consuming given the scope of this research. Therefore, a new session had to be planned after the amTREX tool had been fixed.

A new session was planned and the data was once again collected, this time using the amTREX tool properly. However, the final technical limitation of this research is that some data was missing from this data collection session. For two out of five participants, each of the twelve different search queries, six ‘clean’ and six ‘polluted’, could be collected. For the other three participants, the search results of some of the queries seem to be missing entirely. While the usage of different search queries allows this research to mitigate this problem, ultimately it impacts the conclusiveness and especially the replicability of this research. Nonetheless, enough data was collected for this research to draw conclusions upon. In the next section, the variables defined previously will be used to create

different categories, as is necessary for the method of qualitative data collection. Afterwards, the collected amTREX data will be examined closely.

4.3) Categorising the data

When comparing the search results of the ‘clean’ and ‘polluted’ data scrapes, specific products on the polluted page will either remain the same as on the clean page, or they will differ in various ways. The products can appear in a different order than on the clean results page. Additionally, products could also be replaced entirely by different products. Certain brands can be shown excessively to certain participants, while being completely absent in the results of others. The more differences there are between the two pages, the higher becomes the degree of contingency that is introduced by the various browsing behaviors. When a difference occurs in product choice, it is considered to be more salient than a difference in product location, therefore differences in product location will be disregarded for now, but might be reintroduced if results prove inconclusive. In other words, if there are differences in specific products being shown to specific participants, this is considered to be a more significant finding than if the same products were shown in a different order. If there are certain differences that persist when comparing different users with different browsing behaviors, this research will quantify those differences as ‘browsing-related contingencies’. If there are differences that are unique to specific users unrelated to browsing behaviors, they will be quantified as ‘general contingencies’. Since the amTREX tool distinguishes between products as being offered normally and being offered through personalized advertising, specific differences that only occur in the advertisement sections will be marked as ‘advertising-related contingencies’. This is done because the advertisement section of Amazon is known to already be contingent in and of itself. This is because Amazon sells its ad space to different companies for different time slots, similar to how Google AdWords functions. (DisruptiveAdvertising.com, 2018)

Aside from the abovementioned contingencies, this analysis will also look for specific patterns when comparing the clean and polluted results of specific queries. If a pattern occurs between the two queries, then it can be observed as a sign of device fingerprinting. All other possible variables are eliminated before conducting the ‘clean’ search, since all cookies and trackers are deleted from the browser at that time. Device recognition is the only possible explanation for such a scenario. In order

to observe this pattern, specific search results that are unique to one single pseudonym will be marked as unique. The same will be done for products that are unique to two specific pseudonyms, while being absent in all others. Each time a unique product persists across the clean and polluted version of the same query, it will be marked as ‘Persistent unique’. Each time a product is unique to two specific pseudonyms across the clean and polluted version of the same query, it will be marked as ‘Persistent Semi-unique’. Since it is impossible to make specific quantifications of the amount of unique and semi-unique products a query must have made before the result is considered significant, variance between these categories will be observed across all queries, before such quantifications are made.

In conclusion, there are three categories related to the variable of user behavior. These are ‘browsing-related contingencies’, ‘general contingencies’, and ‘advertisement-related contingencies’. In addition there are two categories related to the variable of using different devices. These are ‘persistent unique’ and ‘persistent semi-unique’.

4.3.1) Phone case

The first query that was analyzed using this method was ‘phone case’. For this query, the data of a total 3 out of 5 participants was successfully collected, as shown in the table below.

Pseudonym	**Behavior**
souffle-endive-cranberry	News sites
tea-icecream-pistachio	YouTube
corn-souffle-hamburger	News sites + YouTube

Figure 1: pseudonyms and behaviors for the query ‘phone case’

In order to identify the different categories and quantify them, specific adjustments were made into the spreadsheets to identify if a pattern emerges, or if results are contingent. First, each product that appears only once within the ‘clean’ scrape was marked with a highlighter to denote their uniqueness. Then, the same was done for the results of the ‘polluted’ scrape. Additionally, products that appeared on the ‘clean’ scrape but were removed in the ‘polluted’ scrape were marked in italics, whereas products that first appear in the ‘polluted’ scrape were marked with bold text. After this process was done, the two spreadsheets were combined into one so that the results could be compared between them.

Figure 2.1 records the contingencies between the ‘clean’ and ‘polluted’ scrape for the query ‘phone case’. Since each of the three pseudonyms used a different browsing behavior, browsing-related contingencies could not be recorded for this query.

pseudonym	Total products shown	General contingencies	%	Advertising-related contingencies	%
souffle-endive-cranberry	60	2	3.33%	6	10%
tea-icecream-pistachio	65	1	1.53%	5	7.69%
corn-souffle-hamburger	60	1	1.6%	7	11.6%

Figure 2.1 Number of contingencies for query ‘phone case’

Figure 2.2 records the amount of times that a unique product appears, and the amount of times one persists across the ‘clean’ and ‘polluted’ search results.

pseudonym	Total products shown	Unique total	Unique persistent	% of total	% of unique
souffle-endive-cranberry	60	13	11	18.3%	84.62%
tea-icecream-pistachio	65	16	12	18.46%	75%
corn-souffle-hamburger	60	32	26	43.33%	81.25%

Figure 2.2 Number of uniques for query ‘phone case’

Figure 2.3 records the same values for semi-unique products

pseudonym	Total products shown	Semi-unique total	Semi-unique persistent	% of total	% of semi-unique
souffle-endive-cranberry	60	25	13	21.67%	52%
tea-icecream-pistachio	65	24	15	23.08%	62.5%
corn-souffle-hamburger	60	8	1	1.67%	12.5%

Figure 2.3 Number of semi-uniques for query ‘phone case’

Some notable observations can be made from this first query. First, there were many more advertisement-related contingencies than there were general contingencies, as hypothesized previously. Since for each pseudonym there are only one or two products that differ outside of products that fall within Amazon's advertising frameworks, so far it seems that the cookies and trackers collected through the various browsing behaviors did not have the expected amount of impact on the contingency of the search results.

However, the results regarding persistence of unique and semi-unique products are much more salient. For each of the three pseudonyms, at least 40% of products displayed within the search results were shown to be either unique or semi-unique. For the corn-souffle-hamburger pseudonym this result was especially skewed towards unique products, reaching 43.33% of total products. (Figure 2.2) Additionally, out of all unique products displayed, at least 75% of products were shown to be persistent between all three pseudonyms. Out of all semi-unique products displayed, more than 50% of them were shown to be persistent for two different pseudonyms, with the semi-unique persistent products for pseudonym corn-souffle-hamburger being the only outlier. (Figure 2.3)

While this is only the first query being analyzed so far, these results already imply heavily that Amazon makes use of device fingerprinting techniques in order to track users and collect their data. This is further emphasized by the fact that the pseudonym corn-souffle-hamburger, for which these results were the most salient, has already participated in an amTREX data collection project previously. Since the cookies and trackers were cleared beforehand, as was done for all other pseudonyms, it is implied that the relatively high number of unique products for this pseudonym can be explained by Amazon recognizing the device being used. This is because the result suggests that Amazon is using previously collected data to improve search results towards them. Analyzing the remaining queries should give insight into whether this pattern keeps occurring, or if the 'phone case' query was merely an outlier in this regard.

4.3.2) Remaining queries

The same method of analysis was applied to the remaining five queries that were made. For most of the queries, the data of the same 3 out of 5 participants was collected successfully. For the ‘sunscreen’ and ‘hand soap’ queries, the data of only 3 out of 5 participants was fully collected. Therefore, their results have been accounted for when considering whether a result is unique or semi-unique, but they have not been added to the table since the full data couldn’t successfully be retrieved. This was done because these results still impact the semi-uniqueness of products. However, including them in the tables would be confusing since there would be an asymmetrical amount of participants for certain queries. Finally, for the ‘face mask’ query, only the polluted data could be collected for the ‘souffle-endive-cranberry’ pseudonym. As with the other cases where part of the data is missing, the results have been accounted for when considering uniqueness and semi-uniqueness. The results can be found below.

4.3.3) Sunscreen

pseudonym	Total products shown	General contingencies	%	Advertising-related contingencies	%
souffle-endive-cranberry	60	0	0%	4	6.66%
tea-icecream-pistachio	57	0	0%	1	1.75%
corn-souffle-hamburger	55	0	0%	1	1.81%

Figure 3.1 Number of contingencies for query ‘sunscreen’

For the sunscreen query, as shown in figure 3.1, no general contingencies could be found. Additionally, the advertising-related contingencies are relatively uncommon compared to the phone case query.

pseudonym	Total products shown	Unique total	Unique persistent	% of total	% of unique
souffle-endive-cranberry	60	2	2	3.3%	100%
tea-icecream-pistachio	57	0	0	0%	0%
corn-souffle-hamburger	55	18	17	30.1%	94.4%

Figure 3.2 Number of uniques for query ‘sunscreen’

pseudonym	Total products shown	Semi-unique total	Semi-unique persistent	% of total	% of semi-unique
souffle-endive-cranberry	60	3	0	0%	0%
tea-icecream-pistachio	57	2	1	3.5%	50%
corn-souffle-hamburger	55	1	1	1.81%	100%

Figure 3.3 Number of semi-uniques for query ‘sunscreen’

Here there is a clear difference in the amount of unique results between the three pseudonyms. While two out of three show barely any unique results, corn-souffle-hamburger has a much higher degree of personalization than even the results of the phone case query. (Figure 3.2 as compared to Figure 2.2) Already in that query the difference between corn-souffle-hamburger and the other participants was remarkable, and this difference is emphasized even more in the sunscreen query. However, in the phone case query the other two participants had a significantly higher amount of semi-unique products shared between them, whereas this pattern is absent in the sunscreen query. (Figure 3.3 as compared to Figure 2.3)

4.3.4) Smart watch

pseudonym	Total products shown	General contingencies	%	Advertising-related contingencies	%
souffle-endive-cranberry	21	0	0%	5	23.8%
tea-icecream-pistachio	26	0	0%	2	7.69%
corn-souffle-hamburger	21	0	0%	4	19.04%

Figure 4.1 Number of contingencies for query ‘smart watch’

Like the previous query, the ‘smart watch’ query displays no general contingencies as shown in Figure 4.1. In this query the total amount of products shown is much smaller than the two previous queries, and the advertising-related contingencies account for a relatively high percentage of the total amount of products shown as a result.

pseudonym	Total products shown	Unique total	Unique persistent	% of total	% of unique
souffle-endive-cranberry	21	1	0	0%	0%
tea-icecream-pistachio	26	3	2	7.69%	66.67%
corn-souffle-hamburger	21	5	4	19.04%	80%

Figure 4.2 Number of uniques for query ‘smart watch’

pseudonym	Total products shown	Semi-unique total	Semi-unique persistent	% of total	% of semi-unique
souffle-endive-cranberry	21	7	4	19.04%	57.14%
tea-icecream-pistachio	26	8	5	19.23%	62.5%
corn-souffle-hamburger	21	1	1	4.76%	100%

4.3 Number of semi-uniques for query ‘smart watch’

In the smart watch query, the pattern that was first shown with the phone case query re-emerges. The corn-souffle-hamburger has the highest amount of unique products by far, and the pattern of unique products remains for 80% the same between the clean and polluted queries. (Figure 4.2) Additionally, the remaining two pseudonyms show a similar pattern of semi-unique products shared between them, although they are relatively less persistent between queries compared to the phone case query. (Figure 4.3)

4.3.5) Hand soap

pseudonym	Total products shown	General contingencies	%	Advertising-related contingencies	%
souffle-endive-cranberry	56	6	10.71 %	5	8.92%
tea-icecream-pistachio	60	3	5%	3	5%
corn-souffle-hamburger	57	2	3.51%	3	5.26%

Figure 5.1 Number of contingencies for query ‘hand soap’

The ‘hand soap’ query gives the highest amount of general contingencies of all queries so far, having a similar share of total products shown as the advertising-related contingencies. (Figure 5.1) For souffle-endive-cranberry the total percentage of contingencies out of products shown amounts to 19.63% which is the highest result out of any pseudonym in any query so far.

pseudonym	Total products shown	Unique total	Unique persistent	% of total	% of unique
souffle-endive-cranberry	56	0	0	0%	0%
tea-icecream-pistachio	60	6	4	6.66%	66%
corn-souffle-hamburger	57	15	14	24.56%	93.3%

5.2 Number of uniques for query ‘hand soap’

pseudonym	Total products shown	Semi-unique total	Semi-unique persistent	% of total	% of semi-unique
souffle-endive-cranberry	56	10	0	0%	0%
tea-icecream-pistachio	60	3	0	0%	0%
corn-souffle-hamburger	57	5	1	1.75%	20%

5.3 Number of semi-uniques for query ‘hand soap’

The pattern shown in the unique and semi-unique results for the hand soap query is very similar to the pattern that was shown from the sunscreen query. Corn-souffle-hamburger still has the highest amount of unique results out of all three pseudonyms, which has been observed for every query so far. (Figure 5.2) It is highly likely that this pseudonym has a higher degree of personalisation since it is the same pseudonym being used on the same device and with the same browser as during the DMI Winter School. This result is very promising in proving that Amazon makes use of device fingerprinting. Unlike the result for the sunscreen query, there is a high amount of semi-unique products for souffle-endive-cranberry, however this is explained by the fact that the 'clean' data was successfully collected for all five participants in the hand soap query. This means that since this query has more data than the others, a higher amount of semi-unique products is expected. What makes these results so similar to the sunscreen query is the fact that there is no persistence for the semi-unique results for both souffle-endive-cranberry and tea-icecream-pistachio. (Figure 5.3)

For the results so far, it is important to observe that there are two different patterns that emerge for the amount of semi-unique products. Moreover, one pattern is shown for the 'smart watch' and 'phone case' queries, which are both related to technology, while the other pattern is shown for the 'sunscreen' and 'hand soap' queries, which are both related to hygiene. It would be interesting to see if this pattern holds up, however the last remaining queries are both related to hygiene so this result can't be made more salient through observation of further results.

4.3.6) Toilet paper

pseudonym	Total products shown	General contingencies	%	Advertising-related contingencies	%
souffle-endive-cranberry	46	2	4.34%	2	4.34%
tea-icecream-pistachio	50	3	6%	0	0%
corn-souffle-hamburger	39	3	7.69%	1	2.56%

Figure 6.1 Number of contingencies for query ‘toilet paper’

For the ‘toilet paper’ query, both the total amount of products shown and the amount of contingencies are quite average. (Figure 6.1) Important to note here is the fact that the result for tea-icecream-pistachio in this query is the first result so far without any advertising-related contingencies. Since this would mean that all advertisements shown are identical for this participant even though the two searches were made more than forty minutes apart, further details for this query must be analyzed. Upon closer inspection of the order that the products appear in for tea-icecream-pistachio in this query, it appears that the only difference in the advertisements displayed between the first and second search is the order that the advertisements appear in.

This can also be argued to be an indication of the fact that Amazon only uses device fingerprinting to identify users, and that users are not being identified based on data collected from AWS. There has to be some factor that explains why cookies and trackers, which are commonly used to target users with personalized advertising, make no discernible impact in the advertisements shown for this participant. It appears the introduction of cookies and trackers to the research browser is not in all cases as significant as previously theorized. However, while the advertising-related contingencies are low in each query that was made, there is not enough of a pattern visible to make this claim based on these results.

pseudonym	Total products shown	Unique total	Unique persistent	% of total	% of unique
souffle-endive-cranberry	46	2	1	2.17%	50%
tea-icecream-pistachio	50	16	12	18.46%	75%
corn-souffle-hamburger	39	3	1	2.56%	33%

Figure 6.2 Number of uniques for query ‘toilet paper’

pseudonym	Total products shown	Semi-unique total	Semi-unique persistent	% of total	% of semi-unique
souffle-endive-cranberry	46	9	7	15.21%	77.77%
tea-icecream-pistachio	50	11	7	14%	63.63%
corn-souffle-hamburger	39	5	0	0%	0%

Figure 6.3 Number of semi-uniques for query ‘toilet paper’

The pattern that emerges in the unique and semi-unique results for the toilet paper query is different than all other queries so far, which stems for the fact that corn-souffle-hamburger has by far the lowest amount of personalization in this query. While this is noteworthy due to the fact that corn-souffle-hamburger has the highest amount of personalization in all other queries, due to the opaque nature of Amazon’s search algorithm no reasonable explanation could be found for this particular result. This means that using the amTREX tool for collecting and analyzing data from Amazon can not always explain certain results conclusively, as the result shown here in Figure 6.2 is incongruent with all other queries.

4.3.7) Face mask

pseudonym	Total products shown	General contingencies	%	Advertising-related contingencies	%
tea-icecream-pistachio	51	1	1.96%	2	3.92%
corn-souffle-hamburger	50	2	4%	2	4%

Figure 7.1 Number of contingencies for query ‘face mask’

pseudonym	Total products shown	Unique total	Unique persistent	% of total	% of unique
tea-icecream-pistachio	51	5	2	3.92%	40%
corn-souffle-hamburger	50	19	13	26%	68.42%

Figure 7.2 Number of uniques for query ‘face mask’

pseudonym	Total products shown	Semi-unique total	Semi-unique persistent	% of total	% of semi-unique
tea-icecream-pistachio	51	24	22	43%	91.6%
corn-souffle-hamburger	50	4	2	4%	50%

Figure 7.3 Number of semi-uniques for query ‘face mask’

For the ‘face mask’ query, as mentioned earlier, the data could only be collected successfully for two out of five participants. Nevertheless, between the two remaining pseudonyms, there are some recurring patterns. The amount of contingencies is once again low, even when compared to most other queries. (Figure 7.1) For the unique and semi-unique products, there is a high degree of personalization for both participants. It is important to note the especially high amount of semi-unique products for the tea-icecream-pistachio pseudonym. (Figure 7.3) For this query, the polluted results for souffle-endive-cranberry were successfully collected and as mentioned these results have been accounted for when considering unique and semi-unique products. In this case, there is a very high

amount of semi-unique products between tea-icecream-pistachio and souffle-endive-cranberry, higher than in any other query. It would be valuable to look at the degree of persistence between the clean and polluted search for souffle-endive-cranberry if these results had been collected. Nevertheless, this final query confirms the one pattern that has been most consistent across all queries: there is a clear difference in the amount and kinds of personalization between corn-souffle-hamburger and the other participants. This result would not nearly be as significant if it showed up in only some of the queries that were made. However, since the pattern is only absent for one out of six queries, this result can be seen as very significant. This difference is the direct result of the only possible explanation for such a variance in personalization, namely the usage of device recognition.

4.3.8) Conclusion

While the results would be much more salient if all data could be collected for all five participants, corn-souffle-hamburger is a clear outlier between all participants. Almost every single factor in setting up the research was exactly the same between all five participants: the queries were made in the same research browser, at the same time, in the same physical location, using the same IP-address. One factor that is different between all participants is the device used, however if this factor was the most crucial, corn-souffle-hamburger would not be the only outlier, and instead all five participants would have unique results.

The other factor that is different between corn-souffle-hamburger and all other participants, is the fact that corn-souffle-hamburger is the pseudonym that I used, and that I had previously used on the same device, using the same research browser, six months earlier when I first used the amTREX tool to collect data from Amazon. Since I used the same laptop and research browser, the fact that corn-souffle-hamburger is such an outlier can be adequately explained by the notion that Amazon used device fingerprinting to access the data profile they made of me months earlier. Since quite a lot of time had already been spent doing this research within the research browser back then, one would expect a higher degree of personalization for someone who has already spent a significant amount of time on Amazon.com in this case.

This is the exact result that was shown for corn-souffle-hamburger in five out of six queries. In all queries except for the toilet paper query, corn-souffle-hamburger showed the highest amount of unique products by far. This result was equally noticeable in the queries that had more than three participants. These queries had a much higher total amount of products shown, yet the amount of unique products shown to corn-souffle-hamburger remained significant even in these cases.

In conclusion, there still seems to be enough proof that alludes to the fact that Amazon uses device fingerprinting to make data profiles of users that visit their website. This is despite the fact that there are severe limitations to the salience, significance and replicability of these results due to the fact that not all data could be successfully collected. As explained earlier in this research, on

Amazon.com there is no prompt that warns the user about data privacy and data usage, even though such a prompt is required by the GDPR if any data is collected or used on the website. The results of this research all show degrees of personalization between participants that remains consistent between the clean and polluted searches. This degree of personalization was even higher for the one participant who could have been profiled by this method of device fingerprinting in the past. From these results it is plausible to conclude that Amazon makes use of device fingerprinting for the creation of data profiles in a way that is noncompliant with article 6 and article 7 of the GDPR. In the next chapter, the implications of this result will be discussed.

Chapter 5: Discussion

In this chapter, the finding of Amazon's GDPR violation will be discussed in a number of different contexts. First this chapter will re-examine the related GDPR articles in conjunction with additional articles which are related to the consequences of violating the GDPR. This includes any fines that Amazon might receive in light of its violation of GDPR. The next section will draw from examples of other companies that have been fined for GDPR violation. This will be done both to look at overlaps with Amazon's case, and to highlight why user privacy is both important and very fragile.

5.1) GDPR Policy

In the GDPR there is an article that is concerned with the fines that may be imposed upon a company when certain GDPR articles are violated. This is article 83 titled 'General conditions for imposing administrative fines'. In order to review what the potential outcome for Amazon will be, we must first define exactly which GDPR articles are being violated, and in what way.

This research has shown that Amazon makes use of device recognition to track users through different browsing sessions and collect data about them. This process is shown to not be reliant on a cookie or tracker being present from any previous browsing sessions, since cookies and trackers were deleted before testing commenced, as shown in section 4.3. Instead, Amazon appears to store this data about the user on their servers, where it can be accessed again if the user visits Amazon from the same device and using the same browser. Deleting the cookie that Amazon gives the user is therefore not a viable way to opt out of this data collection. More precisely, there is no straightforward way to opt out of one's user data being processed by Amazon. It is possible that contacting Amazon's customer service enables users to request for their data to be deleted, however this takes a lot of effort on the part of the user. It should also be possible to request for a copy of all data that Amazon has gathered from the user, but as earlier research by the Tracking Exposed team has shown, it is very likely that this request gets denied. (Tracking Exposed, 2019)

Given this analysis, we need to look closely at the first subparagraph of article 6 of the GDPR, titled 'Lawfulness of processing'. The first sentence reads: "Processing shall be lawful only if and to the extent that each of the following applies", and then names six different conditions in which the processing of data is considered to be lawful. This section will analyze each of these six different conditions that are related to Amazon.

a) 'The data subject has given consent to the processing of his or her personal data for one or more specific purposes.' (GDPR, 2018)

As explained earlier, at no point does Amazon ask for a user's consent when they visit the Amazon.com website on any of its pages. (GDPR, 2018)

b) 'Processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract.'

The definition of a contract in juridical terms is much broader than an agreement which is signed explicitly by both parties. Amazon's lawyers would probably argue that the collection and processing of data is necessary to provide the service of running a webstore that uses recommendation algorithms. However, if this were the case, this would still be in violation of Article 7 which states that consent needs to be given explicitly and unambiguously, since the data subject needs to be informed about this practice before it occurs.

c) 'Processing is necessary for compliance with a legal obligation to which the controller is subject.' (GDPR, 2018)

This condition does not apply to this scenario. Amazon does not provide a public service, thus there can be no legal obligation that they have to fulfill in this scenario.

d) 'Processing is necessary in order to protect the vital interests of the data subject or another natural person.'(GDPR, 2018)

Amazon's data collection practices for the purpose of gaining profits can not be argued to be of vital interest to the data subject.

e) ‘Processing is necessary for the performance carried out in the public interested or in the exercise of official authority vested in the controller.’ (GDPR, 2018)

There are currently no sources that indicate that any such official authority has been invested in Amazon for the processing of this data.

f) ‘Processing is necessary for the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.’ (GDPR, 2018)

This usage of data by Amazon very much falls into the category of legitimate interests by the controller or a third party. Yet, this is immediately voided by the rights and freedoms of the data subject. The data subject is protected by the fact that they have to explicitly consent for this usage of user data, and the rules for this are detailed in the aptly named Article 7: ‘Conditions for consent’. There are four subparagraphs to this article which read:

1) Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.’ (GDPR, 2018)

Since Amazon does not ask the user for consent for their usage of data, they are in direct violation of this rule since they are unable to demonstrate consent.

2) ‘If the data subject’s consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.‘ (GDPR, 2018)

Amazon does have a privacy policy in which they explicitly state exactly which kinds of tracking techniques they use including cookies and device recognition. This policy being publicly available however does not constitute users giving consent to data collection once they visit Amazon.com. This is because of the next rule.

3) ‘The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent. (GDPR, 2018)

As shown earlier, it is not straightforward for users to have their ‘consent’ withdrawn since consent is never explicitly given at all, which is in clear violation of this ruling.

4) ‘When assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.’ (GDPR, 2018)

Although Amazon’s algorithms help to optimize search results based on personal user data, they are not strictly necessary for the performance of Amazon.com’s search function. Therefore, Amazon’s data usage on their website is not being exempted by the ruling given here.

Now that we have shown that Amazon is in clear violation of several rulings within articles 6 and 7 of the GDPR, this thesis will look at the contents of Article 83: ‘General conditions for imposing administrative fines.’ The most important ruling for this thesis is the fifth subparagraph. This is because this paragraph refers to paragraph 2, which explains certain variables that are taken into account when determining the exact amount that is being fined, but is too long to list here. The fifth subparagraph reads: ‘Infringements of the following provisions shall, in accordance with paragraph 2, be subject to administrative fines of up to 20,000,000 EUR, or in the case of an undertaking, up to 4% of the total worldwide annual turnover of the preceding financial year, whichever is higher: a) The basic principles for processing, including conditions for consent, pursuant to articles 5, 6, 7 and 9;’

In order for the GDPR to be enforced for Amazon in this case, it needs to be reviewed by the Data Protection Commission based in each specific EU member country. As stated in the GDPR, they

are the only regulatory body that can make rulings and pass fines based on the GDPR. While it is noteworthy that they have a ‘major backlog’ in cases like this being solved (Lomas, 2020), there have already been some significant cases made. The most noteworthy of this is the case made against Google in 2019.

5.2) Comparison to Google

In 2019, French data protection watchdog CNIL, or the National Commission of Informatics and Civil Liberties, issued a fine of 50 million euros against Google. This was done after a complaint was made in 2018 by two different organizations, named ‘None Of Your Business’ and ‘La Quadrature du Net’. (Dillet, 2019) The complaint was made about the setup process for Android phones, during which the user is prompted to make a Google account in order to set up their device. It was ruled that the action of creating an account should be separated from setting up a device as consent bundling is illegal under the GDPR. (Dillet, 2019) Additionally, during this setup process there are certain boxes where the user can select whether or not they comply with Google’s privacy policy. Some of these boxes are checked by default, and the feature to opt out of personal ads is hidden under a ‘More Options’ link. Both of these practices are also considered to be in violation of the GDPR as the consent given by the user is not unambiguous in this case.

This case shows that even when the user is being prompted to give consent for the usage of their data, in this case for the purpose of showing personalized advertisements, this could still be in violation of the GDPR if it is ruled to be too ambiguous. Compare this to Amazon’s case where consent is not even being asked for in the first place. Further, Amazon certainly prompts the user to make an account on their main page but this is not required for the usage of their services. The user has no way of knowing that their data is being processed in any way just by using the site normally.

Recently Google’s appeal regarding the case made against them has been denied by France’s State Council, the country’s highest body of administrative law. In response to this, a Google spokeswoman sent the following statement to news website Techcrunch: “People expect to understand

and control how their data is used, and we've invested in industry-leading tools that help them do both. This case was not about whether consent is needed for personalised advertising, but about how exactly it should be obtained. In light of this decision, we will now review what changes we need to make." (Lomas, 2020) As the spokeswoman points out, this case was very much about the way that consent should be obtained from the user, rather than the question of whether consent is needed which is made very explicit by the GDPR. Though Amazon might consider that detailing their data processing activities in their Privacy Policy is sufficient to garner consent from users, the GDPR is unambiguous about the fact that this is simply not enough. Since it is very hard for the user to find out whether their data is being processed, not even considering its purpose or the possibility of the data being shared with further advertising partners, companies like Amazon need to be as explicit as possible in this regard.

The importance of research such as this one is highlighted by the fact that there is a major backlog of work needing to be done by the organizations that check whether companies process user data in a way that is in compliance with the GDPR. As of February 7, 2019, there had already been 95,000 unique complaints filed against companies for violation of the GDPR that went into effect in May 2018. (Richard Chapo, Esq. YouTube, 2019) Many companies indeed use business models that fall under the branches of data capitalism, surveillance capitalism, and platform capitalism as detailed in chapter 2 of this thesis. The number of cases, and the fact that global powerhouses such as Google and Amazon are amongst these cases, show that there is a direct public need for the enforcement of the GDPR. If left unchecked, these companies are in no rush to hold themselves to the GDPR's rules unprompted, since this usage of personalized browsing on their web store is surely a very lucrative business for them. Since our data has become the most valuable commodity we create and share, usage of this data has to be fair in order to ensure the privacy of everyone who could be affected.

Chapter 6: Conclusion

This thesis closely examined Amazon and its data collection practices from a data capitalism perspective and the possible ways in which these practices violate the GDPR. This was done through the research question: '*How can we conceptualize the data collection practices of Amazon in relation to the General Data Protection Regulation?*'

Through utilizing the amTREX tool, data was collected from Amazon.com's web store during various sessions. In these sessions, participants deliberately browsed various different websites with the goal of acquiring cookies and trackers from Amazon. This was done in order to examine whether the contents of these websites would in any way influence the search results of the queries that were made on the Amazon web store.

While this particular result was not shown to be significant enough to draw conclusions upon, a different finding arose from examining the data. It appeared that one particular user had a higher degree of personalisation in search results, including personalised advertisements, than any other user. The only difference between this user and all other users is the fact that this user had participated in a similar data research experiment which was conducted on the same device and using the same browser. This difference is already present in the data before additional variables are introduced by collecting cookies and trackers from various different websites. Therefore, the difference can not be explained by the usage of cookies but instead by the usage of device fingerprinting. Using this technique involves analyzing the specific nature of the user's hardware as well as their browser version and settings. Profiling users through device fingerprinting in such a way as well as storing their data for the purpose of enhancing search results in a webshop requires unambiguous consent on part of the user as stated by the GDPR. However, on Amazon.com no prompt of any kind regarding consent of data usage is given to the user. Therefore it is concluded that this way of collecting and using personal data is in violation of articles 6 and 7 of the GDPR.

This thesis further showed that not only this usage of data but particularly the usage of data in such a way that violates international law is not exclusive to Amazon, but that there are thousands of

companies on the receiving end of GDPR-related lawsuits. This includes smaller companies, but also other tech giants such as Google. This thesis argues that this form of data usage can be attributed to a form of data-oriented capitalism that has become increasingly dominant in recent years. Different scholars emphasize different aspects of this data-oriented capitalism, such as Sarah Myers West's Data Capitalism model which emphasizes the way in which power is redistributed to large companies that generate and analyze massive amounts of data. Shoshanna Zuboff's Surveillance Capitalism perspective draws attention to the intrusiveness of this data collection and the consequences it has for user privacy. Nick Srnicek's Platform Capitalism model emphasizes the data collection capabilities that are being facilitated by companies that follow a platform model, which allows companies to turn each interaction within their business into a data point.

Following this, various articles of the GDPR were examined and analyzed, in order to establish what Amazon would be in violation of should the research prove to be successful. Then the experiment was explained in such a way that can be replicated in order to confirm the findings given in this research. After this, the findings were presented in which Amazon was proven to be in violation of the GDPR as described previously.

There is an opportunity for further research to be done both related to Amazon as well as any other major company that makes the bulk of its profit from data collection practices. Replication of this research over a longer time frame can further prove the extent to which Amazon personalizes the search results of users. This could have implications for the severity of the GDPR violation that Amazon has committed. Specifically, it could be investigated if browsing different websites for a longer time period, thus acquiring more cookies and trackers, could have a more significant impact on the search results given.

In conclusion, to answer the research question: How can we conceptualize the data collection practices of Amazon in relation to General Data Protection Regulation? Amazon's data collection practices are so extensive that they collect the data of users who never signed up to create an account or consent to any amount of personal data usage. The extent to which they use the data of consenting

users who have Amazon accounts remains to be seen. And the severity of their GDPR violation is a question that will hopefully be answered as soon as the formal complaint that this thesis has resulted in will have been reviewed.

References

Angelovska, N. (2019, May 20). Top 5 Online Retailers: 'Electronics And Media' Is The Star Of E-commerce Worldwide. Retrieved from https://www.forbes.com/sites/ninaangelovska/2019/05/20/top-5-online-retailers-electronics-and-media-is-the-star-of-e-commerce-worldwide/#6d5fea3f1cd9

B. (2018, 11 oktober). *The staggering amount of data that Amazon records about its users*. Riccardo Coluccini. https://boter.eu/2018/10/11/amazon-data-access-request-clickstream/

Bellinger, G., Castro, D., & Mills, A. (2004). *Data, Information, Knowledge, & Wisdom*. Systems Thinking. https://www.systems-thinking.org/dikw/dikw.htm

Castells, M. (2000). Materials for an exploratory theory of the network society1. *The British Journal of Sociology*, *51*(1), 5–24. https://doi.org/10.1111/j.1468-4446.2000.00005.x

Choudary, S. P. (2015). *Dissecting Amazon's Platform Play | Platform Strategy – by Sangeet Paul Choudary*. Platformed.info. https://platformed.info/amazon-platform/

Contino. (2020, 20 februari). *Who's Using Amazon Web Services? [2020 Update]*. https://www.contino.io/insights/whos-using-aws

Council of Europe, *European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14*, 4 November 1950, ETS 5, available at: https://www.refworld.org/docid/3ae6b3b04.html

Dillet, R. (2019, 21 januari). *French data protection watchdog fines Google $57 million under the GDPR*. Techcrunch. https://techcrunch.com/2019/01/21/french-data-protection-watchdog-fines-google-57-million-under-the-gdpr/

European Parliament. (2018, 25 mei). *General Data Protection Regulation*. GDPR-Info.EU. https://gdpr-info.eu0 /

Gilpin, J. (2000). *The Challenge of Global Capitalism: The World Economy in the 21st Century*. PRINCETON; OXFORD: Princeton University Press. doi:10.2307/j.ctv36zqhs

Guendelsberger, E., & On the Clock: What Low-Wage Work Did to Me and How It Drives America Insane. (2019, July 18). Amazon Treats Its Warehouse Workers Like Robots: Ex-Employee. Retrieved from https://time.com/5629233/amazon-warehouse-employee-treatment-robots/

Hof, R. D. (2009, 6 augustus). *Betting on the Real-Time Web*. Businessweek. http://www.businessweek.com/magazine/content/09_33/b4143046834887.htm

The HSUS v. Amazon.com, Inc., et al. (Animal fighting materials): The Humane Society of the United States. (n.d.). Retrieved from https://web.archive.org/web/20100925194514/http://www.hsus.org/in_the_courts/docket/amazon.html

ICO. (z.d.). *Right of access*. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-of-access/

Investopedia. (2020a, januari 29). *7 Ways Amazon Uses Big Data to Stalk You*. https://www.investopedia.com/articles/insights/090716/7-ways-amazon-uses-big-data-stalk-you-amzn.asp

Investopedia. (2020b, juni 14). *Vertical Integration*. https://www.investopedia.com/terms/v/verticalintegration.asp

Kohavi, R., & Longbotham, R. (2017). Online Controlled Experiments and A/B Testing. *Encyclopedia of Machine Learning and Data Mining*, 922–929. https://doi.org/10.1007/978-1-4899-7687-1_891

Lomas, N. (2020, 19 juni). *French court slaps down Google's appeal against $57M GDPR fine*. Techcrunch. https://techcrunch.com/2020/06/19/french-court-slaps-down-googles-appeal-against-57m-gdpr-fine/

MacKinnon, R. (2012). *Consent of the Networked*. Adfo Books.

Manifest, T. (2019, 6 februari). *Amazon's User Experience: A Case Study - The Manifest*. Medium. https://medium.com/@the_manifest/amazons-user-experience-a-case-study-fb567f79b51f

Miller, R. (2016, 2 juli). *How AWS came to be*. Techcrunch. https://techcrunch.com/2016/07/02/andy-jassys-brief-history-of-the-genesis-of-aws/

Ng, A. (2019, 9 mei). *Amazon Alexa transcripts live on, even after you delete voice records*. CNET. https://www.cnet.com/news/amazon-alexa-transcripts-live-on-even-after-you-delete-voice-records/

Nikiforakis, N., & Acar, G. (2014, 25 juli). *Browser Fingerprinting and the Online-Tracking Arms Race*. IEEE Spectrum. https://spectrum.ieee.org/computing/software/browser-fingerprinting-and-the-onlinetracking-arms-race

Nocun, K. (2018, 28 april). *Beklemmender Selbstversuch: So Viel Weiß Amazon Nach Jedem Meiner Klicks*. Der Spiegel. https://www.spiegel.de/netzwelt/web/amazon-experiment-was-der-konzern-mit-jedem-klick-erfaehrt-a-1205079-amp.html

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2016). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Schreier, M. (2014). Qualitative Content Analysis. *The SAGE Handbook of Qualitative Data Analysis*, 170–183. https://doi.org/10.4135/9781446282243.n12

Shepard, W. (2018, 16 januari). *Fuse Chicken Vs. Amazon Is The David Vs. Goliath Lawsuit To Watch In 2018*. Forbes. https://www.forbes.com/sites/wadeshepard/2018/01/14/fuse-chicken-vs-amazon-is-the-david-vs-goliath-lawsuit-to-watch-in-2018/#78ce39dd5115

Srnicek, N. (2017). *Platform Capitalism*. Wiley. https://www.wiley.com/en-nl/Platform+Capitalism-p-9781509504862

Stalcup, K. (2019, 25 juni). *How Big is AWS?* ParkMyCloud. https://www.parkmycloud.com/blog/how-big-is-aws/

Thornberg, R., & Charmaz, K. (2014). Grounded Theory and Theoretical Coding. *The SAGE Handbook of Qualitative Data Analysis*, 153–169. https://doi.org/10.4135/9781446282243

Thrift, N. (2005). *Knowing Capitalism*. SAGE Publications.

Tracking Exposed. (2020, 13 januari). *Amazon.Tracking.Exposed*. Github. https://github.com/tracking-exposed/presentation/blob/master/amazon.tracking.exposed%20-%20English%20short%20report%20%20-%20Version%204.pdf

van Doorn, N. (2017). The Parameters of Platform Capitalism. *Krisis*, *2017*(2), 103–107. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiCxufTjMbrAhVMyaQKHaK1D0kQFjAAegQIAhAB&url=https%3A%2F%2Fpure.uva.nl%2Fws%2Ffiles%2F35641975%2FKrisis_2018_1_Full_issue.pdf&usg=AOvVaw0HJPYmmok6q-GgyXmMfh0t

Wang, G., Zhang, X., Tang, S., Wilson, C., Zheng, H., & Zhao, B. Y. (2017). Clickstream User Behavior Models. *ACM Transactions on the Web*, *11*(4), 1–37. https://doi.org/10.1145/3068332

Weltevrede, E., Helmond, A., & Gerlitz, C. (2014). The Politics of Real-time: A Device Perspective on Social Media Platforms and Search Engines. *Theory, Culture & Society*, *31*(6), 125–150. https://doi.org/10.1177/0263276414537318

West, S. M. (2017). Data Capitalism: Redefining the Logics of Surveillance and Privacy. *Business & Society*, *58*(1), 20–41. https://doi.org/10.1177/0007650317718185

Willig, C. (n.d.). Interpretation and Analysis. *The SAGE Handbook of Qualitative Data Analysis* , 136-150. https://doi.org/10.4135/9781446282243

Zak, A., Olthof, E., Koehorst, D., Bonati, O., Heydel, Z., & Khatib, A. (2020, 30 januari). *Amazon's Choice*. Digital Methods Initative. https://wiki.digitalmethods.net/Dmi/WinterSchool2020amazonschoice

Zuboff, S. (2016, 5 maart). *The Secrets of Surveillance Capitalism*. FAZ.NET. https://www.faz.net/aktuell/feuilleton/debatten/the-digital-debate/shoshana-zuboff-secrets-of-surveillance-capitalism-14103616.html

ZURB. (2011, 3 oktober). *Why Amazon Will Sell Kindle Fire At a Loss*. https://zurb.com/blog/why-amazon-will-sell-kindle-fire-at-a-los